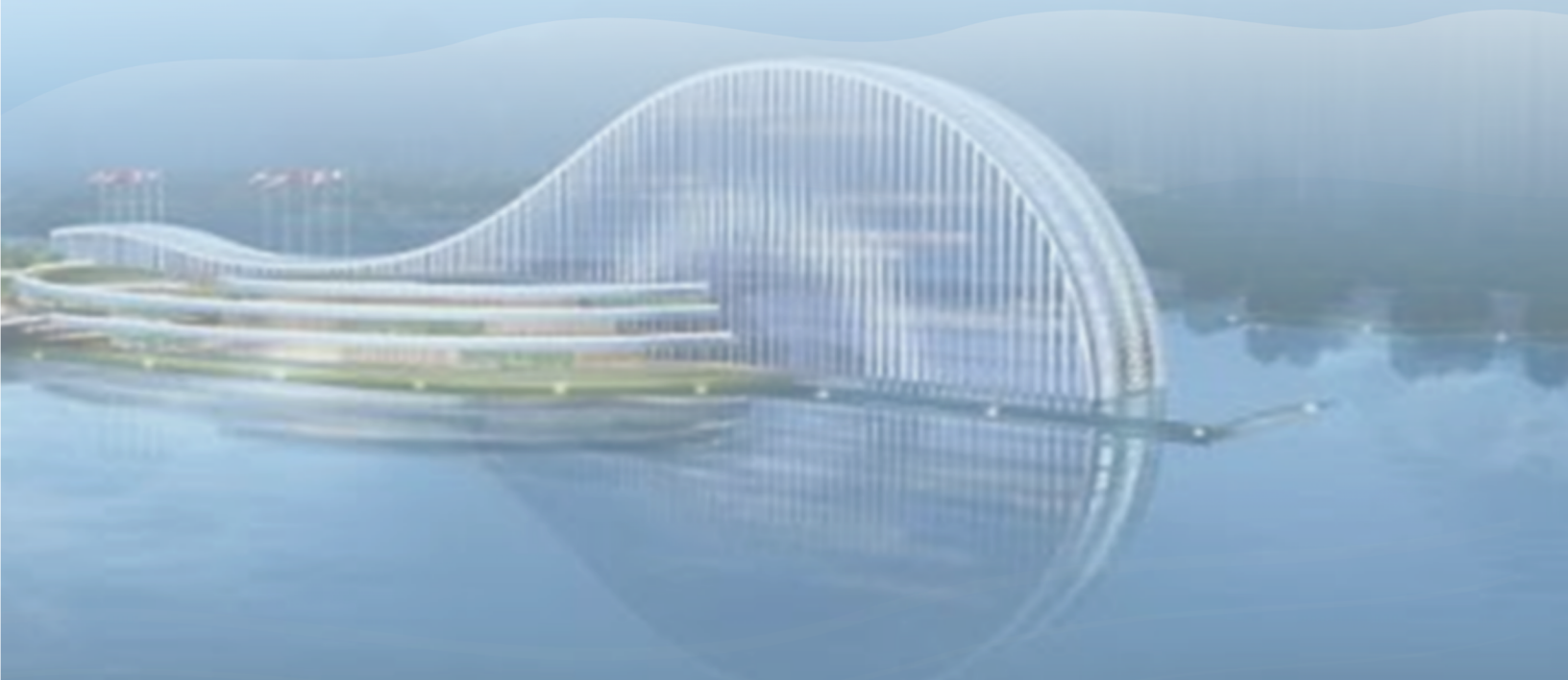
AACL-IJCNLP 2026
Hengqin | China

AFNLP

Conference Handbook

November 6–10, 2026 · **Hengqin, Zhuhai, China**

The **5th** Asia-Pacific Chapter of the Association for Computational Linguistics & the **15th** International Joint Conference on Natural Language Processing



Contents

1. Welcome Messages

2. Conference at a Glance

3. Registration

4. Venue and Maps

5. Program Overview

6. Keynote Talks

7. Tutorials

8. Workshops

9. Main Conference Oral Sessions and Poster Sessions

10. Meals, Reception, and Social Events

11. Travel and Local Information

12. Organizing Committee and Acknowledgments

13. Sponsors

14. Paper Metadata and Proceedings

1. Welcome Messages

Message from the General Chair

Welcome to AACL-IJCNLP 2026 in Hengqin, China!

This joint conference brings together the Asia-Pacific and international natural language processing communities to share research, exchange ideas, and build connections.

AACL's previous editions have been associated with Suzhou in 2020, Taipei as the planned host city for 2022, Bali in 2023, and Mumbai in 2025. The 2020 and 2022 editions were held online. This year, we are pleased to gather in Hengqin.

We are delighted to welcome our four invited speakers: Iryna Gurevych, Yulan He, Tim Baldwin, and Nizar Habash.

The organizing committee includes scholars affiliated with institutions in nine countries and regions. We look forward to conversations that connect languages, disciplines, institutions, and different perspectives on the future of our field.

We hope that the conference will provide a welcoming setting for presenting your work, learning from colleagues, and meeting new collaborators. Whether you are joining us in Hengqin or participating online, thank you for contributing to this community.

We thank the organizing committee, reviewers, tutorial instructors, workshop organizers, invited speakers, volunteers, sponsors, and all participants for the work that makes this meeting possible. We wish you a rewarding conference and an enjoyable stay in Hengqin.

Yue Zhang · General Chair · Westlake University

Message from the Program Chairs

Welcome to 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing and the (AACL-IJCNLP 2026). We are thrilled to host this premier event in the beautiful and vibrant city of Hengqin, Zhuhai, China, from November 6 to 10, 2026. Embracing a hybrid format, the conference welcomes attendees to join us in person or participate virtually from across the globe.

Organizing AACL-IJCNLP 2026 has been a profoundly collaborative endeavor, brought to life by the dedication and tireless efforts of many. We extend our deepest gratitude for the invaluable support and contributions of the following:

- General Chair: Yue Zhang;

- Previous AACL-IJCNLP 2025 Program Committee Co-Chairs: Derek F. Wong, Haofen Wang, Sakriani Sakti;
- ARR Editors-in-Chief : Vincent Ng, Anna Rogers, Jing Jiang, Michael White, Margot Mieskes, Xiaodan Zhu, Sarvnaz Karimi, Xiaojun Wan, Giovanni Da San Martino, Parisa Kordjamshidi, Tanmoy Chakraborty, and Asad Sayeed;
- The OpenReview support team for multiple rounds of technical help in setting up on the OpenReview platform;
- The incredible team of 33 Senior Area Chairs and 1,787 reviewers, whose diligent work ensured a thorough review process;
- Best Paper Committee Chairs: Rui Yan, and all committee members;
- Ethics Chairs: Yangqiu Song, Ekaterina Vylomova;
- Underline Team for their professional services and support in enabling the online component of the conference.

The success of this conference is a testament to the exceptional leadership and collaborative spirit of the AACL-IJCNLP 2026 chairs. We extend our deepest gratitude to the following individuals for their unwavering dedication and invaluable service:

- Demonstration Chairs: Ningyu Zhang, Derry Wijaya;
- SRW Faculty Advisors: Yixin Cao;
- Student Research Workshop Chairs: Shibo Hong, Yuhang Zhou, Wenhe Sun;
- Diversity and Inclusion Chairs: Mamoru Komachi, Emily Ohman;
- Publication Chairs: Junxian He, Trang Vu;
- Handbook Chairs: Leyang Cui;
- Web and App Chairs: Xinchi Chen, Piji Li, Shudong Liu;
- Local Organizing Chairs: Derek F. Wong, Haofen Wang;
- Local Organizing Committee Members: Runzhe Zhan, Zhiwen Xie.

Finally, we extend our deepest appreciation to all the authors for their outstanding scientific contributions, which form the intellectual core of AACL-IJCNLP 2026. We wish all participants a highly rewarding and inspiring conference experience, whether joining us in person or virtually.

Submission and Acceptance

AACL-IJCNLP 2026 utilized the ACL Rolling Review (ARR) as its submission and review platform. Out of a pool of over 16,000 submissions to the May ARR cycle, a total of 776 papers were committed to this conference. These submissions represented a truly global community, with 51.9% from the Asia-Pacific region, 25.6% from North America, 15.2% from Europe, 4.1% from the Middle East, 0.3% from Latin America, and 0.1% from Africa. All manuscripts underwent a rigorous evaluation process across multiple tracks, under the expert guidance of Senior Area Chairs (SACs) and Area Chairs (ACs).

The ARR cycle proceeded seamlessly, ensuring that every committed paper received at least three independent reviews and a meta-review. In strict accordance with the ACL Code of Ethics, we conducted dedicated ethics reviews for submissions raising potential ethical concerns. Authors of papers receiving ethics-related feedback were required to address these comments in their final camera-ready versions.

Following careful deliberation, 355 papers (316 long and 39 short) were accepted to the main conference, yielding an acceptance rate of 45.7%. An additional 265 papers (239 long and 26 short) were accepted to Findings, resulting in an acceptance rate of 34.1%. The overall acceptance rate across both venues was 79.9%.

Design of Tracks

AACL-IJCNLP 2026 incorporates 29 general areas that reflect current trends in the field. Most of these areas align with recent ACL conferences.

- Clinical and Biomedical Applications
- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Efficient Methods for NLP
- Ethics, Bias, and Fairness
- Generation
- Human-Centered NLP and Human-AI Interaction
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Modeling
- LLM Agents
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation
- Multilingualism and Cross-Lingual NLP
- Multimodality and Language Grounding to Vision, Robotics and Beyond
- NLP and Code Models
- NLP and Symbolic Reasoning
- NLP Applications
- Phonology, Morphology, and Word Segmentation

- Question Answering
- Resources and Evaluation
- Semantics: Lexical and Sentence-Level
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech Recognition, Text-to-Speech and Spoken Language Understanding
- Summarization
- Syntax: Tagging, Chunking and Parsing

Best Paper Awards

In accordance with the updated ACL conference awards policy, we convened the Best Paper Award Committee to select this year's recipients. Following the committee's rigorous deliberation, one paper was selected for the Best Paper Award, and three papers were selected for the Outstanding Paper Awards. The winners will be officially announced during a dedicated plenary session on November 8, 2026.

Finally, we hope that AACL-IJCNLP 2026 proves to be an engaging and enriching experience for all participants. We eagerly anticipate the fruitful discussions, innovative ideas, and lasting collaborations that will undoubtedly emerge from this year's program.

Ruifeng Xu (Harbin Institute of Technology, Shenzhen)

Ayu Purwarianti (Institut Teknologi Bandung)

Tong Rong (Singapore Institute of Technology)

Jun Suzuki (Tohoku University)

Program Chairs, AACL-IJCNLP 2026

Message from the Local Organizing Chairs

We are delighted to welcome you to the 5th Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2026) in Hengqin, Zhuhai, China.

Hengqin is a dynamic meeting point within the Guangdong-Hong Kong-Macao Greater Bay Area. We hope the conference offers an enjoyable setting for sharing research, exchanging ideas, and building new collaborations.

We sincerely thank the General Chair, Program Chairs, organizing committee, reviewers, volunteers, sponsors, and all participants whose work makes this conference possible.

We hope your time at AACL-IJCNLP 2026 will be productive, welcoming, and memorable, and that you will enjoy your stay in Hengqin.

AACL-IJCNLP 2026 Local Organization Chairs

Derek F. Wong (University of Macau)

Haofen Wang (Tongji University)

2. Conference at a Glance

Item	Information
Conference	AACL-IJCNLP 2026
Overall dates	November 6–10, 2026
City	Hengqin, Zhuhai, China
Venue	Tianmu Melody Convention Exhibition Center
Program time zone	China Standard Time (UTC+8)
Main conference	November 6–8, 2026
Tutorials	November 9 and 10
Workshops	November 9 and 10, according to the workshop listing

3. Registration

Use the official registration portal at https://2026.aaclnet.org/registration/. On-site registration is available from November 1–10, 2026.

All attendees must hold current 2026 ACL membership and must be registered to access the venue or Underline. All Findings papers must be presented in person or virtually.

For registration questions, contact Wenxuan Zhang at wxzhang@sutd.edu.sg.

4. Venue and Maps

The conference takes place at **Tianmu Melody Convention Exhibition Center**, on the artificial island in Tianmu River, Hengqin Island, Zhuhai, China. Hotel Indigo is in the same building.

Location

Tianmu Melody Convention Exhibition Center, Artificial Island, Tianmu River, Hengqin Island, Zhuhai, China

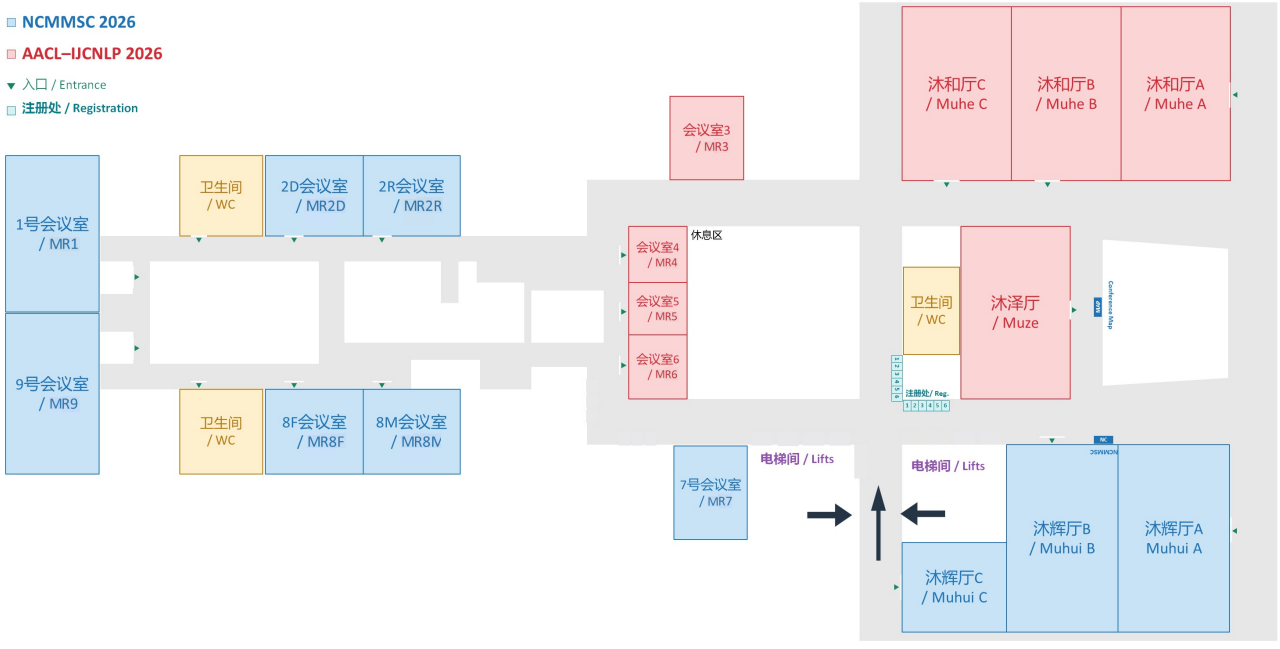
天沐琴台，横琴粤澳深度合作区天沐河中央人工岛

Taxi drop-off point:“请带我去天沐琴台英迪格酒店门口。”



Floor Plan

B1 会场 / B1 Venue



5. Program Overview

AACL-IJCNLP 2026 Program Overview

Day 0 Thurday Nov 5, 2026

14:00-20:00	Registration (Full Conference) @ B1 Floor, Tianmu Qintai Conference and Exhibition Center

Day 1 Friday Nov 6, 2026

Time	Session	Muhe B+C	MR 3	MR 4	MR 5	MR 6	Muhe B	Muhe C	Poster area
08:00-09:00	Registration (Full Conference) @ B1 Floor, Tianmu Qintai Conference and Exhibition Center								
09:00-09:30	Opening @ Muhe B+C								
09:30-10:30	**KS1**	Keynote 1 Iryna Gurevych, Technical University of Darmstadt @ Muhe B+C							
10:30-11:00	Coffee Break @ Foyer								
11:00-12:30	**S1**	Session 1.1 Resources and Evaluation 1	Session 1.2 Interpretability and Analysis of Models for NLP 2	Session 1.3 NLP Applications and Human-AI Interaction 1	Session 1.4 Ethics, Bias, and Fairness	Session 1.5 Generation and Dialogue	/	/	Poster & Demo Session 1 (For seesion 2 oral papers and findings)
12:30-14:00	Lunch @ Muze & Muhe A								
14:00-15:00	**KS2**	Keynote 2 Yulan He, King's College London							
15:00-15:30	Coffee Break @ Foyer								
15:30-17:00	**S2**	Session 2.1 Resources and Evaluation 2	Session 2.2 Multimodality and Language Grounding	Session 2.3 NLP Applications and Human-AI Interaction 2	Session 2.4 Computational Social Science and Cultural Analytics	Session 2.5 Efficient Methods and Machine Learning for NLP 1	/	/	Poster & Demo Session 2 (For session 1 oral papers and findings)
18:30-20:30	Welcome Reception @ Hotel Indigo Restaurant & Function Room								

Day 2 Saturday Nov 7, 2026

Time	Session	Muhe B+C	MR 3	MR 4	MR 5	MR 6	Muhe B	Muhe C	Poster area
08:30-09:00	Registration (Full Conference) @ B1 Floor, Tianmu Qintai Conference and Exhibition Center								
09:00-10:00	**KS3**	Keynote 3 Tim Baldwin, MBZUAI @ Muhe B+C							
10:00-10:30	Coffee Break								
10:30-12:00	**S3**	Session 3.1 Resources and Evaluation 3	Session 3.2 LLM Agents and Learning	Session 3.3 Clinical and Biomedical Applications	Session 3.4 Interpretability and Analysis of Models for NLP 1	Session 3.5 Machine Translation and Phonology	/	/	Poster & Demo Session 3 (For findings)
12:00-13:30	Lunch @ Muze & Muhe A								
13:30-15:00	AACL-NCMMSC Joint Forum								Poster & Demo session 4 (For session 3 oral papers)
15:00-15:30	Coffee Break								
15:30-16:00	Departure to Social Event								
16:00-21:00	Social Event @ Chimelong Ocean Kingdom Hengqin								

Day 3 Sunday Nov 8, 2026

Time	Session	Muhe B+C	MR 3	MR 4	MR 5	MR 6	Muhe B	Muhe C	Poster area
08:30-09:00	Registration (Full Registration&Tutorial/Workshop Registration) @ B1 Floor, Tianmu Qintai Conference and Exhibition Center								
09:00-10:00	**KS4**	Keynote 4 Nizar Habash, New York University Abu Dhabi @ Muhe B+C							
10:00-10:30	Coffee Break								
10:30-12:00	**S4**	Session 4.1 Resources and Evaluation 4	Session 4.2 Language Modeling	Session 4.3 Efficient Methods and Machine Learning for NLP 2	Session 4.4 Semantics and Linguistic Structure	Session 4.5 Question Answering and Retrieval	/	/	Poster & Demo session 5 (For session 5 oral papers and findings)
12:00-14:00	Lunch @ Muze & Muhe A								
14:00-15:30	**S5**	Session 5.1 Resources and Evaluation 5	Session 5.2 Multimodal, Medical, and Embodied Models	Session 5.3 Information Extraction, Retrieval, and Code	Session 5.4 Multilingualism and Cross-Lingual NLP	Session 5.5 Sentiment, Culture, and Model Behavior	/	/	Poster & Demo session 6 (For session 4 oral papers and findings)
15:30-16:00	Coffee Break								
16:00-16:30	Best Paper Session @ Muhe B+C								
16:30-17:00	Closing @ Muhe B+C								

Day 4 Monday Nov 9, 2026

Time	Session	/	MR 3	MR 4	MR 5	MR 6	Muhe B	Muhe C	MR8M	MR8F
08:30-09:00	Registration (Full Registration&Tutorial/Workshop Registration) @ B1 Floor, Tianmu Qintai Conference and Exhibition Center									
09:00-10:30	**AM Session1**	/	WS 01	WS 06	WS 03	WS 04	T01	T02	WS 05	WS02
10:30-11:00	Coffee Break									
11:00-12:30	**AM Session2**	/	WS 01	WS 06	WS 03	WS 04	T01	T02	WS05	WS02
12:30-14:00	Lunch @ Muze & Muhe A									
14:00-15:30	**PM Session 1**	/	WS 01	WS 06	WS 03	WS 04	T03	T04	WS05	WS02
15:30-16:00	Coffee Break									
16:00-17:30	**PM Session 2**	/	WS 01	WS 06	WS 03	WS 04	T03	T04	WS05	WS02

Day 5 Tuesday Nov 10, 2026

Time	Session	/	MR 3	MR 4	MR 5	MR 6	Muhe B	Muhe C	MR8M	/
08:30-09:00	Registration (Full Registration&Tutorial/Workshop Registration) @ B1 Floor, Tianmu Qintai Conference and Exhibition Center									
09:00-10:30	**AM Session1**	/	WS 07	WS 08	WS 09	WS 10	T05	WS 11	WS 12	/
10:30-11:00	Coffee Break									
11:00-12:30	**AM Session2**	/	WS 07	WS 08	WS 09	WS 10	T05	WS 11	WS 12	/
12:30-14:00	Lunch @ Muze & Muhe A									
14:00-15:30	**PM Session 1**	/	WS 07	WS 08	WS 09	WS 10	T06	WS 11	WS 12	/
15:30-16:00	Coffee Break									
16:00-17:30	**PM Session 2**	/	WS 07	WS 08	WS 09	WS 10	T06	WS 11	WS 12	/

The program is also available at the official program page.

6. Keynote Talks

Keynote 1 · Iryna Gurevych



Speaker	Iryna Gurevych
Affiliation	Technical University of Darmstadt
Date & Time	Friday, November 6, 09:30 – 10:30
Venue	Muhe B+C

Talk Title

NLP for Mental Health: Risk, Promise, Responsibility

Abstract

More than a billion people live with a mental health condition, yet care remains scarce and often a poor fit for the person receiving it — so people increasingly turn to unsupervised chatbots to fill the gap. Because language is the richest signal we have into someone's psychological state, natural language processing is the key to doing this better. In this talk, I argue that NLP can genuinely help — but solving it takes clinical grounding, real data, and responsible evaluation, not just larger models. Drawing on work from my group, I show how large language models can turn therapy transcripts into treatment-relevant structure, how clinically grounded synthetic data eases the field's privacy bottleneck, and why evaluation must reflect what clinicians actually care about — not just what's easy to benchmark. I close with an invitation to build mental health AI that is personalized, privacy-aware, and responsibly evaluated: tools that support clinicians, not replace them.

Speaker's Bio

Iryna Gurevych is a full professor of computer science at the Technical University of Darmstadt, Germany, where she founded and heads the Ubiquitous Knowledge Processing (UKP) Lab. She is a pioneer of computational argumentation and cross-document text modeling, with landmark applications to fact verification and misinformation detection, and her research extends to large language models and interdisciplinary NLP for the social sciences and humanities. Prof. Gurevych served as President of the Association for Computational Linguistics (ACL) in 2023 and is an ACL Fellow, a member of the German National Academy of Sciences Leopoldina and Academia Europaea, and a recipient of the 2025 Royal Society Milner Award.

Keynote 2 · Yulan He



Speaker	Yulan He
Affiliation	King's College London
Date & Time	Friday, November 6, 14:00 – 15:00
Venue	Muhe B+C

Talk Title

Beyond Self-Play: Towards Effective Self-Evolution in LLMs

Abstract

Can large language models keep improving through experience they generate themselves? Self-play and self-generated experience offer promising routes. But producing more synthetic data does not necessarily lead to sustained model capability growth. This talk argues that effective self-evolution depends on whether the learning pipeline continues to provide learnable information. Drawing on our recent work, I will examine several ways such information can arise, including generating informative self-play experience, learning directly from environment-generated experience, and actively seeking novel evidence rather than waiting for tasks to be provided. These studies suggest that effective self-improvement requires more than generating its own experience; it needs experience that contains useful information that the model can actually learn from.

Speaker's Bio

Yulan He is a Professor in Natural Language Processing at King's College London and a Turing AI Fellow. Her research focuses on making Large Language Models (LLMs) more reliable, interpretable, and safe, particularly in reasoning and agentic AI, with applications spanning across education, health, and science. She has received several awards for her research. Her five-year Turing AI Fellowship project was awarded the Best Research Project (Research Excellence) at the RAi UK AI & Robotics Awards 2026. She also received a Best Research Paper Award at the same awards ceremony. In addition, she is the recipient of a SWSA Ten-Year Award and a CIKM Test-of-Time Award.

Keynote 3 · Tim Baldwin



Speaker	Tim Baldwin
Affiliation	MBZUAI
Date & Time	Saturday, November 7, 09:00 – 10:00
Venue	Muhe B+C

Talk Title

Uncertainty, Belief, and Trust in Language Models

Abstract

Language models and other AI technologies are being used in an ever-increasing range of applications, including being trialled for decision support or automated decision-making in sensitive applications. A critical component of any such usage is the capability for the model to output a confidence/uncertainty assessment associated with any output, or (statistical) belief underlying a particular action. In this talk, I will present recent work on uncertainty estimation and belief updates, with a particular focus on practical utility in improving model performance and efficiency, and enabling humans to perform tasks more effectively.

Speaker's Bio

Tim Baldwin is Professor of Natural Language Processing at Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), in addition to being a Melbourne Laureate Professor in the School of Computing and Information Systems at The University of Melbourne, and Co-founder and Chief Scientist of startups including LibrAI (focused on AI safety) and Examn AI (focused on dialogue-based candidate profiling). At MBZUAI he has also held positions including Provost, Associate Provost, and Foundation Chair of the Department of Natural Language Processing.

Keynote 4 · Nizar Habash



Speaker	Nizar Habash
Affiliation	New York University Abu Dhabi
Date & Time	Sunday, November 8, 09:00 – 10:00
Venue	Muhe B+C

Talk Title

The Genie of the LLM: What Shall We Wish for the Future of NLP?

Abstract

Large language models have given NLP extraordinary new capabilities. The genie is out of the lamp. The question is no longer only what it can do, but what we should ask it to do, how we know we got what we really wanted, and what it will do to us, our research, and our education.

Speaker's Bio

Nizar Habash is a Professor of Computer Science at New York University Abu Dhabi (NYUAD). He is also the director of the Computational Approaches to Modeling Language (CAMeL) Lab. Professor Habash specializes in natural language processing and computational linguistics. He received his PhD in Computer Science from the University of Maryland College Park in 2003. His research includes extensive work on machine translation, morphological analysis, and computational modeling of Arabic and its dialects. Professor Habash has been a principal investigator or co-investigator on over 30 research grants. And he has over 300 publications including a book entitled “Introduction to Arabic Natural Language Processing”. Professor Habash is one of the inaugural recipients of the King Salman Academy for Arabic Language Award (2022); he is the recipient of the Antonio Zampolli Prize (2024); and he was selected as a Fellow of the Association for Computational Linguistics (2025). His website is www.nizarhabash.com.

7. Tutorials

Monday, November 9, 2026

ID	Title	Instructors	Room	Time
T1	Code-Switching for Multilingual LLMs	Juhyun Oh, Alice Oh, Thamar Solorio, Haneul Yoo, Genta Indra Winata, Ruochen Zhang	Muhe B	9:00-12:30
T2	The Future of NLP Needs Ethical Considerations: An Interactive Tutorial for Researchers and Students	Min-Yen Kan, Barid Xi Ai, Esther Gan, Yisong Miao, Abdoul Jalil Djiberou Mahamadou, Seunghun Lee, Margot Mieskes, Adriana Pagano, Monojit Choudhury, Vamshi Krishna Bonagiri	Muhe C	9:00-12:30
T3	A Tutorial on Measurement Science for Language Model Evaluation	Sang Truong, Rudy Corona, Serena Wang, Sanmi Koyejo	Muhe B	14:00-17:30
T4	Post-Deployment Self-Improving Language Agents: Learning, Memory, and Adaptation	Estevam Hruschka	Muhe C	14:00-17:30

Tuesday, November 10, 2026

ID	Title	Instructors	Room	Time
T5	From Retrieval-Augmented Generation (RAG) to Agentic Deep Research	Luna Dong, Sanat Sharma, Xiao Yang, Jiaqi Wang, Kai Sun, Yinglong Xia, Scott Yih	Muhe B	9:00-12:30
T6	Diffusion Language Models: Foundations and Frontiers of Non-Autoregressive Language Generation	Daisuke Oba, Danushka Bollegala	Muhe B	14:00-17:30

8. Workshops

Monday, November 9, 2026

Code	Workshop Name	Format	Room	Time
WS01	Fifth Workshop on Multimodal Machine Learning in Low-Resource Languages at AACL-IJCNLP 2026	Hybrid	MR3	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS02	The 13th Workshop on Asian Translation (WAT2026)	Hybrid	MR8F	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS03	The 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP)	Virtual	MR05	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS04	The First Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche 2026)	Hybrid	MR06	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS05	Multi-modality and Multi-lingual Language Learning (Multi-LLL 2026)	Hybrid	MR8M	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS06	First Workshop on Language Technologies for Crisis Preparedness and Response (LT4CPR)	Hybrid	MR4	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30

Tuesday, November 10, 2026

Code	Workshop Name	Format	Room	Time
WS07	ARRSI 2026 — First Workshop on Autonomous Research and Recursive Self-Improvement	Hybrid	MR3	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS08	TrustAudio: Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI	Hybrid	MR4	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS09	The First Workshop on Pluralistic Value Alignment of LLMs (PlurVA-LLM)	Hybrid	MR5	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30

WS10	Fourth Workshop on Artificial Intelligence for Scientific Publications (WASP 2026)	Virtual	MR6	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS11	From Reasoning to Agency: Learning, Acting, and Adapting with Foundation Models	Hybrid	Muhe C	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30
WS12	NORA: The 3rd Workshop on KNOwledge GRaphs & Agentic Systems Interplay	Hybrid	MR8M	09:00-10:30, 11:00-12:30 / 14:00-15:30, 16:00-17:30

9. Main Conference Oral Sessions

Session 1.1: Resources and Evaluation 1

Friday, November 6, 2026 · 11:00-12:30 · Muhe B+C

Slot	ID	Title	Authors	Presenter
1	15	Constructive Comment Extraction from Japanese YouTube Shorts: A Dataset and Empirical Study	Siyuan Feng; Luecheng Wang; Koyo Ogura; Yuma Iwamoto; Seinosuke Sato; Yuki Kometani; Masaki Itabashi; Shintaro Ozaki; Yusuke Sakai; Shogo Matsuno; Hidetaka Kamigaito; Katsuhiko Hayashi; Ryoma Obara; Sho Takase; Taro Watanabe	Siyuan Feng
2	51	ARR-Dv2: New Data from ACL Rolling Review and its Potential to Improve Reviewing in the ACL Community	Atnafu Lambebo Tonja; Sheng Lu; Nils Dycke; Dirk Hovy; Koen Dercksen; Lizhen Qu; Xiaodan Zhu; Thamar Solorio; Iryna Gurevych; Margot Mieskes	Thamar Solorio
3	78	Auto-AdvER: Disentangling Condition, History, and Commercial Offers in Automotive Advertisements	Filippos Ventirozos; Ioanna Nteka; Tania Nandy; Jozef Bača; Charlie Roadhouse; Peter A. Appleby; Matthew Shardlow	Filippos Ventirozos
4	90	VIDA: A Dataset for Visually Dependent Ambiguity in Multimodal Machine Translation	Jingheng Pan; Xintong Wang; Longyue Wang; Liang Ding; Weihua Luo; Chris Biemann	Jingheng Pan
5	136	FahMai-TeleBench: A Thai Grounded Tool-Use Benchmark for the Deployable Model Tier	Nutchanon Yongsatianchot; Piyalitt Ittichaiwong; Jitkapat Sawatphol; Tichanat Ruangsiri	Nutchanon Yongsatianchot
6	138	ThaiEditBench: A Mechanism-Typed, Length-Stratified Benchmark for Thai Orthographic Error Correction	Nutchanon Yongsatianchot; Piyalitt Ittichaiwong; Kanyakorn Veerakanjana	Nutchanon Yongsatianchot

Session 1.2: Interpretability and Analysis of Models for NLP 2

Friday, November 6, 2026 · 11:00-12:30 · MR 3

Slot	ID	Title	Authors	Presenter
1	453	Do LLM Embedding Spaces Recover Expert Structure?	Yixuan Zhu; Zhenke Duan; Fanghen Li	Zhenke Duan
2	517	Multidimensional Consistency Improves Reasoning in Language Models	Huiyuan Lai; Xiao Zhang; Malvina Nissim	Huiyuan Lai
3	523	What Changes in Hidden States When Model Merging Hurts Task Performance?	Canasai Kruengkrai; Koichiro Yoshino	Canasai Kruengkrai
4	699	RL-ARC: Calibrating Large Reasoning Models via Reasoning-guided Uncertainty	Gukhyeon Lee; SangKeun Lee	Gukhyeon Lee
5	719	Predicting Mergeability of Language Model Domain Experts	Ervin Dervishaj; Quynh Do; Giannis Karamanolakis; Anna Rumshisky; Salma Sohrabi-Jahromi; Judith Gaspers	Ervin Dervishaj
6	761	Information-Theoretic Design of Sparse Autoencoder-Based Steering	Hirofumi Morise	Hirofumi Morise

Session 1.3: NLP Applications and Human-AI Interaction 1

Friday, November 6, 2026 · 11:00-12:30 · MR 4

Slot	ID	Title	Authors	Presenter
1	122	A Multilingual Evaluation of LLM-Based High-School Mathematics Assistance Across Indian Languages	Yash Kumar; Subhajit Roy	Yash Kumar
2	188	Conversational Intelligent Tutoring Systems: Evolution, Taxonomy, and Pedagogical Frontiers	Zhengyuan Liu; Stella Xin Yin; Li Zhu; Xiaoman Wang; Yunshi Lan; Nancy F. Chen	Zhengyuan Liu
3	261	DeBERTa-ConPara: Attack-Aware and Deployment-Realistic Detection of AI-Generated Text	Mohamed Mady; Yupei Li; Johannes Reschke; Björn Schuller	Mohamed Mady
4	344	MisconceptDialogue: Evaluation Criteria Shape Measured Success in LLM Math Tutoring	Woojin Lee; Woo Yoon Kyu; Jinah Kim; Hyeok-Min Gwon; Jin-Xia Huang	Woojin Lee
5	424	CAFE: Caption-Anchored Feature Enrichment and Evidence Routing for Multimodal Sarcasm Detection	Puchu Zeng; Yun Luo; qianqian li; Yanzhou Ma; Liu Junyi; SongFengxuan; chonghuifang	Yun Luo
6	563	“Do You Truly Love Me?” Benchmarking LLM Capability on Hierarchical Pragmatic Tactic Conversations	Peter Q. Chen; Pangjing Wu; Zhuohang Jiang; Xiaodong Li; Peter H. F. Ng	Peter Q. Chen

Session 1.4: Ethics, Bias, and Fairness

Friday, November 6, 2026 · 11:00-12:30 · MR 5

Slot	ID	Title	Authors	Presenter
1	19	Decomposing LLM Advice Sycophancy: Content, Style, and Framing Drive Distinct Response Dimensions	Jaya Singh; Verena Kaynig-Fittkau	Jaya Singh
2	40	MortalMATH: Evaluating the Conflict Between Reasoning Objectives and Emergency Contexts	Etienne Lanzeray; Stephane Meilliez; Malo Ruelle; Damien Sileo	Damien Sileo
3	61	HALF: Harm-Aware LLM Fairness Evaluation Aligned with Deployment	Omar El Herraoui; Ali Mekky; Preslav Nakov; Yuxia Wang	Omar El Herraoui
4	549	KoEVD: A Korean Benchmark for Metadata-Conditioned Multi-Stage LLM Safety Evaluation	San Kim; Wonjung kim; Heesoo Kwon; Jae Hee Yang	San Kim, Heesoo Kwon
5	604	A2M: Trace-Optimized Agent Hijacking in the MCP Ecosystem	Laizhen Li; Xuan Wang; Peicheng Zhao; Juanjuan Zhao; Kejiang Ye; Cheng-zhong Xu; Xitong Gao	Xitong Gao
6	744	Does Fairness Survive Model Adaptation in Legal Judgment Prediction?	Sarah Kamoun; Parth Patel; Faezeh Ensan	Sarah Kamoun

Session 1.5: Generation and Dialogue

Friday, November 6, 2026 · 11:00-12:30 · MR 6

Slot	ID	Title	Authors	Presenter
1	119	Mitigating the Impact of Glitch Tokens via Targeted Retokenization	Muhammad Alif Al Hakim; Dmitrii Kuzmin; Carlos Javier Galán Zavala; Tatsuya Hiraoka; Munachiso Samuel Nwadike; Ayana Niwa; Marco Cognetta; David Pohl; Kentaro Inui	Muhammad Alif Al Hakim
2	160	DIAM: Discourse Interaction and Addressee Modeling for Malaysia Hansard	Goh Kun Shun; Mohan Raj; Lay-Ki Soon; Wern Han Lim; Ian K. T. Tan	Mohan Raj Chanthran
3	241	Read the Room Before Your Model Generates	Muyao Kong; Zhihao Lin	Muyao Kong; Zhihao Lin
4	302	Grounding What Users Care About: Multi-Aspect Personalized Modeling from Implicit Preferences	Jiin Park; Misuk Kim	Jiin Park

5	511	Hallucinate at the Last in Long Response Generation: A Case Study on Long Document Summarization	Joonho Yang; Seunghyun Yoon; Hwan Chang; Byeongjeong Kim; Hwanhee Lee	Joonho Yang
6	700	Transferable Structural Priors in Continuous Diffusion Language Models	Hao Li; Man Fung Zhuo; Xinqi Liu; Jiayu Wu; Yifan Hong	Hao Li

Session 2.1: Resources and Evaluation 2

Friday, November 6, 2026 · 15:30-17:00 · Muhe B+C

Slot	ID	Title	Authors	Presenter
1	170	JudgeMoE: Distributional Aggregation for LLM-as-a-Judge	Yiqi Liu; Joseph James; Yang Wang; Kun Zhao; Chenghao Xiao; Chenghua Lin	Yiqi Liu
2	194	Are Many Tasks Necessary for General LLM Evaluation? A Matrix Completion View of Benchmark Redundancy	Kohei Tsuji; Yusuke Sakai; Hidetaka Kamigaito; Taro Watanabe	Kohei Tsuji
3	245	T2MBench: Evaluating Text-to-Motion Generation under Diverse Language Instructions	Bin Yang; Rong OU; Weisheng Xu; Jiaqi Xiong; Xintao Li; Taowen Wang; Luyu Zhu; Yuetong Fang; Xu Jiang; Jing Tan; Renjing Xu	Bin Yang
4	258	SG-LegalCite: A Principle-Augmented Benchmark for Legal Citation Retrieval in Singapore Law	LEE YUEH ERN SHANNON; Kaidong Feng; Yingpeng Du; Lee En Jia Chloe; Zhu Sun	Shannon Lee Yueh Ern
5	299	SEA-NLI: Language Inference as a Lens into Southeast Asian Cultural Understanding	Peerawat Chomphooyod; Yosephine Susanto; Jian Gang Ngui; Attapol Rutherford; Alham Fikri Aji; Sarana Nutanong; Can Udomcharoenchaikit; Peerat Limkonchotiwat	Peerat Limkonchotiwat, Can Udomcharoenchaikit
6	341	Cost-Sensitive No-Valid-Option MCQA: Analyzing LLM Abstention under Invalid Choices	Jinhyeok Kim; Hye-Young Jung	Jinhyeok Kim

Session 2.2: Multimodality and Language Grounding

Friday, November 6, 2026 · 15:30-17:00 · MR 3

Slot	ID	Title	Authors	Presenter
1	163	Thinking with Deltas: Incentivizing Reinforcement Learning via Differential Visual Reasoning Policy	Shujian Gao; Yuan Wang; Jiangtao Yan; Zuxuan Wu; Yu-Gang Jiang	Shujian Gao
2	267	MMA: Multimodal Memory Agent	Wanru Cheng; Yihao Lu; Zeyu Zhang; Hao Tang	Yihao Lu

3	455	The Model Knows Where to Look: Visual-Grounding-Guided Head Suppression for Mitigating Hallucination in LVLMs	Mong Yuan Sim; Wei Emma Zhang; Xiang Dai; Biaoyan Fang	Mong Yuan Sim
4	583	CGVR: Contradiction-Guided Visual Reasoning for Mitigating Hallucinations in Multimodal Large Language Models	Haoran Ding; Wanqi Zhou; Lei Xing; Lujuan Dang; Kaizhong Zheng; Badong Chen	Haoran Ding
5	668	CircuitPile: A Data Synthesis Framework for Enhancing Multimodal LLMs' Perception on Circuit Schematics	Hongyu Chang; Xinyue Li; Huanyong Liu; Feng Zhang; Yuhui Yin	Hongyu Chang
6	731	The Expense of Seeing: Attaining Trustworthy Multimodal Reasoning Within the Monolithic Paradigm	Karan Goyal	Karan Goyal

Session 2.3: NLP Applications and Human-AI Interaction 2

Friday, November 6, 2026 · 15:30-17:00 · MR 4

Slot	ID	Title	Authors	Presenter
1	596	A Framework for Evaluating AI-Generated Creativity: Human–LLM Agreement in Story Evaluation	Pharath Sathya; Yin Jou Huang; Fei Cheng	Pharath Sathya
2	602	Structuring the Unstructured: A Systematic Review of Text-to-Structure Generation with a Universal Evaluation Framework	Zheye Deng; Chunkit Chan; Tianshi Zheng; Wei Fan; Weiqi Wang; Yangqiu Song	Zheye Deng
3	648	Reasoning-Based Personalized Generation for Sparse Users	Bo Ni; Branislav Kveton; Samyadeep Basu; Subhojyoti Mukherjee; Leyao Wang; Franck Dernoncourt; Sungchul Kim; Seunghyun Yoon; Zichao Wang; Ruiyi Zhang; Puneet Mathur; Jihyung Kil; Jiuxiang Gu; Nedim Lipka; Yu Wang; Ryan A. Rossi; Tyler Derr	—
4	732	Ontology-Guided Knowledge Graph Construction from Structured Metadata and Climate Science Publications	Aayush Acharya; Qi Zhang; Eduard Dragut; Longin Jan Latecki	Aayush Acharya
5	752	Breaking Bureaucracy: Evaluating open-source LLMs for legal document review	Farrukh Baratov; Niki van Stein; Suzan Verberne	Farrukh Baratov
6	799	From Preferences to Scores: Semi-Supervised Multi-Trait Essay Scoring with LLM Pairwise Preferences and Few Human Anchors	Takumi Shibata; Yuichi Miyamura	Takumi Shibata

Session 2.4: Computational Social Science and Cultural Analytics

Friday, November 6, 2026 · 15:30-17:00 · MR 5

Slot	ID	Title	Authors	Presenter
1	35	WikiOriginQA: Automated Construction of a QA Dataset with Knowledge Origins for Analyzing Cultural Bias	Kento Haneda; Yosuke Kishinami; Ryo Fujii; Keisuke Sakaguchi; Makoto Morishita	Kento Haneda
2	56	Fairness in Prompting: How Demographic Variations Affect Language Model Outputs	Ryandito Diandaru; Gleb Kuzmin; Muhammad Dehan Al Kautsar; Zeerak Talat; Xudong Han; Haonan Li; Tatsuki Kuribayashi; Artem Shelmanov; Timothy Baldwin; Fajri Koto	Ryandito Diandaru
3	219	Learning What Is Said, Not How: Stance-Contrastive Learning for Generalizable Hate Speech Detection	yucheng lin; Dan Kienast; Jiaojiao Jiang	yucheng lin
4	364	Automatic Evaluation of Mental Health Stigma in Online Communication	Naomi Baes; Jemima Kang; Nick Haslam; Chris Groot; Alsa Wu; Luc Raszewski; Yulia Otmakhova	Jemima Kang
5	447	Effects of Theory of Mind and Prosocial Beliefs on Steering Human-Aligned Behaviors of LLMs in Ultimatum Games	Neemesh Yadav; Yihuai Lan; Shan Dong; Mai Hieu Hien; Palakorn Achananuparp; Jing Jiang; Ee-Peng Lim	Palakorn Achananuparp
6	491	Is a Picture Worth a Thousand Words? Adaptive Multimodal Fact-Checking with Visual Evidence Necessity	Jaeyoon Jung; Yejun Yoon; Kunwoo Park	Kunwoo Park

Session 2.5: Efficient Methods and Machine Learning for NLP 1

Friday, November 6, 2026 · 15:30-17:00 · MR 6

Slot	ID	Title	Authors	Presenter
1	16	PRC: An Efficient and Unified Framework for Evaluating Robustness and Controllability of Large Language Models on Few-shot Prompt Variations	Chunpeng Ma	Chunpeng Ma
2	96	Supportive Token Revealing for Fast Diffusion Language Model Decoding	Giries Abu Ayoub; Mario Barbara; Lluís Pastor-Pérez; Tanja Bien; Aneesh Barthakur; Alaa Maalouf; Loay Mualem	Giries Abu Ayoub
3	107	Token-Level On-Policy Contrastive Distillation Without Full Rollouts	Jian Yang; Chern Hong Lim; Thuy-Trang Vu	Jian Yang
4	118	EnComp: Lightweight Encoder-Only Context Compression for Retrieval-Augmented Question Answering	Thao Do; Dinh Phu Tran; An Vo; Seon Kwon Kim; Daeyoung Kim	Thao Do

5	237	Less Random, More Incoherent: Greedy Hadamard Alignment for LLM Quantization	Yun Ye; Zhaoqi Wu; Xueshu Wang; Cong Zhang; Zhu Yongbo; Zhonghua Sun; BERYL XU	Yun Ye
6	342	mSFT: Addressing Dataset Mixtures Overfitting Heterogeneously in Multi-task SFT	Woosung Koh; Jeyoung Jeon; Youngjin Song; Yujin Cheon; Soowon Oh; Jaehyeong Choi; Se-Young Yun	Jeyoung Jeon

Session 3.1: Resources and Evaluation 3

Saturday, November 7, 2026 · 10:30-12:00 · Muhe B+C

Slot	ID	Title	Authors	Presenter
1	361	LKValues: Aligning Large Language Models with Sri Lankan Societal Values	Nethmi Muthugala; Supryadi; Surangika Ranathunga; Nisansa de Silva; Ruijie Tao; Ovindu Gunatunga; Pengyun Zhu; Shaowei Zhang; Jingting Zheng; Deyi Xiong	Nethmi Muthugala (M.W.N.N.Muthugala)
2	381	A Reliability Audit of Aggregated MTEB Scores: Coverage Asymmetry and Language-Specific Rank Inversions Across 17 Languages	Archit Rastogi	Archit Rastogi
3	402	DeepVision-103K: A Visually Diverse, Broad-Coverage, and Verifiable Mathematical Dataset for Multimodal Reasoning	Haoxiang Sun; Lizhen Xu; Bing Zhao; Wotao Yin; Wei Wang; Boyu Yang; Rui Wang; HU WEI	Haoxiang Sun
4	416	VERGE: Verification-Enhanced Generation of Multi-Hop Evaluation Datasets for Task-Specific RAG	Kyosuke Morita	Kyosuke Morita
5	421	Auditing PEFT Claims: A Stability Probe Unifying Three LoRA-Variant Fragility Patterns	Suan Lee; JaeSeong Kim	Suan Lee
6	422	Memorization or Extraction? Auditing LLM-Based Recovery of Scientific Assumptions	Suan Lee; JaeSeong Kim	Suan Lee

Session 3.2: LLM Agents and Learning

Saturday, November 7, 2026 · 10:30-12:00 · MR 3

Slot	ID	Title	Authors	Presenter
1	108	A Survey of the Training Process of LLM-Based Agents	Zhisong Zhang; Lemao Liu; Tianqing Fang; Deng Cai; Jiahao Xu	Zhisong Zhang

2	227	Knowing When to Critique: Task-Adaptive Metacognitive Regulation for Reliable LLM Reasoning	Xinmeng Hou; Ziting Chang; Zhouquan Lu; Bohao Qu; Liang Wan; Wei Feng; Hai Hu; Qing Guo	Ziting Chang
3	345	YouRA: A Persistent-State Architecture for Evidence-Traceable Autonomous Research Agents	Woo Yoon Kyu; Woojin Lee; Jin-Xia Huang	Yoonkyu Woo
4	589	JanusPO: Asymmetric Clip and KL for Entropy-Stable Group Policy Optimization in Tool-Augmented LLM Reasoning	Hao Li; Man Fung Zhuo; Guangxin Wu; Yusen Wu; Jiayu Wu	Hao Li
5	591	HorizonBench: Evaluating Ideation Agents on Future Scientific Discoveries	Markus Flicke; Haoyu He; Kashyap Chitta; Yong Cao; Andreas Geiger	Yong Cao
6	637	Diagnosing and Mitigating Compounding Failures in Agentic Persuasion via Taxonomic Strategy Retrieval	Sana Ayromlou; Purvi Sehgal; Pradyumna Narayana	Pradyumna Narayana

Session 3.3: Clinical and Biomedical Applications

Saturday, November 7, 2026 · 10:30-12:00 · MR 4

Slot	ID	Title	Authors	Presenter
1	5	Generating Linguistically Diverse and Semantically Faithful Patient Reports	Alejandra Lorenzo; Adrien Coulet; Claire Gardent	Adrien Coulet
2	139	EvidenceRL: Reinforcing Evidence Consistency for Trustworthy Language Models	J. Ben Tamo; Yuxing Lu; Benoit Louis Marteau; Micky C. Nnamdi; May Dongmei Wang	Yuxing Lu
3	221	Cleaner Speech, Weaker Generalization: Revisiting Pitt-Derived Benchmarks for Alzheimer's Disease Detection	Luqi Sun; Shreeram Suresh Chandra; Lin Zhang; You-Jin Li; Brian MacWhinney; Yu Tsao; Emily Mower Provost; Berrak Sisman	Luqi Sun
4	222	CARE-Bench: Benchmarking Patient-Facing LLM Triage	Yining Hua; Hongbin Na; Cyrus Ayubcha	Hongbin Na
5	667	Lay-Language and Cross-Lingual Disparities in LLM Medical Diagnosis: A Case Study in Rheumatological Skin Conditions	Grayson Hubbell; Yifei Zhang; Saksham Khatwani; Sriram Sankaranarayanan; Indira Sriram; Maria Leonor Pacheco	Grayson Hubbell
6	703	Query-Conditioned Graph Retrieval for Contextualized LLM Reasoning in Personalized Wearable Data	Zhenyu Lu; Mahyar Abbasian; Amir M. Rahmani	Zhenyu Lu

Note (ID 5): Email request: Nov 7 or 8; scheduled Nov 7

Session 3.4: Interpretability and Analysis of Models for NLP 1

Saturday, November 7, 2026 · 10:30-12:00 · MR 5

Slot	ID	Title	Authors	Presenter
1	65	Sparse Autoencoders Can Capture Language-Specific Concepts Across Diverse Languages	Lyzander Marciano Andrylie; Inaya Rahmanisa; Mahardika Krisna Ihsani; Alfan Farizki Wicaksono; Haryo Akbarianto Wibowo; Alham Fikri Aji	Lyzander Marciano Andrylie
2	144	Signatures of Steerability in Activation Space of Language Models	Prajjwal Bhattarai; Tuka Alhanai	Prajjwal Bhattarai
3	259	Decoupling Internal Representational Changes and Causal Importance in Fine-Tuned Large Language Models	Lingfang Li; Procheta Sen; Shubham Das; Danushka Bollegala	Lingfang Li
4	271	Empathy Is Not the Scapegoat for Declines in Factuality and Reliability	Kanwal Mehreen; Letian Shi; Deheng Zhang; Yuxia Wang	Yuxia Wang
5	318	Mechanistic Evidence Routing in Transformers: Margin-Gated Intervention for Retrieval-Conflict Hallucination	Sridurgesh Ravichandran	Sridurgesh Ravichandran
6	417	Encoded but Not Decoded: Layer-Localized Evidence for a Three-Level Gap in LLM Syntax	Zhenyan Lu; He Wang; Xiaohui Huang	Zhenyan Lu

Session 3.5: Machine Translation and Phonology

Saturday, November 7, 2026 · 10:30-12:00 · MR 6

Slot	ID	Title	Authors	Presenter
1	52	Improving Japanese-English Patent Claim Translation via Reasoning and GRPO with Syntactic Reward	Shogo Tsujimoto; Katsuki Chousa; Masaaki Nagata; Ryohei Sasano	Shogo Tsujimoto
2	191	Reference-Grounded Data Curation for Instruction-Following Thai-English Machine Translation	Thodsaporn Chay-intr; Krittapad Harnchang; Mahannop Thabua; Kobkrit Viriyayudhakorn; Thanaruk Theeramunkong	Thodsaporn Chay-intr
3	201	Video-guided Machine Translation for Indic Languages	Pinaki Das; Srikar Ayyagari; Raj Dabre; Thuy-Trang Vu; Virendra Singh; Pushpak Bhattacharyya; Gholamreza Haffari	Pinaki Das
4	207	Lexically Stable but Semantically Improvable: On the Editability of Machine Translation Drafts for LLM Post-Editing	Tianlin Xiao; WenHan Chao	Tianlin Xiao
5	235	ToneCL:Contrastive Learning for Few-Shot Syllable-Level Tone Classification	Qisheng Liao; Youngah Do	Qisheng Liao

6	571	Does Translation Preserve Who You Are? Quantifying Speaker Personality Consistency Across Languages	Xinger Fu; Masaaki Nagata; Chenhui Chu	Xinger Fu

Session 4.1: Resources and Evaluation 4

Sunday, November 8, 2026 · 10:30-12:00 · Muhe B+C

Slot	ID	Title	Authors	Presenter
1	459	PersonaPath: A Benchmark for Knowledge-Centric Personalized Learning Path Planning	Yu Liu; Zeming Liu; Tianle Zhang; Zihao Cheng; Yuhang Guo; Kehai Chen; Min Zhang; Yunhong Wang; Haifeng Wang	Yu Liu
2	461	Multicultural Spyfall: Assessing LLMs through Dynamic Multilingual Social Deduction Game	Haryo Akbarianto Wibowo; Alaa Elsetohy; Qinrong Cui; Alham Fikri Aji	Haryo Akbarianto Wibowo
3	464	Benchmarking Fine-Grained Readability Explanations for Arabic	Nour Rabih; Khalid N. Elmadani; Hanada Taha; Nizar Habash	Nour Rabih
4	468	Dating the Model: Hidden Dates in System Prompts Affect LLM Evaluation	Mario Sanz-Guerrero; Minh Duc Bui; Manuel Mager; Katharina von der Wense	Mario Sanz-Guerrero
5	476	Beyond Isolated Dots: Benchmarking Structured Table Construction as Deep Knowledge Extraction	Tianyun Zhong; Guozhao Mo; Yanjiang Liu; Yihan Chen; Lingdi Kong; Xuanang Chen; Yaojie Lu; Hongyu Lin; Shiwei ye; Shanshan Jiang; Bin Dong; Le Sun; Ben He; Xianpei Han	Tianyun Zhong
6	519	CriticGen: Generation-Aware Evaluation as Actionable Feedback	Huifang Du; Zecheng Zuo; Sen Wang; Haofen Wang; Yehui Yang	Zecheng Zuo

Session 4.2: Language Modeling

Sunday, November 8, 2026 · 10:30-12:00 · MR 3

Slot	ID	Title	Authors	Presenter
1	109	CHASE: Adversarial Red-Blue Teaming for Improving LLM Safety using Reinforcement Learning	Rahul Markasserithodi; Aditya Joshi; Yuekang Li; Ishmanbir Singh; Chris Yoo; Alan Niu	Rahul Markasserithodi
2	243	Bait-and-Recover: Poisoning Internal Refusal Signals to Defend LLMs against White-Box Editing Jailbreaks	Tian Gao; Zhipeng Xie; Yuhao Wu; Junhua Liu; Xin Fang	Zhipeng Xie
3	559	CLEF: Closed-Loop Error Feedback for Multi-Bit Watermarking in Short LLM-Generated Text	Jing Wang; Kejun Qian; Pengyu Xiong; Qiankun Zhang; Yi Yang; Jing Fan	Kejun Qian

4	567	Safety in Batches? Understanding and Mitigating Safety Failures in Batch Prompting	Kihyun Kim; Hee-Seon Kim; Wonjun Lee; Changick Kim	Kihyun Kim
5	657	LLM-as-RNN: A Recurrent Language Model for Memory Updates and Sequence Prediction	Yuxing Lu; J. Ben Tamo; Jinrui Fang; Weichen Zhao; Nan Sun; Xukai Zhao; Yishan Zhong; Wenqi Shi; Jinzhuo Wang; May Dongmei Wang	Yuxing Lu
6	776	See Every Block Once: Variance-Reduced Post-Training for Masked Diffusion Language Models	Yunzhen He; Hiroaki Yamagiwa; Hidetoshi Shimodaira	Yunzhen He

Session 4.3: Efficient Methods and Machine Learning for NLP 2

Sunday, November 8, 2026 · 10:30-12:00 · MR 4

Slot	ID	Title	Authors	Presenter
1	410	S^{4}R: Selective Sampling, Subspaces, and Sparse Reconstruction for Compressed Long-Context KV Caching	Jialong Han; You Wu; Kewei Tu	Jialong Han
2	471	ROCOB: Dual-Manifold Alignment with Optimal Budgeting for Training-Free PTQ Recovery	Haoqi Yang; Teng Xiao; Zuchao Li; Hao Huang; Lefei Zhang; hai zhao	Haoqi Yang
3	497	When Token Supervision Fails: Segment-Level On-Policy Distillation for Long-Horizon Agents	Jialong Liu; Zuchao Li; Yuling Shi; Lefei Zhang; Baoyuan Qi; hai zhao	JIALONG LIU
4	530	Efficient Task Adaptation in Large Language Models: A Survey of Weight-Based, Prompt-Based, and Embedding-Based Adaptations	Jungwon Park; Changin Choi; Jimyeong Kim; Nojun Kwak; Wonjong Rhee	Jimyeong Kim
5	640	Influcoder: Distilling Decoders' Gradient Influence Rankings into an Encoder for Data Attribution	Dimitri Kachler; Damien Sileo; Pascal Denis	Dimitri Kachler
6	781	Not All Tokens Are Equally Important: Auditing and Compressing Redundancy in LLM Reasoning	Nguyen-Khang Le; Dieu-Hien Nguyen; Truong Dinh Do; Le-Minh Nguyen	Dieu-Hien Nguyen

Session 4.4: Semantics and Linguistic Structure

Sunday, November 8, 2026 · 10:30-12:00 · MR 5

Slot	ID	Title	Authors	Presenter
1	7	Dynamics of Meaning: Towards the Evaluation of Diachronic Semantic Change in Sinhala	Nevidu Jayatilleke; Nisansa de Silva	Nevidu Jayatilleke
2	13	Are Linguistic and Arithmetic Circuits Dissociated in LLMs?	Riku Kisako; Tatsuki Kuribayashi; Ryohei Sasano	Riku Kisako
3	384	WiC is Not WSD: A Study on LLMs and Lexical Ambiguity Resolution	Yi Zhou; Kiamehr Rezaee; Danushka Bollegala; Mohammad Taher Pilehvar; Jose Camacho-Collados	Yi Zhou
4	451	Edit Policy-Aligned Text Simplification	Xuanxin Wu; Yuki Arase; Masaaki Nagata	Xuanxin Wu
5	486	You Can Create Words Rather Than Just Detect Them: Towards A New Paradigm of Chinese Neologisms	Qiliang Liang; Hansi Wang; Yue Wang; Yang Liu	Qiliang Liang
6	575	Cross-Lingual Semantic Fusion: Towards Universal Sentence Representations via LLMs	Zhen Hu; Zhu Liu; Ying Liu	Zhen Hu

Session 4.5: Question Answering and Retrieval

Sunday, November 8, 2026 · 10:30-12:00 · MR 6

Slot	ID	Title	Authors	Presenter
1	81	DBRAG: Multi-Table Retrieval-Augmented Generation for Complex Database Queries	Prince Larbi Ampofo; Ryoji Kubo; Djellel Difallah	Djellel Difallah
2	133	When the Correct Law Still Fails: A Diagnostic Benchmark for Law-in-Force Retrieval in Versioned Legal RAG	Yuchen Yao; Yiran HU; 张凯翔; Yixin Su; Jiani Ge; Yun Liu	Yuchen Yao
3	277	CoMMA: Conflict-Aware Read-Time Control for Conversational Memory QA	Zixiao Wang; Xingbo Du; Duzhen Zhang; Le Song	Zixiao Wang
4	399	Training-Free Residual-Guided Retrieval for Multi-hop KGQA with Single-Step LLM Inference	Keito Sei; Sunil Kumar Maurya; Yohei Kobashi; Yutaka Matsuo	Keito Sei
5	516	CRAFT: A Unified Counterfactual Reasoning Framework for Tabular Question Answering and Fact Verification	Chenshuo Pan; Yu Zhao; Jie Zhang; Changzai Pan; Jiayi Liang; Yujie Mao; Zhenhe Wu; Shuangyong Song; Yongxiang Li; Zhongjiang He	Yu Zhao

6	528	STRIDE: Structural Reasoning for In-context Demonstration Extraction in Multi-hop Question Answering	Shiwei Chen; Bin Liang; Hui Wang; Yue Yu; Kam-Fai Wong; Ruifeng Xu	Keyang Ding

Session 5.1: Resources and Evaluation 5

Sunday, November 8, 2026 · 14:00-15:30 · Muhe B+C

Slot	ID	Title	Authors	Presenter
1	531	ArtHopQA: Diagnosing Hidden-Chain Recovery in Artwork Question Answering	Xuefeng Wei; Xuan Zhou; Yuta Kato; Shintaro Ozaki; Kazuki Hayashi; Yusuke Sakai; Hidetaka Kamigaito; Katsuhiko Hayashi; Taro Watanabe	Yusuke Sakai
2	547	K-Prism: A Korean Diagnostic Benchmark for Multi-faceted Hallucination in Vision-Language Models	Min Hyuk Kim; DONG HYOK SHIN; Jimi Lee; Jin Yea Jang	Min Hyuk Kim
3	548	KoMMERCE: A Korean Multimodal and Multitask Evaluation Benchmark for Reliable Conversational E-commerce Agents	Byunggill Joe; San Kim; Minyoung Jung; DONG HYOK SHIN; Yu-Jeong Sung; Saim Shin	Byunggill Joe
4	584	KBS: A Korean Benchmark Suite for Foundation Model Evaluation	Yeonkyoung So; Jongmin Kim; Sungmok Jung; Gyuseong Lee; Sangho Kim; Jongyeon Park; Joonhak Lee; Seho Pyo; Gyeongje Cho; Seorin Kim; Jisoo Kim; Suyoung Park; Hyunji M. Park; Yelim Ahn; Yeongho Seo; Jaejin Lee	Yeonkyoung So
5	730	What Improves During Pretraining? Temporal Stability Reveals Uneven Learning in LLM Benchmarks	Chang Su; Bin Wu; Emine Yilmaz; Yao Lu	Chang Su
6	778	Back to Basics, Powered by Data: Character-Level Transformers for Robust Chinese Machine-Generated Text Detection	Shengjie Xu; Zhenghao Gao; Rujie Dai; Zijing Li; Xiaoyang Xie; Zehua Lyu	Shengjie Xu; Zhenghao Gao

Session 5.2: Multimodal, Medical, and Embodied Models

Sunday, November 8, 2026 · 14:00-15:30 · MR 3

Slot	ID	Title	Authors	Presenter
1	315	Routing-Based Continual Learning for Multimodal Large Language Models	Jay Mohta; Kenan E. Ak; Gwang Gook Lee; Dimitrios Dimitriadis; Yan Xu; Mingwei Shen	Gwang Gook Lee
2	689	Latent Prediction Needs Alignment: A Controlled Study of Joint-Embedding Predictive Vision--Language Learning	Mohammad Kohankhaki; Daniel Kusuma; Shirin Salehi; Carsten Kamp; Sigrid Brell-Cokcan; Anke Schmeink	Mohammad Kohankhaki

3	321	Universal Boosts, Specific Suppressors: Sparse Autoencoder Steering of Medical Vision-Language Models	Farhad Nooralahzadeh; Benjamin Gundersen; Nicolas Deperrois; Hidetoshi Matsuo; Mizuho Nishio; Thomas Frauenfelder; Ahmed Allam; Christian Bluethgen; Michael Moor; Michael Krauthammer	Farhad Nooralahzadeh
4	613	MedHal: A Synthetic Dataset for Medical Hallucination Detection	Fabrice Lamarche; Gaya Mehenni; Neshat Elhami Fard; Odette Rios-Ibacache; Li Ming Wang; John Kildea; Amal Zouaq	Fabrice Lamarche
5	774	Encoded but Not Express: Probing Subtle Speech-Conditioned Emotion in Large Audio Language Models	Jingxi Liang; Pengze Guo; Qifeng Wang; Zhiwen Xie; Yujia Huo; Derek F. Wong	Pengze Guo
6	777	Amortizing Self-Correction for Continual Embodied Adaptation	Woo Kyung Kim; Jooyoung Kim; Minjong Yoo; Honguk Woo	Woo Kyung Kim

Session 5.3: Information Extraction, Retrieval, and Code

Sunday, November 8, 2026 · 14:00-15:30 · MR 4

Slot	ID	Title	Authors	Presenter
1	44	Escaping the Tokenizer Trap: Probabilistic Semantic Routing for Dense Retrieval in Specialized Domains	Siu Hin Ng	Siu Hin Ng
2	586	DisamKB: Direct Construction of Disambiguated Knowledge Bases from Large Language Models	Yujia Hu; Tuan-Phong Nguyen; Simon Razniewski	Yujia Hu
3	608	Target-Conditioned Gating for Many-to-Many Argument Assignment in Document-Level Event Extraction	Junru Liao; Haoran Wang; Xuanjing Huang; Yatian Shen	Haoran Wang
4	696	Do not be greedy, Think Twice: Sampling and Selection for Document-level Information Extraction	Mikel Zubillaga; Oscar Sainz; Oier Lopez de Lacalle; Eneko Agirre	Mikel Zubillaga
5	705	Policy-Conditioned Constrained Decoding for Column-Level Access Control in Text-to-SQL	Miyamoto Ryoto; Xin Fan; Hayato Yamana	Ryoto MIYAMOTO
6	714	Domain-Agnostic Neural Topic Modeling with Contextual Token-Level Semantic Graph Representation	Seung-Won Seo; Won Ik Cho; Yongmin Yoo	Seung-Won Seo

Session 5.4: Multilingualism and Cross-Lingual NLP

Sunday, November 8, 2026 · 14:00-15:30 · MR 5

Slot	ID	Title	Authors	Presenter
1	137	Composing Within Constraints: A Benchmark for Thai Classical Poetry in Large Language Models	Nutchanon Yongsatianchot; Phuvis Kerdpramote	Nutchanon Yongsatianchot
2	296	DuDi: Dual-Signal Distillation with Cross-Lingual Verbalizer	Patomporn Payoungkhamdee; Tinnakit Udsa; Jian Gang Ngui; Sarana Nutanong; Alham Fikri Aji; Peerat Limkonchotiwat	Tinnakit Udsa
3	388	MGSM-Pro: A Simple Strategy for Robust Multilingual Mathematical Reasoning Evaluation	Tianyi Xu; Kosei Uemura; Alfred Malengo Kondoro; Catherine Nana Nyaah Essuman; Tadesse Destaw Belay; Ifeoma Okoh; Ganiyat Afolabi; Ayodele Awokoya; David Ifeoluwa Adelani	Tianyi Xu
4	389	IndicIFEval: A Benchmark for Verifiable Instruction-Following Evaluation in 14 Indic Languages	Thanmay Jayakumar; Mohammed Safi Ur Rahman Khan; Raj Dabre; Ratish Puduppully; Anoop Kunchukuttan	Thanmay Jayakumar
5	554	CulturALL: Benchmarking Multilingual and Multicultural Competence of LLMs on Grounded Tasks	Peiqin Lin; Chenyang Lyu; Wenjiang Luo; Haotian Ye; Md Mehrab Hossain; Chunlan Ma; Shaoxiong Ji; Younes Samih; Bo Zeng; Fan Jiang; Yuanbin Cao; DILDA DUISENBEK; Adrian Neo Sau Xun; POZDNIAKOVA DARIA; MISEVICH LIUBOU; Marinković Nevena; Ngoc Gia Linh Nguyen; THI KHANH LINH DO; SARAKMATAK SOPHY; Baotian Hu; Guanhua Chen; Gongbo Tang; Alham Fikri Aji; Longyue Wang; Weihua Luo	Peiqin Lin
6	695	Toward LLMs Beyond English-Centric Development	Sho Takase; Ukyo Honda	Sho Takase

Session 5.5: Sentiment, Culture, and Model Behavior

Sunday, November 8, 2026 · 14:00-15:30 · MR 6

Slot	ID	Title	Authors	Presenter
1	566	What Makes User Reactions Useful for Fake News Detection?	Yassine Toughrai; David Langlois; Smaili	Yassine Toughrai
2	34	Argument Summarization: A Unified Task Definition, Taxonomy, and Survey	Zhiqiang Shi; Oana Cocarascu	Zhiqiang Shi
3	478	Explainable End-to-End Argument Mining with Toulmin-Structured Explanations	Zhiqiang Shi; Oana Cocarascu	Zhiqiang Shi

4	14	Steering with Multilingual Lexicons for Sentiment Control in Low-Resource Languages	Lucinda Laurent; Muhammad Dehan Al Kautsar; Alfan Farizki Wicaksono; Fajri Koto	Lucinda Laurent

Poster sessions

In-person Posters

Poster Session 1

Friday, November 6, 2026 · 11:00-12:30 · Poster area

Poster Board Number	PaperID	Title	Authors	Presenter
1	3	YUE-PUB: A Cantonese Pragmatic Understanding Benchmark	Chengyan Wu; Ziwei Gong; Jinzhi Chen; Mengmeng Su; Yun Xue; Bolei Ma	Ziwei Gong
2	4	Beyond Bilingual Transfer: Multilingual Code-Switching in Instruction Tuning	Shunta Asano; Jeonghun Baek; Toshihiko Yamasaki	Shunta Asano
3	18	Investigating Learner-Aware Design of LLM-Generated Educational Feedback	Momoka Furuhashi; Kouta Nakayama; Noboru Kawai; Takashi Kodama; Saku Sugawara; Kyosuke Takami	Momoka Furuhashi
4	24	Multi-Signal Interpretability Identifies Task-Critical Circuits for Efficient Model Compression	MANGADODDI SRIKAR VARDHAN	Srikar Vardhan Mangadoddi
5	27	Multilinguality after Quantization: Low-Bits, Calibration Languages, Mixed-Precisions, and Cross-Linguality	Muhammad Rafif Priyo Waskito; Alfan Farizki Wicaksono; Alham Fikri Aji; Haryo Akbarianto Wibowo	Muhammad Rafif Priyo Waskito
6	37	Can LLMs Write Reliable Rubrics? A Meta-Evaluation for Experiment Reproduction	Hanhua Hong; Yizhi LI; Jiaoyan Chen; Luu Gia Huy; Sophia Ananiadou; Jung-jae Kim; Chenghua Lin	Hanhua Hong
7	46	WAON: A Large-Scale Japanese Image-Text Dataset for Cultural Adaptation in Contrastive Vision-Language Models	Issa Sugiura; Shuhei Kurita; Yusuke Oda; Daisuke Kawahara; Yasuo Okabe; Naoaki Okazaki	Issa Sugiura
8	49	Grounded Revision vs. Prior Injection: Probing Retrieval-Augmented Patent Claim Amendment	Josepha Michiko Leo; Hyun-seok Min; Yehoon Jang; Irvan Zidny; Jin-Woo Chung; Sungchul Choi	JOSEPHA MICHIKO LEO
9	62	Thinking Silently for Patent Claim Generation	Lekang Jiang; Wenjun Sun; Stefan Goetz	Lekang Jiang
10	66	Useful Diversity or Hallucinated Variety? Persona-Conditioned Synthetic Query Generation for E-Commerce Search	Filippos Ventirozos; Tom Armitage; Mahmoud Oshagh; Peter A. Appleby; Matthew Shardlow	Filippos Ventirozos

11	70	Improving Multimodal Contrastive Learning of Sentence Embeddings with Object-Phrase Alignment	Kaiyan Zhao; Zhongtao Miao; Yoshimasa Tsuruoka	Kaiyan Zhao
12	80	MuSe-MI: A Framework for Assessing Multi-Session Motivational Interviewing with AI-Generated Patients	Topaz Freizeit; Moreah Zisquit; Kai Golan Hashiloni; Chaitze Tavens; Doron Friedman; Kfir Bar	Topaz Freizeit
13	82	Do LLMs Understand Context? A Knowledge Graph-Based Evaluation Framework	pragatheeswaran vipulanandan; Subavarshana Arumugam; Mamta Nallaretnam; Chamath Gunapala; Kithuni Wickramasinghe; Kamal Premaratne; Uthayasanker Thayasivam	Mamta N; Subavarshana A; Kithuni W; Chamath G
14	97	Mechanistically Understanding and Mitigating the Anchoring Effect in LLMs via Attention Head Intervention	Yoshiki Takenami; Yin Jou Huang; Yugo Murawaki; Koh Takeuchi; Chenhui Chu	Yoshiki Takenami
15	100	Measuring Iterative Temporal Reasoning with Time Puzzles	Zhengxiang Wang; Zeyu Dong	Zhengxiang Wang
16	101	Anthropogenic Regional Adaptation in Multimodal Vision-Language Model	Samuel Cahyawijaya; Peerat Limkonchotiwat; Tack Hwa Wong; Carlos Rafael Catalan; Manuel Antonio Rufino; Hitesh Laxmichand Patel; Amit Agarwal; Muhammad Reza Qorib; Vicky Feliren; Holy Lovenia; Aye Hninn Khine; Frederikus Hudi; David Anugraha; Alham Fikri Aji; Romrawin Chumpu; Viet Thanh Pham; Minghan Wang; Mohamed Fazli Mohamed Imam; Ruochen Zhang; Joseph Marvin Imperial; Khumaisa Nur'aini; Do Xuan Long; Musa Izzanardi Wijanarko; Joel Ruben Antony Moniz; Patrick Amadeus Irawan; Hanif Muhammad Zhafran; Isaiah Edri W. Flores; Salsabila Zahirah Pranida; Jun Kevin; Jostin Jerico Rosal; Patricia Nicole Monderin; Kun Kerdthaisong; Ahmad Mustafid; My Chiffon Nguyen; Natchapon Jongwiriyanurak; Siva Worajitwannakul; Haochen Li; Adrian Xuan Wei Lim; Bin Wang; Muhammad Ravi Shulthan Habibi; Lynnette Hui Xian Ng; Mithil Bangera; Yeshil Bangera; Priyaranjan Pattnayak; Dun Li Chan; Sherissa Caren Djuniwar; Cho Chan Myei Oo; Ming Shan Hee	Peerat Limkonchotiwat

Poster Session 2

Friday, November 6, 2026 · 15:30-17:00 · Poster area

Poster Board Number	PaperID	Title	Authors	Presenter
1	104	RabakBench: Evaluating Safety Guardrails in Singapore's Multilingual Vernaculars	Gabriel Chua; Leanne Tan; Ziyu Ge; Roy Ka-Wei Lee	Leanne Tan
2	111	Two-Stage LLM-as-a-Judge Using Required and Bonus Criteria	Kohei Tsunefuji; Akihiro Tamura; Daichi Nakatsuji; Hiroaki Oobayashi; Mitsuo Hamada; Tsuneo Kato	Kohei Tsunefuji
3	125	VMMU: A Vietnamese Multitask Multimodal Understanding and Reasoning Benchmark	Vy Tuong Dang; An Vo; Emilio Villa-Cueva; Quang Tau; Duc Dm; Thamar Solorio; Daeyoung Kim	Vy Tuong Dang
4	127	Mirrored Frames: LLM-Based Narrative Framing Analysis of Thai-Cambodian Border Conflict Coverage	Siwakorn Bubphasawan; Attapol Rutherford	Attapol Rutherford
5	130	Hidden Consensus: Preference-Validity Compression in Human Feedback	Dorcas Chia Ern Chua; Karen Myn Hui Lee; Jia Yue Tan; Zhen Xue Gue; Han Ying Lim; Norzalena Abdul Hamid; Azima Binti Azmi; Keat Mei Yeong; Aizat Izyani binti Mujab; Hafsah Noor Azam; Chee Guo Khoo; Chee Seng Chan	Karen Myn Hui Lee
6	132	Replacing the Human in the Loop: Towards Automated Evaluation of LLM Mathematical Proof Generalization	Carolina Dias-Alexiou; Edison Marrese-Taylor; A. Solano; Yutaka Matsuo	Carolina Dias-Alexiou
7	135	STEERLSN: Training-Free Representation Alignment through Language-Specific Neuron Subspace Interpolation	Minjun Kim; Dongyeon Kim; Hoyun Song; Jihyo Kim; KyungTae Lim	Kyungtae Lim
8	146	Synthesizing Dialogues from Social Media Data for Mental Health Support: A Case Study on Gambling Disorders	Zhouqing Zhang; Lis Pereira; Wan Jou She; Benjamin Adams; Tham Piumsomboon; Kongmeng Liew	Zhouqing Zhang
9	147	Beyond SFT: Reinforcement Learning for Safer Large Reasoning Models with Better Reasoning Ability	Jinghan Jia; Nathalie Baracaldo; Sijia Liu	Jinghan Jia
10	152	Towards the Roof of the World: NLP Progress in Tibetan Languages	Xi Cao; Yijie Li; Yuan Sun; Shiwei Dong; Zeguang Huang; Nuo Qun	Xi Cao
11	168	When Tool-Backed Skill Retrieval Fails: Source-Style Collapse in Executable Capability Retrieval	Yiqi Liu; Joseph James; Yang Wang; Chenghao Xiao; Chenghua Lin	Yiqi Liu
12	169	Feedback as an Interface: Controlling Form and Detail in LLM Translation Refinement	Yiqi Liu; Flavjo Tomori; Nafise Sadat Moosavi; Steffen Eger; Chenghua Lin	Yiqi Liu

13	173	Greedy Matching Aggregation for Soft Contrastive Learning in Multi-Label Text Classification	Ryoma Kobayashi; Jiyi Li	Ryoma Kobayashi
14	182	Dependency-Aware Trajectory Refinement for Efficient Multi-Turn Agent Fine-Tuning	Zhuo Chen; Zhen Zhang; Xinyu Wang; Kewei Tu	Zhuo Chen
15	183	ReThink: Adversarially Robust Arithmetic Reasoning via Neuro-Symbolic Soft Constraints	Soikat Das; Sparsh Mittal; Rekha Singhal	Soikat Das
16	193	Measuring the Impact of Big Five Personality Traits on Suicide Severity Level Detection with Large Language Models	Kit May Shum; Yusuke Sakai; Masahiro Kaneko; Taro Watanabe	Kit May Shum

Poster Session 3

Saturday, November 7, 2026 · 10:30-12:00 · Poster area

Poster Board Number	PaperID	Title	Authors	Presenter
1	202	Plan–Write–Refine: An Iterative Framework for Argumentative Essay Generation	Zhijun Zhang; Qianlong Wang; Jianzhu Bao; Keyang Ding; Genan Dai; Bowen Zhang; Ruifeng Xu; Yongsheng Liang	Zhijun Zhang
2	206	Adversarial Prompt Distillation: Efficient Jailbreak Attacks from Large to Small Language Models	Xiang Li; Chong Zhang; Jia Wang; Fangyu Wu; Ben Li; Xiaobo Jin	Xiang Li
3	209	RESCUE-BENCH: Towards Relation-Aware Multi-Party Emotional Support Conversation Systems	Haichuan Hu; Jiashuo Wang; Yang Xiao; Mingni Tang; Jiawen Duan; Quanjun Zhang; Congqing He; Hao Zhang; Johan F. Hoorn; Wenjie Li	Haichuan Hu
4	216	GuideWeb: A Benchmark for Automatic In-App Guide Generation on Real-World Web UIs	Chengguang Gan; Yunhao Liang; Yoshihiro Tsujii; Tatsunori Mori; Shiwen Ni; Hiroki Itoh	GAN CHENGGUANG
5	220	Beyond Accuracy: ARIA-Rubrics for Evaluating Audio Reasoning in Large Audio Language Models	Yupei Li; Qiyang Sun; Mohamed Mady; Chenxi Wang; Zhengwei Gong; Berrak Sisman; Björn Schuller	Mohamed Mady
6	228	Large Language and Vision-Language Models as World Models: A Survey	Azmine Toushik Wasi; Sadia Tasnim Meem; Riashat Islam; Md Rizwan Parvez	Azmine Toushik Wasi
7	230	Is There a Dr. House in the Machine? A Survey of Clinical and Diagnostic Reasoning in LLM-based Medical Agents	Azmine Toushik Wasi; Abdur Rahman; Riashat Islam; Md Rizwan Parvez	Azmine Toushik Wasi
8	233	When Choices Become Risks: Safety Failures of Large Language Models under Multiple-Choice Constraints	Yuheng Chen; Zhiyu Wu; Bowen Cheng; Yihang Wu; Tetsuro Takahashi	Yuheng Chen

9	248	RegMix-D: Dynamic Data Mixing via Proxy Training Trajectories	Kaiyan Zhao; Zhongtao Miao; Akiko Aizawa; Yoshimasa Tsuruoka	Kaiyan Zhao
10	251	CCTU: A Benchmark for Tool Use under Complex Constraints	Junjie Ye; Guoqiang Zhang; Wenjie Fu; Zelin Li; Tao Gui; Qi Zhang; Xuanjing Huang	Junjie Ye
11	253	LOTUS: Language-vision Optimal Transport for Unified report Synthesis	Xuhui Li; Jiachen Liu	Jiachen Liu
12	256	From Lexical to Entity-Level: Predicting Individual Knowledge Boundaries via Knowledge Graph Label Propagation	Shuyi Yu; Yusuke Sakai; Hidetaka Kamigaito; Taro Watanabe	Shuyi Yu
13	279	BHAID: A Dataset of LLM-Generated Text in Indian Languages with a Dual Contrastive Learning Framework for Detection	Shifali Agrahari; SUJIT KUMAR; Ranbir Singh Sanasam	Shifali Agrahari
14	280	LecTrans: A Multimodal Translation Benchmark for Academic Lectures	Frederikus Hudi; Nina Nanakorn; Justin Vasselli; Yusuke Sakai; Hidetaka Kamigaito; Akihiro Tamura; Taro Watanabe	Frederikus Hudi
15	311	DunbaaBERT: From Sacrifice to Semantics	Iffat Maab; Waleed Jamil; Raphael Schmitt	Iffat Maab
16	332	TileBench: A Controlled Benchmark for Performance Evaluation and Bottleneck Diagnosis of Tile-Based Programming Models	Bowen Cui; Zhongchun Zhou; Hao Wu; Tejas Ramesh; Junyu Yin; Jialiang Gu; Keren Zhou	Zhongchun Zhou

Poster Session 4

Saturday, November 7, 2026 · 13:30-15:00 · Poster area

Poster Board Number	Poster ID	Title	Authors	Presenter
1	354	HakushoBench: A Japanese Chart and Table VQA Benchmark from Governmental White Papers	Issa Sugiura; Shuhei Kurita; Yusuke Oda; Naoaki Okazaki	Issa Sugiura
2	360	Scale-Aligned Reinforcement Learning for Adaptive Computation Allocation in Reasoning Language Models	Yilin Liu; Xuan Wang; Zheng Yang; Zhuoran Li; FAN WAN; Wu Chen; Xiaojie Wang	Yilin Liu
3	369	BAEmpEval: A Framework for Probing Large Audio Language Models on Multidimensional Empathetic Response Assessment	Jingxi Liang; Pengze Guo; Qifeng Wang; Zhiwen Xie; Yujia Huo; Derek F. Wong	Pengze Guo
4	370	SQLStructEval: Structural Evaluation of LLM Text-to-SQL Generation	Yixi Zhou; Fan Zhang; Zhiqiao Guo; YU CHEN; Haipeng Zhang; Preslav Nakov; Zhuohan Xie	Yixi Zhou

5	379	Style-Aware Translation for Multi-Word Expressions in Low-Resource Languages	Antonios Dimakis; Antonios Anastasopoulos	Antonios Dimakis
6	387	ReleaseEval: A Benchmark for Evaluating Language Models in Automated Release Note Generation	Qianru Meng; Zhaochun Ren; Joost Visser	Qianru Meng
7	390	EXAM^2: $\underline{Ex}tending$ $\underline{A}udio$ $Understanding$ in $\underline{M}ultilingual$ and $\underline{M}ultimodal$ $Analysis$	Jiawen Wang; Xiaoxue Gao; Zi Haur Pang; Nancy F. Chen	Jiawen Wang
8	397	The Image Reconstruction Game: Drawing Common Ground Through Iterative Multimodal Dialogue	Sherzod Hakimov; Mattia D'Agostini; Ivan Samodelkin; David Schlangen	David Schlangen
9	407	Activation Residual Hessian Quantization for Low-Bit Large Language Model Inference	Yifeng wang; Zhun Sun; Hakaze Cho; Linlin Wang; Keisuke Sakaguchi	Yifeng Wang
10	423	When Quality Filters Backfire: Agreement-Based Latent Filtering Hurts Bootstrap Pretraining	Suan Lee; JaeSeong Kim	Jaeseong Kim
11	438	An Empirical Study of Many-Shot In-Context Learning for Machine Translation of Low-Resource Languages	Yinhan Lu; Gaganpreet Jhajj; Chen Zhang; Anietie Andy; David Ifeoluwa Adelani	Yinhan Lu
12	448	When more supervision hurts: reward design for structured biomedical reasoning in 4B models	Batuhan Nursal; Christophe Ye; Xinyu Guo; Xin He; Cassie S. Mitchell	Batuhan Nursal
13	452	CULTURETALK-ID: A Multi-Task Dialogue Benchmark for Cultural Commonsense in Indonesian Local Languages	Muhammad Dehan Al Kautsar; Salsabila Zahirah Pranida; Bilal Elbouardi; Fajri Koto	Muhammad Dehan Al Kautsar
14	458	Pretraining Builds Cooperation, Not Components: Tracking Inter-Layer Coordination via Lie Brackets	JaeSeong Kim; Suan Lee	Jaeseong Kim
15	466	Corrective Retrieval Augmented Generation	Shi-Qi Yan; Jia-Chen Gu; Yun Zhu; Zhen-Hua Ling	Shi-Qi Yan
16	492	Copy Less, Ground More: Selective and Confidence-Aware Decoding for Grounded Legal Generation	Tan-Minh Nguyen; Truong Dinh Do; Hiep K. V. Nguyen; Le-Minh Nguyen	Tan-Minh Nguyen
17	601	GLM-RAG: Graph Language Models for Graph-Based Retrieval-Augmented Generation	Maya Arseven; Anette Frank; Beni Egressy; Johann Higl; Moritz Plenz	Moritz Plenz

Poster Session 5

Sunday, November 8, 2026 · 10:30-12:00 · Poster area

Poster Board Number	Poster ID	Title	Authors	Presenter
1	533	OMIT the Action: Measuring Framing-Invariant Omission Bias under Philosophical Disagreement	Sihyeon Lee; Jihun Song; Chanwoo Kim; Jiwoo Kum; Chanjun Park	Sihyeon Lee

2	534	Structured but Silent: Probing Capability Requirements in LLM Hidden States	Kyojun Choo; Minsoo Song; Yunju Kang; Chanjun Park	Kyojun Choo
3	539	CASCADE: Cognitive Assessment via Sentence Context and Distortion Extraction	Archit Narumanchi; Azamat Rashidov; Jong C. Park	Archit Narumanchi
4	543	Entailed Opinion Matters: Improving the Fact-Checking Performance of Language Models by Relying on their Entailment Ability	Gaurav Kumar; Ayush Garg; Debajyoti Mazumder; Aditya Kishore; Babu kumar; Jasabanta Patro	Gaurav Kumar
5	546	Mass-Tail SHAP Entropy: A Regime-Conditional Decoder for Vision-Language Model Hallucination	Jaegwang Shin; Suan Lee	Jaegwang Shin
6	553	IndexRAG: Bridging Facts for Cross-Document Reasoning at Index Time	Zhenghua Bao; Yi Shi	Zhenghua Bao
7	556	GPC-E4: Generalized Consistency and Entropic Evidence for Backdoor Detection in Vision Language Models	Xuan Wang; Xuehu Yan; Yu-liang Lu; Kejiang Ye; Cheng-zhong Xu; Xitong Gao	Xitong Gao
8	570	Efficient Many-Shot In-Context Learning with Task-Level Compression and Instance-Level Retrieval	Masayuki Kawarada; Ukyo Honda; Sho Takase	Masayuki Kawarada
9	578	Polar: A Benchmark for Evaluating Political Bias in LLMs	Heejin Kim; Sangho Kim; Hyunggeun Jeon; Yoonhee Park; Jaejin Lee	Sangho Kim
10	580	Reducing Logprob Without Causal Verification: Diagnosing SAE-Based PII Suppression in LLMs	CHAERIN LEE; Wonho Lee; Sohee Park; Daeseon Choi	CHAERIN LEE
11	594	Encoded but Not Routed: Explaining the Table-Chart Gap in Scientific Claim Verification	Sunisth Kumar; Xanh Ho; Tim Schopf; Andre Greiner-Petter; Florian Boudin; Akiko Aizawa	Sunisth Kumar
12	603	Psychometric Benchmark and Bias Assessment in Open-Source Pre-trained NLI Models	Fadi Amun; Ofir Cohen; Aviad Elyashar; Odeya Cohen; Rami Puzis	Rami Puzis
13	606	Not All Graphs Are Maps: A Task-Agnostic Exploration Framework for Mobile GUI Navigation	Adamo Orsini; Seyed Mehdi Ayyoubzadeh; Sadman Rakib Pinon; Hannuo Zhang; Koosha Gholipour Baradari; YUANHAO YU; Xinxin Zuo; Yang Wang	Adamo Orsini
14	611	Skin-Deep: A Geometric Diagnostic for Alignment Fragility in Large Language Model Representations	Dongyub Jude Lee; Jungseob Lee; Seungyoon Lee; Seongtae Hong; Suhyune Son; Sugyeong Eo; Jaehyung Seo; Heuiseok Lim	Jungseob Lee
15	618	DiagBench: A Medical-Text-Adjudicated Benchmark for Staged Diagnosis-Revision Calibration in Chinese Inpatient EHRs	Huawei Feng; Dongsheng Shi; Linlin Wang	Huawei Feng
16	624	Molecule Generation through Reasoning with Large Language Models	Congfeng Cao; Di Wu; Zhi Zhang; JIA-HONG HUANG; Jelke Bloem; Khalil Sima'an	Jelke Bloem
17	635	FreakOut-LLM: The Effect of Emotional Stimuli on Safety Alignment	Daniel Kuznetsov; Ofir Cohen; Karin Shistik; Asaf Shabtai; Rami Puzis	Rami Puzis

Poster Session 6

Sunday, November 8, 2026 · 14:00-15:30 · Poster area

Poster Board Number	PosterID	Title	Authors	Presenter
1	642	Behaviorally Grounded User Profiles from the Wild for Personalized Alignment and Multi-Perspective Reasoning	Yuxuan Li; Victor Zhong; Ehsan Kamalloo	Yuxuan Li
2	676	From Chains to Graphs: Self-Structured Reasoning for General-Domain LLMs	Yingjian Chen; Haoran Liu; Yinhong Liu; Sherry T. Tong; Aosong Feng; Jinghui Lu; JTzhang; Yusuke Iwasawa; Yutaka Matsuo; Irene Li	Yingjian Chen
3	677	CORA: A Protocol for Diagnosing Boundary Robustness in Text-to-Audio Retrieval under Query Reformulations	WOO JAE MIN; Kyongmin Kong; Bogyung Jeong; Minjeong kim; HaeJun Yoo; Du-Seong Chang	Jae Min Woo
4	678	It Hears, It Sees too: Multi-Modal LLM for Depression Detection By Integrating Visual Understanding into Audio Language Models	Xiangyu Zhao; Yaling Shen; Yiwen Jiang; Zimu Wang; Jiahe Liu; Maxmartwell H Cheng; Guilherme C Oliveira; Robert Desimone; Dominic Dwyer; Zongyuan Ge	Xiangyu Zhao
5	683	UniToR: Towards Universal Token-level Routing for Efficient Reasoning in Large Language Models	Nguyen-Khang Le; Truong Dinh Do; Le-Minh Nguyen	Nguyen-Khang Le
6	711	No Masking Needed: Predicting Latent Semantics for Efficient Pre-training of Encoder-based Language Models	Gábor Berend	Gábor Berend
7	715	POLAR: Ontology-Guided Risk Prevention for Tool-Calling LLM Agents	Yunju Kang; Seonghyeon Cho; Irene Li; Yeo Chan Yoon; Chanjun Park	Seonghyeon Cho
8	751	Cultural Awareness is Represented but Not Decoded: Tracing Mythological Knowledge across 18 Open-Source LLMs	Iaroslav Chelombitko; Mika Hämäläinen	Iaroslav Chelombitko
9	757	Superficial Reflection or Genuine Thought? A Fine-Grained Cognitive Analysis of Large Reasoning Models	Yuxiang Chen; Zuohan Wu; Ziwei Wang; Xiangning Yu; Xujia Li; Linyi Yang; Mengyue Yang; Jun Wang; Lei Chen	Zuohan Wu
10	759	Beyond Diagnosis: Evaluating Recommendation Stability under Counterfactual Skin-Tone Perturbations	Hyunwoo Yoo; Kyusik Kim; Gail Rosen	Hyunwoo Yoo
11	760	Revisiting Co-Evolution for Multi-Turn Safety Alignment: Limitations and Alignment Tax Mitigation	Yoshiaki Sata; Kouta Nakayama	Kouta Nakayama
12	764	Optimizing Reasoning LLMs for Automated Feature Engineering via Multi-Stage Direct Preference Optimization	Mingzhi Hu; Menghai Pan; Yuzhong Chen; Junpeng Wang; Yingtong Dou; Vineeth Rakesh; Hoang Anh Duy Le; Uday Singh	Yuzhong Chen

			Saini; Xin Dai; Xiran Fan; Yuqing Liu; Yiwei Cai	
13	786	Efficient Multimodal Planning Agent for Visual Question-Answering	Zhuo Chen; Xinyu Geng; Xinyu Wang; Yong Jiang; Zhen Zhang; Pengjun Xie; Kewei Tu	Zhuo Chen
14	788	YaPO: Learnable Sparse Activation Steering Vectors for Domain Adaptation	Abdelaziz BOUNHAR; Rania Elbadry; Hadi Abdine; Preslav Nakov; Michalis Vazirgiannis; Guokan Shang	Guokan Shang
15	790	Reading Between the Layers: Cross-Modal Alignment and Representational Hierarchy in Vision Transformers	Sugam Karki; Raúl Vázquez; Yova Kementchedjhieva	Sugam Karki
16	796	SeEM: Selective Dictionary-based Translation on Large Language Models	Hongyuan Lu; Zixuan Li; Wai Lam	Hongyuan Lu
17	797	Beyond Consensus: Diagnosing Coverage, Majority Bias, and Grounding Equity in High-Stakes Multi-Document Diversity Summarization	Jane Arleth dela Cruz; Iris Hendrickx; Martha Larson	Jane Arleth dela Cruz

Virtual Posters

ID	Title	Authors	Presenter
1	Towards Probabilistic Question Answering Over Tabular Data	Chen Shen; Estevam Hruschka	Estevam Hruschka
2	An Empirical Study of On-Device Translation for Real-Time Live-Stream Chat on Mobile Devices	Jeiyoon Park; Leedaehwan; ChangminYeo; Yongshin Han; Minseop Kim	Jeiyoon Park
6	Relational Priors as Convergence Pressure in LLM-Based Multi-Agent Systems	Ming Shen; Chao Shang; Sadat Shahriar; Devang Kulshreshtha; Yi Zhang; Sandesh Swamy; Yanjun Qi	Ming Shen
8	ImmunoFOMO: Are Language Models missing what oncologists see?	Aman Sinha; Bogdan-Valentin Popescu; Xavier Coubez; Marianne Clausel; Mathieu Constant	Aman Sinha
9	ECQL: An Executable Benchmark for Auditable Corpus Query Semantic Parsing	Yuxi Zheng; Yeqing Liu; Ziming Zhou; Xiaodong Wang; Yongwei Zhang; Jingsong Yu	Yuxi Zheng
10	ChartTableViVQA: A Vietnamese Benchmark Dataset for Visual Question Answering with Reasoning Instruction over Charts and Tables Images	Quang Nhat Truong; Pham Quoc Thang; Son T. Luu	Quang Nhat Truong
11	Navigating the Reality Gap: Privacy-Preserving On-Device Continual Adaptation of ASR for Clinical Telephony	Darshil Chauhan; Vansh Patel; Adityasinh Solanki; Kanav Kapoor; Dhruv Kumar; Pratik Narang; Ritvik Jain; Aditya Bansal	1. Darshil Chauhan; 2. Vansh Patel

12	SPECI: Fine-grained Evaluation of Specificity for Image Captioning	Son T. Luu; Hiep K. V. Nguyen; Trung Bui; Viet Dac Lai; Le-Minh Nguyen	Son Thanh Luu
17	JUMPSTARTER: Human-AI Planning with Task-Structured Context Curation	Xuanming Zhang; Sitong Wang; Jenny Ma; Alyssa Hwang; Zhou Yu; Lydia Chilton	Xuanming Zhang
20	Extended-SVD Low-Rank Adaptation: PLORA for Parameter-Efficient Fine-Tuning	RATAN KUMAR GULIPILLI	___
21	HybridScholar: Personalized Paper Recommendation with Cross Attention-Gated Fusion and Verified Explanation Generation	Joytun Nessa Tonny; Sajidur Rahman; Saddam Mukta; Swakkhar Shatabda; Salekul Islam	Joytun Nessa Tonny
22	Human Preference Guided Evaluation of RAG-based QA Systems	Shakila Mahjabin Tonni; Xiang Dai; Sarvnaz Karimi; Xiaomei Li; Samuel C. Andrew; Stephen Wan	Shakila Mahjabin Tonni
23	Beyond Factual Accuracy: A Benchmark for Evaluating Perspective Alignment in LLMs	Sunghoon Kim; Sangyeop Kim	Sunghoon Kim
26	How Robust Is Multimodal Claim Verification to LLM Rewriting?	Yun-Ang Wu; Xanh Ho; Andre Greiner-Petter; Sunisth Kumar; Tian Cheng Xia; Florian Boudin; Akiko Aizawa	Yun-Ang Wu
31	When Does Knowledge Distillation Hurt? Reliability-Aware Distillation for Low-Resource Language Summarization	Dipto Sumit; Ankan Kumar Roy Srizon; Atia Haque Asha; Mourchona Afrin; Sadia Khair Rodela; Niloy Farhan; Farig Sadeque	Dipto Sumit
32	A Systematic Analysis of Linguistic Features in AI-Generated English Text Detection Across Domains and Models	Yassir El Attar; Esra Dönmez; Maximilian Maurer; Agnieszka Falenska	Yassir El Attar
33	Generate, Verify, and Refine: Event Timelines for Robust Video-Grounded Dialogue Systems	Wiradee Imrattanatrai; Masaki Asada; Kimihiro Hasegawa; Susan Holm; Ken Fukuda; Teruko Mitamura	Masaki Asada
38	Are Language Models Script-Aware?	David Kletz; Sandra Mitrovic; Ljiljana Dolamic; Fabio Rinaldi	David Kletz
39	Evaluating Contextually Mediated Factual Recall in Multilingual Large Language Models	Yihong Liu; Bingyu Xiong; Hinrich Schuetze	Yihong Liu
41	PAL-UI: Planning with Active Look-back for Vision-Based GUI Agents	Zikang Liu; Junyi Li; Xin Zhao; Dawei Gao; Yaliang Li; Ji-Rong Wen	Zikang Liu
53	PLAUSIEVE:A Distribution Based Framework for Measuring Model Belief Beyond Logits	Arghodeep Nandi; Tanmoy Chakraborty	ARGHODEEP NANDI
54	Attention-Guided Layer Selection for Contrastive Decoding in Large Language Models	Yusuke Sakai; Natthawut Kertkeidkachorn; Kiyoaki Shirai	Natthawut Kertkeidkachorn
57	Episodic Memory from Compression Boundaries in Latent Representation Space	David Oneil Campos Ferreira; Priscila Rocha Maia Freitas Ribeiro; EMANUEL BORGES PASSINATO; Diogo Fernandes Costa Silva; Arlindo Rodrigues Galvão Filho	David O'Neil Campos Ferreira

59	Protocode: Prototype-Driven Interpretability for Code Generation in LLMs	bodla Krishna Vamshi; Haizhao Yang	Bodla Krishna Vamshi
60	North or South? Diagnosing LLMs' Understanding of North Korean Language, Society, and Knowledge	Kiwoong Park; Seyoung Song; Donghyun Woo; Alice Oh	Seyoung Song
63	Judging the judge: Can LLMs track attitude-holders in predicates of personal taste?	Yan Cong	Yan Cong
67	Think When Unsure: Leveraging Model Confidence to Decide When to Use Chain-of-Thought	Samuel Lewis-Lim; Xingwei Tan; Zhixue Zhao; Nikolaos Aletras	Sam Lewis-Lim
69	Ask Twice: Bidirectional Cross-Encoder Scoring for Faithfulness Evaluation	Arnav Mantro; Leslie Barrett	Leslie Barrett
71	Multi-Perspective LLM Annotations for Valid Analyses in Subjective Tasks	Navya Mehrotra; Adam Visokay; Kristina Gligorić	Navya Mehrotra
72	CIL: Causal-Intervention and LLM-Synergized Multimodal Fake News Detection	HONGJIE WANG; Hongyu Chen; Peilin Jiang	Hongyu Chen
73	Agent+P: Guiding UI Agents via Symbolic Planning	Shang Ma; Xusheng Xiao; Yanfang Ye	Shang Ma
74	Does Continued Pretraining on a Learner Corpus Improve Automated Essay Scoring on English Proficiency Tests? Evidence from EFCAMDAT	Duy Anh Nguyen	Duy Anh Nguyen
75	Twenty-Five Years of Inter-Annotator Agreement in NLP: A Systematic Survey of Trends, Metrics, and Reporting Practices	Hend Al-Khalifa	Hend Al-Khalifa
79	Are You Paying Too Many Tokens? Cost-Efficient Multimodal LLM Annotation with Simple Heuristics	Zhixi Zhu; Kristina Gligorić	Zhixi Zhu
83	SelfGrader: LLM Jailbreak Detection via Anchored Token-Level Logits	Zikai Zhang; Rui Hu; Olivera Kotevska; Jiahao Xu	Zikai Zhang
84	Self-Routing RAG: Binding Selective Retrieval with Knowledge Verbalization	Di Wu; Jia-Chen Gu; Kai-Wei Chang; Nanyun Peng	Di Wu
85	Towards Low-Resource Manchu Text Generation: Corpus Construction and Multi-Granularity Augmentation	Meiqi Wang; Zhiyong Xu; Xiaoxin Sun; Ruixin Yu; Zhiyong Guo; Zifan Wang; Shuqi You; Siyu Liu; Xiayu Qin; Suhua Wang	Ruixin Yu
86	Breaking the Silence: A Dataset and Benchmark for Bangla Text-to-Gloss Translation	Sharif Mohammad Abdullah; Abhijit Paul; Shubhashis Roy Dipta; Zarif Masud; Shebuti Rayana; Ahmedul Kabir	Abhijit Paul
87	BusMA: A Bus Communication Substrate for Multi-Agent Systems	Yanwen Peng; Delvin Ce Zhang; Xi Wang; Nikolaos Aletras	Yanwen Peng
88	DOSA: A Tree-Guided, Self-Regressive Framework for Long Document Structure Analysis	Bohou Li; Benjamin Sowell; Mehul A Shah; Mark Lindblad; Henry Lindeman	Bohou Li

92	PeMA-KGQA: A Multi-Agent LLM Framework for Persian Knowledge Graph Question Answering	Kimia Hosseini; Mohammad Hosseini; Saeedeh Momtazi	Kimia Hosseini
93	CODE-PROMPT: Evaluating Code-LLMs for their NLP Abilities via Code Aligned Prompts	Yuxuan Chen; Philippe Thomas; Robert Oehring; Felix Krenzien; Sebastian Möller; Roland Roller	Yuxuan Chen
95	Evaluating Multi-Dimensional Generalization of Large Language Models in Temporal Extraction Tasks	FAHMID SHAHRIAR IQBAL; Ritam Dutt; Soumitra Das; Arnav Verma; Sagnik Ray Choudhury	FAHMID SHAHRIAR IQBAL
98	Beyond Top Words: Interpretable Topic Modeling with Monosemantic Features from Sparse Autoencoders	Una Joh; Bei Yu	Una Joh
103	Right for the Wrong Reasons: Investigating the Unrobust Nature of High Accuracy Models in Discourse Relation Prediction	Yi Fan; Wei Liu	Yi Fan
105	Understanding Knowledge Distillation in Post-Training: When It Helps and When It Fails	Xin Liu; Simin Ma; Shujian Liu; Song Wang; Sathish Reddy Indurthi; Haoyun Deng; Lu Wang; Kaiqiang Song	Xin Liu
110	Iterative Critique-Refine Framework for Enhancing LLM Personalization	Durga Prasad Maram; Dhruvin Gandhi; Zonghai Yao; Gayathri Akkinapalli; Franck Dernoncourt; Yu Wang; Ryan A. Rossi; Nesreen K. Ahmed	Durga Prasad Maram
114	Stylistic Evolution and LLM Neutrality in Singlish Language	Linus Tze En Foo; Weihan Angela Ng; Wenkai Li; Lynnette Hui Xian Ng	Tze En Linus Foo
117	Linguistic Cues in Political Debates: Lexical Choices and Evaluative Framing in GPT-Generated Commentary	Yu-Yu Chang; Chung-Chi Chen; Ming-Hung Wang	Yu-Yu Chang
120	\textit{Learn To Say NO!} Incorporating Model Abstention In Medical Vision-Language Models for Pharmacovigilance	Sofia Jamil; Sooraj Veer R; Chinmay Purushottam Bhat; Sriparna Saha	Sofia Jamil
121	Reasoning under Vision: Understanding Visual-Spatial Cognition in Vision-Language Models for CAPTCHA	Python Song; Luke Tenyi Chang; Yun-Yun Tsai; Penghui Li; Junfeng Yang	Python Song
123	CLFEC: A New Task for Unified Linguistic and Factual Error Correction in paragraph-level Chinese Professional Writing	Jian Kai; Zidong Zhang; Jiwen Chen; Zhengxiang Wu; Songtao Sun; Fuyang Li; Yang Cao; Qiang Liu	Jian Kai
128	PRADA: Fostering Culturally Adaptive Dialogue Agents via Personality	Lingzhi Shen; Xiaohao Cai; Yunfei Long; Imran Razzak; Shoaib Jameel	Lingzhi Shen
129	PrivacyEvalClaw: Towards Deliberative Tool-Augmented Agent for LLM Privacy Jailbreak Evaluation	Krishna Kanth Nakka; Yujun Wang; Shijie Zhou; Tu Nguyen; Gabriele Gelardi; Xuebing Zhou	Yujun Wang
131	EffComp: Efficient Prompt Compression via Hybrid Reinforcement & Supervised Learning	Romtham Tungsuntham; Piyawat Lertvittayakumjorn; Atiwong Suchato; Peerapon Vateekul	Romtham Tungsuntham
141	BreakFun: Jailbreaking LLMs via Object Instantiation under Simulated Code Execution	Amirkia Rafiei Oskooei; Mehmet Aktas	Amirkia Rafiei Oskooei

143	Beyond the Login: Continuous Behavioral Authorship Verification for LLM Agent Sessions	Muskan Sharma; Seba Susan	Muskan Sharma
145	AgentArch: A Benchmark for Evaluating Agent Architectures in Enterprise Workflows	Tara Bogavelli; Hari Subramani; Roshnee Sharma	Tara Bogavelli
151	CROP: Token-Efficient Reasoning in Large Language Models via Regularized Prompt Optimization	Deep Shah; Sanket Badhe; Nehal Kathrotia; Priyanka Tiwari	Deep Shah
155	Don’t Forget the Last Trade: Integrating Episodic Recall into Memory-Driven Agents for High-Frequency Trading	Nuoyue Xu; Xiang Zhang; Jiang Liu; Juntai Cao; Wenxuan Huang; Jiaqi Wei	Nuoyue Xu
156	ProcrastiCoT: A Dual-Encoder Cross-Attention Framework for Multi-Dimensional Procrastination Motives Assessment	Chia-Ni Chiang; Chia-Yun Lee; Hsin-Hsuan Chen; Jheng-Long Wu	Chia-Ni Chiang
157	J-ToRL: Tool-Integrated LLM-as-a-Judge via Reinforcement Learning	Bryan Wilie; Tiezheng YU; Gerry Windiarto Mohamad Dunda; Jierun Chen; Chaofan Tao; Yan Xu; Junxian He; Pascale Fung	Bryan Wilie
162	Comparative Personalization for Multi-document Summarization	Haoyuan Li; Snigdha Chaturvedi	Haoyuan Li
165	Position: Real-World Clinical NLP Robustness Requires Messy, Multimodal, Longitudinal, Privacy-Preserving Corpora	Azmine Toushik Wasi; Shahriyar Zaman Ridoy	Azmine Toushik Wasi
172	HPO: Hysteretic Policy Optimization for Stable and Efficient Training under Sparse-Reward Regime	Mohamed SANA; Nicola Piovesan; Antonio De Domenico; Fadhel Ayed; Haozhe Zhang	___
174	IBPS: Indian Bail Prediction System	Uddeshya Raj; Puspesh Kumar Srivastava; Praveen Patel; Shubham Kumar Nigam; Noel Shallum; Arnab Bhattacharya	Uddeshya Raj; Praveen Patel; Puspesh Srivastava
176	Method Ranking Inverts Across Targets: A Matched-Grid Evaluation of Jailbreak Attacks on Open-Weight Vision-Language Models	Venkat Aniruddh Thamatam; Vasudev Majhi; Aaryan Nagda; Jagat Sesh Challa; Murari Mandal; Yash Sinha; Dhruv Kumar	Venkat Aniruddh Thamatam
177	The Language of Bargaining: Linguistic Effects in LLM Negotiations	Stuti Sinha; Himanshu Kumar; Aryan Raju Mandapati; Rakshit Sakhuja; Dhruv Kumar	Stuti Sinha
179	When Reject Turns into Accept: Quantifying the Vulnerability of LLM-Based Scientific Reviewers to Indirect Prompt Injection	Devanshu Sahoo; Manish Prasad; Vasudev Majhi; Jahnvi Singh; Vinay Chamola; Yash Sinha; Murari Mandal; Dhruv Kumar	Devanshu Sahoo, Manish Prasad
181	Exploring Agentic Multimodal Large Language Models: A Survey for AIScientists	Jinglin Jian; Yi R. Fung; Denghui Zhang; Yiqian Liang; Qingyu Chen; Zhiyong Lu; Qingyun Wang	Jinglin Jian
184	Enhancing Dynamic Text-Attributed Graph Modeling via Instruction Tuning	Weili Shi; Haoming Li; Sheng Li	Weili Shi
187	The Social Cost of Intelligence: Emergence, Propagation, and Amplification of Stereotypical Bias in Multi-Agent Systems	Thi-Nhung Nguyen; Linhao Luo; Amardeep Kaur; Rollin Omari; Tamas Abraham; Junae Kim; Thuy-Trang Vu; Dinh Phung	Thuy-Trang Vu

190	How Good Are LLM Agents? Benchmarking Opinion Dynamics Against Indonesian Electoral Discourse on Social Media	Musa Izzanardi Wijanarko; Derry Tanti Wijaya	Musa Izzanardi Wijanarko
192	From Early Encoding to Late Suppression: Interpreting LLMs on Character Counting Tasks	Ayan Datta; Mounika Marreddy; Alexander Mehler; Zhixue Zhao; Radhika Mamidi	Ayan Datta
195	SignSim: A Landmark Driven Model for Phonological Sign Similarity Prediction	Abhishek Bharadwaj Varanasi; Manjira Sinha; Tirthankar Dasgupta; Dr. Charudatta Jadhav	Abhishek Bharadwaj Varanasi
196	BioCAT5: Concept Mention-Aware Text-to-Text Transfer Transformer for Biomedical Literature	Mohammad Golam Sohrab	Mohammad Golam Sohrab
199	Investigating LLM Capabilities on Long Context Comprehension for Clinical Question Answering	Feras AlMannaa; Talia Tseriotou; Jenny Chim; Maria Liakata	Feras AlMannaa
212	Bahnar–Vietnamese Cross-Lingual Sentence Retrieval: A Low-Resource Case Study with Lexical, Neural, and Hybrid Methods	Thi Ty Nguyen; Long S. T. Nguyen; Phat T. Tran-Truong; Tan Sang Nguyen; Tho Quan	Thi Ty Nguyen
217	Audio Editing in the Era of Foundation Models: A Survey	Changhao Pan; YiFei Fan; Fan Zhuo; Yifu Chen; Wenxiang Guo; Yu Zhang; Zhiyuan Zhu; Rui Yang; Shengpeng Ji; Chenyuhao Wen; Jiayang Xu; Ke Lei; Xiaoda Yang; Jingyu Lu; Zhou Zhao	Changhao Pan
218	Atomic and Holistic LLM Judges for Reference-Grounded Support Labels: A Prompt-Controlled Comparison	Xinran Zhang	Xinran Zhang
223	OpenQTS: A Metadata-Rich Open Dataset for Tang Dynasty Poetry	Robin Young	Robin Young
224	Geo-spatial and Geo-temporal Reasoning in Vision-Language and Large Language Models : A Review	Shahriyar Zaman Ridoy; Azmine Toushik Wasi; Koushik Ahamed Tonmoy; Md Rizwan Parvez	Mohammed Eunus Ali
225	Is Brevity the Soul of Wit? An Empirical Test Using 570,000 Reddit Jokes	Robin Young	Robin Young
229	Can LLMs Threaten Human Survival? Benchmarking Potential Existential Threats from LLMs via Prefix Completion	Yu Cui; Yifei Liu; Hang Fu; Sicheng Pan; Haibin Zhang; Cong Zuo; Licheng Wang	Yu Cui
232	Spatial Reasoning in Multimodal Foundation Models: A Survey of Spatial Understanding in LLMs and VLMs	Azmine Toushik Wasi; Md. Masudur Rahman; Abdur Rahman; Mohsin Mahmud Topu; Mohammed Eunus Ali; Riashat Islam; Md Rizwan Parvez	Mohammed Eunus Ali
234	Agent-Environment Alignment via Automated Interface Generation	Kaiming Liu; Xuanyu Lei; Ziyue Wang; Peng Li; Yang Liu	Kaiming Liu
238	DAMG-RAG: Document-Aware Multi-Granularity RAG for Visually Rich Document Understanding	Jiahao Gu; Ningjing Liu; Yuanzhuo Ding; Siheng Chen	Jiahao Gu

239	AstroReason-Bench: Evaluating Agentic Planning on Space Mission Planning Tasks	Weiyi Wang; Xinchi Chen; Jingjing Gong; Xuanjing Huang; Xipeng Qiu	Weiyi Wang
246	Recoverability Is Not Use: Readout-Local Audits for Concept Erasure	Gagan Aryan	Gagan Aryan
249	State-Anchored Complete-View Distillation for Robust Conversational Multimodal Emotion Recognition	Zhaoyan Pan; Xiangdong Li; Wenke Wu; Mengting Ma; Ye Lou; Ji Zhou; Jiatong Pan; Wei Zhang	Zhaoyan Pan
252	Can LLMs Follow the Pulse of a Crisis? Evaluating Crisis Sentiment in Bangladesh's July Uprising	Md. Samiul Alim; Mahir Shahriar Tamim; Tanvir Ahmed Khan; Sharjil khan; Rafia Ferdous Duti; Shahriyar Zaman Ridoy; Mohammad Ali Moni	Mahir Shahriar Tamim
255	Hierarchy Aware Preference Optimization for the Safety of Korean Small Language Models	Soo Yong Kim; Junyoung Koh; Kyeonghun Kim; Seunghyeok Hong	Soo Yong Kim
257	CCD-CBT: Multi-Agent Therapeutic Interaction for CBT Guided by Cognitive Conceptualization Diagram	Chang Liu; Changsheng Ma; Yongfeng Tao; Bin Hu; Minqiang Yang	Chang Liu
262	Cantus: Toward Pedagogical Feedback Generation for Sight-Singing Learning	Jiale Chen; Xuelian Dong; Yubing Chen; Yongxin Hu; Wenxiu Xie; Tianyong Hao	Jiale Chen
264	ChartHal: A Fine-grained Framework Evaluating Hallucination of Large Vision Language Models in Chart Understanding	Xingqi Wang; Yiming Cui; Xin Yao; Shijin Wang; Guoping Hu; Xiaoyu Qin	Xingqi Wang
265	Do LLMs Respect Concept Boundaries? A Bilingual Diagnostic Benchmark	Renjun Lai; Yixuan Si	Yixuan Si
266	CoSelector: Multi-Agent Collaboration for Multilingual Retrieval-Augmented Generation	Long Yuan; Rui Qi; Xue Zhang; Jinan Xu; Kaiyu Huang	Long Yuan
274	FACTA: Structured Fact Alignment for Multi-Hop Retrieval-Augmented Generation	Jiahao Li; Yongxiu Xu; XUMENG ZHANG; Hongbo Xu; Minghao Tang; Yubin Wang; Gaopeng Gou	Jiahao Li
275	AGOPNs: Null-Space Activation Steering for Jailbreak Defense via Average Gradient Outer Product Concept Extraction	Minh-Nguyen Bui; Binh-Phuong Phan; Thanh-Khang Doan; Bao Huynh-Gia; Phung-Anh Nguyen	Minh-Nguyen Bui
281	Predicting Emerging Topics from Outliers: A Prospective Study of Weak Signals in Embedding Space	Evangelia Zve; Gauvain Bourgne; Jean-Gabriel Ganascia	Evangelia Zve
282	LogicLure: Toward Evidence-Grounded Safe Decision-Making under Hidden Risks	Kyungho Kim; JunCheol Shin	Kyungho Kim
283	Initialisation Determines the Basin: Efficient Codebook Optimisation for Extreme LLM Quantization	Ian W Kennedy; Nafise Sadat Moosavi	Ian W. Kennedy
285	PhaseLoRA: Role-Specialised Low-Rank Adaptation through Prompt-Completion Phase Splitting	Rishi Ravikumar; Riza Batista-Navarro	Rishi Ravikumar
286	Edit Robustness Under Fine-Tuning in Text-to-Image Models	Feng He; Marco Valentino; Zhixue Zhao	Feng He

287	Neuron-Level Emotion Control in Speech-Generative Large Audio-Language Models	Xiutian Zhao; Ismail Rasim Ulgen; Philipp Koehn; Björn Schuller; Berrak Sisman	Xiutian Zhao
288	FrenchNews-7: Benchmarking Cross-Publisher French News Editorial Desk Classification	Amr Sobhy	Amr Sobhy
290	Explore-on-Graph: Hybrid Embedding–LLM Reasoning for Knowledge Graph Question Answering under Incompleteness	Ola El Khatib; Djellel Difallah	Ola El Khatib
291	SweSpecQA: A Swedish Benchmark for Specialized Question Answering	Murathan Kurfali	MURATHAN KURFALI
293	Iterative Topic Taxonomy Induction with LLMs: A Case Study of Electoral Advertising	Alexander Brady; Tunazzina Islam	Tunazzina Islam
294	TokenMapper: A Step Toward Interoperable Speech Token Translation	Tal Kozakov; Tal Rosenwein; Eliya Nachmani	Tal Kozakov
295	Modeling Triplet Interactions Beyond Pairwise Attention	Mafizur Rahman; Lijun Qian	Mafizur Rahman
297	A Multimodal Text- and Graph-Based Approach for Open-Domain Event Extraction from Documents	Praval Sharma	Praval Sharma
300	Progress Drift: Auditing False Progress in Long-Horizon LLM Agents	Xiangbo Zhang; Xiaoxu Ma	Xiangbo Zhang
303	Scalable Methods for Long-Range Argumentation in Distributed Discourse	Alina Littek; John Lawrence	Alina Littek
304	STEVE: Stabilizing Textual Gradient-Based Prompt Optimization via Error-Driven Refinement and Regularized Verification	Yifan Xu; Yixuan Li; Xinzhuo Li; Yixin Gu; Yifan Shen; Lijun Yu; Haohan Wang	Yifan Xu
305	Incorporating Abstract Meaning Representation and Uniform Meaning Representation into In-Context Learning Strategies for Machine Translation	Thu Hoang; Ellie Yang; Weixin Lin; Shira Wein	Thu Hoang
309	Set-Theoretic Compositionality of Sentence Embeddings	Yash mahajan; Naman Bansal; Faria Binte Kader; Santu Karmaker	Yash Mahajan
313	Are Reasoning Vision-Language Models Robust to Semantic Visual Distractions?	Yizheng Sun; Mochuan Zhan; Yanan Ma; Jia Tong See; Yifan Wang; Ziyi Wang; Hao Li; Yang Cui; Wenhao Cai; Jingyu Sun; Chenghua Lin; Riza Batista-Navarro; Jingyuan Sun	Yizheng Sun
314	CLUE: A Contrastive Learning Framework for Unsupervised Event Type Induction and Annotation	Pratik Maitra; Lauren Kwon; Qi Li	Pratik Maitra
316	The Validity–Fidelity Gap: A Controlled Study of Representation Drift in Structured Extraction Under Redundant Contexts	Gokul Srinath Seetha Ram; Anirudh Vasudevan; Rashmi Elavazhagan	Anirudh Vasudevan
320	Prompting Underestimates LLM Capability for Time Series Classification	Dan Schumacher; Erfan Nourbakhsh; Rocky Slavin; Anthony Rios	Daniel Schumacher

323	Retrieval-Induced Mirage: When Similar Cases Become Fabricated Evidence in Medical VLMs	Xiaoxu Ma; Xiangbo Zhang	Xiaoxu Ma
324	Language Models Should be Used to Surface the Unwritten Code of Science and Society	Honglin Bao; Siyang Wu; Jiwoong Choi; Yingrong Mao; James Evans	Honglin Bao
325	Absolute Error Is Not Enough: Ranking Model Performance Without Labels to Estimate Disparities at Scale	Veronica Rammouz; Aaron O. Gonzalez; Carlos Cruzportillo; Adrian Tan; Nicole Lang Beebe; Ke Yang; Anthony Rios	Veronica Rammouz
326	Modeling Iterative Text Revisions Through Chain of Revisions	Zhexiong Liu; Diane Litman	Zhexiong Liu
327	Persona Non Grata: LLM Persona-Driven Generations in MCQA are Unstable in Distinct Dimensions	César Guerra-Solano; Xiang Lorraine Li	César Guerra-Solano
328	Large Language Models Approach Expert Pedagogical Quality in Math Tutoring but Differ in Instructional and Linguistic Profiles	Ramatu Oiza Abdulsalam; Segun Aroyehun	Segun Aroyehun
330	Text-Preserving Lossy Text Compression: A Study of Strategic Deletion and LLM Reconstruction	Yuchun Zou; Junhong Tong; Jun Li	Yuchun Zou
331	INMS: Memory Sharing for Large Language Model based Agents	Hang Gao; Yongfeng Zhang	Hang Gao
333	ATOD: An Evaluation Framework and Benchmark for Agentic Task-Oriented Dialogue Systems	Yifei Zhang; Hooshang Nayyeri; Rinat Khaziev; Emine Yilmaz; Gokhan Tur; Dilek Hakkani-Tür; Hari P Thadakamalla	Yifei Zhang
335	MF-QAT: Multi-Format Quantization-Aware Training for Elastic Inference	Zifei Xu; Sayeh Sharify; Hesham Mostafa	Zifei Xu
337	Embarrassingly Simple Access Control for Large Language Models	Veronica Rammouz; Đorđe Klisura; RAM KRISHNAN; Anthony Rios	Veronica Rammouz
339	A Unified Ambiguity Signal for Calibrated Multimodal Emotion Recognition	Ahmad Tahir; Raza Mukhtar; Chanjun Park; Yeo Chan Yoon	Tahir Ahmad
340	Bi-Phase Training: Learning Efficiently From Scratch	Abel Gurung; Joseph Campbell	Abel Gurung
343	NEXUS: Structured Runtime Safety for Tool-Using LLM Agent	Elias Hossain; Md Mehedi Hasan Nipu; Tasfia Nuzhat Ornee; Rajib Rana; Niloofar Yousefi	Elias Hossain
347	TRIAGE: Severity-Ranked Multi-Agent Failure Attribution	Xizhi Wang	Xizhi Wang
348	Tracing the Genesis of Logic: Causal Data Attribution of Mechanistic Features via Counterfactual Poisoning	Xiaokun Wang; Siyu Song; Hongbo Zhao; Wentao Liu; Jiaqi Jiang; Qingqing Mao	Xiaokun Wang
349	Three Numbers and a Precondition: Auditing Multi-Source LLM Benchmarks for Source-Identity Leakage	Tian Xia; Sarp Aykent	Tian Xia
350	SFMP: Fine-Grained, Hardware-Friendly and Search-Free Mixed-Precision Quantization for Large Language Models	Xin Nie; Haicheng Zhang; Liang Dong; Beining Feng; Jinhong Weng; Guiling Sun	Xin Nie

351	PatchHolmes: Agentic Patch Retrieval via Listwise Selection	Guanqun Yang; Yingming Zhou; Jiangrui Zheng; Shudong Hao; Xueqing Liu	Jiangrui Zheng
352	Markhor: A framework for Adapting LLMs to Urdu via Alignment-Free Tokenizer Replacement	Muhammad Zain Ali; Yuxia Wang; Muhammad Arslan Manzoor; Tony C Smith; Bernhard Pfahringer	Muhammad Zain Ali
355	YFPO: Yoked Feature Preference Optimization with Neuron-Guided Rewards	YifanLe	Yifan Le
356	Disentangling Perception and Reasoning in Multimodal LLMs via Reward Design	Omar Sharif; Eftekhar Hossain; Nikhil Singh; Patrick Ng	Omar Sharif
357	Schema-Key Wording as an Instruction Channel in Structured Generation under Constrained Decoding	YifanLe	Yifan Le
358	SkillSeek: Revisiting Agent Skill Retrieval at Marketplace Scale	Guanqun Yang; Wenlong Zhang; TIAN SHI; Ping Wang	Wenlong Zhang
359	Anchored Turns: A Length-Driven Gradient in Chain-of-Thought Causal Necessity Across Multi-Turn Conversations	DHRUV GUPTA; Vasudev Majhi; Jagat Sesh Challa; Murari Mandal; Dhruv Kumar	Dhruv Gupta
362	UniGuardian: A Unified Defense for Detecting Prompt Injection, Backdoor Attacks and Adversarial Attacks in Large Language Models	Huawei Lin; Yingjie Lao; Tony Geng; Tan Yu; Weijie Zhao	Huawei Lin
363	EcoStatKG: A Domain-Specific Statistical Knowledge Graph and Benchmark for Evaluating Numerical Hallucination in LLMs	Ratnesh Kumar Joshi; Sagnik Sengupta; Asif Ekbal	Ratnesh Kumar Joshi
365	Domain Knowledge Annealing: A Structured Local-to-Global Paradigm for Constructing Expert LLMs	Yikang Zhang; Zuchao Li; Ping Wang; Lefei Zhang; hai zhao; Bo Du	Yikang Zhang
366	The Geometry of Uncertainty in Language Model Representations	Zy Li	Zy Li
367	Language Models Detokenize, But Only for Themselves	Qingyun Li; Ruoyun Li	Qingyun Li
368	ProgMem: Programmatic Belief-State Memory for Long-Horizon Conversational Agents	Pengchao Zheng; Xinan wang; Ziyi Wei; Haoran Liu; Lei Wei; Jiajun Niu; Bohan Guo; Han Wu	Pengchao Zheng
373	NüshuVoice: Reviving the Voice of Endangered Nüshu with Pitch-Aware Text-to-Speech	Hongkun Yang; YI Xinhui; Zhao Xiyan; YIBO MENG; Lionel Z. WANG; Lixu Wang; Yaqi Zhang; Ruiqi Chen; Xuanyue Zhao; Lanxin Zhang; Yu Zeng; CHU WEIJIA; Yiming Ma; Chenyu Liu; Jianghao Lin; Xu Xin	Lionel Z. Wang
374	GenState-AI: State-Aware Dataset for Text-to-Video Retrieval on AI-Generated Videos	Minghan Li; Tongna Chen; Tianrui Lv; Yishuai Zhang; Suchao An; Guodong Zhou	Tongna Chen
375	PAIRSAFE: Proxy-Resistant Paired Safety Evaluation for Refusal Boundaries	Wenzhang Du	Wenzhang Du

376	Semantic Coverage: Measuring LLM Uncertainty Through Coverage in Latent Semantic Space	Gabriele Carrino; Brown Ebouky; Katja Hose; Mattia Rigotti	Gabriele Carrino
377	KaiToM: A Benchmark for Culturally Informed Theory-of-Mind in Thai Context	Kun Kerdthaisong; Pasin Buakhaw; Wayupuk Sommuang; Nutchanon Yongsatianchot	Kun Kerdthaisong
378	Selective Reflection for Context-Sensitive Emotion Recognition in Conversation	Xinyu Wang; Weijie Feng; Hongchuang Wang; Zhiyong Cheng	Xinyu Wang
382	Diagnosing Prompt Sensitivity and Interaction Instability in LLM-Based Humor Judgment	Dohyeon Kim; Jihoon Yang	Dohyeon Kim
385	MM-NeuroOnco: A Multimodal Benchmark and Instruction Dataset for MRI-Based Brain Tumor Diagnosis	Feng Guo; Jiaxiang Liu; Yang Li; Qianqian Shi; Mingkun Xu	Feng Guo
393	Student-Centered AI Requires Auditable Evidence of Learning	Xu Aofeng; Sichao Li	Aofeng Xu
394	Rethinking Cross-Modal Attention for Visual Token Pruning in Multimodal LLMs	Diego Weng Qiu; Lin Shang	Diego Weng Qiu
398	Corrective Feedback-Driven Humor Generation for Tiny Language Models	Ramakrishna Pinninti; Lalit Chandra Routhu; Dibyanayan Bandyopadhyay; Gaël Dias; Alexis Lechervy; Jeremie Pantin; Mohammed Hasanuzzaman; Asif Ekbal	Ramakrishna Pinninti
400	PINGALA: Prosody-Aware Decoding for Sanskrit Poetry Generation	Manoj Balaji Jagadeeshan; Atul Singh; Nallani Chakravartula Sahith; Amrith Krishna; Pawan Goyal	Manoj Balaji Jagadeeshan
401	REMIND: Enhancing CoT Reliability with Self-Supervised Introspection	Souvik Maji; Dibyanayan Bandyopadhyay; Asif Ekbal	Dibyanayan Bandyopadhyay
404	SMARTS: A Small-to-large Model Coordinated Retrieval and Response Framework for Task-Oriented Dialogue	Jiale Chen; Xuelian Dong; Shengcheng Ye; Wenxiu Xie; Kun Zeng; Tianyong Hao	Jiale Chen
405	SurgAgents: A Multi-agent Framework for Embodied Cataract Surgical Intent Prediction	Shengyuan Wang; Yanzhe Ji; Yue Yao; Zhenyue Qin; Yang Liu	Yanzhe Ji
406	GROWTHEVAL: Evaluating Stage-Conditioned Developmental Realism in Role-Playing LLMs	Danping Yan	Danping Yan
409	Beyond Correctness: A Framework for Analyzing Reasoning and Faithfulness of Multi-hop Question Answering	Jiyuan Liu; Yi Xie; Yanghui Rao; Fu Lee Wang; Jianxing Yu; Li Qing	Yi Xie
411	SEAL-PR: Scientific Evidence-Audited and Locally Isolated Peer Review under Untrusted Manuscripts	Zida Yang; Guquan You; Weiqing Zeng; Yu Zheng	Zida Yang
412	QUILT: Query–User Interaction Linguistic Taxonomy	Yuyang Dai	Yuyang Dai
413	You Don’t Understand: Benchmarking Experiential Comprehension Gaps in LLMs	Yuyang Dai	YUYANG DAI
414	Prompt Rewrites Are Semantic Actions: Testing Whether Language Models Obey Typed Operator Laws	Zida Yang; Weiqing Zeng; Guquan You; Yu Zheng	Zida Yang

425	A Survey of Neurosymbolic Approaches to Propaganda and Persuasion Technique Detection	Alysa Ziying Tan; Daniel Kainovan Handoyo	Alysa Ziying Tan
426	FairJudge: Abstention-Aware Multimodal Judges for Fairness and Alignment Evaluation in Text-to-Image Models	Zahraa Al Sahili; Maimuna Nowaz; Maryam Fetanat; Ioannis Patras; Matthew Purver	Zahraa Al Sahili
427	ILRR: Inference-Time Steering Method for Masked Diffusion Language Models	Eden Avrahami; Eliya Nachmani	Eden Avrahami
428	How Far Do Auto-Interpretation Labels Generalize: A Controlled Study Across Languages, Scripts, and Rewordings	Sripad Karne	Sripad Karne
429	The Missing "I Don't Know": Why Three Reasoning-Reliability Findings Converge on Calibrated Abstention	Srijith Ravikumar	Srijith Ravikumar
433	WK-Drift: Benchmarking Procedural Memory Maintenance under World Knowledge Drift	Pratyay Banerjee; Masud Moshtaghi; Ankit Chadha	Pratyay Banerjee
434	LLM-Ideoplasticity: Measuring Ideological Plasticity in the Political Behavior of LLMs as a Context-Conditioned Distribution	Adib Sakhawat; Syed Rifat Raiyan; Tahsin Islam; Takia Farhin; Hasan Mahmud; Md Kamrul Hasan	Adib Sakhawat
435	ProMoral-Bench: Comparing Prompting Strategies for Moral Judgment and Jailbreak Robustness in LLMs	Rohan Subramanian Thomas; Shikhar Shiromani; Abdullah Chaudhry; Ruizhe Li; Vasu Sharma; Kevin Zhu	Rohan Subramanian Thomas
437	DeScribe: Discourse-Driven Extraction and Structuring of Claim Verification for Better Faithfulness Evaluation	Yang Zhong; Yang Janet Liu; Diane Litman	Yang Zhong
439	Probing for Negation: Extracting Compositional Negation Signals from Pretrained Language Representations for Image Retrieval	Animesh Singh; Brejesh Lall	Animesh Singh
440	Enhancing Safety Alignment by Universal Adversarial Prompts	Yuan Zhou; Xiangyu Zhang	Yuan Zhou
442	RL-MR: Reinforcement Learning based Budget-Aware Sequential Pretrained Model Ranking	Saddam Mukta; Md. Ahashan Habib; Mir Tafseer Nayeem; Md Adnanul Islam	Md. Ahashan Habib
444	Bridging the Long-Tail Gap: Robust Retrieval-Augmented Relation Completion via Multi-Stage Paraphrase Infusion	Fahmida Alam; Mihai Surdeanu; Ellen Riloff	Fahmida Alam
446	DKCD: Domain Knowledge-Enhanced Causal Discovery from Unstructured Data	Xin Li; Jin Li; Shoujin Wang; Kun Yu; Fang Chen	Xin Li
449	Modeling Discourse Profiles in Climate Misinformation Discourse	Ge Gao; Zhengyang Shan; Bryan A. Plummer; Margrit Betke; Derry Tanti Wijaya	Ge Gao
450	From Relation-Specific Heads to Shared Interventions: Contextual Entrainment in Language Models	Jinghan Zheng; Jiajun Shi	Jiajun Shi
456	Evaluating Automated Research: A Survey of Protocols, Evidence, and Validity for AI/NLP/ML Research Workflows	Liulei Zhang; Dejing Zhou; Chuyue Huang; Guanhua Chen; Yutong Yao; Lidia S. Chao; Chi Man VONG; Derek F. Wong	Liulei Zhang

457	Prompt-Based Abstention Fails Under Misleading Context: A Controlled Study of Small Frozen RAG Models	Yohanes Andre Setiawan	Yohanes Andre Setiawan
460	CogMem: A Cognitive Memory Architecture for Sustained Multi-Turn Reasoning in Large Language Models	Yiran Zhang; Jincheng Hu; Mark Dras; Usman Naseem	Yiran Zhang
462	Counting Clues: A Lightweight Probabilistic Baseline Can Match an LLM	Furong Jia; Yuan Pu; Finn Guo; Monica Agrawal	Furong Jia
463	Plan Validity Lives in the Grammar: An Induced Grammar Mask for LLM Planning	Nijesh Upreti	Nijesh Upreti
467	AI-Sustainability Discourse Across Social Media: Sentiment, Topics, and Engagement on LinkedIn, X, and Bluesky	Kholoud Khalil Aldous; Md. Rafiul Biswas; Mabrouka Bessghaier; Shimaa Ibrahim; Kais Attia; Wajdi Zaghouani	Md. Rafiul Biswas
470	Multi-Task Instruction Tuning via Data Scheduling for Low-Resource Arabic SpeechLLMs	Hunzalah Hassan Bhatti; Firoj Alam; Shammur Absar Chowdhury	Firoj Alam, Hunzalah Hassan Bhatti
472	Synergizing Large Language Models and Theory for Human-like Causal Reasoning	Yanxi Zhang; Xin Cong; Zhong Zhang; Xiao Liu; Dongyan Zhao; Yesai Wu	Yanxi Zhang
473	WhisperX-LLM-TermFixer: Enhancing Whisper with LLM-Guided Post-Decoding for Domain-Specific Terminology	Ronghao Pan; José Antonio García-Díaz; Rafael Valencia-García	Ronghao Pan
482	WebWorld: Web-Simulated World Model	Ting Huang; Zeyu Zhang; Hao Tang	Ting Huang
483	DramaAgent: Agentic Storytelling Video Generation	Ting Huang; Biao Wu; Ronghao Chen; Zeyu Zhang; Tengfei Cheng; Qizhen Lan; Huacan Wang; Hao Tang	Ting Huang
484	BlockBench: Contamination-Controlled Evaluation of LLM Code Analysis	Paul Osemudiame Oamen; Courage Ochuko Mene; Abraham Ugbedeojo Daniel; Ananya Mukherjee; Pius Onobhayedo; Wei Zhao	PAUL OSEMUDIAME OAMEN
485	SAFE-KV: Causal Suffix Repair for Behavior-Preserving Non-Prefix KV Cache Reuse in RAG	Xingjian Zhang	Xingjian Zhang
488	From Construction to Injection: Edit-Based Fingerprints for Large Language Models	Yue Li; Xin Yi; Dongsheng Shi; Yongyi Cui; Gerard de Melo; Linlin Wang	Yue Li
489	InLegalLaySum: A Novel Benchmark for Lay Summarization of Indian Legal Documents	Trishita Saha; Debtanu Datta; SHUBHAM TIWARI; MOHIT KAUSHAL; Priyank Rathore; Mohammad Khalid Pandit; Pawan Goyal; Saptarshi Ghosh	Trishita Saha
490	WISE-Flow: Workflow-Induced Structured Experience for Self-Evolving Conversational Service Agents	Yuqing Zhou; Zhuoer Wang; JK; Hong Wang; Samson Koelle; Ziwei Zhu; Wei Niu	Zhuoer Wang
499	Private Evaluation Protocol: Contamination-Resistant Black-Box Evaluation via Public-Slice Surrogates	Md. Mahadi Hassan; Sazzadur Rahaman; Yusuf Sarwar Uddin; Santu Karmaker	Md Mahadi Hassan
501	Llm agent memory: a survey from a unified representation–management perspective	Zhenheng Tang; Xin He; Tiancheng Zhao; Fanjunduo Wei; Xiang Liu; Peijie Dong; Qian Wang; Qi Li; Huacan Wang; Ronghao	Fanjunduo Wei

		Chen; Sen Hu; Weidong Guo; Yu Xu; Haolan Chen; Kunfeng Lai; Kaiyong Zhao; Keyan Ding; Ivor Tsang; Yew-Soon Ong; Bo Li; Xiaowen Chu	
502	Emotion Is Not External Knowledge: Internal Attention Signals for Affective State Reasoning	Wenqiu Tang; Zhen Wan; Yihui Wang; Takahiro Komamizu; Ichiro Ide	WENQIU TANG
503	When Do Supervised UQ Ensembles Improve LLM Hallucination Detection? A Robustness Study	Mohit Singh Chauhan; Vipin Gyanchandani; Dylan Bouchard	Mohit Singh Chauhan
504	IndicRipple: Measuring Cross-Lingual Ripple Effects of Knowledge Editing in Indic Languages	Gunjan Rawat; Sneha Venkatesh; Shashidhar Reddy Javaji; Zining Zhu	Shashidhar Reddy Javaji
507	FAIR: Balancing Identity-Associated Training Signals to Mitigate Personalization Bias	Anvesh Rao Vijjini; Snigdha Chaturvedi	Anvesh Rao Vijjini
509	A Benchmark for Controllable Speaking-Style Captioning in Audio-Language Models	Shreeram Suresh Chandra; Ashwyn Sharma; Liang Lu; Xiutian Zhao; Leda Sari; Björn Schuller; Berrak Sisman	Shreeram Suresh Chandra
513	SAVER: Selective Auditing of Verbal Evidence for Error Recovery in VLM Change Reasoning	Youdi Li	Youdi Li
514	What's Worth Asking? Evaluating Preference Elicitation in Agents	Kandarp Joshi; Tushar Mittal; Chirag Soni; Lillian Zhou; Petar Aleksic	Kandarp Joshi
515	Cross-Domain Audit Distinguishes Conservative Defaulting from Capability Loss in ≤8B Medical Fine-Tuned LLMs	Abhineet Rajendra Kulkarni; Pranav Madhav Kuber; Anurag Bihani	Abhineet Rajendra Kulkarni
518	When and How to use In-context Learning for modern LVLMs? An Extensive Investigation	Pengxiang Li; Yu Zhao; Jiaxin Yu; Liang Xie; Erwei Yin	Pengxiang Li
521	Beyond Human–LLM Rating Agreement: A Preregistered Downstream Validation of LLM-Generated Social Desirability Ratings for Forced-Choice Personality Scale Construction	Yui Furukawa; Kyosuke Bunji; Kensuke Okada	Kensuke Okada
522	Analyzing Multilingual Parametric Knowledge Retrieval with Sparse Autoencoders	Marc E. Canby; Ikhyun Cho; Julia Hockenmaier	Marc E. Canby
525	DECK: A Consistency × Confidence Taxonomy of LLM Hallucinations	Mohit Singh Chauhan	Mohit Singh Chauhan
527	DEAF: A Benchmark for Diagnostic Evaluation of Acoustic Faithfulness in Audio Language Models	Jiaqi Xiong; Yunjia Qi; Qi Cao; Yu Zheng; Yutong Zhang; Ziteng Wang; Ruofan Liao; Weisheng Xu; Sichen Liu	Jiaqi Xiong
529	Interactive and Scalable Evaluation of LLM Math Tutoring via Simulated Teacher-Student Dialogues	Gihyeon Choi; Harksoo Kim	Gihyeon Choi
532	Role-aware Heuristic Episodic Attention for Conversational LLMs	hongwanyang; Zhaoning Zhang; Yi Chen; Libo Zhang; Baihui Liu; Linbo Qiao; Zhiliang Tian; Dongsheng Li	Wanyang Hong

536	AtomicCheck: Learning Structured Reasoning for Fact Verification via Protocol-Guided Reinforcement Learning	Xingquan Zhang; Jianghao Chen; Yang Zhao; Pu Jian; Haoran Gao; Fanyu Meng; Junlan Feng; Chengqing Zong	Xingquan Zhang
540	One Prompt Does Not Fit All: Self-Meta-Evolve for Personalized Information Extraction	Hongliang Li; Lu Wang; Yong Xu; Hanyang Chen; Zhitao Hou; Xiaoting Qin; Song Ge; Qingwei Lin; Dongmei Zhang	hongliang li
542	Models Know Their Shortcuts: Deployment-Time Shortcut Mitigation	Jiayi Li; Shijie Tang; Gün Kaynar; Shiyi Du; Carl Kingsford	Jiayi Li
545	COAT: An Orthogonal Benchmark for Diagnosing Chinese Adversarial Toxicity Detection Failures	Hu Jiaqi; Mingshuai Liu; Haobo Wu; Xin Li	Jiaqi Hu
551	ConceptWeave: Concept-Driven Prompting through Adaptive Semantic Infilling	Atul Kumar Singh; Rahul Mishra	Atul Kumar Singh
555	Implicit User Context Inference and Adaptation via Self-Distillation with Privileged Information	___	Xinshuang Liu
557	Mitigating Cross-Image Information Leakage in Multi-Image Understanding with Large Vision-Language Models	Yeji Park; Minyoung Lee; Sanghyuk Chun; Junsuk Choe	Yeji Park
558	Self-Distillation for Multi-Token Prediction	Guoliang Zhao; Ruobing Xie; An Wang; Kai Zhang; Shuaipeng Li; Huaibing Xie; Samm Sun	Guoliang Zhao
560	Closing the Bilingual Gap: Cross-lingual Semantic Scaffolding for Fine-Grained Machine Translation Evaluation in Taiwanese Hokkien	You-Wei Wu; En-Shiun Annie Lee; Richard Tzong-Han Tsai	You-Wei, Wu
562	TripPattern: A Pattern-based Text Watermarking Method for Large Language Models	Sangjun Moon; Dasom Choi; Jingun Kwon; Hidetaka Kamigaito; Taro Watanabe; Manabu Okumura	Sangjun Moon
564	Contextual Cue Sensitivity in LLM Scoring of Project Proposals: A Controlled ERC-Style Study	Marco Rospocher	Marco Rospocher
565	IMFD: End-to-end Multi-Face Forgery Detection through Instruction-based Large Vision-Language Models	Dasom Choi; Sangjun Moon; Hyeongchan Im; Jaeeon Park; Jingun Kwon; Hidetaka Kamigaito; Taro Watanabe; Manabu Okumura	Dasom Choi
568	Positive-Unlabeled Contrastive Retrieval for Graduate Program Recommendation	Shiyan Liu; Rui Qu	Shiyan Liu
573	Beyond EOS: Revealing the Internal Monologue of Language Models	Aishwarya Maheswaran; Maunendra Sankar Desarkar	Aishwarya Maheswaran
585	GRAIN: Gradient-Based Interaction Graph of Neurons for Explainable Misinformation Detection	Rathod Darshan D; Rahul Mishra	Rahul Mishra
587	SIEVE: Slot-wise Evaluation of Information-Role Retention in Fixed-Length Meeting Summarization	Jumpei Nagasawa; SEIYA SHIBAYAMA	Jumpei Nagasawa
590	Explainable Multimodal Aspect-Based Sentiment Analysis with Dependency-guided Large Language Model	Zhongzheng Wang; Yuanhe Tian; Hongzhi Wang; Yan Song	Zhongzheng Wang

595	Language Triggers Hijack Language Circuits: A Mechanistic Analysis of Backdoor Behaviors in Large Language Models	Théo Lasnier; Wissam Antoun; Francis Kulumba; Benoît Sagot; Djamé Seddah	Théo Lasnier
597	Chained Cross-Modal Learning for Modality-Agnostic Embeddings	Naoki Tamura; Taiki Sekii	Naoki Tamura
598	SASU: Structure-Anchored Semantic Unification for Code Search	Qian Zhang; Hao Li; Hao Jiang	Qian Zhang
599	Evaluating Large Reasoning Models on Belief Revision	Nurul Fajrin Ariyani; Zied Bouraoui; Richard Booth; Steven Schockaert	Nurul Fajrin Ariyani
600	EMR: Self-Evolving Medical Multi-Agent System via Experience Mining and Reuse	Dongsheng Shi; Yue Li; Xin Yi; Linlin Wang	Dongsheng Shi
605	All Verdicts are Not Equal: Rethinking LLM Judge Reliability	Vineet Kumar; Darshita Rathore; ANINDYA MOITRA	Vineet Kumar, Darshita Rathore
609	The Guardrail Trinity: Procedural Safety Schemas, Reasoning Chain Decomposition, and Dual-Mode Inference for LLM Guardrails	Huy H. Nguyen; Pride Kavumba; Jiaxuan Li; MinhDuc Vo; Koki Wataoka	Huy H. Nguyen
614	ArabMisinfo: A Decade-Long Arabic Facebook Corpus for Misinformation Discourse and Engagement Analysis	Wajdi Zaghouani; Shimaa Ibrahim; Mabrouka Bessghaier; Md. Rafiul Biswas	Wajdi Zaghoauni
619	XChoice: Explainable Evaluation of AI–Human Alignment in LLM-based Constrained Choice Decision Making	Weihong Qi; Fan Huang; Rasika Muralidharan; Jisun An; Haewoon Kwak	Weihong Qi
620	Structure Tax: How Structured Output affects LLMs Performance	Vineet Kumar; Kanishka; Bhuvanesh Mandora	Vineet Kumar
622	Grounded or Guessing? LVLM Confidence Estimation via Blind-Image Contrastive Ranking	Reza Khanmohammadi; Erfan Miahi; Simerjot Kaur; Charese Smiley; Ivan Brugere; Kundan S Thind; Mohammad M. Ghassemi	Reza Khanmohammadi
623	Exploring Cross-Scenario Generality of Agentic Memory Systems: Diagnostics and a Strong Baseline	Zhikai Chen; Jialiang Gu; Junyu Yin; Xianxuan Long; Shenglai Zeng; Xiaoze Liu; Kai Guo; Keren Zhou; Jiliang Tang	Zhikai Chen
626	Forgetting Only What Matters: Layer-Selective Unlearning toward Robust LLMs	Ravi Ranjan; Olivera Kotevska; Agoritsa Polyzou	Ravi Ranjan
627	Breaking the Speech-Act Contract: Uptake Failures and Void Commitments in Large Language Models	Sahana Varadaraju; Bharathwaj Vijayakumar; Lathish Balaji Baskaran	Sahana Varadaraju
628	A Pre-Training Analogue of Grokking in Language Models: Tracing Delayed Grammatical Generalization	Sherin Muckatira; Namrata Shivagunde; Vijeta Deshpande; Anna Rumshisky	Sherin Muckatira
632	Towards Robust Numerical Claim Verification	Peter Røysland Aarnes; Vinay Setty	Peter Røysland Aarnes
634	GEAR: Training-Free Rule Distillation for Advanced and Efficient Tool-Augmented Reasoning	Bhiman Kumar Baghel; Fanyou Wu; Suraj Maharjan; Peiran Yao; Srinivasan H. Sengamedu	Fanyou Wu

636	HARPA: A Testability-Driven, Literature-Grounded Framework for Research Ideation	Rosni Vasu; Peter Jansen; Pao Siangliulue; Cristina Sarasua; Abraham Bernstein; Peter Clark; Bhavana Dalvi Mishra	Rosni Vasu
638	Fast Listwise Passage Reranking via Semantic Tokens	Omkar Prabhune; Xijun Wang; Nan Bi; Haiyan Wang; Shima Alizadeh; Anoop Jain; Yu Lou; Junbang Liang	Omkar Prabhune
639	Personalization Matters: Long-Horizon Conversation Agent with User-Centric Information in Shopping Interactions	Rena Wei Gao; Yue Dai; Hao Guan; Shengxiang Gao; Wangyang Wu; Yixin Shen; Jey Han Lau	Rena Gao
641	From Textual Columns to Query Plans: A Unified Relational-Semantic Execution Framework for Hybrid Query Processing	Nima Shahbazi; Seiji Maekawa; Nikita Bhutani; Estevam Hruschka	Nima Shahbazi
644	PPL-Factory: Task-Aware and Budget-Aware Data Selection from Language Modeling to Reasoning	Hang Zhang; Warren J. Gross	Hang Zhang
646	TeleSWEBench: A Commit-Driven Benchmark for Evaluating LLM-Powered Software Engineering in Telecommunications	Pranshav Gajjar; Ali Mamaghani; Dinesh Bharadia; Vijay K Shah	Pranshav Gajjar
649	Training free Cross-Modal Alignment for Multilingual Image Retrieval	Usman Shahid; Manan Jain; Natalie Parde	Usman Shahid
650	User-Inference-Based Out-of-Character Behavior in LLM-Based Role-Playing Agents	Shinji Iwata; Shiki Sato; Jun Baba; Koya Ihara; Michimasa Inaba	Shinij Iwata
652	No Convo! The Illusion of Reliability in Online Toxicity Detection	Rachita Rajesh; Christina A Chance; Ashima Suvarna; Kai-Wei Chang; Saadia Gabriel	Ashima Suvarna
654	BetXplain: An Explanation-Annotated Dataset for Detecting Manipulative Betting Advertisements on Social Media	MSVPJ Sathvik; Parmitha Vangapandu; Nishit Rahul Rane; Sathwik Narkedimilli; Mark G. Lee; Akrati Saxena	Akrati Saxena
655	The Error Is the Signal: Misconception Discovery via Step-Error Annotation Clustering	Robinson Vasquez Ferrer; Mrinmaya Sachan; Shashank Sonkar	Robinson Vasquez Ferrer
658	Epistemic Erosion: A New Error Class Beyond Hallucination in LLM Text Generation	Micky C. Nnamdi; Shaunna Wang; Chinaza Nnamdi; Jieru Li; Benoit Louis Marteau; J. Ben Tamo; Yining Yuan; May Dongmei Wang	Shaunna Wang
662	The bias is in the generator: locating medical chain-of-thought subgroup gaps upstream of the verifier	Micky C. Nnamdi; Benoit Louis Marteau; Shaun Qien Yeau Tan; J. Ben Tamo; Yining Yuan; May Dongmei Wang	Benoit Marteau
664	HYDRA: Evaluating Logical Thinking Abilities of LLMs through Formulating Hypotheses and Verification Methods	Shin-nosuke Ishikawa; Hirotsugu OHBA; Masato Todo; Atsuhi Yoshino; Masahiro Daimon	Shin-nosuke Ishikawa
666	GCoTD: A Non-Intrusive Diagnostic Pipeline for Black-Box LLM Backdoor Detection via Graph Chain-of-Thought	Chengyu Jia; AprilPyone MaungMaung; Huy H. Nguyen; Isao Echizen; Jinyin Chen	Chengyu Jia

670	Efficient Batched Mixture-of-Experts Inference with Joint Expert Selection	Aradhana Mohan Parvathy; Sarada Krithivasan; Swagath Venkataramani; Anand Raghunathan; Viji Srinivasan	Aradhana Mohan Parvathy
672	Not All Verifiable Preference Pairs Are Equal: Support-Aware Filtering for Stable Iterative DPO	Viet Anh Ngo; Long Tri Thai Son; Nguyen Viet Duc; Dong Minh Duc; Quang Minh Tran; Nguyen Thi Thuy Linh; Nguyen Viet Ha	Viet Anh Ngo
674	Do Correctness Probes Transfer? A Cross-Family Study of LLM Hidden States	Evan Duan	Evan Duan
675	Eliciting Bias through Conversation: Multi-Turn Interviewers for LLM Auditing	Da Gao; Ziyu Chen	Da Gao
680	SWE-Adept: An LLM-Based Agentic Framework for Deep Codebase Analysis and Structured Issue Resolution	Kang He; Kaushik Roy	Kang He
681	VTBench: Evaluating Visual Tokenizers for Autoregressive Image Generation	Huawei Lin; Tony Geng; Zhaozhuo Xu; Weijie Zhao	Huawei Lin
684	CuBEs: Culturally-Situated Behavioral Evaluations and the Limitations of Culture-Blind LLM Judges	Hoda Ayad; Tanu Mitra; Abhishek Mukherji	Hoda Ayad
685	User Feedback to Actionable Insights: A Fine-grained Multilabel Approach for Classifying App Reviews	Chirag Jain; Preethu Rose Anish	Preethu Rose Anish
688	Beyond Idealized Patients: Evaluating LLMs under Challenging Patient Behaviors in Medical Consultations	Yahan Li; Xinyi Jie; Wanjia Ruan; Xubei Zhang; Huaijie ZHU; Yicheng Gao; Chaohao Du; Ruishan Liu	Yahan Li
690	SaFeR-ToolKit: Structured Reasoning via Virtual Tool Calling for Multimodal Safety	Zixuan Xu; Tiancheng He; Huahui Yi; Kun Wang; Xi Chen; Gongli Xi; Qiankun Li; Kang Li; Yang Liu; Zhigang Zeng	Zixuan Xu
692	K-FORGE: Fisher-Guided Initialization for Preference-Based LLM Unlearning	Daniil Moskovskiy; Alexander Panchenko	Alexander Panchenko
693	Improving LLM-Based Entity Detection with Annotation Guidelines	Marcel Geromel; Philipp Cimiano	Marcel Geromel
701	HeRE: A Novel Benchmark for Hebrew Relation Extraction via LLM-Guided Denoising of Knowledge Graph Alignments	Ron Keinan; Amir David Nissan Cohen; Reut Tsarfaty	Ron Keinan
702	MOOC-MESE: Dataset and Framework for Multimodal Entity Set Expansion in MOOCs	Shichao Yue; Jing Wan; yubin chen	Shichao Yue
706	Self-Route: Automatic Mode Switching via Capability Estimation for Efficient Reasoning	Yang He; Xiao Ding; Bibo Cai; Yufei Zhang; Kai Xiong; Zhouhao Sun; Bing Qin; Ting Liu	Yang He
709	A Systematic Review of Meta-Reasoning for Large Reasoning Models	Haoran Yu; Zilong Wang; Kang Pan; Xiao Zhang; Erli Meng; Shuai Li	Haoran Yu

712	Long-Term Agent Memory through Complementary Curation and Consolidation	Abhishek Gupta; Dongge Han; Daniel Madrigal; Wei-Ning Chen; Mayukh Das; Victor Rühle; Saravan Rajmohan	Donnge Han
713	RASER: Recoverability-Aware Selective Escalation Router for Multi-Hop Question Answering	Yuyang Li; Zihe Yan; Tobias Käfer	Yuyang Li
717	Surfacing Minor Nuances in Subjective Expressions: Initial Evidence for Improved Output Quality in Collaborative Caption Revision	Takeru Isaka; Jiro Nagao; Iwaki Toshima	Takeru Isaka
718	PoliSpace: Political Opinion Prediction using Shared Latent Space	Johanna D'Ciofalo Khodaverdian; Frédéric Berdoz; Roger Wattenhofer	Johanna D'Ciofalo
720	RAGCap-Bench: Benchmarking Capabilities of LLMs in Agentic Retrieval Augmented Generation Systems	Jingru Lin; Chen Zhang; Yong Liu; Haizhou Li	Jingru Lin
721	Beyond Retrieval: Deterministic Memory Governance with Pairwise Constraints and Certified Repair for Long-Term Language Agents	Panayiotis Christou; Miguel Palma; Zefan Du; Kaixun Hua; Ying Mao	Panayiotis Christou
722	Outgrowing the Curse of Multilinguality: Scaling Strategies for Efficient Language Adaptation	Kevin Glocker; Kätriin Kukk; Romina Oji; Marcel Bollmann; Marco Kuhlmann; Jenny Kunz	Kevin Glocker
723	Auditing Bias Detection Under Source Shift: A Source-Aware Benchmark and Evaluation Protocol	Ana Pinto; Alvaro Figueira	Ana Pinto; if Ana is unavailable, Álvaro Figueira
724	HDSL: A Hierarchical Domain-Specific Language for Structured 3D Indoor Scene Generation and Localized Editing with LLM Agents	Letian Li; Chao Shen; Shuzhao Xie; Chenghao Gu; ZhengXiao He; Yu Meng; Xin Yang; Wenyuan Jiang; Zhi Wang	Letian Li
726	Vouch: Counterfactual Contracts for Multi-Agent Coordination	Abhay Bhandarkar; Vedanth Nanesha	Abhay Bhandarkar
727	Training LLMs Beyond Next Token Prediction - Filling the Mutual Information Gap	Chun-Hao Yang; Bo-Han Feng; Tzu-Yuan Lai; Yan-Yu Chen; Yin-Kai Dean Huang; Shou-De Lin	Chun-Hao Yang, Bo-Han Feng
728	Structural vs Cultural Reasoning: Reliability Limits of LLM-as-a-Judge in Multilingual Evaluation	Shehenaz Hossain; Houda Bouamor; Haithem Afli	Shehenaz Hossain
729	Decisions by Discrimination: Differential Evidence for Medical Reasoning	Yunpeng Xiao; Yuzhang Xie; Junhao Huang; Carl Yang; Xiao Hu; Kai Shu	Yunpeng Xiao
740	Towards Smart Exploration: Reinforcement Learning for LLM Agents Guided by Meta-Actions	Hanwen Tong; Zhengqi Zhang; Ruixuan Lu; Jincheng Dai; Bin Yang; Bo Zheng; Feng Wei	Hanwen Tong
741	Sampling via Decision-Flow: Training-Free Extraction of Improved Latent Reasoning Paths in Large Language Models	Zhendong Mi; Shaoyi Huang	Zhendong Mi
742	Towards Universal Calibration for In-Context Learning	Awik Dhar; Sunita Sarawagi; Arpit Agarwal	Awik Dhar

743	Data Augmentation for Improving Contextual Faithfulness Under Knowledge Conflict	Mohith Rajesh; Raj Maheshwari; Vashisth Tiwari; Yonatan Bisk; Emma Strubell; Jeremiah Milbauer; Jared Fernandez	Mohith Rajesh
745	Adaptation, Not Algorithm: PEFT and Full Fine-Tuning Degrade Jailbreak Safety Equivalently Across Five Open-Weight LLMs	Taleef Tamsal; Jonathan Rusert	Taleef Tamsal
750	Activation-Efficient Fine-Tuning via Sequence-Compressed Low-Rank Subspaces	Irene Tenison; Stella Ahn; Miriam Kim; Ebtisam Alshehri; Lalana Kagal	Irene Tenison
754	Detect Before You Disambiguate: Perturbation-Based Input Ambiguity Detection in LLMs	Tiejin Chen; Xiaoou Liu; Ahmadreza Moradipari; Michael Clifford; Kyungtae Han; Yitao Chen; Nejib Ammar; Hua Wei	Tiejin Chen
755	Leveraging Articulatory Phonetic Group Recognition for Robust Speech Enhancement for Cochlear Implants in Complex Acoustic Environments	Sohini Saha; Leslie M. Collins; Boyla Mainsah	Sohini Saha
756	Diverse Reasoning Paths Matter: Symbolic Equivalence Augmentation for Enhancing RL Exploration	Xiaoou Liu; Tiejin Chen; Hanyang Chen; Yuan Ling; Hua Wei	Xiaoou Liu, Tiejin Chen
762	The Cost of Compression: A Rate-Distortion Limit on Factual Hallucination	Wang Xi; Shijia Xu; Rongfeng Guo	Shijia Xu
763	An Empirical Study of Position Bias in Visualized Text Retrieval	Cuiqiaoshu Pan; Ziyang Zeng; Dun Zhang; Yuqing Yang	Cuiqiaoshu Pan
765	Synthesizing Code Completion Trajectories for Mid-training in Software Engineering	Siwei Xu; Qiyuan Liu; Ruifan Li	Siwei Xu
766	A Geometric Perspective on Stabilizing Value Conflict Resolution	Saket Reddy; Andy Liu	Saket Reddy
767	Do LLMs Have Social Sense? Evaluating Alignment of Conversation Understanding and Reasoning	Zhaoqing Wu; Dan Goldwasser	Zhaoqing Wu
768	Parameter-Efficient Domain Adversarial Training for Lightweight Framewise Phoneme Classification	Sohini Saha; Leslie M. Collins; Boyla Mainsah	Sohini Saha
769	M^4olGen: Multi-Agent, Multi-Stage Molecular Generation under Precise Multi-Property Constraints	Yizhan Li; Florence Cloutier; Sifan Wu; Ali Parviz; Boris Knyazev; Yan Zhang; Glen Berseth; Bang Liu	Bang Liu
771	BG-RAG: Concept-Mediated Bipartite Graphs for Scalable RAG	Hang Gao; Yangmin Ding; Shaobo Han; Zhuocheng Jiang	Hang Gao
773	Topo-Narrator: Heterogeneous Topological Memory for Long-form Narrative Generation	Kai Chen; Haoran Zhang; Tian Lan; Zehua Duo; Jiang Li; Guanglai Gao; Xiangdong Su	Kai Chen
779	PULSE-CARE: A Measurement-Validity Audit of Simulated-Patient Clinical Agent Evaluation	Srimonti Dutta; Dinky Mishra; Aditi Singh; Akshata Kishore Moharir	Srimonti Dutta
780	DensePath REINFORCE: Lightweight Policy Improvement from SFT-Induced Token-Level Reward Signals	Jiangnan Li; Thuy-Trang Vu; Ehsan Abbasnejad; Gholamreza Haffari	Jiangnan Li

782	When Should Language Models Search? A Topology-Aware Analysis of Reasoning Strategies	Dibyanayan Bandyopadhyay; MD Tanzeel Adam Khan; Asif Ekbal	Dibyanayan Bandyopadhyay
784	More Modalities, More Problems: When Multimodal Agentic Debate Hurts Time-Series Reasoning	Patara Trirat	Patara Trirat
785	Evaluating LLM-as-a-Judge Beyond Score Alignment: A Psychometric Analysis of Residual Judging Difficulty	Longwei Cong; Sonja Hahn; Sebastian Gombert; Leon Camus; Fabian Zehner; Hendrik Drachsler; Ulf Kroehne	Longwei Cong
787	ICCU: In-Context Continual Unlearning via Pattern-Induced Refusal Rules	Ruihao Pan; Suhang Wang	Ruihao Pan
789	Evaluation of Multilingual LLMs Personalized Text Generation Capabilities Targeting Groups and Social-Media Platforms	Dominik Macko	Dominik Macko
791	PerQ: Efficient Evaluation of Multilingual Text Platform-Adaptation Quality	Dominik Macko; Andrew Pulver	Dominik Macko
792	BrainMoE: A Brain-Inspired Modular Mixture-of-Experts Model for Multimodal Reasoning	Siyao Song; Yunhao Zhang; Jingyuan Sun; Yaping Zhang; Chengqing Zong	Siyao Song
793	Large Language Models for Solid-Earth Geoscience: Applications, Modalities, and Memory-Reason-Act Techniques	Yisu Chen; Yiran Zhang; Yifeng Dai; Puzhen Wu; Sirui Li; Wei Liu; Yihao Ding	Yiran Zhang
794	Probing Reasoning Length Control in Language Models via Neurons	Dongwook Min; Jeonhwi Kang; Cheoneum Park; Seung-Hoon Na	Dongwook Min
795	DALSA: Difficulty-Aware Long-to-Short LLM Reasoning with Model Merging	Dongwook Min; Jeonhwi Kang; Donghyeon Jeon; Inho Kang; Seung-Hoon Na	Dongwook MIn
800	DPDisc: From Factoid Questions to Data Product Requests for Open-World Data Product Discovery over Tables and Text	Liangliang Zhang; Nandana Mihindukulasooriya; Niharika S. D'Souza; Sola Shirai; Sarthak Dash; Yao Ma; Horst Samulowitz	Liangliang Zhang
802	Orthogonal Style Injection: Mechanistic Persona Steering for Zero-Shot Cross-Lingual Subtitle Translation	Tarun Chintada; Kshetrimayum Boynao Singh; Asif Ekbal	Tarun Chintada
803	Cleansing Fine-Tuning Data to Mitigate Backdoors in LLMs via the Base Model's Inherent Generative Abilities	Qi Zhao; Yan-Keon Elster; Christian Wressnegger	Qi Zhao

10. Meals, Reception, and Social Events

Please always take your badge with you.

Lunch

Date	Time	Location
Friday, November 6, 2026	12:30-14:00	Muze & Muhe A
Saturday, November 7, 2026	12:00-13:30	Muze & Muhe A
Sunday, November 8, 2026	12:00-14:00	Muze & Muhe A
Monday, November 9, 2026	12:30-14:00	Muze & Muhe A
Tuesday, November 10, 2026	12:30-14:00	Muze & Muhe A

Welcome Reception

Date	Time	Location
Friday, November 6, 2026	18:30-20:30	Hotel Indigo Resaurant & Function Room

Social Event

Date	Time	Arrangement / Location
Saturday, November 7, 2026	15:30	Shuttle buses depart from the conference venue
Saturday, November 7, 2026	16:00-21:00	Chimelong Ocean Kingdom, Hengqin

11. Travel and Local Information

Accommodation

Official hotels listed on the venue page include Hotel Indigo Hengqin in the venue building, Sumjoy Hengqin Tianmu Hotel, and Zhuhai Hengqin Port Xingledu Xinglv Hotel. Please consult the official venue page for current availability and booking arrangements.

Arriving in Hengqin

The official transport guidance describes routes via Zhuhai Airport, the Hong Kong–Zhuhai–Macao Bridge, and Hengqin Port from Macao. Consult the current official guidance when arranging travel.

12. Organizing Committee and Acknowledgments

The following roster is based on the official organizing committee.

Role	Names
General Chair	Yue Zhang
Program Chairs	Ruifeng Xu; Ayu Purwarianti; Tong Rong; Jun Suzuki
Local Organizing Chairs	Derek F. Wong; Haofen Wang
Local Organizing Committees	Runzhe Zhan; Zhiwen Xie
Workshop Chairs	Xuefeng Bai; Ratish Puduppully
Tutorial Chairs	Zhongqing Wang; Naoaki Okazaki
Demonstration Chairs	Ningyu Zhang; Derry Wijaya
SRW Faculty Advisor	Yixin Cao
SRW Chairs	Shibo Hong; Yuhang Zhou; Wenhe Sun
Best Paper Award Chair	Rui Yan
Publication Chairs	Junxian He; Trang Vu
Handbook Chair	Leyang Cui
Publicity Chairs	Aditya Joshi; KyungTae Lim
Web and App Chairs	Xinchi Chen; Piji Li; Shudong Liu
Finance Chair	Hou Pong Chan
Registration Chair	Wenxuan Zhang
Ethics Chairs	Yangqiu Song; Ekaterina Vylomova
Diversity & Inclusion Chairs	Mamoru Komachi; Emily Ohman
Student Volunteer Chair	Juntao Li
Sponsorship Chairs	Yan Song; Yuanhe Tian

We thank all organizers, reviewers, volunteers, and supporting organizations for their contributions.

13. Sponsors

We gratefully acknowledge the organizations supporting AACL-IJCNLP 2026.

The conference homepage announced Baidu as a Platinum sponsor on October 6, 2026.

Co-organizer

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

信息學院
Faculdade de Ciência da Informação e Computação
Faculty of Information Science and Computing

Venue Supporter

珠海大横琴星乐度文化旅游发展有限公司
ZHUHAI DA HENG QIN SUMLODOL CULTURE & TOURISM DEVELOPMENT CO.,LTD

Platinum

Baidu 百度

Gold

復旦大學
FUDAN UNIVERSITY

记忆张量
MemTensor

Silver

荔枝

14. Paper Metadata and Proceedings

Proceedings ISBNs

The supplied ISBN list contains 19 proceedings titles. Titles and ISBNs are reproduced below as supplied.

Proceedings title	ISBN
Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)	979-8-89176-439-2
Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)	979-8-89176-440-8
Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)	979-8-89176-441-5
Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)	979-8-89176-442-2
Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)	979-8-89176-443-9
Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026	979-8-89176-444-6
Proceedings of the 5th Workshop on Multimodal Machine Learning in Low-resource Languages (MMLow 2026)	979-8-89176-445-3
Proceedings of the 13th Workshop on Asian Translation (WAT 2026)	979-8-89176-446-0

Proceedings title	ISBN
Proceedings of the 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP)	979-8-89176-447-7
Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche)	979-8-89176-448-4
Proceedings of the 1st Workshop on Multi-modality and Multi-lingual Language Learning	979-8-89176-449-1
Proceedings of the 1st Workshop on Language Technologies for Crisis Preparedness and Response (LT4CPR)	979-8-89176-450-7
Proceedings of the 1st Workshop on Data, Agents, Systems, and Human-in-the-Loop (DASH)	979-8-89176-451-4
Proceedings of the 1st Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio)	979-8-89176-452-1
Proceedings of the 1st Workshop on Pluralistic Value Alignment of LLMs (PlurVA-LLM)	979-8-89176-453-8
Proceedings of the 6th Workshop on Scholarly Document Processing (SDP)	979-8-89176-454-5
Proceedings of the 1st Workshop on Grounding Documents with Reasoning, Agents, Retrieval, and Attribution (RARA)	979-8-89176-455-2
Proceedings of the 1st Workshop on AI Agents for Robust Natural Language to SQL Generation (NL2SQL)	979-8-89176-456-9
Proceedings of the 1st Workshop on KNOwledge GRaphs & Agentic Systems Interplay (NORA)	979-8-89176-457-6

Accepted Paper Metadata

Imported from the supplied papers.yml file: 620 records, including 355 Main papers and 265 Findings papers. Author order and abstracts are preserved from the supplied metadata.

Main Papers

Paper 1 Towards Probabilistic Question Answering Over Tabular Data

Authors: Chen Shen, Estevam Hruschka

Abstract Progress in question answering (QA) over tabular data has enabled reliable factual retrieval from relational tables. However, many real-world questions are probabilistic, requiring reasoning under uncertainty and latent conditional dependencies that are not directly stored in individual cells. We introduce LUCARIO, a large-scale benchmark for probabilistic QA over real-world tabular datasets, covering diverse domains, dependency patterns, and natural language variations. We propose Auto-BN, a fully automatic neuro-symbolic framework that induces a Bayesian Network from raw tables, translates natural-language questions into formal probabilistic queries, and performs exact inference. Experiments across multiple LLM backbones show that Auto-BN consistently outperforms NL2SQL, retrieval-based, and premise-based baselines, while maintaining a near-zero error rate. LUCARIO fills a relevant gap and provides a practical and scalable testbed for advancing probabilistic reasoning over structured data.

Paper 2 An Empirical Study of On-Device Translation for Real-Time Live-Stream Chat on Mobile Devices

Authors: Jeiyoon Park, Leedaehwan, Changmin Yeo, Yongshin Han, Minseop Kim

Abstract Despite its efficiency, there has been little research on the practical aspects required for real-world deployment of on-device AI models, such as the device's CPU utilization and thermal conditions. In this paper, through extensive experiments, we investigate two key issues that must be addressed to deploy on-device models in real-world services: (i) the selection of on-device models and the resource consumption of each model, and (ii) the capability and potential of on-device models for domain adaptation. To this end, we focus on a task of translating live-stream chat messages and manually construct LiveChatBench, a benchmark consisting of 1,000 Korean–English parallel sentence pairs. Experiments on five mobile devices demonstrate that, although serving a large and heterogeneous user base requires careful consideration of highly constrained deployment settings and model selection, the proposed approach nevertheless achieves performance comparable to commercial models such as GPT-5.1 on the well-targeted task. The code, trained models, and LiveChatBench will be made publicly available at our GitHub.

Paper 5 Generating Linguistically Diverse and Semantically Faithful Patient Reports

Authors: Alejandra Lorenzo, Adrien Coulet, Claire Gardent

Abstract To protect patient privacy, medical models are frequently trained on synthetic datasets. Focusing on 13 clinical attributes commonly found in breast cancer records, we propose a synthetic data generation framework in which each generated sample is a data–report pair: the structured data capture the 13 attributes of interest, while the accompanying (French) text simulates a corresponding breast cancer pathology report. Data privacy and coverage are obtained by generating the structured data using a tabular data synthesiser. In this work, we focus on enhancing the diversity and quality of the generated texts through a combination of data-driven knowledge distillation, data filtering and data combination. Our findings suggest that the quality of the training data used for the student model outweighs its quantity; that combining data from multiple teachers helps improve

performance; and that a compact model (Mistral 7B) can match the zero-shot Information Extraction performance of a significantly larger model (MedGemma 27B) when fine-tuned on high-quality synthetic data.

Paper 6 Relational Priors as Convergence Pressure in LLM-Based Multi-Agent Systems

Authors: Ming Shen, Chao Shang, Sadat Shahriar, Devang Kulshreshtha, ~Yi_Zhang13, ~Sandesh_Swamy2, ~Yanjun_Qi1

Abstract Large language model-based multi-agent systems (LLM-MAS) are designed through roles, debate protocols, and aggregation rules. These choices create implicit expectations of trust, skepticism, deference, or collaboration. We make inter-agent relations explicit as signed pairwise priors, rendered in natural language and added to system prompts while keeping the task protocol fixed. Across commons governance and multi-agent debate, these priors change how readily agents coordinate or agree, a pattern we call convergence pressure. More positive relations generally improve sustainability within the GovSim relational-prior sweep and increase consensus on subjective questions. These improvements over negative relations differ from gains over no-prior prompting. On objective QA, fully positive priors usually yield lower final-answer accuracy than the no-prior baseline, and some conditions produce more frequent but less accurate consensus. Effects depend on model backbone, relation type, and topology; explicitly neutral relations also produce different outcomes from omitting relational framing. Relational priors are therefore task-specific interventions and diagnostic probes of sensitivity to social framing. Evaluate them against a no-prior baseline using the task's primary metric, report accuracy and consensus correctness alongside agreement on objective tasks, and keep no-prior prompting as the default for accuracy-centric tasks unless validation supports a relational prior.

Paper 7 Dynamics of Meaning: Towards the Evaluation of Diachronic Semantic Change in Sinhala

Authors: ~Nevidu_Jayatilleke1, ~Nisansa_de_Silva1

Abstract Tracking semantic change in low-resource languages across extensive historical timelines presents significant challenges due to data scarcity and the limitations of static embedding alignments. This study investigates the diachronic evolution of the Sinhala language from the 13th to the 20th century using a multi-stage computational framework. We first align century-specific Word2Vec and FastText embeddings using Similarity Matrix Based Alignment (SMA) and Orthogonal Procrustes (OP) techniques, finding that OP alignment provides more stable neighbourhood tracking for identifying temporal similarity dips. To move beyond aggregate measures, we introduce a Bidirectional Semantic Impact Pruning approach using contextualised embeddings from a fine-tuned Llama-3.1-8B. By applying Leave-One-Out (LOO) diagnostics, we attempt to isolate influential sentences to distinguish between systemic semantic shifts and transient polysemic expansion. Our results show that semantic drift in the fine-tuned Llama-3.1-8B is not evenly distributed across all usages. Instead, a significant part of the change is driven by a smaller set of high-impact contextual instances, rather than gradual and uniform change across all occurrences. This work provides a preliminary framework for low-resource Sinhala diachronic analysis, highlighting the trade-offs between model sensitivity and data availability.

Paper 8 ImmunoFOMO: Are Language Models missing what oncologists see?

Authors: Aman Sinha, Bogdan-Valentin Popescu, Xavier Coubez, Marianne Clausel, Mathieu Constant

Abstract Language model (LM) capabilities have rapidly advanced over the past decade, driving their adoption across fields, including biomedical research. While domain-specific pretrained language models (PLMs) have long supported biomedical NLP, recent focus has shifted to medical large language models (LLMs) and their language understanding capabilities. In this paper, we investigate this capability by probing the medical conceptual alignment of various LMs against clinical expertise for identifying hallmarks of immunotherapy in breast cancer abstracts. Our results show a clear trade-off: LLMs achieve stronger performance and higher consistency on high-level concepts, whereas PLMs are more reliable for fine-grained, low-level concept identification.

We find that hallucinations are the primary factor limiting LLM performance in these settings; once mitigated, LLMs surpass PLMs across all tiers of biomedical concepts. Despite this, PLMs remain competitive, particularly in closely aligning with medical expert judgments on detailed conceptual distinctions.

Paper 10 ChartTableViVQA: A Vietnamese Benchmark Dataset for Visual Question Answering with Reasoning Instruction over Charts and Tables Images

Authors: Quang Nhat Truong, Pham Quoc Thang, Son T. Luu

Abstract Visual Question Answering (VQA) over charts and tables requires joint reasoning over language, visual structure, and numerical information, yet no standard benchmark currently exists for Vietnamese. We introduce \textbf{ChartTableViVQA}, a Vietnamese VQA dataset with 13,003 question--answer--instruction triplets over 868 chart and table images. Questions cover multiple reasoning types, ranging from simple value lookup to multi-step aggregation, comparison, and trend analysis, and are accompanied by concise instructions that explicitly guide the reasoning process. The dataset is built via a semi-automatic human-in-the-loop pipeline, combining large language model generation with manual verification by Vietnamese annotators. We evaluate several vision-language models in zero-shot and instruction-based settings. Results show strong performance on simple questions but clear weaknesses in complex reasoning, particularly on charts, while instruction-based prompting consistently improves accuracy.

Paper 11 Navigating the Reality Gap: On-Device Continual Adaptation of ASR for Clinical Telephony

Authors: Darshil Chauhan, Adityasinh Solanki, Vansh Patel, Kanav Kapoor, Ritvik Jain, Aditya Bansal, Pratik Narang, Dhruv Kumar

Abstract Automatic Speech Recognition (ASR) can ease clinical documentation in resource-constrained regions, but deployment is hindered by a "Reality Gap" between laboratory performance and noisy, real-world clinical telephony, compounded by strict data residency and compute constraints. We study this gap using Gram Vaani, a telephonic Hindi corpus spanning rural healthcare and agricultural helplines, as the closest publicly available proxy for clinical telephony speech, and show that a robust multilingual model (IndicWav2Vec) degrades from 11.60\% WER on clean read Hindi to 41.72\% WER on this data. We evaluate a progression of adaptation regimes, from full fine-tuning and offline Low-Rank Adaptation (LoRA) upper bounds to an on-device, stream-based continual adaptation framework in which raw audio never leaves the local device, and characterize the trade-offs between data-driven and parameter-driven stabilization strategies. Our evaluation covers both lexical accuracy (WER and CER) and semantic fidelity (BERTScore) on the target domain, alongside the retention of general-domain knowledge. Multi-domain Experience Replay (ER) yields the primary gains, improving target WER by 18.2\% relative and reducing catastrophic forgetting by 54\% compared to naive adaptation, with BERTScore reflecting consistent gains in semantic fidelity. Combining replay with Elastic Weight Consolidation based on a stabilized importance estimate (Absolute Fisher) yields the strongest retention at a small cost in plasticity. Finally, a language model spot check empirically verifies that the core mismatch lies at the acoustic level and cannot be resolved by language models alone.

Paper 12 SPECI: Fine-grained Evaluation of Specificity for Image Captioning

Authors: Son T. Luu, Hiep K. V. Nguyen, Trung Bui, Viet Dac Lai, Le-Minh Nguyen

Abstract Image Captioning lies at the center of synthetic data generation in computer vision. The quality of the caption defines the behavior of the resulting model trained on image-caption pairs generated by Image Captioning. Measuring the specificity of image captioning plays an important role in supplying high-quality captions for the downstream tasks. To address the challenge

of automatically assessing the quality of the caption, we propose SPECI, an automatic metric that decomposes dense captions into atomic facts and evaluates specificity across visual criteria such as granularity, spatial detail, color, and linguistic focus using WordNet-based hyponym levels. Empirical results on the benchmark datasets demonstrate significant correlation with human judgments, and its application in image captioning and image reconstruction tasks highlights the practical utility of SPECI. The source code of SPECI is available at https://github.com/sonlam1102/SPECI.

Paper 13 Are Linguistic and Arithmetic Circuits Dissociated in LLMs?

Authors: Riku Kisako, Tatsuki Kuribayashi, Ryohei Sasano

Abstract Large language models (LLMs) exhibit both knowledge of language, rendered by, for example, grammatical judgment ability, and abilities to perform tasks through language, such as math-word problem solving. Motivated by neurolinguistic studies of their neural dissociation in human brains, in this study, we investigate the overlap of their circuits in large language models (LLMs). Specifically, we examine the overlap and causal effects between the circuits extracted for linguistic and arithmetic tasks, using controlled materials and circuit discovery methods from the mechanistic interpretability field. Our experiments with five LLMs show that circuit similarity is mostly explained by task domains (linguistic vs. arithmetic), rather than by their input and output vocabularies or by the syntactic similarities between natural-language sentences and arithmetic expressions. This finding suggests that LLMs may process linguistic and arithmetic information using largely separate neural circuits, highlighting an ability-based modular organization of circuits within these models.

Paper 14 Steering with Multilingual Lexicons for Sentiment Control in Low-Resource Languages

Authors: Lucinda Laurent, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono, Fajri Koto

Abstract Activation steering offers a lightweight mechanism for controlling model behavior, but existing methods depend on sentence-level contrastive data that is unavailable for most of the world's languages. We show that multilingual sentiment lexicons are a viable substitute across three model families: aggregated multilingual lexicons outperform English-derived steering vectors for mid- and low-resource languages, while monolingual lexicons remain stronger for languages a model has specialized in during pre-training. The steering efficacy is mechanistically dictated by both representational density and pre-training priors. Large-magnitude English vectors overwhelm the compact representations of low-resource languages, causing "linguistic collapse." This fragility is mitigated by the norm-dampening of multilingual aggregation. Architectural biases, including bilingual specialization penalty and latent numerical rigidity, also heavily constrain cross-lingual steering responsiveness. Ultimately, our work establishes lexicons as a practical steering signal and offers a comprehensive mechanistic account of cross-lingual steering dynamics.

Paper 15 Constructive Comment Extraction from Japanese YouTube Shorts: A Dataset and Empirical Study

Authors: Siyuan Feng, Luecheng Wang, Koyo Ogura, Yuma Iwamoto, Seinosuke Sato, Yuki Kometani, Masaki Itabashi, Shintaro Ozaki, Yusuke Sakai, Shogo Matsuno, Hidetaka Kamigaito, Katsuhiko Hayashi, Ryoma Obara, Sho Takase, Taro Watanabe

Abstract Identifying constructive comments from user-generated content is important for improving information discovery and discussion quality on platforms such as YouTube. However, existing resources focus predominantly on English news comments, leaving short-form video platforms underexplored despite their distinct characteristics: comments are shorter, more reactive, and not well captured by popularity signals. In this work, we introduce the first Japanese YouTube Shorts dataset for constructive comment extraction, featuring multi-annotator labels that preserve naturally occurring disagreement. We also develop a real-time system that surfaces constructive comments from a noisy comment stream, evaluated in a preliminary user study. To train under noisy annotation conditions, we combine self-training with high-confidence pseudo-labels and active learning. Experiments

show that iterative data expansion improves performance, while our compact task-specific model achieves the highest PR-AUC point estimate among the evaluated systems. These findings demonstrate the value of dataset construction and targeted training for constructive comment extraction in realistic settings. Our dataset and code are available at https://github.com/nlp-hayashi/japanese-youtube-constructive-comments

Paper 16 PRC: An Efficient and Unified Framework for Evaluating Robustness and Controllability of Large Language Models on Few-shot Prompt Variations

Authors: Chunpeng Ma

Abstract Understanding the impact of prompt variations on large language model (LLM) outputs is critical. Despite extensive research, two challenges persist: conflated terminology stemming from loosely defined concepts, and inefficient evaluations reliant on output generation. To address these, we first formalize prompt categorization to clearly define prompt robustness and controllability. We then introduce the prompt vector (PV), a compact internal representation that reformulates these properties at the representation level. Building on PV, we propose PRC, an efficient and unified framework to systematically evaluate LLMs under few-shot prompt variations without requiring costly autoregressive decoding. Experiments across diverse instruction-tuned LLMs demonstrate that PRC tightly correlates with traditional output-based metrics while substantially reducing computational overhead. Overall, PRC offers a principled and highly efficient tool for diagnosing LLM behavior.

Paper 17 JUMPSTARTER: Human-AI Planning with Task-Structured Context Curation

Authors: Xuanming Zhang, Sitong Wang, Jenny Ma, Alyssa Hwang, Zhou Yu, Lydia Chilton

Abstract Human-AI collaboration on complex planning goals is bottlenecked by how LLM interfaces handle context: users must manually curate and re-surface relevant information across long and unstructured chat histories. Despite advances in long-context prompting and memory-augmented retrieval, this burden remains unresolved: users still have to identify and supply the right context at each decision point, regardless of how much the model can store or surface. We propose JumpStarter, a system that enables LLMs to collaborate with humans on complex goals by dynamically decomposing tasks to help users manage context. We specifically introduce task-structured context curation, a framework that breaks down a user's goal into a hierarchy of actionable subtasks and scopes context to localized decision points, enabling finer-grained personalization and reuse. The framework is realized through three core mechanisms: context elicitation, selection, and reuse. In a within-subjects user study, plans produced with JumpStarter were rated substantially higher in quality than those produced with ChatGPT. A complementary automatic simulation study shows that JumpStarter consistently outperforms ChatGPT baselines, planning and memory agents, and workflow ablations. These findings show that effective human-AI planning depends not on the volume of context provided, but on attaching the right context to the right subtask at the right time.

Paper 19 Decomposing LLM Advice Sycophancy: Content, Style, and Framing Drive Distinct Response Dimensions

Authors: Jaya Singh, Verena Kaynig-Fittkau

Abstract Large language models are widely used for personal advice (relationship conflicts, financial decisions, mental-health support, etc.), but they often align their response with the user's position rather than offering independent judgment, even when users need pushback rather than validation. Mitigating this, whether through training objectives, system-level guardrails, or evaluation design, requires knowing what triggers it. But what in the user's message triggers this behavior is unknown. Natural advice-seeking messages vary simultaneously in semantic content (what is disclosed), linguistic style (how it is written), and

narrative framing (whose side is told), but existing benchmarks cannot attribute model behavior to any single cue. Without this attribution, mitigations cannot be targeted, for example, suppressing warmth may harm users with genuine distress but might not fix the model's tendency to endorse whichever side of a conflict speaks. In this work, we isolate which of these cues drive which aspect of sycophantic behavior across 787 advice-seeking scenarios and observe three patterns emerge: emotional warmth is content-driven; stance endorsement is framing-driven: endorsement drops by ${\sim}50$% when the same conflict is retold from the opposing perspective; and recommendation directness is style-driven. We also find that hedging reduces rather than increases model endorsement; models do not simply validate vulnerable-sounding users.

Paper 21 HybridScholar: Personalized Paper Recommendation with Cross Attention-Gated Fusion and Verified Explanation Generation

Authors: Joytun Nessa Tonny, Sajidur Rahman, Saddam Mukta, Swakkhar Shatabda, Salekul Islam

Abstract The exponential growth of scientific literature has made personalized, factually grounded paper discovery increasingly difficult. Existing systems suffer from limited personalization, cold-start fragility, or hallucinated explanations. We introduce HybridScholar, an integrated architecture combining a CrossAttention-gated fusion of collaborative filtering and RAG learning candidate-specific gate weights per candidate with a Chain-of-Verification (CoVe) pipeline for factually grounded explanations. Evaluated on the Scholar Inbox dataset, HybridScholar substantially outperforms (NDCG@10=0.2991) strong baselines including LightGCN and Scholar-Inbox. Cold-start analysis shows the RAG fallback prevents hard failure for sparse users (NDCG@10=0.283), with cold NDCG rising monotonically as interactions grow. The CoVe pipeline reduces hallucination rate by 23.8\%, confirmed by a human evaluation study showing CoVe-revised explanations are preferred by experts on factual accuracy (76.7% vs. 36.0%) and by end-users on helpfulness (78.8\%, $p{<}0.0001$). Overall, HybridScholar improves ranking quality, cold-start robustness, and factual reliability for large-scale scholarly research discovery, demonstrated on 15,881 users across 239,618 papers.

Paper 26 How Robust Is Multimodal Claim Verification to LLM Rewriting?

Authors: Yun-Ang Wu, Xanh Ho, Andre Greiner-Petter, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, Akiko Aizawa

Abstract LLMs are known to introduce stylistic changes into generated text, yet how these stylistic shifts affect model decisions on scientific tasks remains underexplored. In this paper, we focus on multimodal claim verification, where the goal is to determine whether a textual claim is grounded in a given piece of evidence. We apply two rewriting strategies: natural rewriting, which simulates how researchers routinely use LLMs to polish academic text, and controlled injection, which inserts a single LLM-associated word to isolate the effect of vocabulary choice. We evaluate 11 open-weight models spanning five VLM families and ranging from 2B to 38B parameters. We find that models are robust to these modifications: most show no significant drop in accuracy, and compared to prior work on review-score manipulation, verification appears far more stable. However, consistent probability shifts do occur. Hedging-oriented conditions produce significant shifts across nearly all models, while boosting conditions show a weaker effect and general polishing conditions (e.g., grammar correction, fluency improvement) have little effect.

Paper 31 When Does Knowledge Distillation Hurt? Reliability-Aware Distillation for Low-Resource Language Summarization

Authors: Dipto Sumit, Ankan Kumar Roy Srizon, Sadia Khair Rodela, Atia Haque Asha, Mourchona Afrin, Niloy Farhan, Farig Sadeque

Abstract Knowledge distillation (KD) is a standard approach for compressing sequence-to-sequence models, but its per-sample effects are rarely examined. On the BanSum Bangla summarization benchmark, we find that standard KD improves ROUGE-L by only +0.0003 over a cross-entropy baseline, and that approximately 51.3% of training samples are estimated to actively harm student validation loss under standard KD. We propose two complementary reliability-aware distillation methods. CHAD (Counterfactual Harm-Aware Distillation) measures per-sample KD usefulness via gradient alignment with the validation loss direction and trains a lightweight gate that generalizes this counterfactual judgment to the full training set. EWAD+CPDP combines token-level entropy-weighted adaptive distillation with a capacity-proportional geometric constraint from a second, vocabulary-incompatible teacher. On BanSum, both methods substantially outperform standard KD: CHAD by +0.0173 ROUGE-L and EWAD+CPDP by +0.0219 ROUGE-L, where standard KD itself improves ROUGE-L by only +0.0003; despite using only 60M parameters, both outperform a fine-tuned Qwen 2.5-3B model (50× larger). We further evaluate the stronger method, EWAD+CPDP, across 15 typologically diverse XL-Sum languages organised into three sets, beating the CE-only baseline on 10/15 languages; gains are most reliable where the two teachers contribute complementary signal, and weakest where they have saturated or jointly weak target-language coverage. We release code and trained models to support reproducibility and further research on selective distillation.

Paper 33 Generate, Verify, and Refine: Event Timelines for Robust Video-Grounded Dialogue Systems

Authors: Wiradee Imrattanatrai, Masaki Asada, Kimihiro Hasegawa, Susan Holm, Ken Fukuda, Teruko Mitamura

Abstract Video-grounded dialogue (VGD) requires models to generate coherent responses by tracking and reasoning over temporally evolving events in video content. However, existing approaches offer limited ability to capture temporal dependencies, particularly in complex scenarios. In this paper, we present a framework that integrates structured event timelines with verification-aware reasoning to improve temporal consistency in VGD. We represent video content as structured states and temporally ordered actions, enabling models to both infer event timelines from video segments and leverage them to guide response generation. In addition, to address the inherent noise in predicted structure timelines, we introduce a Generate-Verify-Refine (GVR) mechanism that explicitly evaluates and revises unreliable event predictions, thereby reducing the propagation of upstream errors. Experiments on the VDAct benchmark show that our approach consistently improves dialogue quality across models, achieving gains of up to +13.0 LAVE and accuracy over unstructured grounding. Human evaluation further demonstrates improvements of 20.7-31.9 points in response correctness. Additional analysis shows that while better event predictions correlate with stronger dialogue performance, naive reliance on structured outputs can degrade results, highlighting the need for verification-aware reasoning. Our implementation of GVR on event timelines for VGD is available at https://github.com/aistairc/GVR-VGD.

Paper 34 Argument Summarization: A Unified Task Definition, Taxonomy, and Survey

Authors: Zhiqiang Shi, Oana Cocarascu

Abstract Argument summarization seeks to generate concise representations of argumentative content while preserving argumentative meaning, yet the field remains conceptually fragmented. In this survey, we provide a unified task definition of argument summarization that encompasses both argumentative text summarization and key point analysis, and introduce a

taxonomy along input structure, output representation, and the role of argumentative structure. Using this framework, we analyze the evolution of the field and systematically review existing datasets, methods, and evaluation practices, revealing misalignment that hinders comparability and progress. We further situate argument summarization within the broader context of generic text summarization, identifying practical limitations and intrinsic challenges. Finally, we outline promising future research directions.

Paper 35 WikiOriginQA: Automated Construction of a QA Dataset with Knowledge Origins for Analyzing Cultural Bias

Authors: Kento Haneda, Yosuke Kishinami, Ryo Fujii, Keisuke Sakaguchi, Makoto Morishita

Abstract Existing datasets on cultural knowledge have been typically created by human annotators familiar with each culture. However, the process requires considerable human effort, presenting a significant challenge to scalability. To address this challenge, we introduce WikiOriginQA, a QA dataset automatically constructed by objectively representing the origins of knowledge through the creation dates of different Wikipedia language editions. Experimental results with the created dataset confirm that LLM performance tends to be biased toward Anglophone cultures. These findings demonstrate the validity of our automated approach, enabling cultural bias evaluations with superior scalability across diverse languages without extensive manual annotation.

Paper 39 Evaluating Contextually Mediated Factual Recall in Multilingual Large Language Models

Authors: Yihong Liu, Bingyu Xiong, Hinrich Schuetze

Abstract Large language models (LLMs) can recall a wide range of factual knowledge across languages. However, existing factual recall evaluations primarily assess fact retrieval in isolation, where the queried entity is explicitly named and the fact is requested directly. In natural language use, facts are often accessed through context, where the relevant entity is introduced only indirectly. In this work, we study contextually mediated factual recall, asking whether LLMs can reliably retrieve factual knowledge when the target entity is embedded in a naturalistic context rather than queried explicitly, across languages. We construct controlled prompts that preserve the underlying fact while introducing referential mediation through contextual sentences. To disentangle contextual effects from name-specific associations, we further compare performance using synthetic names and real names across languages. Evaluating multiple model families in five languages, we find that contextual mediation consistently degrades factual recall, with substantial variation across relations. Larger models are more robust to contextual mediation, exhibiting a reduced performance gap relative to direct queries, while the effect of real names and name origin is mixed and unsystematic. These findings highlight a gap between isolated factual recall and context-dependent language understanding in multilingual LLMs.

Paper 40 MortalMATH: Evaluating the Conflict Between Reasoning Objectives and Emergency Contexts

Authors: ~Etienne_Lanzeray1, ~Stephane_Meilliez1, ~Malo_Ruelle1, ~Damien_Sileo2

Abstract Large Language Models are increasingly optimized for deep reasoning, often through post-training objectives that strongly reward correct task completion. We ask whether this pressure toward correct completion can interfere with recognizing when a task should no longer be pursued. We introduce \textit{MortalMATH}, a benchmark of 150 scenarios in which users request algebra help while describing increasingly urgent situations, from mild distraction to life-threatening emergencies. We find a striking contrast across models: several reasoning-oriented models, including Qwen3-32B and GPT-5 variants, continue solving the mathematical task with high accuracy even under extreme emergency contexts, while other models increasingly interrupt the task or acknowledge the danger. Reasoning models also spend substantial computation on the original problem,

delaying safety-relevant information by several seconds. These results expose a failure mode of highly optimized task completion: training models to persist until they reach a verifiably correct answer may also make them less willing to stop when the surrounding context demands it.

Paper 42 FinDex: Detecting Help-Seeking Intent and Psychological-Economic Impact in Financial Distress Self-Disclosures

Authors: ~Priyansh_Singhal1, ~Piyush_Joshi1

Abstract Financial distress expressed on social media often signals genuine need for intervention, yet existing approaches treat such posts monolithically, using broad labels like "stress" or "no stress" that fail to distinguish between why individuals share their struggles (help-seeking intent) and what harms they experience (multidimensional impact). This paper introduces FinDex (Financial Distress Expression), a dataset of 2,408 Reddit posts annotated with a dual taxonomy: Intent (seven help-seeking behaviors) and Impact (six consequence domains). Grounded in digital intervention literature, this separation enables support systems to match interventions to both an individual's readiness for help and their specific constellation of needs. We benchmark 19 transformer models with 3-fold cross-validation and statistical significance testing, finding that domain-specific financial models significantly underperform ($p<0.05$) and Intent detection remains substantially harder than Impact (12-19% F1 gap). Evaluation of state-of-the-art LLMs (GPT-5, Claude, Gemini, Llama 4) reveals systematic over-labeling, demonstrating that FinDex enables capabilities that neither scaling nor prompting can achieve. We release the dataset and code publicly.

Paper 44 Escaping the Tokenizer Trap: Probabilistic Semantic Routing for Dense Retrieval in Specialized Domains

Authors: Siu Hin Ng

Abstract Dense retrievers can blur the distinctions between short, jargon-heavy records that share most of their vocabulary. We study Lex-TM, a training-free layer that interpolates dense cosine similarity with the Jensen–Shannon similarity of topic distributions from an LDA model whose topic–word prior is scaled by a Herfindahl–Hirschman (HHI) exclusivity score. On NTCIR-13 MedWeb, cast as retrieving tweets that share a query's exact symptom-label set, Lex-TM reaches an English MRR of 0.8151 against 0.7081 for BM25+dense reciprocal rank fusion, a significant paired difference; in Japanese and Chinese the differences from fusion are not significant. With Qwen3-Embedding-0.6B, routing improves on dense retrieval significantly in English and Chinese. Three findings bound these gains. SPLADE++ and BGE-M3's multi-vector mode match or exceed Lex-TM. In English and Chinese the gain depends on the label strings in the indexed tweets: without them, routing no longer improves on dense retrieval. The HHI prior brings no benefit: on short texts it becomes a rarity weight, and a symmetric prior performs as well across five Gibbs seeds. Outside MedWeb, routing matches dense retrieval on Chinese medical QA and lowers nDCG@10 on ChatDoctor, FiQA and long biomedical abstracts. Code, seeds and a changelog of fixed bugs are released.

Paper 51 ARR-Dv2: New Data from ACL Rolling Review and its Potential to Improve Reviewing in the ACL Community

Authors: ~Atnafu_Lambebo_Tonja1, ~Sheng_Lu1, ~Nils_Dycke1, ~Dirk_Hovy2, ~Koen_Dercksen1, ~Lizhen_Qu2, ~Xiaodan_Zhu1, ~Thamar_Solorio2, ~Iryna_Gurevych1, ~Margot_Mieskes1

Abstract We present a new dataset to support AI-assisted and human studies of the peer reviewing process, based on the ARR Open Review Platform. Unlike previous similar data collections, ours, ARR-Dv2, not only contains the reviews but also author-reviewer discussions (aka rebuttal phase) and meta-reviews. We present statistical analyses of the data with respect to the

overall score and the correlations among the overall score, confidence, and soundness scores. Additionally, we extend the task of Review Score Prediction also to include rebuttal data and analyze its effect on score prediction. Finally, we introduce a new task, Meta Review Score Prediction, based on a set of up to three reviews rather than a single review. Our initial results in a zero-shot setting indicate that the rebuttal data adds valuable information to the score prediction and enables reliable predictions.

Paper 52 Improving Japanese-English Patent Claim Translation via Reasoning and GRPO with Syntactic Reward

Authors: Shogo Tsujimoto, Katsuki Chousa, Masaaki Nagata, Ryohei Sasano

Abstract Patent claims are legal documents with distinctive formats and complex syntactic dependencies. Translating these claims requires not only accurately understanding the complex syntactic structure of the source claims but also ensuring the translation follows the structural conventions of claims in the target language. In this study, we leverage reasoning models to disentangle the complex syntactic structures of source claims. Furthermore, to ensure the translation reflects the structural requirements in the target language, we incorporate FastKASSIM---a syntactic similarity metric---into the reward of reinforcement learning, alongside conventional metrics such as BLEU and COMET. Experimental results confirm that our model, trained on Qwen3-8B, outperforms Qwen3-235B in syntactic similarity while simultaneously achieving comparable translation quality, despite the latter having approximately 30 times more parameters. Automatic and human evaluations show that models trained with rewards combining FastKASSIM and translation quality metrics such as COMET or COMETKiwi achieve strong performance, suggesting that syntactic rewards help improve patent-specific structural conformity while maintaining translation quality.

Paper 54 Attention-Guided Layer Selection for Contrastive Decoding in Large Language Models

Authors: Yusuke Sakai, Natthawut Kertkeidkachorn, Kiyoaki Shirai

Abstract Contrastive decoding methods such as DoLa improve the factuality of Large Language Models (LLMs) by contrasting the output distributions of mature and premature layers. However, DoLa's dynamic layer selection relies solely on divergences in output vocabulary distributions. In this work, we propose three attention-guided strategies—Attention-JSD, Attention-Entropy-Max, and Attention-Entropy-Min—that leverage structural information carried by internal self-attention mechanisms as a signal for layer selection. Experimental results on TruthfulQA demonstrate that our strategies, particularly Attention-JSD and Attention-Entropy-Min, consistently outperform the original DoLa. We observe significant gains on multi-answer metrics (MC2 and MC3), suggesting that attention distributions can provide a more sensitive signal for resolving factual knowledge than output vocabulary distributions.

Paper 56 Fairness in Prompting: How Demographic Variations Affect Language Model Outputs

Authors: Ryandito Diandaru, Gleb Kuzmin, Muhammad Dehan Al Kautsar, Zeerak Talat, Xudong Han, Haonan Li, Tatsuki Kuribayashi, Artem Shelmanov, Timothy Baldwin, Fajri Koto

Abstract LLMs are expected to respond consistently across demographic groups, yet this assumption remains largely untested due to the absence of demographic information in existing instruction datasets. To address this gap, we introduce PromptDial, a collection of 2,289 English prompts written by real users and annotated with seven demographic attributes: sex, race, education, age, language, employment sector, and nationality. We evaluate state-of-the-art generative models on 39 datasets, including machine translation, summarization, grammar correction, knowledge and reasoning, semantics, and question answering, and observe performance disparities of up to 7.7% between demographic groups, with statistically significant differences in over half

of the datasets. Our linguistic analysis points to variation in prompt tone and linguistic features as potential drivers of these disparities. Our findings suggest that current instruction tuning practices overlook key aspects of linguistic diversity, and we call for the inclusion of demographic metadata and more representative prompt data to support fairer and more robust language model behavior.

Paper 57 Episodic Memory from Compression Boundaries in Latent Representation Space

Authors: ~David_Oneil_Campos_Ferreira1, ~Priscila_Rocha_Maia_Freitas_Ribeiro1, ~EMANUEL_BORGES_PASSINATO1, ~Diogo_Fernandes_Costa_Silva1, ~Arlindo_Rodrigues_Galvão_Filho1

Abstract Long-term memory in Large Language Model (LLM) agents requires selective persistence: only a subset of interactions should be consolidated beyond the current context window. Existing memory systems rely on heuristic importance rules or similarity-based novelty, which remain external to the model's internal computation. We propose a geometric principle for memory formation: episodic memory can emerge from compression failure in latent representation space. We approximate the manifold of routine LLM activations using Sparse Autoencoders (SAEs) and define \emph{representational surprise} as reconstruction error relative to this learned manifold. Deviations from routine structure yield elevated residuals, providing an unsupervised, model-internal signal for memory writing. We demonstrate this principle through ReSuME, a surprise-gated memory mechanism that commits turns to memory only when normalized reconstruction error exceeds a calibrated threshold. In long-horizon multi-turn dialogue settings, representational surprise separates routine, critical, and out-of-distribution states, and achieves a superior performance–memory trade-off against established memory-agent systems (MemGPT, EM-LLM) and unsupervised anomaly detectors, while operating without labels and at substantially lower inference cost than LLM-judge controllers. Covariance-aware normalization further enables robust cross-domain calibration. These results suggest that episodic memory gating in neural agents can be grounded in intrinsic latent geometry rather than externally engineered rules.

Paper 59 Protocode: Prototype-Driven Interpretability for Code Generation in LLMs

Authors: ~bodla_Krishna_Vamshi1, ~Haizhao_Yang1

Abstract Large Language Models (LLMs) are now widely used for tasks such as summarization, question answering, and speech-to-text translation, and increasingly for code generation, with tools such as Cursor and Windsurf analyzing massive code repositories and recommending changes. Big tech companies have acknowledged growing reliance on LLM-generated code. Although this improves developer productivity, it also raises the risk of suboptimal and insecure code. We introduce a retrieval framework for code generation that connects geometry-aware prototype sampling with syntax-grounded interpretability analysis. By conditioning AST-based explanations on retrieved demonstrations, our method evaluates retrieval quality through both functional correctness and interpretability alignment. In our experiments, high-quality in-context learning (ICL) demonstrations make outputs easier to interpret and improve P@10, whereas poorly chosen demonstrations reduce P@10 relative to the base model. Overall, our approach highlights the importance of efficient sampling strategies for ICL, which can affect model performance on any given task.

Paper 60 North or South? Diagnosing LLMs' Understanding of North Korean Language, Society, and Knowledge

Authors: Kiwoong Park, Seyoung Song, Donghyun Woo, Alice Oh

Abstract Large Language Models (LLMs) have made substantial progress in multilingual question answering, yet their ability to understand low-resource sociolinguistic contexts remains underexplored. We study North Korea as a unique testbed where

decades of division have produced distinct differences from South Korea in vocabulary, orthography, norms, and ideological discourse. We introduce LENS (Language and sociocultural Evaluation for North Korean Society), a human-validated benchmark of 1,765 questions evaluating LLMs across three dimensions: linguistic robustness, sociocultural perspective alignment, and domain-specific knowledge. Evaluating diverse recent LLMs, we find a modest performance drop on North Korean language inputs, a larger gap in sociocultural perspective alignment, and severe limitations on domain-specific knowledge. Qualitative analysis reveals that models often conflate formal institutions with lived realities, project liberal-democratic assumptions, or rely on coarse stereotypes. Our results demonstrate that linguistic similarity alone does not guarantee sociocultural or domain-specific understanding, highlighting the critical need for grounded evaluation resources for underrepresented cultural and political contexts.

Paper 61 HALF: Harm-Aware LLM Fairness Evaluation Aligned with Deployment

Authors: Omar El Herraoui, Ali Mekky, Preslav Nakov, Yuxia Wang

Abstract Large language models (LLMs) are increasingly deployed across high-impact domains, from clinical decision support and legal analysis to hiring and education, making fairness and bias evaluation before deployment critical. However, existing evaluations lack grounding in real-world scenarios and do not account for differences in harm severity, e.g., a biased decision in surgery should not be weighed the same as a stylistic bias in text summarization. To address this gap, we introduce HALF (Harm-Aware LLM Fairness), a deployment-aligned framework that assesses model bias in realistic applications and weighs the outcomes by harm severity. HALF organizes evaluation datasets into harm tiers (Severe, Moderate, Mild) based on the specific task and bias type they measure. Harm severity is assigned according to deployment-relevant criteria grounded in regulatory and risk assessment frameworks. Our evaluation results across eight LLMs show that (1) LLMs are not consistently fair across datasets, (2) model size or performance do not guarantee fairness, and (3) reasoning models perform better in medical decision support but worse in education. We conclude that HALF exposes a clear gap between previous benchmarking success and model fairness. All data and code are available at https://github.com/mbzuai-nlp/HALF

Paper 65 Sparse Autoencoders Can Capture Language-Specific Concepts Across Diverse Languages

Authors: Lyzander Marciano Andrylie, Inaya Rahmanisa, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, Alham Fikri Aji

Abstract Understanding the multilingual mechanisms of large language models (LLMs) provides insight into how they process different languages, yet this remains challenging. Existing studies often focus on individual neurons, but their polysemantic nature makes it difficult to isolate language-specific units from cross-lingual representations. To address this, we explore sparse autoencoders (SAEs) for their ability to learn monosemantic features that represent concrete and abstract concepts across languages in LLMs. While some of these features are language-independent, the presence of language-specific features remains underexplored. In this work, we introduce $\textit{SAE-LAPE}$, a method based on feature activation probability, to identify language-specific features within the feed-forward network. We find that many such features predominantly appear in the middle to late layers of the model and are interpretable. These features influence the model's multilingual performance and language output, and can be used for language identification with performance comparable to fastText, along with more interpretability. Our code and complete figures are available [here](https://github.com/LyzanderAndrylie/language-specific-features).

Paper 69 Ask Twice: Bidirectional Cross-Encoder Scoring for Faithfulness Evaluation

Authors: Arnav Mantro, Leslie Barrett

Abstract Current LLM-based question answering systems provide cited evidence for their claims. However, determining whether each claim within the generated responses is fully supported by its cited evidence is still an ongoing research problem. The task of evaluating attribution quality, or faithfulness to the source text, goes beyond the traditional idea of lexical matching used for relevance in early QA tasks, and the use of lexical models based on semantic distance or word-overlap has declined in favor of NLI-based models and large fine-tuned models. The larger fine-tuned models tend to perform best on current special-purpose benchmarks such as AttributionBench despite struggling with OOD data. Still, exactly why human judgments of faithfulness should align better with LLM evaluators than those based on lexical entailment has never been fully explained. We address the problem with current entailment models by introducing bidirectionality and codifying that idea into a new evaluation modeling approach. We show that this approach is both lightweight and extensible yet competitive with SOTA LLM-based methods, avoiding the pitfalls of both standard NLI and word-overlap approaches. We further show that our model architecture is motivated by a deeper understanding of the gap between lexical distance and faithful claim attribution, exposing the conceptual mechanism behind attribution quality evaluation.

Paper 71 Multi-Perspective LLM Annotations for Valid Analyses in Subjective Tasks

Authors: Navya Mehrotra, Adam Visokay, Kristina Gligorić

Abstract Large language models are increasingly used to annotate texts, but their outputs reflect some human perspectives better than others. Existing methods for correcting LLM annotation error assume a single ground truth. However, this assumption fails in subjective tasks where disagreement across demographic groups is meaningful. Here we introduce Perspective-Driven Inference, a method that treats the distribution of annotations across groups as the quantity of interest, and estimates it using a small human annotation budget. We contribute an adaptive sampling strategy that concentrates human annotation effort on groups where LLM proxies are least accurate. We evaluate on politeness and offensiveness rating tasks, showing targeted improvements for harder-to-model demographic groups relative to uniform sampling baselines, while maintaining coverage.

Paper 73 Agent+P: Guiding UI Agents via Symbolic Planning

Authors: Shang Ma, Xusheng Xiao, Yanfang Ye

Abstract Large Language Model (LLM)-based UI agents show great promise for UI automation but of- ten hallucinate in long-horizon tasks due to their lack of understanding of the global UI transition structure. To address this, we in- troduce AGENT+P, a novel framework that leverages symbolic planning to guide LLM- based UI agents. Specifically, we model an app's UI transition structure as a UI Transi- tion Graph (UTG), which allows us to refor- mulate the UI automation task as a pathfinding problem on the UTG. This further enables an off-the-shelf symbolic planner to generate a provably correct and optimal high-level plan, preventing the agent from redundant explo- ration and guiding the agent to achieve the au- tomation goals. AGENT+P is designed as a plug-and-play framework to enhance existing UI agents. Evaluation on the AndroidWorld benchmark demonstrates that AGENT+P im- proves the success rates of state-of-the-art UI agents by up to 14.32% and reduces the ac- tion steps by 37.70%. Our code is available at: https://github.com/KiteFlyKid/agentp.

Paper 75 Twenty-Five Years of Inter-Annotator Agreement in NLP: A Systematic Survey of Trends, Metrics, and Reporting Practices

Authors: Hend Al-Khalifa

Abstract Large language models are now deployed as annotators across NLP, generating training labels and evaluation scores, yet the field has no established standard for verifying whether LLM-generated labels are reliable, internally consistent, or comparable across studies. This methodological gap does not begin with LLMs: a systematic corpus study of 665 ACL Anthology papers from 1999 to 2026 reveals that even classical human annotation in NLP suffers from systematic measurement and reporting failures that have left a growing body of work scientifically incomparable. Papers were identified through keyword search and organised via topic modelling into a thematic taxonomy spanning eight annotation domains from named entity recognition to disagreement modelling. Full-text analysis of 549 papers yields three principal findings. First, Cohen's kappa dominates across all task types (259 papers, 38.9%), regardless of whether its assumptions are satisfied; ICC, appropriate for continuous-scale tasks such as MT evaluation, appears in only 16 full-text papers (2.4%). Second, 226 of the 665 papers (34.0%) name no inter-annotator agreement (IAA) coefficient anywhere in the full text, which is a methodology-reporting concern; a further 50.1% of papers name a coefficient in the body but not in the title or abstract, so 84.1% of papers omit the metric from the abstract layer. We frame the body-only gap as a discoverability concern (the methodology is reported but is not visible at the abstract layer readers and reviewers encounter first), distinct from the full-text omission. Third, annotator disagreement modelling and LLM-as-annotator constitute the fastest-growing area of the corpus (124 papers, 18.6%): 26 of these appeared between 1999 and 2022 and 98 between 2023 and 2026, the only dimension whose volume rises in that window. Fewer than 5% of papers report confidence intervals. We provide a large-scale empirical analysis of metric reporting practice within the ACL Anthology and a seven-point agenda for standardised reporting, complementing prior normative guidance with field-scale descriptive evidence.

Paper 78 Auto-AdvER: Disentangling Condition, History, and Commercial Offers in Automotive Advertisements

Authors: Filippos Ventirozos, Ioanna Nteka, Tania Nandy, Jozef Bača, Charlie Roadhouse, Peter A. Appleby, Matthew Shardlow

Abstract This paper presents a case study on the development of Auto-AdvER, a specialised span-labelling schema, along with its corresponding dataset for the textual car advertisement domain. Developed with industry needs in mind, Auto-AdvER is designed to enhance text mining analytics in this domain and contributes a linguistically unique sequence-labelling dataset. We introduce a schema consisting of three labels: 'Condition', 'Historic', and 'Sales Options'. We outline the guiding principles for annotation, describe the methodology for schema development, and present the results of an annotation study demonstrating an inter-annotator agreement of 78%. Furthermore, we compare the performance of various models, including transformer encoder models (BERT, DeBERTaV3) and transformer decoder-based open- and closed-source large language models: Llama, Qwen, GPT, and Gemini, as well as conditional random fields. Our results show that a fine-tuned DeBERTaV3 encoder achieves the highest F1-score, followed closely by conditional random fields, while the large language models reach the highest precision but lower recall. We present this work as a stepping stone towards more fine-grained analysis and discuss Auto-AdvER's potential impact on advertisement analytics, pricing, and customer insights. Potential applications include the analysis of market dynamics and data-driven predictive maintenance.

Paper 79 Paying for Too Many Tokens? Valid and Cost-Efficient Multimodal LLM Annotation with Simple Heuristics

Authors: Zhixi Zhu, Kristina Gligorić

Abstract Vision-Language Models (VLMs) enable video annotation at scale, but costs accumulate quickly: processing a typical 60-second short-form video at one frame per second requires millions of tokens. To reduce costs, researchers rely on heuristics such as sampling a subset of frames, compressing videos into image grids, or using only a single modality. However, it remains unclear which heuristics save cost, and whether they preserve the downstream conclusions these annotations enable. To address this gap, we conduct a systematic evaluation of these heuristics using short-form videos, on two computational social science (CSS) tasks: sentiment and topic classification. We evaluate each configuration along three axes the literature typically treats separately: classification accuracy, validity of downstream inference, and per-video token cost. First, we find that accuracy and validity diverge: the highest-accuracy configuration can produce wrong conclusions. Second, modality value is not guaranteed: text alone can yield strong performance, indicating that adding modalities can add cost without adding signal. Finally, we find that cost can be decoupled from video length when annotating short-form videos: a single 2x8 image grid built via simple shot-transition detection approaches full-video understanding (kappa within~.05), at ~15% of the token cost. Based on these findings, we derive guidelines that can enable cost-aware VLM annotation in CSS.

Paper 81 DBRAG: Multi-Table Retrieval-Augmented Generation for Complex Database Queries

Authors: Prince Larbi Ampofo, Ryoji Kubo, Djellel Difallah

Abstract Recent advancements in large language models have introduced new capabilities for reasoning over structured data, particularly through program-aided tools that can analyze tables. However, many existing methods address single-table scenarios or assume that the relevant tables are already provided. In practice, users often issue complex data exploration queries over entire databases, where relevant information may be distributed across multiple relations. In this work, we introduce DBRAG, a retrieval-augmented generation framework tailored for multi-table question answering. DBRAG first retrieves candidate tables using an offline table index, enriches their summaries with query-relevant rows, and uses an LLM to rerank the candidates. A program-aided reasoner then selects the required tables and executes operations over their full contents, keeping the initial prompt context compact. Experiments on the Spider, GeoQuery, and ATIS datasets used in this study demonstrate improvements in table retrieval and multi-table question answering.

Paper 86 Breaking the Silence: A Dataset and Benchmark for Bangla Text-to-Gloss Translation

Authors: Sharif Mohammad Abdullah, Abhijit Paul, Shubhashis Roy Dipta, Zarif Masud, Shebuti Rayana, Ahmedul Kabir

Abstract Gloss is a written approximation that bridges Sign Language (SL) and its corresponding spoken language. Despite a deaf and hard-of-hearing population of at least 3 million in Bangladesh, Bangla Sign Language (BdSL) remains largely understudied, with very limited prior work on Bangla text-to-gloss translation and no publicly available parallel dataset at the time of this work. To address this gap, we construct the first publicly available Bangla text-to-gloss dataset, consisting of 1,000 manually annotated and 4,000 synthetically generated Bangla sentence–gloss pairs, along with 159 expert human-annotated pairs used as a test set. Our experimental framework performs a comparative analysis between several fine-tuned open-source models and a leading closed-source LLM to evaluate their performance in low-resource BdSL translation. GPT-5.4 achieves the best automatic metric scores, while open-source models perform competitively despite being much smaller. This work introduces the first publicly available expert-annotated Bangla text-to-gloss dataset, benchmark and trained model for Bangla text-to-gloss

translation. It also demonstrates the effectiveness of systematically generated synthetic data for addressing challenges in low-resource sign language translation.

Paper 87 BusMA: A Bus Communication Substrate for Multi-Agent Systems

Authors: Yanwen Peng, Delvin Ce Zhang, Xi Wang, Nikolaos Aletras

Abstract Multi-Agent (MA) systems are effective at solving complex tasks that demand planning, tool use, and the synthesis of evidence from multiple sources. Existing systems typically adopt Hierarchical Manager-Worker (HMW) or Router-based Message Passing (RMP) structures as their communication protocol. However, these designs restrict agent autonomy: Worker agents cannot directly consult specific ``peers'', and misrouted messages can propagate errors. Inspired by bus architectures in computer systems, we propose BusMA, a communication framework that allows any agent to address other agents through a shared channel, i.e., the Bus. It consists of agent registration, message routing, and shared memory management components. Worker agents, each equipped with tools, have their own local memory and can reason, act (tool usage), and communicate by posting shared messages with specific intents. We introduce four intents: discussion, challenge, guidance, and request for explanation, which support fine-grained communication among agents. A Chair agent monitors the shared memory to coordinate interactions and facilitate convergence among Workers. To evaluate the effectiveness of BusMA, we conduct extensive experiments with two frontier LLMs across 13 tasks spanning visual reasoning, mathematical reasoning, and knowledge retrieval. The results demonstrate that BusMA consistently outperforms state-of-the-art HMW and RMP methods.

Paper 88 DOSA: A Tree-Guided, Self-Regressive Framework for Long Document Structure Analysis

Authors: Bohou Li, Benjamin Sowell, Mehul A Shah, Mark Lindblad, Henry Lindeman

Abstract In visually-rich documents, information is encoded not only in individual page objects such as tables, headers, and text blocks, but also in the structural relations among them, making document structure analysis fundamental to information retrieval and document understanding. However, accurately inferring such relations remains challenging in multi-page documents with long-range dependencies and heterogeneous layouts. To address this, we propose a tree-guided and self-regressive framework, termed $\textbf{DOcument Structure Analyzer (DOSA)}$, for inferring relations among page objects and reconstructing document-level semantic trees. DOSA processes documents chunk-by-chunk, fusing visual, textual, and layout features for each page object and predicting hierarchical and ordering relations. The predicted relations are used to incrementally construct a semantic tree, which is then leveraged as structural context to guide inference on subsequent chunks. Experimental results on five benchmarks demonstrate the effectiveness of DOSA, with improvements of up to 4 F1 points and 19 TEDS points on DocHieNet, the most challenging multi-page hierarchy benchmark.

Paper 90 VIDA: A Dataset for Visually Dependent Ambiguity in Multimodal Machine Translation

Authors: Jingheng Pan, Xintong Wang, Longyue Wang, Liang Ding, Weihua Luo, Chris Biemann

Abstract Ambiguity resolution is a key challenge in multimodal machine translation (MMT), where models must genuinely leverage visual input to map an ambiguous expression to its intended meaning. Although prior work has proposed disambiguation-oriented benchmarks probing the role of vision, we observe that existing benchmarks remain limited by task-format mismatch, narrow ambiguity coverage, or insufficient visual-dependency validation. Moreover, existing ambiguity evaluations are not well suited to diverse ambiguity types in open-ended translation. To address these limitations, we present VIDA (Visually-Dependent Ambiguity), a dataset of 2,500 carefully curated instances in which resolving an annotated source span requires visual evidence. We further propose Disambiguation-Centric Metrics that use an LLM-as-a-judge classifier to verify whether annotated ambiguous expressions are resolved correctly at the span level. Evaluations with stronger recent LVLMs

show that visual disambiguation remains challenging. Using chain-of-thought supervised fine-tuning as a diagnostic setting, we observe stronger out-of-distribution disambiguation than with SFT, with robust gains on collective-noun ambiguities and model-dependent gains on sentence-level ambiguities.

Paper 96 Supportive Token Revealing for Fast Diffusion Language Model Decoding

Authors: Giries Abu Ayoub, Mario Barbara, Lluís Pastor-Pérez, Tanja Bien, Aneesh Barthakur, Alaa Maalouf, Loay Mualem

Abstract Discrete diffusion language models can generate text efficiently by updating multiple masked positions in parallel, but this parallelism introduces a quality-latency trade-off. Aggressive decoding may commit mutually dependent tokens too early, while conservative decoding requires many denoising steps. Existing methods address this tension by deciding which tokens are safe to reveal using confidence or dependency criteria. However, avoiding unsafe commits does not necessarily make the remaining masked sequence easy to decode, since uncertain tokens may depend on masked tokens, creating a bottleneck for denoising steps. We propose AXON, a training-free module that can be added on top of existing parallel decoding strategies for diffusion language models. Rather than replacing the base decoder, AXON monitors the remaining uncertain masked tokens and intervenes only when their current state suggests that additional context is needed. It then shifts the criterion from which tokens are safest to reveal to which confident reveals would best support later denoising. AXON selects anchors, confident masked tokens that uncertain positions attend to, using attention, uncertainty, and confidence signals. Experiments on reasoning and code-generation benchmarks across multiple diffusion language models show that AXON improves the quality-latency trade-off of existing parallel decoders, often reducing the number of function evaluations while maintaining or improving accuracy.

Paper 107 Token-Level On-Policy Contrastive Learning for Black-box Distillation Without Full Rollouts

Authors: ~Jian_Yang43, ~Chern_Hong_Lim1, ~Thuy-Trang_Vu1

Abstract Black-box distillation trains student models using only teacher-generated outputs, making it practical when teacher models are accessible only through APIs. However, existing methods either rely mainly on positive imitation or require costly full-response rollouts to obtain corrective signals. We show that such corrective signals can be obtained locally from teacher trajectories alone. At each position, the teacher token provides a positive anchor, while a high-probability alternative sampled from the student provides an informative negative. Based on this insight, we propose Token-level On-policy contrastive Learning for blackbox Distillation (TOLD), which samples a single negative token from the student's distribution and optimizes a Bradley-Terry objective anchored by negative log-likelihood. This enables the student to reinforce teacher tokens while suppressing its own likely mistakes, without teacher logits, reward models, or full-response rollouts. Experiments on reasoning and instruction-following tasks show that TOLD outperforms or matches strong black-box distillation baselines and achieves up to 1.88 times training speedup over rollout-based on-policy baselines.

Paper 108 A Survey of the Training Process of LLM-Based Agents

Authors: Zhisong Zhang, Lemao Liu, Tianqing Fang, Deng Cai, Jiahao Xu

Abstract Autonomous agents based on large language models (LLMs) are becoming an increasingly prevalent paradigm for tackling complex and real-world tasks. Despite the remarkable zero-shot capabilities of modern LLMs, specialized agent training is often essential to obtain reliable and improved performance for specific target tasks. In this survey, we provide a process-centric review of the training of LLM-based agents, with a primary focus on methods that update the parameters of the main agent policy. Rather than organizing prior work only by application domain, agent capability, or learning algorithm, we examine the end-to-end pipeline through which LLM-based agents are trained: environment setup, data and trajectory construction,

learning-signal design, and optimization objectives and schemes. This perspective highlights how design choices at different stages interact with one another. Finally, we conclude with a discussion on potential future directions.

Paper 109 CHASE: Adversarial Red-Blue Teaming for Improving LLM Safety using Reinforcement Learning

Authors: Rahul Markasserithodi, Aditya Joshi, Yuekang Li, Ishmanbir Singh, Chris Yoo, Alan Niu

Abstract Despite advances in safety alignment, prompt-rewriting attacks such as persona modulation, fictional framing and persuasion-based reformulation, can bypass safety filters even on frontier models. Existing defenses either rely on non-scalable human curation or white-box optimisation that overfits to specific model internals, leaving aligned models brittle against the very class of adaptive black-box adversaries they will face in deployment. To address this gap, we introduce CHASE (Co-evolutionary Hardening through Adversarial Safety-Escalation), a closed-loop red-blue teaming framework in which a black-box attacker and a safety-aligned defender co-evolve. The attacker is trained via Group Relative Policy Optimization (GRPO) under a multiplicative reward that jointly enforces bypass effectiveness and intent fidelity, while the defender is hardened on the harvested adversarial rewrites through a two-stage GRPO + rejection-sampled SFT pipeline balanced with benign data. Evaluated on BeaverTails and JailbreakBench against five held-out attack families (PAIR, TAP, AutoDAN, PAP, Translation), CHASE cuts mean StrongREJECT score by 43.2\% on Llama-3.1-8B-Instruct with 0\% false refusal on benign (Alpaca) prompts, at a measured helpfulness cost of a 55.6\% refusal rate on XSTest safe prompts and a 1.92-point MT-Bench reduction. A controlled ablation that varies only the training attack distribution shows that this generalisation comes from the template-free attack distribution rather than the pipeline, and the result replicates on Qwen2.5-7B-Instruct. Across the two similarly sized open models we study, these results suggest that template-free RL exploration recovers latent attack primitives that transfer across mechanistically distinct attack families.

Paper 110 Iterative Critique-Refine Framework for Enhancing LLM Personalization

Authors: Durga Prasad Maram, Dhruvin Gandhi, Zonghai Yao, Gayathri Akkinapalli, Franck Dernoncourt, Yu Wang, Ryan A. Rossi, Nesreen K. Ahmed

Abstract Personalized text generation requires models not only to produce coherent text but also to align with a target user's style, tone, and topical focus. Existing retrieval-augmented approaches such as LaMP and PGraphRAG enrich profiles with user and neighbor histories, but they stop at generation and often yield outputs that drift in tone, topic, or style. We present $\textbf{PerFine}$, a unified, training-free critique–refine framework that enhances personalization through iterative, profile-grounded feedback. In each iteration, an LLM generator produces a draft conditioned on the retrieved profile, and a critic LLM - also conditioned on the same profile - provides structured feedback on tone, vocabulary, sentence structure, and topicality. The generator then revises, while a novel knockout strategy retains the stronger draft across iterations. We further study additional inference-time strategies such as Best-of-N and Topic Extraction to balance quality and efficiency. Across Yelp, Goodreads, and Amazon datasets, PerFine consistently improves personalization over PGraphRAG, with G-Eval gains of +8.5–12\%, steady improvements over 3–5 refinement iterations, and scalability with increasing critic size. These results highlight that post-hoc, profile-aware feedback offers a powerful paradigm for personalized LLM generation that is both training-free and model-agnostic.

Paper 115 The Curious Case of Curiosity across Cultures: Evaluating Cross-Cultural Information-Seeking Questions in Humans and LLMs

Authors: Angana Borah, Zhijing Jin, Rada Mihalcea

Abstract Recent advances in Large Language Models (LLMs) have expanded their role in human interaction, yet information-seeking behavior, particularly across cultural contexts, remains underexplored in these systems. In this work, we study cross-cultural variation in information-seeking using question-asking as a mechanism. We introduce IQUEST (Information-seeking QUestion Evaluation across SocieTies), an evaluation framework that measures human-model alignment in information-seeking question behavior through linguistic (style), topic preference (content) analysis and grounding insights in social science constructs. Across open- and closed-source models, we find that LLMs flatten cross-cultural diversity, aligning more closely with question-asking patterns expressed in Western countries. We also explore fine-tuning strategies, which narrow the human-model alignment gap by 43% overall. Finally, we demonstrate that stronger information-seeking question behavior improves downstream cultural adaptability, highlighting its importance for inclusive NLP research.

Paper 118 EnComp: Lightweight Encoder-Only Context Compression for Retrieval-Augmented Question Answering

Authors: Thao Do, Dinh Phu Tran, An Vo, Seon Kwon Kim, Daeyoung Kim

Abstract Efficient context compression is critical for retrieval-augmented question answering in resource-constrained settings, where long retrieved contexts increase latency, memory use, and LLM reader cost. We propose a lightweight encoder-only framework for query-driven sentence pruning that preserves answer-critical evidence while aggressively reducing irrelevant context. Our method learns marginal contribution scores for sentences using counterfactual training signals and optimizes a contrastive ranking objective that separates critical evidence from noncritical context. Our approach scores all sentences from a single full-context encoding, enabling fast inference with low computational overhead. Experiments show that it maintains accuracy comparable to the strongest baseline while using 3.7$\times$ less peak memory and achieving nearly 3$\times$ lower compression latency, demonstrating an

Paper 119 Mitigating the Impact of Glitch Tokens via Targeted Retokenization

Authors: Muhammad Alif Al Hakim, Dmitrii Kuzmin, Carlos Javier Galán Zavala, Tatsuya Hiraoka, Munachiso Samuel Nwadike, Ayana Niwa, Marco Cognetta, David Pohl, Kentaro Inui

Abstract This paper examines whether avoiding undertrained or “glitch” tokens and maximizing tokenization quality during inference can improve downstream task performance without retraining the model. We evaluate this question using a retokenization method that selects the highest-quality tokenization based on the trainedness of tokens, which is defined through the norm or entropy of token embeddings. We compare the performance difference between the tokenization that includes undertrained tokens and the one that avoids such tokens on multiple benchmarks. Our findings reveal a key nuance: undertrained tokens degrade performance primarily when they occupy semantically salient positions that are critical for solving the task. Consequently, retokenization preserves performance on ordinary inputs, while yielding consistent performance improvements when such tokens appear in high-importance contexts.

Paper 120 \textit{Learn To Say NO!} Incorporating Model Abstention In Medical Vision-Language Models for Pharmacovigilance

Authors: Sofia Jamil, Chinmay Purushottam Bhat, Sooraj Veer R, Sriparna Saha

Abstract Abstention is the intentional refusal of models to provide an answer. Adverse drug reactions (ADRs), which range from mild to life-threatening, are often underreported by patients in hospitals and to doctors because patients assume these are unavoidable. With the growing adoption of LLMs and vision language models (VLMs) in healthcare, patients are increasingly reaching out to these models for solutions. In such high-risk scenarios, incorrect or overconfident responses can cause unnecessary anxiety, lead patients to discontinue essential medications, or encourage inappropriate self-medication. In this work, we propose the $\textbf{\textit{MAD-MED}}$, a $\textbf{M}$odel $\textbf{A}$bstention approach for $\textbf{D}$rug Adverse Reaction for $\textbf{MED}$ical-VLMs. It is a training strategy to encourage models to abstain when answering would contradict the available input, exceed the model's knowledge, or violate model ethics. In addition, we release an augmented $\textbf{A}$dverse $\textbf{D}$rug $\textbf{R}$eaction $\textbf{ABS}$tention ($\textit{ADR-ABS}$) dataset comprising image-answerable, text-answerable, and unanswerable questions to support systematic evaluation of abstention behavior. Quantitative and human evaluations confirm that our proposed $\textit{MAD-MED}$ approach improved the abstention capabilities of models considerably while keeping the generative capabilities intact in cases where the model is expected to answer.

Paper 122 A Multilingual Evaluation of LLM-Based High-School Mathematics Assistance Across Indian Languages

Authors: Yash Kumar, Subhajit Roy

Abstract Large Language Models (LLMs) are increasingly used by students as informal assistants for problem solving and concept clarification. Despite growing adoption, the multilingual reliability of these systems for educational mathematics remains under-explored, particularly across Indian regional languages. In this work, we present a multilingual evaluation of contemporary LLMs on high-school mathematics problems across six Indian languages: English, Hindi, Bengali, Urdu, Gujarati, and Malayalam. To guide the design of lightweight prompting conditions, we conduct a small exploratory prompt elicitation exercise with Grade 11--12 students and observe that students predominantly use concise direct-solution requests and lightweight expertise-oriented phrasing. Guided by these observations, we evaluate GPT-4o-mini, LLaMA 3-8B, and DeepSeek-R1 using simplified student-inspired prompting strategies on a parallel corpus of numerically perturbed textbook mathematics problems. Our results reveal consistent multilingual performance disparities, with accuracy strongest in English and lower across several regional languages. We further analyze the effects of prompting strategy and mathematical topic, and provide qualitative error analysis identifying recurring arithmetic, reasoning, and regional-language comprehension failures. These findings highlight the importance of evaluating multilingual educational reliability as LLM-based learning assistance becomes increasingly accessible across diverse linguistic settings.

Paper 123 CLFEC: A New Task for Unified Linguistic and Factual Error Correction in paragraph-level Chinese Professional Writing

Authors: Jian Kai, Zidong Zhang, Jiwen Chen, Zhengxiang Wu, Songtao Sun, Fuyang Li, Yang Cao, Qiang Liu

Abstract Chinese text correction has traditionally focused on spelling and grammar, while factual error correction is usually treated separately. In paragraph-level Chinese professional writing, linguistic (word/grammar/punctuation) and factual errors can co-occur and interact. Moreover, editorial review leaves few observable errors in published professional texts, making controlled

benchmark construction important. This paper introduces **CLFEC (Chinese Linguistic \& Factual Error Correction)**, a new task for joint linguistic and factual correction. We construct a mixed, multi-domain Chinese professional writing dataset spanning current affairs, finance, law, and medicine. We then conduct a systematic study of LLM-based correction paradigms, from prompting to retrieval-augmented generation (RAG) and agentic workflows. The analysis reveals practical challenges, including limited transfer of a specialized corrector, the need for evidence grounding for factual repair, the difficulty of mixed-error paragraphs, and over-correction on clean inputs. Results further show that handling linguistic and factual errors within the same context outperforms decoupled pipelines, and that agentic workflows can be effective with suitable backbone models. Overall, CLFEC provides a new benchmark for Chinese text correction research and practical guidance for proofreading systems. The dataset is publicly available at https://github.com/jiu2021/CLFEC-Dataset.

Paper 133 When the Correct Law Still Fails: A Diagnostic Benchmark for Law-in-Force Retrieval in Versioned Legal RAG

Authors: Yuchen Yao, Yiran HU, Kaixiang Zhang, Yixin Su, Jiani Ge, Yun Liu

Abstract Legal RAG systems often retrieve passages by semantic relevance, lexical score, or citation match. In versioned legal corpora, however, a passage can be topically relevant and citation-matched while being legally invalid for the operative date. We study law-in-force retrieval: retrieving evidence that matches the authority, is valid at the operative date, and supports the answer. We build LawInForce-eCFR from 805 human-validated eCFR/CFR amendment pairs and 3,220 controlled answer-flipping yes/no questions. We also present AuthorityFirst-RAG, a transparent diagnostic retriever that decomposes retrieval into authority identification, temporal scoping, and rule-support ranking. AuthorityFirst-RAG improves Recall@1 from 2.80% for BM25-date-filter to 30.90%, but generation remains difficult even with GoldOracle evidence. These results show a dual bottleneck in versioned legal RAG: retrieving the law in force and applying numeric legal rules.

Paper 136 FahMai-TeleBench: A Thai Grounded Tool-Use Benchmark for the Deployable Model Tier

Authors: Nutchanon Yongsatianchot, Piyalitt Ittichaiwong, Jitkapat Sawatphol, Tichanat Ruangsiri

Abstract Multilingual agentic benchmarks remain rare, especially for under-resourced languages such as Thai, where nicknames, unspaced script, and fused honorifics change how an agent must query its tools. We present FahMai-TeleBench, a bilingual (Thai/English) benchmark for grounded tool use. A model answers employee-directory questions over a synthetic company directory by issuing tool calls, or refuses when the answer is absent. All 626 items are synthetic but modeled on a real Thai enterprise-directory deployment, and they are graded deterministically without an LLM judge. Our central design choice is to treat tool availability as an experimental variable: each item runs under four tool configurations (grep, structured search, both, and a Python REPL), so the benchmark measures not just whether a model answers but how it uses tools. The task is near-solved at the frontier (GPT-5.4 97–99%), yet sharply discriminating for the Thai-capable models an organization would host itself (Typhoon-2.5 58–74%, Typhoon-S-8B 25–37%). The Thai penalty is model-specific, and Typhoon-2.5 answers 90% of items correctly once the relevant records are placed in its context, so its failures are failures of tool use. The tool axis also exposes differences at the frontier: GPT-5.5 drops to 73% with a wide-schema structured search tool while scoring 97–98% with grep and the REPL.

Paper 137 Composing Within Constraints: A Benchmark for Thai Classical Poetry in Large Language Models

Authors: Nutchanon Yongsatianchot, Phuvis Kerdpramote

Abstract Thai is spoken by tens of millions, yet its literary tradition is largely absent from LLM evaluation. We present ThaiProsody, a benchmark for composing Thai classical poetry. Composing classical Thai verse means choosing words that jointly satisfy several hard formal constraints (fixed syllable counts, written tone-mark positions, heavy/light syllable patterns, and an interlocking web of rhymes) while staying on a given topic. ThaiProsody crosses four canonical verse forms with 40 topics and grades each poem with an open, deterministic rule checker for tone, weight, and rhyme, a syllable counter validated against hand-segmented gold, and a separate content judge. Across 17 models in 24 configurations, the task is far from solved. The best model passes 94.7% of rule checks, yet one stanza in four still breaks a rule; the best open-weight model writes a fully valid stanza only a quarter of the time, and no Thai-trained model more than once in 160 tries. Rhyme is the most frequently broken constraint, and khlong and chan, the forms that add tone or weight constraints, are the hardest. Thai-trained 8B models stay close to their general base model, while a general 31B model scores more than three times as high as a Thai 30B model.

Paper 138 ThaiEditBench: A Mechanism-Typed, Length-Stratified Benchmark for Thai Orthographic Error Correction

Authors: Nutchanon Yongsatianchot, Piyalitt Ittichaiwong, Kanyakorn Veerakanjana

Abstract Thai is written without spaces between words, and its spelling depends on tone marks, silent consonants, and the transliteration of many loanwords. Each of these is a common source of misspellings. We present ThaiEditBench, a benchmark for correcting Thai spelling errors at the character level. It has three components: (i) a six-class taxonomy that types each error by the part of the spelling that goes wrong, plus an etymology tag; (ii) three length tiers (sentence, paragraph, and page) that test whether a model stays accurate over long, mostly correct text; and (iii) a contamination-robust construction, in which naturally attested misspellings are injected into clean text, so the corrupted passages do not appear verbatim in the source corpora. Because the errors are injected, the gold corrections are exact, and a blind two-annotator study measures how reliably the taxonomy can be applied. Grading is at the level of Thai character clusters, with span-based matching adapted from ERRANT. We evaluate 16 models and find that (1) under the benchmark's conditions, a model that locates an error almost always fixes it correctly, so what separates models is detection; (2) the evaluated Thai-adapted models do not lead the strongest general models; and (3) in long, mostly correct passages, weaker models accumulate over-edits that pull their scores far below their sentence-level results. On real errors mined from Wikipedia edit history, scores are somewhat lower, but the model ranking largely holds.

Paper 139 EvidenceRL: Reinforcing Evidence Consistency for Trustworthy Language Models

Authors: J. Ben Tamo, Yuxing Lu, Benoit Louis Marteau, Micky C. Nnamdi, May Dongmei Wang

Abstract Large Language Models (LLMs) are fluent but prone to hallucinations, producing answers that appear plausible yet are unsupported by available evidence. This failure is especially problematic in high-stakes domains where decisions must be justified by verifiable information. We introduce \textbf{EvidenceRL}, a reinforcement learning framework that enforces evidence adherence during training. EvidenceRL scores candidate responses for grounding (entailment with retrieved evidence and context) and correctness (agreement with reference answers) and optimizes the generator using Group Relative Policy Optimization (GRPO). We evaluate across two high-stakes domains, cardiac diagnosis and legal reasoning, where EvidenceRL consistently improves evidence grounding and faithfulness without sacrificing task accuracy. On cardiac diagnosis, F1@3 increases from 37.0 to 54.5 on Llama-3.2-3B while grounding ($G_{max}@3$) rises from 47.6 to 78.2; hallucinations drop nearly

5x and evidence-supported diagnoses increase from 31.8% to 61.6%. On legal reasoning, EvidenceRL raises Faithfulness from 32.8% to 67.6% on Llama-3.1-8B, demonstrating consistent behavioral change across domains.

Paper 140 Diagnosing and Mitigating Script-Based Representational Divergence in Multilingual Encoders

Authors: Henry Gagnier

Abstract Pretrained multilingual models are assumed to learn representations invariant to surface form, yet the script-based representational divergence and its consequences for downstream tasks remain poorly understood. We systematically diagnose and mitigate this divergence in six typologically diverse language pairs, which are Kazakh, Serbian, Chinese, Uyghur, Japanese, and Hindi--Urdu, and probe five multilingual encoders, which are XLM-R, ByT5, mT5, LaBSE, and Qwen3-Embedding, through layer-wise cosine similarity, script cluster analysis, masked language model probing, and cross-script retrieval. XLM-R cross-script retrieval Mean Reciprocal Rank (MRR) falls below 0.05 for Kazakh, Uyghur, and Hindi--Urdu, and a consistent gap in silhouette and cosine similarity scores for Kazakh and Serbian is consistent with pretraining data volume as a contributor to divergence beyond tokenization. LaBSE's retrieval-specialized training achieves a mean MRR of 0.905 across all six language pairs, with Hindi--Urdu remaining the hardest case at 0.722. Script normalization and whitening greatly improve retrieval and are computationally inexpensive, although normalization requires a reliable transliterator and fails on natural Urdu text. We release parallel corpora, code, and a token overlap diagnostic that correlates with cross-script retrieval failure without model inference. We hope to encourage future work in NLP on low-resource and digraphic languages.

Paper 144 Signatures of Steerability in Activation Space of Language Models

Authors: Prajjwal Bhattarai, Tuka Alhanai

Abstract Steering language models using a set of contrastive representations has been a canonical and computationally efficient method for controlling model behavior. Despite this success in controlling certain model behaviors, the effectiveness of activation steering varies markedly across concepts; the generalization properties of steering vectors are often considered a function of the dataset used to construct them. We make this dataset-dependence claim more rigorous and show that simple separation metrics strongly correlate with the downstream steerability of language models across diverse settings, even after controlling for layers and dataset effects. Beyond prediction, we provide evidence from a synthetic superposition experiment that separation metrics are strongly correlated with alignment between the empirical and true feature direction. Our results suggest that simple separability statistics can serve as practical diagnostics for when steering vectors are likely to work.

Paper 145 AgentArch: A Benchmark for Evaluating Agent Architectures in Enterprise Workflows

Authors: ~Tara_Bogavelli1, ~Hari_Subramani1, ~Roshnee_Sharma1

Abstract While individual components of agentic architectures have been studied in isolation, we lack empirical evidence on how design dimensions interact in realistic multi-agent systems. This study aims to address these gaps by providing a comprehensive enterprise-specific benchmark. We introduce AgentArch: a benchmark that evaluates 18 agentic configurations across state-of-the-art large language models. The study spans four dimensions: orchestration strategy, agent style (ReAct versus function calling), memory architecture, and thinking tool integration. Results reveal strong model-specific architectural preferences that challenge one-size-fits-all assumptions in agentic AI. They also expose notable performance gaps on enterprise tasks: the best models reach only 35.3% success on the complex task and 70.8% on the simpler task. We aim to inform architecting choices for enterprise agents by grounding model selection and component design in systematic evidence.

Paper 151 CROP: Token-Efficient Reasoning in Large Language Models via Regularized Prompt Optimization

Authors: Deep Shah, Sanket Badhe, Nehal Kathrotia, Priyanka Tiwari

Abstract Large Language Models utilizing reasoning techniques improve task performance but incur significant latency and token costs due to verbose generation. Existing automatic prompt optimization(APO) frameworks target task accuracy exclusively at the expense of generating long reasoning traces. We propose Cost-Regularized Optimization of Prompts (CROP), an APO method that introduces regularization on response length by generating textual feedback in addition to standard accuracy feedback. This forces the optimization process to produce prompts that elicit concise responses containing only critical information and reasoning. We evaluate our approach on complex reasoning datasets, specifically GSM8K, LogiQA and BIG-Bench Hard. We achieved an average 68\% reduction in token consumption across diverse reasoning benchmarks (and up to 80.6\% on algorithmic tasks) while maintaining competitive accuracy, seeing only a nominal decline in performance. This presents a pragmatic solution for deploying token-efficient and cost-effective agentic AI systems in production pipelines.

Paper 157 J-ToRL: Tool-Integrated LLM-as-a-Judge via Reinforcement Learning

Authors: Bryan Wilie, Tiezheng YU, Gerry Windiarto Mohamad Dunda, Jierun Chen, Chaofan Tao, Yan Xu, Junxian He, Pascale Fung

Abstract LLM-as-a-Judge frameworks offer scalable alternatives to human evaluation but inherit the limitations of text-only reasoning, particularly in tasks requiring precise computations. We introduce J-ToRL, a framework that trains LLM judges to leverage computational tools via code generation through end-to-end reinforcement learning (RL). Unlike prior approaches relying on inference-time integration, J-ToRL internalizes tool-use strategies as part of its policy, enabling exploration of diverse reasoning pathways. We identify a key challenge in applying RL with verifiable rewards to judgment: the coarse reward signals inherent in multiple-choice verdicts can incentivize shortcut learning. J-ToRL mitigates this via an auxiliary verifiable reward mechanism based on reference-answers, that enforces stricter criteria beyond simple verdict accuracy alone. With a sophisticated pairwise comparison dataset across five domains, J-ToRL-trained LLMs improve the base models, outperform prior judges, and close the gap to frontier LLMs. Further, J-ToRL-trained models exhibit beneficial cognitive behaviors such as self-error-recovery and self-refinement of tool use strategies.

Paper 160 DIAM: Discourse Interaction and Addressee Modeling for Malaysia Hansard

Authors: Goh Kun Shun, Mohan Raj, Lay-Ki Soon, Wern Han Lim, Ian K. T. Tan

Abstract Parliamentary debates are vital records of democratic processes, yet official transcripts such as the Hansard provide only a sequential account of speeches, omitting the underlying interactional structure. Although conducted in a formal setting, debates can become highly dynamic, making it difficult to determine who addresses whom and how utterances relate within the broader discussion. To address these gaps, we introduce DIAM, the first manually annotated discourse resource for the Malaysian Parliament, jointly supporting EU-level Dialogue Discourse Parsing (DDP) and multi-addressee Addressee Recognition (AR) over long, formal, unscripted spoken multi-party debate. Unlike existing DDP datasets, DIAM provides substantially longer conversations and explicit multi-addressee labels. Our benchmark experiments validate that DIAM supports both DDP and AR, while showing that existing discourse parsing methods cannot be directly applied without adaptations such as context-window strategies, and that addressee recognition is challenging for prompted LLMs, though fine-tuning a smaller model with LoRA substantially improves robustness to the large number of participants per sitting. We release DIAM and our code to facilitate future research on formal discourse parsing and addressee modeling: https://github.com/xKHUNx/DIAM

Paper 162 Comparative Personalization for Multi-document Summarization

Authors: Haoyuan Li, Snigdha Chaturvedi

Abstract Personalized multi-document summarization (MDS) is essential for meeting individual user preferences of writing style and content focus for summaries. In this paper, we propose that for effective personalization, it is important to identify fine-grained differences between users' preferences by comparing preferences of different users. Motivated by this, we propose ComPSum, a personalized MDS framework. It first generates a structured analysis of a user by comparison between users. The generated structured analysis is then used to guide the generation of personalized summaries. To evaluate ComPSum, we propose AuthorMap, a fine-grained reference-free evaluation framework for personalized MDS. It evaluates the personalization of a system based on the authorship attribution between two personalized summaries generated for different users. For robust evaluation of personalized MDS, we construct PerMSum, a personalized MDS dataset in the review and news domain. We evaluate the performance of ComPSum, showing that it outperforms strong baselines. The code is available at https://github.com/leehaoyuan/comparative_personalize_sum.

Paper 163 Thinking with Deltas: Incentivizing Reinforcement Learning via Differential Visual Reasoning Policy

Authors: Shujian Gao, Yuan Wang, Jiangtao Yan, Zuxuan Wu, Yu-Gang Jiang

Abstract Reinforcement Learning with Verifiable Rewards (RLVR) has significantly advanced reasoning capabilities in Large Language Models. However, adapting RLVR to multimodal domains suffers from a critical \textit{perception-reasoning decoupling}. Existing paradigms, driven by text-centric outcome rewards, reasoning in language medium, inadvertently encourage models to bypass visual perception. We empirically validate this through blind experiments: paradoxically, state-of-the-art policies maintain or even improve performance even when visual inputs are entirely removed. This reveals that these models degenerate into \textit{blind reasoners}, exploiting linguistic priors to generate plausible answers instead of attending to visual evidence. In response, we propose \textbf{Thinking with Deltas}, a framework driven by a \textbf{Differential Visual Reasoning Policy (DVRP)}. DVRP introduces intrinsic supervision via visual triplets, comprising original, masked, and perturbed inputs. It optimizes the model to maximize reasoning divergence from masked inputs (enforcing \textit{visual sensitivity}) while minimizing divergence from perturbed inputs (ensuring \textit{visual robustness}). By aligning reasoning variations strictly with the \textit{Delta} of visual information, DVRP inherently bolsters visual understanding capabilities and significantly outperforms state-of-the-art methods on both general and medical benchmarks, without requiring external annotations or auxiliary tools.

Paper 170 JudgeMoE: Distributional Aggregation for LLM-as-a-Judge

Authors: Yiqi Liu, Joseph James, Yang Wang, Kun Zhao, Chenghao Xiao, Chenghua Lin

Abstract When an LLM judge scores an output, its score distribution retains uncertainty and disagreement information that is lost after scalar compression. We introduce JudgeMoE, a lightweight aggregator that assigns example-specific weights to cached judge score distributions and fuses them before computing a final score. A protocol study shows that score-range choice is unstable across judge--dataset settings and that soft scoring usually outperforms hard decoding. On the original 10-cell benchmark, JudgeMoE improves mean Spearman over uniform log pooling by +0.079. Applying the same configuration to six additional cells yields a +0.0393 mean gain over the strongest local single judge across 16 cells, with positive differences in 12/16 cells and a one-sided Wilcoxon signed-rank p=0.0091. Validation-based analyses further show that the preferred aggregation method depends on the task and judge pool.

Paper 176 Method Ranking Inverts Across Targets: A Matched-Grid Evaluation of Jailbreak Attacks on Open-Weight Vision-Language Models

Authors: ~Venkat_Aniruddh_Thamatam1, ~Vasudev_Majhi1, ~Aaryan_Nagda1, ~Jagat_Sesh_Challa1, ~Murari_Mandal1, ~Yash_Sinha1, ~Dhruv_Kumar4

Abstract Jailbreak attacks are usually compared by their attack success rate (ASR), with a higher ASR read as a stronger attack. But an ASR mixes an attack's general strength with its fit to the particular model it was measured on, and it also depends on the judge that scored it. We separate these factors with a matched evaluation grid: five attacks on eight open-weight vision--language models, 200 HarmBench behaviours per cell, every turn scored by three judges (Qwen2.5-7B, Qwen2.5-72B and Llama-Guard-3), and each ASR decomposed into general strength, target susceptibility and target-specific fit. Attack rankings invert across targets: FigStep leads PAIR on InternVL2.5-8B (13.0% vs.\ 3.5%, paired $p{=}4{\times}10^{-10}$) and trails it on Qwen2.5-VL-7B, in the same direction under all three judges. Across the grid, the target-specific component is statistically significant under every judge ($p{<}10^{-12}$). The judge matters as well: agreement between judges is low (Cohen's κ between 0.03 and 0.29), and a larger judge with the same rubric changes which attack ranks first on three of eight targets. A higher ASR therefore shows that an attack succeeded more often on a given model under a given judge, but not whether its advantage is general or specific to that model. We release the protocol and ${\sim}20{,}400$ trajectories with per-turn scores under all three judges

Paper 177 The Language of Bargaining: Linguistic Effects in LLM Negotiations

Authors: Stuti Sinha, Himanshu Kumar, Aryan Raju Mandapati, Rakshit Sakhuja, Dhruv Kumar

Abstract Negotiation is a core component of social intelligence, requiring agents to balance strategic reasoning, cooperation, and social norms. Recent work shows that LLMs can engage in multi-turn negotiation, yet nearly all evaluations occur exclusively in English. We systematically isolate language effects across English and three Indic framings (Hindi, Punjabi, Gujarati) by holding game rules, model parameters, and incentives constant for multi-agent simulations for Ultimatum, Buy-Sell, and Resource Exchange games. Our results indicate that language choice is correlated to agreement and process dynamics, while final payoffs are statistically unchanged, and that this correlation persists across model architectures. Crucially, effects are task-contingent: Indic languages lower acceptance and longer negotiations in the Ultimatum Game, and higher trade volume in Resource Exchange. Further, they suggest that findings obtained under English-only conditions may not generalize to other linguistic settings. Even within one related language family, we observe task-contingent divergence from English, which cautions against treating English-only evaluation as representative. Whether this extends to typologically distant languages remains open.

Paper 181 Exploring Agentic Multimodal Large Language Models: A Survey for AIScientists

Authors: Jinglin Jian, Yi R. Fung, Denghui Zhang, Yiqian Liang, Qingyu Chen, Zhiyong Lu, Qingyun Wang

Abstract Agentic Multimodal Large Language Models (MLLMs) are reshaping scientific discovery. They do more than answer questions: they read multimodal evidence, reason over scientific constraints, use tools, and close parts of the research loop. This survey explains how such systems are built for scientific discovery, moving from multimodal perception to training, inference, evaluation, and human-AI collaboration. Across this pipeline, scientific agents face bottlenecks that generic assistants often hide, including scarce expert annotations, evolving domain structures, complex reasoning, immature tool interfaces, safety and toxicity risks, and unfaithful results. We highlight design choices for turning agentic MLLMs into trustworthy, evidence-grounded collaborators. Curated resources are available at https://github.com/JianJinglin/awesome-agentic-AIScientists.

Paper 188 Conversational Intelligent Tutoring Systems: Evolution, Taxonomy, and Pedagogical Frontiers

Authors: Zhengyuan Liu, Stella Xin Yin, Li Zhu, Xiaoman Wang, Yunshi Lan, Nancy F. Chen

Abstract Conversational Intelligent Tutoring Systems (CITS) represent a convergence of artificial intelligence, learning sciences, and human-computer interaction, aiming to provide the adaptive and dialogic learning as expert teachers. In the past decades, the advancement of machine learning and natural language processing reshaped CITS's design, implementation, and evaluation. In this survey, we analyze the evolution of computational paradigms in shifting pedagogical affordances, interaction styles, and methodological development. We frame CITS not merely as dialogue systems, but as proactive agents that reason about learner cognition, affect, and context while enacting pedagogical strategies grounded in educational theories. From domains, populations, and instructional goals, we identify key components, evaluation approaches, and unresolved challenges. Our paper highlights the promise and limitations of conversational tutoring in the generative AI era, offering a foundation for future research that seeks to maximize their pedagogical value.

Paper 191 Reference-Grounded Data Curation for Instruction-Following Thai-English Machine Translation

Authors: Thodsaporn Chay-intr, Krittapad Harnchang, Mahannop Thabua, Kobkrit Viriyayudhakorn, Thanaruk Theeramunkong

Abstract Instruction-following machine translation (IF-MT) requires respecting prompt-level rules on terminology, formatting, and register. Rule compliance typically trades off against translation quality, a tension that general-purpose IF data augmentation methods do not address. We propose Reference-Grounded Data Curation, a two-phase pipeline that extracts every supervised constraint from a reference translation that already satisfies it, ensuring feasibility by construction. Phase 1 applies Instruction-Following Difficulty (IFD) scoring to retain the hardest-but-learnable instances from an English-Thai parallel pool. Phase 2 extracts constraints from each reference target and keeps only generations satisfying every constraint, yielding the 1.97M-record Grounded dataset. We fine-tune open-weight bases on Grounded to produce ChindaMT, a Thai-English translation family at 4B, 2B, and 0.8B parameters. Under length-controlled pairwise judging, ChindaMT outperforms or matches every same-size baseline at every tier on both plain translation and under explicit rules, reaching up to a 68.4% win rate against the strongest baseline. The recipe transfers cleanly across Qwen generations. We release model weights, the Grounded dataset, and evaluation suites.

Paper 194 Are Many Tasks Necessary for General LLM Evaluation? A Matrix Completion View of Benchmark Redundancy

Authors: Kohei Tsuji, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe

Abstract Large language model (LLM) evaluation usually averages scores across many benchmark tasks, implicitly assuming broad task coverage is necessary for reliable general evaluation. This practice also implicitly treats all benchmark tasks as equally important for general evaluation. We revisit this assumption by treating benchmark results as a model-task score matrix under the performance prediction task. We find that a broad range of benchmarks are redundant with scores for 100 LLMs on 13 aggregate tasks and 523 subtasks. Specifically, we recovered model rankings with high rank correlation with the ALS matrix completion method after masking out several task scores. This means that no tasks dominated the recovery process, and both low-scoring and high-scoring tasks shared the same trends. Further analyses show that multiple-choice scores allow us to recover model rankings on free-form generation tasks. For general model ranking, free-form generation evaluation is not necessarily more informative than multiple-choice evaluation. These findings suggest that evaluating general-purpose LLMs can be accomplished with fewer tasks than the current common practice. Our analysis also estimates the number of models required to capture the overall trends in the benchmarks.

Paper 199 Investigating LLM Capabilities on Long Context Comprehension for Clinical Question Answering

Authors: Feras AlMannaa, Talia Tseriotou, Jenny Chim, Maria Liakata

Abstract Despite the impressive reported performance of Large Language Models (LLMs) on medical Question-Answering (QA), evaluation still overrelies on multiple-choice and extractive settings primarily assessing textbook knowledge, with limited work on open-ended QA over patient records. We present a comprehensive study investigating LLM long-context (LC) comprehension on longitudinal, patient-centric Electronic Health Records (EHRs) across multiple datasets, models and metrics, covering extractive, multiple-choice, and open-ended generative QA under different context formulation and Retrieval Augmented Generation (RAG) settings. We show that medically fine-tuned models offer no measurable benefit over base counterparts, smaller models exhibit stronger memorization, and longitudinal supplementary notes aid generative QA. Reasoning models gain on complex multi-note QA under LLM Correctness but not stricter reference-aligned metrics, while hybrid RAG with reranking outperforms Full-Context (FC) prompting. A question-category analysis and multi-faceted metric evaluation further reveal where RAG and FC excel and struggle, with RAG falling behind in longitudinal reasoning, thus uncovering systematic failures pointing to the need for reasoning-appropriate metrics.

Paper 201 Video-guided Machine Translation for Indic Languages

Authors: Pinaki Das, Srikar Ayyagari, Raj Dabre, Thuy-Trang Vu, Virendra Singh, Pushpak Bhattacharyya, Gholamreza Haffari

Abstract Video-guided machine translation (VMT) translates text by leveraging the context provided by the accompanying video. This additional contextual information helps improve translation quality by disambiguating ambiguous words or explicitly indicating attributes such as the colour or gender of objects and people. Prior works have extensively explored VMT; however, most approaches train models from scratch, which requires large amounts of data. Additionally, the majority of VMT research has focused on the English–Chinese language pair. In this paper, we explore training strategies for VMT using only a limited amount of data and extend the task to Indic languages, specifically Hindi, Marathi and Odia. We also propose a synthetic data framework that is effective for training VMT models in low-resource language settings. Our results demonstrate an improvement of +1.75 BLEU compared to baseline models.

Paper 207 Lexically Stable but Semantically Improvable: On the Editability of Machine Translation Drafts for LLM Post-Editing

Authors: Tianlin Xiao, WenHan Chao

Abstract Large language models (LLMs) perform strongly in machine translation, but their effectiveness as translation refiners depends on the draft source. We compare direct translation, self-refinement, and cross-system refinement on English-to-Chinese WMT test sets. Sockeye drafts obtain higher lexical-overlap scores but lower COMET than the stronger Qwen3 drafts, and receive larger gains after Qwen3 refinement. Additional controls qualify this pattern. Qwen3-0.6B drafts obtain the largest raw gain but lower final quality, while Qwen3-1.7B drafts start near Sockeye in COMET yet benefit less. Cross-family GLM-4-9B-Chat drafts also improve with Qwen3-8B, but achieve smaller gains and lower final quality than refined Sockeye drafts. Descriptive bucket, common-support, and regression analyses suggest that initial quality alone is an incomplete account of these differences. xCOMET-XL score and span diagnostics further characterize the contrast with Qwen3-8B self-refinement. We frame translation editability as a property of the evaluated draft–editor pairing, assessed jointly through quality gain, improvement probability, degradation risk, and final quality.

Paper 212 Bahnar–Vietnamese Parallel-Sentence Retrieval: A Low-Resource Case Study with Lexical, Neural, and Hybrid Methods

Authors: Thi Ty Nguyen, Long S. T. Nguyen, Phat T. Tran-Truong, Tan Sang Nguyen, Tho Quan

Abstract For very low-resource language pairs such as Bahnar-Vietnamese, the relative effectiveness of generic multilingual, language-pair-specific lexical, and task-adapted retrieval remains unclear. We study Bahnar-Vietnamese parallel-sentence retrieval (bitext retrieval) as a controlled candidate-ranking task and compare lexical, alignment-based, neural, and hybrid methods under a development-based selection protocol. The strongest off-the-shelf encoder, LaBSE, reaches 24.24\% Recall@1, while symmetric IBM1 reaches 49.33\%. Among four in-domain modern dense retrievers, development selection chooses DPR-XM, which reaches 47.43\% Recall@1 and 72.21\% Recall@10. XLM-R LoRA reaches 64.02\% Recall@10, and the development-selected XLM-R-first IBM1 reranker gives the highest Recall@1 point estimate, at 51.62\%, with an MRR of 0.5579. Paired tests do not establish a significant difference between the hybrid and IBM1 at the 0.05 level, although reranking substantially improves over its neural first stage. Under pair-specific training, LoRA projection with CSLS also performs strongly on Khmer-Vietnamese and Lao-Vietnamese, reaching 80.14\% and 84.58\% Recall@1. Overall, the results show complementary strengths between lexical alignment and lightweight neural adaptation in low-resource bitext retrieval. Data and code are available at \url{https://github.com/NgThTy/translation-vn-bahna}.

Paper 214 Temporal Attacks on LLM Agents: Exploiting Monitoring Windows in Persistent Agent State

Authors: ~Yinzhi_Jin1, ~Wuyang_Zhang1, ~Chuqiao_Gu1, ~Zhen_Luo3, ~Calvin_Chang_Liu1, ~Jianming_Ma3, ~Yebo_Cao1, ~Wangming_Yuan1

Abstract In agentic AI systems, memories, cached responses, retrieved documents, and tool configurations persist across sessions and later guide responses and tool calls. This persistence creates a timing problem that ordinary poisoning evaluations miss. If a poisoned entry acts on first retrieval, output and tool-call monitors can observe the harmful behavior during an initial inspection window. We study the opposite case: a payload that looks like routine stored content, but contains a natural-language gate and becomes harmful only after a date, count, memory-decay event, later session, or runtime context is satisfied. We call these **temporal attacks**. Our evaluation inserts one payload into LangChain, AutoGPT, or Claude/MCP agents, monitors the first $K{=}50$ benign sessions, and then measures post-window activation across four LLM backbones. In matched-payload ablations, temporal gates lower raw ASR by 22.6 points relative to ungated payloads, but reduce monitor catch rate from 81.6\% to 2.8\%. Across 21{,}000 trials, attacks reach 64.1\% raw ASR and 77.8\% conditioned ASR. A four-defense stack gives only partial mitigation, reducing ASR to 24.7\% (35.9--58.1\% under adaptation). Persistent state therefore needs defenses at storage, retrieval, and action time, not only output monitoring.

Paper 218 Atomic and Holistic LLM Judges for Reference-Grounded Support Labels: A Prompt-Controlled Comparison

Authors: Xinran Zhang

Abstract When an LLM judge only has to assign a three-way support label to a candidate answer given a reference, does asking it to decompose the answer into atomic claims help, and at what cost? We compare four single-call designs that share the judge model, the inputs, and the level of instruction detail: candidate-side atomic decomposition, a matched holistic rubric, reference-side decomposition that checks whether each reference claim is covered, and a bidirectional combination. Support labels are constructed from TruthfulQA, ASQA, and QAMPARI references (200 questions and 400 rows per dataset). All four designs are

run with Opus-4.6, GPT-4.1, and Gemini Flash Lite; Sonnet-4.6 is added for the candidate-side and holistic designs. Candidate-side decomposition is weak where the label depends on completeness: the holistic rubric is more accurate on ASQA and QAMPARI for every judge while using fewer tokens. Reference-side decomposition is 12.5-21.3 points more accurate than holistic on ASQA for 28-29% more tokens, and stays near the holistic ceiling on QAMPARI at 63-69% more. On TruthfulQA misconceptions, candidate-side decomposition is competitive and significantly better for two judges. On a 60-row single-author subset whose labels are looser than strict reference completeness, candidate-side decomposition edges ahead of holistic for every judge, reversing the construction-label order. What a judge decomposes should follow what the label measures.

Paper 219 Learning What Is Said, Not How: Stance-Contrastive Learning for Generalizable Hate Speech Detection

Authors: yucheng lin, Dan Kienast, Jiaojiao Jiang

Abstract Hate speech classifiers often misclassify counter speech, including condemnation and quotation of hateful content, as hateful because they fail to distinguish propositional content from speaker stance. We identify this issue as content stance conflation and address it as a representation learning problem. We learn stance invariant representations of the same proposition by constructing contrastive pairs that share content but differ in stance. We implement this idea through factorized annotation, stance controlled augmentation, and Stance Contrastive Learning (SCL). On the Implicit Hate Corpus, stance variant augmentation reduces counter speech false positives by 60%, and SCL further improves cross domain F1 by up to 6.5 points on top of the augmented signal. On SBIC, where use mention conflation is less prominent in the training labels, SCL maintains a low counter speech false positive rate. Semantic similarity pairing on the same augmented data yields no cross domain gain over the model trained without contrastive learning (0.623 vs. 0.625), which identifies the pair construction signal rather than contrastive learning itself as the determining factor.

Paper 221 Cleaner Speech, Weaker Generalization: Revisiting Pitt-Derived Benchmarks for Alzheimer's Disease Detection

Authors: Luqi Sun, Shreeram Suresh Chandra, Lin Zhang, You-Jin Li, Brian MacWhinney, Yu Tsao, Emily Mower Provost, Berrak Sisman

Abstract Speech-based Alzheimer’s disease (AD) detection increasingly relies on speech-enhanced and curated versions of the Pitt Corpus, where speech enhancement, sample selection, and demographic balancing are often treated as beneficial preprocessing steps. However, whether these transformations improve real-world AD detection or instead affect model generalization and prediction behavior remains unclear. In this work, we revisit the role of speech preprocessing and dataset curation across widely used benchmarks for speech-based AD detection. We evaluated the speech quality of different datasets, the cross-dataset generalization of multiple deep learning models under matched and mismatched enhancement settings, and the behavior of several recent large audio-language models (LALMs). Experimental results show that across multiple supervised speech models, speech-enhanced datasets often improve in-domain performance while reducing robustness in cross-domain evaluation. Matched enhancement between training and test data alleviates, but does not eliminate, this degradation. LALMs show a similar sensitivity: enhanced datasets induce stronger class imbalance and prediction shifts than unprocessed data. These results suggest that speech preprocessing and dataset curation can substantially influence downstream AD detection behavior, indicating that "cleaner" speech datasets are not necessarily more reliable for real-world AD detection.

Paper 222 CARE-Bench: Benchmarking Patient-Facing LLM Triage

Authors: Yining Hua, Hongbin Na, Cyrus Ayubcha

Abstract Patient-facing medical LLMs and agents increasingly answer symptom questions before clinician contact, where the key safety question is what action the user should take next. We introduce CARE-Bench, a source-grounded benchmark that evaluates sequential patient-facing triage as a four-label per-turn current-action task. CARE-Bench contains 500 cases and 1,059 evaluated patient-disclosure prefixes reconstructed from medical dialogue, consultation, and follow-up-question sources. We evaluate 11 models on 269 held-out rounds under unprompted and minimally prompted open-ended protocols, using a fixed gpt-5.5 mapper to code each response into the four-label action space. Unprompted macro-F1 remains low, ranging from 31.2 to 50.4. Prompting improves 10 of 11 models, with prompted macro-F1 ranging from 46.9 to 63.4, but substantial threshold errors remain. Prompted models often recommend care before needed clarification is obtained; when the correct action was to ask for more information, only 33.5% of prompted outputs preserved the step. The persistence of these errors after prompting suggests that patient-facing triage is not a simple prompting problem and supports explicit evaluation of action timing before deployment.

Paper 223 OpenQTS: A Metadata-Rich Open Dataset for Tang Dynasty Poetry

Authors: ~Robin_Young1

Abstract Tang Dynasty poetry (618--907 CE) represents one of the most significant corpora in Chinese literature, yet no openly accessible, computationally-suitable resource exists with comprehensive author metadata. We present OpenQTS, a complete machine-readable edition of the \textit{Quan Tang Shi} (全唐詩), containing 48,300 poems from 2,602 authors across 900 volumes. Unlike existing resources, OpenQTS provides systematic biographical metadata through algorithmic inference and authority file integration. Each poem receives a unique identifier, with proper handling of multi-part poems (15.5\% of corpus) and collaborative works. All metadata includes provenance tracking, documenting whether information derives from Wikidata, biographical text inference, or historical markers.

Paper 224 Geo-spatial and Geo-temporal Reasoning in Vision-Language and Large Language Models : A Review

Authors: Shahriyar Zaman Ridoy, Azmine Toushik Wasi, Koushik Ahamed Tonmoy, Md Rizwan Parvez

Abstract Geo-spatial and geo-temporal reasoning require models to determine not only what information is present but also where and when it applies, which is critical for applications such as crisis response, environmental monitoring, and urban planning. This review surveys recent advances in vision–language models (VLMs) and large language models (LLMs) that address geographically and temporally grounded reasoning across tasks including image geolocation, spatiotemporal question answering, language-guided navigation, and Earth observation. We organize the literature through a structured analysis covering task families, input modalities, model architectures, and reasoning paradigms, highlighting how perception, retrieval, tool use, and intermediate reasoning steps contribute to system performance. The survey further synthesizes existing datasets and benchmarks, identifying common evaluation pitfalls such as spatial granularity mismatch, temporal drift, geographic bias, and potential data leakage. Finally, we outline key research challenges and opportunities for developing robust, verifiable, and time-aware geo-spatial AI systems while addressing broader concerns related to fairness, bias, and location privacy.

Paper 225 Is Brevity the Soul of Wit? An Empirical Test Using 570,000 Reddit Jokes

Authors: Robin Young

Abstract The assertion that "brevity is the soul of wit" has become proverbial in discussions of effective communication. We test this hypothesis using 570,634 jokes from Reddit's r/Jokes community and a cross-platform comparison of 22.1 million posts across 10 subreddits. Longer posts receive more upvotes across all of Reddit, not just jokes, indicating the baseline length-score correlation reflects platform dynamics rather than humor-specific effects. However, we find a pattern unique to humor: the length-score correlation disappears entirely for top-performing jokes, while every non-humor subreddit retains a positive correlation for its best content. For genuine wit, length is irrelevant, which is a property that distinguishes humor from other content types tested. Temporal analysis reveals winner-take-all consolidation. As Reddit matured, long jokes' median scores dropped 71\% while mean scores increased 6.8$\times$, indicating most long jokes now fail while rare successes dominate.

Paper 227 Knowing When to Critique: Task-Adaptive Metacognitive Regulation for Reliable LLM Reasoning

Authors: Xinmeng Hou, Ziting Chang, Zhouquan Lu, Bohao Qu, Liang Wan, Wei Feng, Hai Hu, Qing Guo

Abstract Large language models (LLMs) reason fluently but do not regulate their reasoning: they apply uniform scrutiny to every input, which leaves them vulnerable to adversarial and counterfactual prompts, while indiscriminate critique over-corrects answers that were already sound. We propose MetaCrit, a multi-agent framework grounded in Nelson and Narens' metacognitive regulation theory that calibrates how much critique each task receives. MetaCrit separates regulation into four agents: an object-level generator, a monitoring agent that assesses response validity, a control agent that critiques logical soundness, and a meta-level synthesizer that reconciles their signals into a regulated response. Adaptivity here is input-conditioned intervention strength within a fixed pipeline: all four agents run on every input and what varies is the direction and magnitude of the correction they produce, not which stages execute. Across reasoning, safety, and bias benchmarks, MetaCrit improves truthfulness and logical soundness and reaches zero toxicity on BOLD and HONEST without a reasoning trade-off, whereas the same critique applied indiscriminately degrades performance. The cost is four calls per query, about one sixth of the cost of a dedicated reasoning model of similar accuracy. A writing study shows that MetaCrit is preferred for critical-thinking support, and its agents transfer to existing frameworks without architectural change. Code is available at https://github.com/Paparare/EduThink4AI.

Paper 231 Reward-Induced Transcription Mediation for Speech-to-Text Translation

Authors: ~YONGJIAN_CHEN1, ~Xiao_Zhang44, ~Juan_Antonio_Pérez-Ortiz1, ~Antonio_Toral1

Abstract Speech-to-text translation has evolved from cascaded pipelines to end-to-end models that avoid error propagation, but sacrifice intermediate representations. We present a GRPO-based reinforcement learning approach that induces transcription-form intermediates within a structured reasoning space, optimising exclusively for translation quality without transcription supervision. Evaluation across Japanese, German, and Chinese, from and into English, shows substantial improvements over foundation models and supervised fine-tuning (up to 38% BLEU gain). We further show that the intermediate transcriptions function as translation-useful mediators whose role varies by translation direction. Finally, we show that lexical reward is a robust and efficient translation quality signal whose choice shapes the nature of intermediate representations.

Paper 232 Spatial Reasoning in Multimodal Foundation Models: A Survey of Spatial Understanding in LLMs and VLMs

Authors: Azmine Toushik Wasi, Md. Masudur Rahman, Abdur Rahman, Mohsin Mahmud Topu, Mohammed Eunus Ali, Riashat Islam, Md Rizwan Parvez

Abstract Spatial reasoning, the ability to represent and reason about spatial relations, geometry, topology, and embodied environments, is increasingly vital for evaluating and deploying large language models (LLMs) and vision–language models (VLMs). While multimodal foundation models show strong performance on spatial benchmarks, navigation, and embodied planning, much of this stems from shallow pattern matching rather than grounded spatial understanding. This review offers a comprehensive, framework-driven analysis of spatial reasoning in LLMs and VLMs, covering tasks, representations, datasets, benchmarks, architectures, training, and failure modes. We also link spatial reasoning to embodied AI, emphasizing challenges in generalization, compositionality, and grounding. Our survey unifies diverse fields, NLP, vision, robotics, and cognitive science, to guide the development of truly spatially grounded foundation models.

Paper 234 Agent-Environment Alignment via Automated Interface Generation

Authors: Kaiming Liu, Xuanyu Lei, Ziyue Wang, Peng Li, Yang Liu

Abstract Large language model agents have shown impressive capabilities in interactive decision-making tasks. These agents interact with environment through intermediate interfaces, such as predefined action spaces and interaction rules, which mediate the perception and action. However, mismatches often happen between the internal expectations of the agent regarding the influence of its issued actions and the actual state transitions in the environment, a phenomenon referred to as **agent-environment misalignment**. While prior work has invested substantially in improving agent strategies and environment design, the critical role of the interface still remains underexplored. In this work, we empirically demonstrate that agent-environment misalignment poses a significant bottleneck to agent performance. To mitigate this issue, we propose **ALIGN**, an Auto-Aligned Interface Generation framework that alleviates the misalignment by enriching the interface. Implemented as a lightweight wrapper, ALIGN-generated interface achieves the alignment without modifying either the agent logic or the environment code. Experiments across multiple domains including embodied tasks, web navigation and tool-use, achieve consistent performance improvements, with a 45.67% success rate improvement in ALFWorld.

Paper 235 ToneCL:Contrastive Learning for Few-Shot Syllable-Level Tone Classification

Authors: Qisheng Liao, Youngah Do

Abstract Tone languages constitute over 50-70% of the world's languages, but the vast majority are low-resource, lacking the large transcribed corpora needed for automatic tone classification. Existing datasets are typically collected at the sentence level, whereas field linguists require fine-grained syllable-level annotations. We propose ToneCL, a lightweight contrastive learning framework for few-shot syllable-level tone classification. We simulate low-resource conditions on Mandarin and Vietnamese, limiting labeled data to tens of examples per tone class. ToneCL is pretrained on unlabeled speech with augmentations that preserve tonal identity, then fine-tuned on few-shot examples. Experiments show our method consistently outperforms baselines, achieving 91.6% on six-speaker Mandarin at 10 shots. Cross-lingual transfer is also effective: pretraining on Vietnamese and fine-tuning on Mandarin reaches 91.0% accuracy at 10 shots. Ablation confirms that frequency band rejection is the most critical augmentation.

Paper 237 Less Random, More Incoherent: Greedy Hadamard Alignment for LLM Quantization

Authors: Yun Ye, Zhaoqi Wu, Xueshu Wang, Cong Zhang, Zhu Yongbo, Zhonghua Sun, BERYL XU

Abstract Hadamard preprocessing is widely used in large language model (LLM) quantization to reduce activation coherence and suppress peaks. Existing methods typically rely on random sign flips before the Hadamard transform, which do not use information from the target data distribution. We propose Greedy Hadamard Alignment (GHA), a simple training-free method to build a signed-permutation preconditioner from calibration data. GHA uses a pooled representative activation, a greedy search over flat targets, and an efficient alignment step that is analytically optimal once the target is fixed. We also study when one shared GHA preconditioner can transfer across LLM activations. GHA works best when activations are directionally aligned and stable across calibration sources. Overall, GHA improves post-quantization model quality over randomized Hadamard baselines, with the strongest gains in KV-cache quantization, more moderate gains for hidden states, and little compute overhead during inference.

Paper 238 DAMG-RAG: Document-Aware Multi-Granularity RAG for Visually Rich Document Understanding

Authors: Jiahao Gu, Ningjing Liu, Yuanzhuo Ding, Siheng Chen

Abstract Real-world documents are visually rich and structurally organized, with text, tables, figures, and captions interleaved across pages. Although retrieval-augmented generation (RAG) is widely used for document question answering, existing pipelines often treat such documents as fixed page- or chunk-level collections. This document-agnostic construction either dilutes evidence in dense pages or fragments semantically related layout elements. Moreover, cross-page questions often require iterative evidence exploration, but existing systems typically follow fixed retrieval paths that do not adapt to the evidence already found. We propose DAMG-RAG, a Document-Aware Multi-Granularity RAG framework for visually rich document understanding. During ingestion, DAMG-RAG constructs layout-aware atomic retrieval units and links them to page- and section-level nodes, forming a multi-granularity document tree. During inference, a state-aware agentic workflow maintains explored regions and validated evidence, dynamically selecting exploration and navigation tools over this tree. On MMLongBench, DAMG-RAG obtains 64.7\% accuracy with Claude Sonnet 4.6, yielding a 3.6-point gain over the best-performing agentic baseline in our experiments while using 18.8\% of its token budget.

Paper 239 AstroAgentBench: Evaluating Agentic Planning on Space Mission Planning Tasks

Authors: Weiyi Wang, Xinchi Chen, Jingjing Gong, Xuanjing Huang, Xipeng Qiu

Abstract Recent LLM-for-Space systems address mission planning, scheduling, operations support, simulator control, and autonomy, but their evaluations use different task contracts, control settings, simulators, and success criteria. We introduce AstroAgentBench, a seven-family benchmark for executable space mission planning in the domains of scheduling, observation planning, constellation design, and relay support. For each case, an agent submits a planning artifact that is checked by an external verifier for schema, timing, geometry, resources, and mission value. Results report validity and normalized scores, with comparisons to task-specific solver references. Across five LLM agent systems and 35 held-out cases, the strongest systems approach or exceed solver-reference scores on several families, while weaker systems often fail to produce high-value valid plans and even strong systems lose quality on geometric, product-level, or design-heavy tasks. Trace analyses separate two failure points: task-contract misformulation and weak solution construction. Successful runs instead calibrate agent-written implementations against verifier feedback and adapt search to case-specific structure. Ablations show that procedure injection and memory accumulation help selectively, when they supply the missing formulation, calibration, or search support.

Paper 241 Read the Room Before Your Model Generates

Authors: Muyao Kong, Zhihao Lin

Abstract Large language models (LLMs) generally rely on sampling temperature to regulate the amount of randomness in their outputs. Mainstream methods for selecting temperature typically involve fixed, heuristic global settings and context or token-level calibration, rather than a geometrically interpretable representation of the communicative scenario. In this work, we introduce a simple and interpretable temperature selection framework, the **Scenario-Actor**, which adapts decoding to the inferred situational characteristics. Scenario-Actor extracts scenario-relevant structures from Transformer hidden states and determines the temperature from scenario embedding, a structured geometric representation of the communicative setting. Empirical results show that on Llama-3.1-8B-Instruct, Scenario-Actor achieves strong goal-completion performance in interactive social scenarios, significant gains on open-ended instruction following, and competitive performance on multi-turn conversation compared to multiple baselines.

Paper 243 Bait-and-Recover: Poisoning Internal Refusal Signals to Defend LLMs against White-Box Editing Jailbreaks

Authors: Tian Gao, Zhipeng Xie, Yuhao Wu, Junhua Liu, Xin Fang

Abstract Open-weight large language models face a low-cost white-box threat from representation engineering attacks. Attackers can estimate refusal directions and search for projection-matrix edits that suppress safety alignment while preserving general capabilities, within minutes on a single GPU and without gradient-based training. We propose Bait-and-Recover, a weight-level defense that places a bait adapter where attackers read activations and a paired recovery adapter at the subsequent layer. Trained via gradient routing, this decouples the observation path from the behavior path. By actively poisoning the residual signal used for measurement, Bait-and-Recover disrupts the attacker's edit search, while the recovery layer restores clean downstream computation. Across four open-weight models, our defense raises the minimum refusal rate against white-box edit searches from 16.25% to 71.75% under a strict behavior-preservation budget (KL ≤ 0.10), with negligible impact on general benchmarks. By invalidating the core measurement assumption of these attacks, observation-path poisoning offers a practical complement to behavior-level safety training.

Paper 245 T2MBench: Evaluating Text-to-Motion Generation under Diverse Language Instructions

Authors: Bin Yang, Rong OU, Weisheng Xu, Jiaqi Xiong, Xintao Li, Taowen Wang, Luyu Zhu, Yuetong Fang, Xu Jiang, Jing Tan, Renjing Xu

Abstract As text-to-motion (T2M) models move beyond short action captions toward real-world applications, they are increasingly expected to follow diverse instructions. However, existing evaluation protocols remain insufficient for systematically assessing instruction following, physical plausibility, and fine-grained controllability under such diverse language instructions. To address this gap, we introduce T2MBench, a benchmark with 1,025 prompts across four categories—Dynamics, Complexity, Interaction, and Accuracy—covering 14 representative T2M models and three evaluation dimensions: LLM-Based Evaluation, Multi-Factor Motion Evaluation, and Fine-grained Accuracy Evaluation. Our results reveal substantial gaps in current T2M models' ability to consistently follow diverse language instructions while maintaining motion plausibility and fine-grained control.T2MBench establishes a reusable diagnostic benchmark for advancing research toward more reliable T2M generation under diverse language instructions. Data and materials are available at \url{https://github.com/Courteous121/T2MBench-release}.

Paper 246 When Does Concept Erasure Change Behavior?

Authors: ~Gagan_Aryan1

Abstract Concept erasure can remove linearly recoverable information from a representation without showing that this information influenced the model output being evaluated. For linear edits followed by explicit linear output maps, we characterize the resulting logit change using the output map and representation covariance. We then compare downstream effects with control edits matched for validation-side damage to test whether any benefit depends on the targeted concept. On Bios, Least-squares Concept Erasure (LEACE) reduces gender-probe accuracy to chance but changes only about 0.4% of occupation-logit energy and yields no detectable improvement in true positive rate (TPR) gaps beyond shuffled-label controls. Iterative Nullspace Projection (INLP) changes more occupation-logit energy at higher ranks, but its small, uncertain fairness gains are accompanied by occupation-accuracy loss. A controlled text shortcut yields a large target-specific gain when the classifier relies on the erased marker. In final-state, next-token audits of TinyLlama and Qwen2.5-7B, gender erasure substantially reduces he/she margins but leaves occupation-token choices intact. These results show that successful erasure, output sensitivity, and downstream benefit are distinct and should be evaluated separately.

Paper 249 State-Anchored Complete-View Distillation for Robust Conversational Multimodal Emotion Recognition

Authors: Zhaoyan Pan, Xiangdong Li, Wenke Wu, Mengting Ma, Ye Lou, Ji Zhou, Jiatong Pan, Wei Zhang

Abstract Conversational multimodal emotion recognition MER requires reliable prediction when language acoustic or visual observations are missing or unreliable. Many missing-modality methods reconstruct absent inputs yet such recovery can be non-unique in dialogue context and nonverbal cues may conflict with the target utterance. To this end we propose CoRe-KD Complete-view Reference-guided Knowledge Distillation a state-anchored conflict-regularized complete-view distillation framework for robust conversational MER. A complete-view teacher provides structured references including prediction-level references fused states and modality-specific states. Complete-view State Anchoring CSA aligns incomplete-view student predictions and states with these references while Nonverbal Conflict Exposure NCE trains on target-preserving nonverbal conflict views to reduce donor-label bias. Experiments on IEMOCAP and MELD with CMU-MOSEI as a supplementary utterance-level check show consistent gains under fixed- and random-missing protocols. Comprehensive ablation studies and further analyses support the role of CSA and the complementary effect of NCE.

Paper 252 Can LLMs Follow the Pulse of a Crisis? Evaluating Crisis Sentiment in Bangladesh's July Uprising

Authors: Md. Samiul Alim, Mahir Shahriar Tamim, Tanvir Ahmed Khan, Sharjil khan, Rafia Ferdous Duti, Shahriyar Zaman Ridoy, Mohammad Ali Moni

Abstract Sentiment analysis during political crises is difficult because language, context, and public reaction can change rapidly. This challenge is especially underexplored for low-resource languages such as Bangla. We introduce **UnrestSent200K**, a Bangla crisis sentiment dataset containing approximately 200K Facebook and YouTube comments from the July–August 2024 Bangladesh uprising. The dataset covers five event-aligned phases, ranging from early escalation and internet blackout to regime transition and a later flood crisis. Each comment is linked to its parent post, allowing evaluation both with and without discourse context. All comments are annotated through a fully human process involving 14 native Bangla-speaking annotators and senior validation. The annotation process achieves substantial agreement, with Cohen's kappa of 0.73, Krippendorff's alpha of 0.71, and 94.2% blind-audit agreement. We benchmark fine-tuned encoders, prompted LLMs, and LoRA-tuned LLMs. Results show that adding parent-post context consistently improves performance, while temporal shifts across crisis phases cause large

performance drops. Strong LLMs perform well, but they still struggle with sarcasm, implicit political references, and meanings that depend on the event phase. UnrestSent200K provides a benchmark for studying context-aware and temporally robust sentiment analysis in low-resource crisis discourse.

Paper 255 Hierarchy Aware Preference Optimization for the Safety of Korean Small Language Models

Authors: Soo Yong Kim, Junyoung Koh, Kyeonghun Kim, Seunghyeok Hong

Abstract Safety alignment of small language models (SLMs) typically incurs an alignment tax: benign prompts are refused along with truly harmful ones. The problem is sharper for Korean SLMs, where limited capacity meets a language whose intent signal is blurred by honorifics. We propose KoSCoPe, which combines a Hierarchical Safety Curriculum, a level-conditional adaptive margin (HARM-DPO), and Counterfactual Consistency Regularization, with a reasoning internalization schedule that recovers 1.5 to 3.2× inference speed. Across five Korean SLMs of 1.7B to 4B, KoSCoPe achieves the strongest Safety/Over-refusal trade-off among the baselines under our matched protocol and remains competitive with general-purpose models several times larger. Under paired-bootstrap tests, its Safety gains over the matched preference-optimization baselines survive multiple-comparison correction for all but the closest competitor, Cat-DPO, where the gain is smaller but directionally consistent. Preliminary cross-lingual validation on Japanese, Mandarin Chinese, and English suggests the same design can transfer beyond Korean, with larger gains where surface politeness markers are more salient.

Paper 257 CCD-CBT: Multi-Agent Therapeutic Interaction for CBT Guided by Cognitive Conceptualization Diagram

Authors: Chang Liu, Changsheng Ma, Yongfeng Tao, Bin Hu, Minqiang Yang

Abstract Large language models show potential for scalable mental-health support by simulating Cognitive Behavioral Therapy (CBT) counselors. However, existing methods often rely on static cognitive profiles and omniscient single-agent simulation, failing to capture the dynamic, information-asymmetric nature of real therapy. We introduce CCD-CBT, a multi-agent framework that shifts CBT simulation along two axes: 1) from a static to a dynamically reconstructed Cognitive Conceptualization Diagram (CCD), updated by a dedicated Control Agent, and 2) from omniscient to information-asymmetric interaction, where the Therapist Agent must reason from inferred client states. We release CCDChat, a synthetic multi-turn CBT dataset generated under this framework. Evaluations with clinical scales and expert therapists show that models fine-tuned on CCDChat outperform strong baselines in both counseling fidelity and positive-affect enhancement, with ablations validating dynamic CCD guidance and diagnostic analysis supporting asymmetric interaction design. Our work offers a new paradigm for building theory-grounded, clinically-plausible conversational agents.

Paper 258 SG-LegalCite: A Principle-Augmented Benchmark for Legal Citation Retrieval in Singapore Law

Authors: LEE YUEH ERN SHANNON, Kaidong Feng, Yingpeng Du, Lee En Jia Chloe, Zhu Sun

Abstract Legal citation in common-law systems depends not only on factual similarity, but also on the legal principle for which a precedent is invoked. However, existing benchmarks for legal citation retrieval use case facts, citation context or full judgments as inputs, where the governing legal principle is often missing or only implicitly expressed and entangled with broader context. As a result, models may retrieve precedents that are factually similar yet doctrinally irrelevant. This limitation is particularly consequential in Singapore, where the legal system has evolved independently: only domestic precedents are binding, while foreign authorities serve merely as persuasive references. Thus, we propose a new retrieval paradigm that ranks cited cases based on queries integrating case facts and explicit legal principle, inspired by the real-world legal reasoning workflows. To

support this paradigm, we further introduce SG-LegalCite, a dataset of 100,890 case–principle pairs extracted from 8,523 Singapore Supreme Court judgments spanning 2000–2025. Experiments across 11 baselines demonstrate the effectiveness of our principle-augmented retrieval paradigm, showing that explicit legal principles provide strong discriminative signals for legal citation retrieval. Our dataset and code is at https://github.com/ShannonLeeYuehErn/SG-LegalCite

Paper 259 Decoupling Internal Representational Changes and Causal Importance in Fine-Tuned Large Language Models

Authors: Lingfang Li, Procheta Sen, Shubham Das, Danushka Bollegala

Abstract Fine-tuning has emerged as a widely adopted approach for adapting Large Language Models (LLMs) to a variety of downstream tasks. However, how it reshapes their internal mechanisms remains poorly understood. To address this, we investigate how fine-tuning alters internal representations in LLMs, including attention patterns and layer-wise activations, and examine whether these changes are linked to task-relevant components identified by Edge Attribution Patching (EAP) (e.g., attention heads and logit-level activations) that drive task performance. We find that EAP-identified components are concentrated within specific layers, indicating a degree of functional localisation in how models internalise task-specific behaviour. Notably, the distribution of these components across layers is largely uncorrelated with the layers undergoing the most substantial representational changes during fine-tuning. Furthermore, we observe that overlap in EAP-identified components across tasks does not translate into cross-task performance transfer if the tasks are different in nature (e.g., classification vs. generative tasks). More specifically, fine-tuning on one task can lead to a degradation of performance on another when the two tasks exhibit a high degree of overlap in their EAP-identified components.

Paper 261 DeBERTa-ConPara: Attack-Aware and Deployment-Realistic Detection of AI-Generated Text

Authors: Mohamed Mady, Yupei Li, Johannes Reschke, Björn Schuller

Abstract Robust detection of AI-generated text under deployment conditions is challenging: distribution shifts across domains and generators, adversarial perturbations of the input surface, and the absence of target-domain labels for threshold calibration all degrade detectors that perform well in-domain. We present DeBERTa-ConPara, a deployment-oriented detector combining attack-aware Unicode preprocessing with a contextual transformer encoder trained over HC3 Plus, M4, MAGE and RAID. Our central finding is that preprocessing acts in opposite directions depending on where it is applied: normalising the training corpus deduplicates it, collapsing 35.4% of RAID rows into copies of their clean siblings and deleting the adversarial supervision, whereas normalising at inference is an effective defence. A factorial varying the two placements independently identifies raw training with normalised inference as the best configuration, reaching 99.61% AUROC, 99.01% TPR@5% FPR and 96.57% TPR@1% FPR on the official RAID hidden test, alongside 93.14% average balanced accuracy across HC3 Plus and MAGE under a fixed threshold. The gain is confined to two of twelve attack classes: homoglyph and zero-width-space insertion rise from 11.05% and 1.12% to 96.98%. The same signature reproduces in a zero-shot detector of different architecture, showing the effect belongs to the attacks rather than to our model. We additionally report two negative results: semantic-invariance augmentation through paraphrasing and supervised contrastive learning (ConPara) does not improve the best configuration, and the handcrafted feature-fusion branch is inert in distribution and harmful outside it.

Paper 262 Cantus: Toward Pedagogical Feedback Generation for Sight-Singing Learning

Authors: Jiale Chen, Xuelian Dong, Yubing Chen, Yongxin Hu, Wenxiu Xie, Tianyong Hao

Abstract Sight-singing is a fundamental yet challenging task in music education, where timely and fine-grained feedback is crucial for helping learners identify mistakes and improve their performance efficiently. However, existing automatic methods mainly focus on isolated subtasks such as alignment or scoring, and are unable to provide end-to-end diagnostic feedback with explicit identification of pitch, rhythm, and musical fluency errors. In this paper, we propose Cantus, a novel multimodal large language model framework for sight-singing assessment and exercise recommendation generation. Cantus jointly models vocal audio and symbolic musical representations, and leverages a pedagogically grounded error representation module together with a large language model to diagnose performance errors, generate structured assessment comments, and provide personalized exercise recommendations. To support this task, we construct a new dataset MIVO (MIDI-Voice) containing 1,003 real-world sight-singing samples with aligned MIDI notes and vocal recordings collected from authentic teaching scenarios. Experimental results demonstrate the effectiveness of Cantus in sight-singing assessment, achieving strong performance across multiple metrics. Human evaluation of feedback quality shows the potential of Cantus in supporting student-centered sight-singing learning.

Paper 266 CoSelector: Multi-Agent Collaboration for Multilingual Retrieval-Augmented Generation

Authors: Long Yuan, Rui Qi, Xue Zhang, Jinan Xu, Kaiyu Huang

Abstract General Retrieval-Augmented Generation (RAG) systems would suffer from performance degradation in multilingual scenarios due to the complexity of integrating different retrieval results from multilingual collections. In this paper, we propose CoSelector, a multi-agent collaborative framework for multilingual retrieval selection, which is optimized by the centralized training with decentralized execution. Under this paradigm, agents share the same goal but execute decisions using only local observations. CoSelector consists of a Selector agent and a Generator agent. The Selector dynamically filters and composes the most pertinent multilingual documents into a coherent context, while the Generator processes this context to understand and integrate information for answer generation. Across the 15 languages evaluated on MKQA and XOR-TyDi QA, CoSelector achieves competitive improvement in response language consistency, demonstrating its effectiveness in both linguistic alignment and the selection of retrieved results.

Paper 267 MMA: Multimodal Memory Agent

Authors: Yihao Lu, Wanru Cheng, Zeyu Zhang, Hao Tang

Abstract Long-horizon multimodal agents depend on external memory. However, similarity-based retrieval often surfaces stale, low-credibility, or conflicting items, which can trigger overconfident errors. We propose Multimodal Memory Agent (MMA), which assigns each retrieved memory item a dynamic reliability score by combining source credibility, temporal decay, and conflict-aware network consensus, and uses this signal to reweight evidence and abstain when support is insufficient. We also introduce MMA-Bench, a programmatically generated benchmark for belief dynamics with controlled speaker reliability and structured text–vision contradictions. Using this framework, we uncover the "Visual Placebo Effect," revealing how RAG-based agents inherit latent visual biases from foundation models. On FEVER, MMA matches baseline accuracy while reducing accuracy standard deviation by 35.2% and improving selective prediction utility; on LoCoMo, a safety-oriented configuration improves actionable accuracy and reduces wrong answers; on MMA-Bench, MMA reaches 41.18% Type-B accuracy in Vision mode, while the baseline collapses to 0.0% under the same protocol. Code is available at https://github.com/AIGeeksGroup/MMA.

Paper 271 Empathy Is Not the Scapegoat for Declines in Factuality and Reliability

Authors: Kanwal Mehreen, Letian Shi, Deheng Zhang, Yuxia Wang

Abstract Language models are increasingly expected to support users in stressful or uncertain situations, where warmth and empathy are desirable. While recent work links warmer models to lower factuality and reliability, we show that this conclusion is incomplete: prior studies often use warmth as a global persona and conflate tone effects with SFT data distribution shift. We revisit this issue through prompting, fine-tuning, factual robustness, and empathy evaluation, using an 11-dataset benchmark suite and EmoFact, a benchmark for factual accuracy under emotional context. Across three large language models, we find that warmth effects are task- and model-dependent, and that task-aligned adaptive tone can improve empathy while preserving reliability. We further train human-validated empathy evaluators to evaluate the output of our finetuned model. Our results suggest that models can be made more supportive in human relationships and social interactions without compromising overall reliability.

Paper 276 RouterKGQA: Specialized–General Model Routing for Constraint-Aware Knowledge Graph Question Answering

Authors: Bo Yuan, Hexuan Deng, Long Bang, Xuebo Liu, Zheng Shuai, Zhou Mingzhen, Min Zhang

Abstract Knowledge graph question answering (KGQA) is a promising approach for mitigating LLM hallucination by grounding reasoning in structured and verifiable knowledge graphs. Existing approaches fall into two paradigms: retrieval-based methods utilize small specialized models, which are efficient but often produce unreachable paths and miss implicit constraints, while agent-based methods utilize large general models, which achieve stronger structural grounding at substantially higher cost. We propose RouterKGQA, a framework for specialized–general model collaboration, in which a specialized model generates reasoning paths and a general model performs KG-guided repair only when needed, improving performance at minimal cost. We further equip the specialized model with constraint-aware answer filtering, which reduces redundant answers. In addition, we design a more efficient general agent workflow, further lowering inference cost. Experimental results show that RouterKGQA outperforms the previous best by 3.57 points in F1 and 0.49 points in Hits@1 on average across benchmarks, while requiring only 1.15 LLM calls per question on WebQSP (1.23 on CWQ). Code and models are available at https://github.com/Oldcircle/RouterKGQA.

Paper 277 CoMMA: Conflict-Aware Read-Time Control for Conversational Memory QA

Authors: Zixiao Wang, Xingbo Du, Duzhen Zhang, Le Song

Abstract Memory-augmented assistants must answer questions from long-horizon conversations in which facts are not only missing, but also revised, corrected, or mutually inconsistent. Existing methods largely optimize retrieval, storage, or updating, and many controllers still treat failure primarily as insufficient evidence. We argue that conversational memory QA also requires an explicit model of inconsistency: when remembered claims disagree, a system should not simply retrieve more or guess, but may need to resolve versions, return a qualified answer, ask for clarification, or abstain. We introduce CoMMA, a transparent read-time controller with explicit conflict state for conversational memory QA. CoMMA extracts provenance-aware claims from retrieved memories, groups potentially incompatible claims into conflict clusters, infers lightweight question intent, and chooses among direct answer, qualified answer, clarification, and abstention. In an evaluation centered on the full LongMemEval-Oracle benchmark, a LongMemEval conflict-focused diagnostic slice, and FactConsolidation-20, CoMMA performs best on the main LongMemEval answer-quality comparison, is the only method with non-zero conflict F1 on the conflict-focused diagnostic slice,

and ties the strongest write-time overwrite baseline on FactConsolidation-20. These results suggest that making conflict explicit at read time improves both decision quality and the auditability of memory-aware behavior.

Paper 282 LogicLure: Toward Evidence-Grounded Safe Decision-Making under Hidden Risks

Authors: ~Kyungho_Kim1, ~JunCheol_Shin1

Abstract Safety evaluations of large language models (LLMs) have largely focused on whether models refuse explicitly harmful requests. However, in high-stakes decision-support settings, seemingly legitimate requests may conceal missing information or hidden risk signals. We propose LogicLure, a benchmark for evaluating whether LLMs avoid premature decisions and actively seek evidence under such plausible scenarios with hidden risks. The benchmark spans eight domains and pairs persuasive visible narratives with hidden evidence maps accessible only through verification. Our evaluation measures not only the decision of models, but also whether they ask verification questions and whether their final decision is grounded in the revealed evidence. In our evaluation, the strongest model reaches only 26.4% evidence-grounded refusal despite 94.7% overall refusal. Furthermore, we analyze how scenario surface cues, explicit scrutiny setting, and prompt-induced stance shape these decisions. Together, these analyses characterize the key behavioral factors underlying LLM performance on LogicLure. The benchmark and code will be made publicly available.

Paper 283 Initialisation Determines the Basin: Efficient Codebook Optimisation for Extreme LLM Quantization

Authors: ~Ian_W_Kennedy1, ~Nafise_Sadat_Moosavi1

Abstract Additive quantization enables extreme LLM compression while preserving lookup-table inference with free-form codebooks. At 2 bits per weight, however, greedy sequential initialisation can cause large quality losses that wider assignment search and PV-tuning do not fully erase. We show that this failure reflects an interaction between representational pressure and sequential codebook coupling. We use $\rho=N/K^M$ as a structural pressure measure and propose OA-EM, a hard generalised-EM initialisation procedure using Hessian-weighted Mahalanobis distance to minimise output-aware reconstruction error. Across four models from 3B to 14B parameters, OA-EM consistently lowers matched post-PV perplexity. Controlled experiments at a fixed bitrate show that ρ alone does not explain sensitivity to initialisation; the number and size of the additive codebooks also matter. A Euclidean-EM ablation and curvature diagnostics show that the gain is not explained by EM refinement alone: OA-EM reduces error in high-curvature directions. On Llama 3.2 3B, OA-EM also improves the measured quality-compute trade-off. Overall, the final quantized solution remains persistently dependent on initial codebook geometry. Code: https://github.com/kenno94-IK/aqlm-oaem.

Paper 288 FrenchNews-7: Benchmarking Cross-Publisher French News Editorial Desk Classification

Authors: Amr Sobhy

Abstract We present FrenchNews-7, a cross-publisher France-based French-language news editorial desk classification benchmark combining a large multi-outlet corpus, a URL-derived seven-class taxonomy, and a fine-tuned CamemBERT classifier. Labels are assigned via a hybrid pipeline combining publisher URL slugs with LLM annotation for structurally ambiguous cases, audited through an inter-rater study (2 humans + 2 LLMs; pairwise $\kappa \geq 0.766$, human–human $\kappa = 0.806$). We evaluate lexical, multilingual, and French-specific trained classifiers under both in-distribution and held-out-publisher settings, with additional comparison against zero-shot LLM baselines (GPT-OSS-120B, Mistral Small 3.2, Llama-3.3-70B) on the held-out pool. The strongest model, CamemBERT-base on full article text, outperforms headline-only input, generalizes to unseen outlets, and exceeds all three zero-shot LLM baselines on overall recall (0.799), with the gap concentrated in the ambiguous editorial-boundary categories Économie and Société. Cross-publisher evaluation reveals uneven boundary stability: Sport,

Culture & Loisirs, and International transfer cleanly, while Économie (recall = 0.517) is statistically indistinguishable from blinded human agreement on the same articles (0.55, n = 40), and Société (precision = 0.577) absorbs boundary ambiguity, both suggesting editorial conventions rather than recoverable classifier headroom. The fine-tuned CamemBERT-base model is released on Hugging Face (https://huggingface.co/LeFrenchNewsLab/camembert-base-frenchnews7) and the labeled manifest on (https://huggingface.co/datasets/LeFrenchNewsLab/frenchnews-7), alongside reference collection scripts to assist re-fetching and a reliability-tier guidance table.

Paper 291 SweSpecQA: A Swedish Benchmark for Specialized Document-Native Question Answering

Authors: Murathan Kurfali

Abstract We introduce \textit{SweSpecQA}, to our knowledge, the first Swedish benchmark for specialized institutional and professional QA with paired text and full-page PDF evaluation. It contains 1,500 questions from public Swedish examination and certification materials across six domains: EU aptitude, boating, antibiotic stewardship, interpreter licensing, physician licensing, and pharmacist licensing. Of these, 234 questions require an original visual or layout-dependent source element; the remaining 1,266 are text-eligible. All benchmark questions are drawn directly from Swedish-language source materials, avoiding translation artifacts in multilingual adaptations of English benchmarks. We evaluate 17 models under a closed-book protocol, reporting accuracy by domain, exam subtype, visual/layout dependence, modality, and prompting strategy, as well as whether models follow the requested answer format. On the text-eligible subset, every multimodal model performs worse with full source-page input than extracted text; PDF penalties range from 4.0 to 38.2 points under direct prompting and from 1.7 to 36.4 points under chain-of-thought prompting. Reasoning prompts often improve PDF performance for several open-weight VLMs, while sometimes reducing clean-text accuracy. These results show that document-native evaluation exposes failures hidden by text-only evaluation and that preprocessing remains important for reliable Swedish exam-style QA.

Paper 296 DuDi: Dual-Signal Distillation with Cross-Lingual Verbalizer

Authors: Patomporn Payoungkhamdee, Tinnakit Udsa, Jian Gang Ngui, Sarana Nutanong, Alham Fikri Aji, Peerat Limkonchotiwat

Abstract Small language models (SLMs) are efficient and scalable, but their multilingual capabilities degrade severely at sub-billion scales, especially for Southeast Asian (SEA) languages. We introduce DuDi, a dual-signal multilingual distillation framework that combines an online sequence-level signal with off-policy and on-policy token-level signals. DuDi further uses a cross-lingual verbalizer to refine teacher feedback and improve teacher-student transferability in multilingual settings. Experiments on SEA-HELM across multiple model families, scales, and teacher-student settings show that DuDi consistently outperforms competitive distillation baselines. Ablations and analyses confirm that sequence-level optimization, token-level supervision, and cross-lingual verbalization provide complementary and transferable learning signals for multilingual SLMs.

Paper 299 SEA-NLI: Language Inference as a Lens into Southeast Asian Cultural Understanding

Authors: Peerawat Chomphooyod, Yosephine Susanto, Jian Gang Ngui, Attapol Rutherford, Alham Fikri Aji, Sarana Nutanong, Can Udomcharoenchaikit, Peerat Limkonchotiwat

Abstract Frontier LLMs perform well in Western contexts, but remain poorly tested on underrepresented cultures such as those in Southeast Asia (SEA). Existing NLI benchmarks are largely Western-centric, translation-derived, or monolingual, limiting their ability to measure culturally grounded reasoning. We introduce SEA-NLI, a native, culturally grounded NLI benchmark covering eight SEA countries in English and native regional languages, verified by native speakers. Across 17 encoder and decoder

models, we observe a low performance from all models, especially for knowledge-intensive categories such as Languages and Science and Technology. Our analysis shows that failure cases mainly stem from missing SEA cultural knowledge: SEA-adapted models and culture-aware prompting improve performance, while CoT prompting offers limited gains.

Paper 300 Progress Drift: Auditing False Progress in Long-Horizon LLM Agents

Authors: Xiangbo Zhang, Xiaoxu Ma

Abstract Long-horizon language agents increasingly report intermediate progress while searching, editing, executing tools, and producing artifacts. Users rely on these reports to decide whether to intervene, continue, or accept an output, yet final task success does not reveal whether the agent described its work faithfully along the way. We study Progress Drift, a failure mode in which reported task state diverges from observable evidence. We introduce ProgressBench, a claim-level benchmark that aligns progress reports with milestones, tool observations, and artifacts across coding, web research, and research-replication tasks. We further propose Progress Ledger, an evidence-linked reporting protocol in which material progress claims carry evidence pointers, timestamps, verification status, and unresolved risks. In the reported evaluation across 1,280 tasks, ledger-constrained reporting is associated with higher Progress Claim Accuracy (80.6% to 89.7%) and lower False Completion Rate (12.7% to 6.9%) under shared tool budgets. These descriptive results suggest that agent evaluation should measure not only task completion, but also whether agents faithfully report what has and has not been achieved during execution.

Paper 301 SID-V: Disentangling Specification- and Implementation-Alignment in Code Verifiers for Reward-Hacking-Resistant RLVR

Authors: Zhao ChuanPeng, Xiaofei Bian, Liu Peng, Pan Xiaoli, Yang YuXiao, Wang Yu

Abstract Verifiers and reward models trained on programming tasks underpin nearly every recent advance in code generation, agentic software engineering, and reinforcement learning with verifiable rewards (RLVR). Yet a verifier observes three correlated artifacts - a natural-language specification, an implementation, and an assertion or test - and is universally asked to collapse their joint correctness into a single scalar. We show that this conflation is precisely what enables reward hacking: an assertion can perfectly mirror what a buggy implementation does while violating what the specification requires, and a single-scalar verifier cannot distinguish the two failure modes. We introduce SID-V (Spec-Implementation Disentangled Verifier), which takes a triple (spec, code, assertion) and produces two independent entailment scores: spec-alignment and implementation-alignment. The gap between them is a direct, calibratable measure of implementation-overfitting. We train SID-V via shared-encoder dual-head contrastive learning over approximately 10K (buggy, fixed) program pairs from Defects4J and SF110, augmented with mutation-injected implementation variants that change behavior while holding specification constant. SID-V probes show that strong scalar verifiers including CodePRM, Aletheia, and frontier LLM judges score above 0.78 on spec-impl conflict cases that are by construction contradictions. When deployed as the reward signal for a GRPO test-generator, SID-V cuts reward-hacked test rate from 31.4% to 6.2% while preserving Pass@1 on HumanEval-Plus and SWE-bench Verified. The spec-impl gap also yields a strong oracle-warning signal, achieving 0.74 precision at 0.61 recall for buggy-yet-passing patches.

Paper 302 Grounding What Users Care About: Multi-Aspect Personalized Modeling from Implicit Preferences

Authors: Jiin Park, Misuk Kim

Abstract Large Language Model (LLM) personalization is challenging because user preferences are rarely stated explicitly; instead, they emerge as implicit, multifaceted signals in users' past interaction histories. Existing methods often compress user histories into a single representation or retrieve and inject past records without explicitly linking them to supporting evidence,

which can lead to hallucinated or overly generic outputs. In this work, we propose Multi-Aspect PersonaLized Modeling with Evidence (MAPLE). MAPLE structures implicit user preferences through a MULTI-ASPECT PREFERENCE PROFILE and connects them to target-specific grounding evidence through a PERSONALIZED EVIDENCE PLAN. By selecting only generation points that are both relevant to the user and supported by available evidence, MAPLE improves personalization and grounding faithfulness simultaneously. Experiments on the Games and Books domains with two backbone LLMs show that MAPLE consistently outperforms strong baselines and generalizes robustly under limited-history and unseen-user settings.

Paper 305 Incorporating Abstract Meaning Representation and Uniform Meaning Representation into In-Context Learning Strategies for Machine Translation

Authors: Thu Hoang, Ellie Yang, Weixin Lin, Shira Wein

Abstract While in-context machine translation approaches lead to notable progress for low-resource languages, it is not yet clear what kinds of linguistic information in prompts can consistently and reliably lead to performance gains. We evaluate whether English semantic representations, specifically Abstract Meaning Representation (AMR) and its extension Uniform Meaning Representation (UMR), can improve LLM-based MT via prompting. Across 12 languages, 7 models from 5 families, and multiple prompting strategies, we find that conclusions about their effectiveness are highly sensitive to the setting. We further observe that prompts using UMR tend to outperform those using AMR, particularly in few-shot scenarios and for certain models. These results highlight the need for comprehensive evaluation for in-context learning work.

Paper 306 MagiC: Evaluating Multimodal Cognition Toward Grounded Visual Reasoning

Authors: Chengfei Wu, Ronald Seoh, Bingxuan Li, Liqiang Zhang, Fengrong Han, Dan Goldwasser

Abstract Recent advances in large vision-language models have led to impressive performance in visual question answering and multimodal reasoning. However, it remains unclear whether these models genuinely perform grounded visual reasoning or rely on superficial patterns and dataset biases. In this work, we introduce MagiC, a comprehensive benchmark designed to evaluate grounded multimodal cognition—assessing not only answer accuracy but also the quality of step-by-step reasoning and its alignment with relevant visual evidence. Our benchmark includes 5,534 weakly supervised QA examples generated from strong model outputs and 896 human-curated examples with fine-grained annotations, including answers, rationales, and bounding box groundings. We evaluate 15 vision-language models ranging from 7B to 70B+ parameters across four dimensions: final answer correctness, reasoning validity, grounding fidelity, and self-correction ability. MagiC further includes diagnostic settings to probe model robustness under adversarial visual cues and assess their capacity for introspective error correction. We introduce new metrics such as MagiScore and StepSense, and provide comprehensive analyses that reveal key limitations and opportunities in current approaches to grounded visual reasoning.

Paper 309 Beyond Downstream Task Performances: A Set-Theoretic Evaluation of Sentence Embeddings

Authors: Yash mahajan, Naman Bansal, Faria Binte Kader, Santu Karmaker

Abstract Sentence encoders are foundational to a wide range of NLP applications, yet they are predominantly evaluated through performance on downstream tasks. Such task-specific evaluations provide limited insight into whether the resulting sentence representations capture fundamental compositional relationships in a task-independent manner. To address this gap, we introduce a set-theoretic framework for evaluating the compositional properties of sentence embeddings. Drawing on classical set theory, we define six evaluation criteria based on three core set-like text operations: TextOverlap, TextDifference, and TextUnion. We use these criteria to systematically evaluate 19 sentence encoders, including 7 classical models and 12 Large

Language Model (LLM)-based encoders. Our results reveal substantial differences in their compositional behavior: notably, SBERT exhibits the strongest overall alignment with the proposed set-theoretic criteria, outperforming even recent LLM-based encoders. We further examine whether these intrinsic compositional properties are associated with downstream performance across five NLP tasks. Among them, summarization exhibits the strongest correlation with the TextOverlap and TextDifference criteria, suggesting that our framework may also provide useful signals about downstream capability without relying exclusively on task-specific evaluation. Finally, we release a benchmark comprising ~192K samples to support systematic evaluation and future research on the compositional properties of sentence embeddings.

Paper 312 Beyond Correctness: Enhancing Architectural Reasoning in Code LLMs via Scalable Labeling with Agentic Judgment

Authors: Kirill Vasilevski, Ximing Dong, Benjamin Rombaut, Ruochen Deng, Jiahuei Lin, Arthur Leung, Dayi Lin, Boyuan Chen, Shaowei Wang, Ahmed E. Hassan

Abstract LLMs have substantially improved software engineering. Yet real-world development requires architectural understanding. Such understanding is prohibitively expensive to label manually and impossible to verify through tests alone. We propose an agentic judging pipeline using a strong LLM as a scalable proxy for expert architectural evaluation, comprising two judges: the Architecture Complexity Judge (ACJ), which estimates codebase-specific architectural understanding a task demands, and the Architecture Quality Judge (AQJ), which evaluates patch conformance to repository-specific architectural conventions via source-grounded rubrics. Fine-tuning Qwen3-8B/14B/32B on 3,360 curated instances achieves resolved rates of up to 27.2% on SWE-bench Verified — up to 540% over the base model and 256% over unfiltered fine-tuning. Meanwhile, the trained models achieve strong cross-language generalization and consistent improvements in architectural patch quality.

Paper 313 Are Reasoning Vision-Language Models Robust to Semantic Visual Distractions?

Authors: ~Yizheng_Sun1, ~Mochuan_Zhan1, ~Yanan_Ma2, ~Jia_Tong_See1, ~Yifan_Wang82, ~Ziyi_Wang59, ~Hao_Li25, ~Yang_Cui4, ~Wenhao_Cai3, ~Jingyu_Sun2, ~Chenghua_Lin1, ~Riza_Batista-Navarro1, ~Jingyuan_Sun3

Abstract Reasoning Vision-Language Models (VLMs) achieve strong performance on complex multimodal tasks, but reliable real-world application requires handling visual inputs that are messier than clean, curated benchmarks. Existing works mainly evaluate such reliability of VLMs through input corruptions, such as noise, blur and weather effects, which make visual evidence harder to perceive. This leaves a critical reliability failure mode underexplored: a model may perceive the evidence correctly, yet reason from plausible but irrelevant and distracting evidence and propagate this mistake to its final answer. To address this gap, we introduce **Distract-Bench**, a benchmark for evaluating VLM robustness to **semantic visual distractions**, defined as meaningful but task-irrelevant visual cues added to inputs while preserving the ground-truth answer. We comprehensively evaluate eight leading open-source and two closed-source VLMs across conventional vision corruptions and Distract-Bench. Our results show that Distract-Bench exposes a robustness failure distinct from vision corruptions: reasoning VLMs largely track their non-reasoning base models under perceptual degradation, but show consistently lower robustness to semantic distractions. Further analysis shows that these distractions often enter the reasoning process of VLMs, are treated as evidence, and lead to incorrect answers. Together, these findings reframe robustness evaluation for reasoning VLMs, shifting the focus from degraded perception to distractions for reliable real-world visual reasoning. Our data and code are available at _https://github.com/Yizheng-Sun/Distract-Bench_.

Paper 315 Routing-Based Continual Learning for Multimodal Large Language Models

Authors: Jay Mohta, Kenan E. Ak, Gwang Gook Lee, Dimitrios Dimitriadis, Yan Xu, Mingwei Shen

Abstract Multimodal Large Language Models (MLLMs) struggle with continual learning, often suffering from catastrophic forgetting when adapting to sequential tasks. We introduce a routing-based architecture that integrates new capabilities while robustly preserving foundational knowledge. While Multi-Task Learning (MTL) is a strong performance reference, its training cost grows linearly with the number of tasks. In contrast, our method maintains fixed per-task training data and compute requirements regardless of the task sequence length. Across models ranging from 2B to 8B parameters, we demonstrate that our routing approach performs on par with MTL while retaining the training efficiency of sequential fine-tuning. Beyond merely mitigating forgetting, we observe that token-level routing facilitates cross-modal transfer, leveraging knowledge from one modality to bolster performance in another. We also show that routing remains robust even with large expert pools and effectively capitalizes on task relatedness. Finally, we show that our method scales favorably, with larger models exhibiting minimal degradation compared to fully specialized fine-tuning.

Paper 318 Mechanistic Evidence Routing in Transformers: Margin-Gated Intervention for Retrieval-Conflict Hallucination

Authors: Sridurgesh Ravichandran

Abstract When a language model ignores retrieved evidence and answers from memory instead, this is not a knowledge failure, the correct information is present. It is a routing failure; the model commits to its parametric answer despite the evidence signal. We show this distinction is mechanistically meaningful, localizable, and correctable at inference time without training data or weight modification. A single prefill-margin computation detects routing failures (AUC ≥ 0.93 across five model dataset pairs) and gates a format-appropriate logit-bias correction. On MedQA multi-option conflicts, this achieves 77.5% restoration at zero harm, exceeding the strongest prompting baseline by 67.5 pp and matching domain-matched ITI without training data. On open-ended PopQA (3,141 items, three models), balanced alias promotion achieves 80-88% restoration at near-zero harm, Pareto-dominating prompt-only baselines in five of six cross-model conditions. The central mechanistic finding is that effective late-layer intervention sites are structurally distinct from the causal commitment locus (a five-head cluster, joint flip rate 88.6%, $p < 10^{-9}$), suggesting that steering can succeed through residual-stream write access without exactly identifying the full decision circuit.

Paper 320 Prompting Underestimates LLM Capability for Time Series Classification

Authors: Dan Schumacher, Erfan Nourbakhsh, Rocky Slavin, Anthony Rios

Abstract Prompt-based prediction pipelines suggest that large language models (LLMs) perform poorly on time-series classification, raising doubts about whether they encode meaningful temporal structure. We show that this conclusion reflects limitations of prompt-based generation rather than the model's representational capacity by directly comparing prompt outputs with linear probes over the same internal representations while controlling for important confounding factors. While prompting performs near chance, linear probes improve average F1 from 0.15–0.26 to 0.61–0.67, often matching or exceeding specialized state-of-the-art (SoTA) time-series models. Layer-wise analyses further show that class-discriminative time-series information emerges in early transformer layers and is amplified with multimodal inputs and the presence of pretrained weights. Together, these results demonstrate a systematic mismatch between what LLMs internally represent and what prompt-based prediction pipelines reveal, leading current prompt-based prediction pipelines to underestimate the degree to which these models function as useful time-series feature extractors.

Paper 321 Universal Boosts, Specific Suppressors: Sparse Autoencoder Steering of Medical Vision-Language Models

Authors: Farhad Nooralahzadeh, Benjamin Gundersen, Nicolas Deperrois, Hidetoshi Matsuo, Mizuho Nishio, Thomas Frauenfelder, Ahmed Allam, Christian Bluethgen, Michael Moor, Michael Krauthammer

Abstract Medical vision-language models (VLMs) commit clinical errors when generating chest X-ray reports: they fabricate findings that are not present in the image, miss important ones, or locate them incorrectly. We steer the error composition of generated reports \emph{without} weight updates by decoding-time residual steering on a per-token sparse autoencoder (SAE) basis: Top-K SAEs on late layers, per-error-type causal feature screening against clinical errors, then combined \emph{suppress/boost} intervention at inference time. On the MIMIC-CXR test split, our inference-only method improves the quality of generated reports for three radiology VLMs (RadVLM, LLaVA-Rad, and CheXOne), with relative improvements of +5.4\%, +7.2\%, and +17.0\% in the clinical composite metric, and statistically significant GREEN gains on all backbones. The gain is an explicit, controllable trade-off rather than uniform error reduction: steering shifts models toward \emph{more complete reporting}, substantially fewer missing findings, at the cost of some additional false findings, and we analyse and quantify this redistribution per error type. A cross-model feature alignment shows that the quality-promoting (boost) directions have functionally similar error-type profiles across architectures, whereas hallucination-linkedm (suppress) directions are model-specific. Census-motivated steering should therefore treat suppression per-backbone, rather than sharing a universal suppress list. The same recipe applied zero-shot to IU-Xray (GREEN $+7.7\%$~rel., on RadVLM) suggests that the identified features are properties of the model rather than of the training corpus. The causal feature sets and an interactive feature dashboard are shared.

Paper 324 Language Models Should Be Used to Surface the Unwritten Code of Science

Authors: Honglin Bao, Siyang Wu, Jiwoong Choi, Yingrong Mao, James Evans

Abstract This position paper argues that researchers should not only study how human biases enter large language models (LLMs), but actively leverage those biases as diagnostic tools to surface society's "unwritten code"---the tacit norms and hidden heuristics that govern high-stakes decisions yet resist explicit articulation. We present a case study in peer review: a framework that iteratively pushes LLMs to generate self-consistent hypotheses explaining score differences across 26,731 submissions to 46 conferences. LLMs' normative priors (cues extracted by self-talk) emphasizing theoretical rigor are systematically updated toward posteriors (cues actually used in the judgment) emphasizing storytelling and contextualization---a finding cross-validated by independent human survey evidence. Human reviewers explicitly reward prior-aligned aspects (r=0.49) but rarely articulate the storytelling posteriors (r=-0.14), despite implicitly rewarding them. These findings generalize to out-of-sample judgment and different models. We argue this framework applies to hiring, admissions, and criminal justice, offering a scalable path to making hidden evaluative codes visible for public scrutiny and equitable reform.

Paper 325 Absolute Error Is Not Enough: Ranking Model Performance Without Labels to Estimate Disparities at Scale

Authors: Veronica Rammouz, Aaron O. Gonzalez, Carlos Cruzportillo, Adrian Tan, Nicole Lang Beebe, Ke Yang, Anthony Rios

Abstract Estimating model performance without labeled data is important for deploying NLP systems across diverse domains. Prior work has proposed proxy measures based on distribution shift or predicted correctness, but it remains unclear whether these methods reliably preserve the \emph{ranking} of domains, which is critical for model selection and risk triage. We study

this problem from an ordinal perspective using a two-step framework that pairs a base classifier with an error predictor to estimate instance-level failures and disparities and aggregate them into label-free performance estimates. Across multiple sample sizes from the \textit{GeoOLID} and \textit{Amazon Reviews} datasets, large language model–based error predictors substantially outperform drift-based baselines in ρ and τ rank correlations. We further show that lower absolute estimation error ($RMSE$) does not necessarily imply better ranking, highlighting the importance of evaluating ordinal consistency directly. \textbf{WARNING:} This paper contains offensive and hateful speech in the appendix.

Paper 326 Modeling Iterative Text Revisions Through Chain of Revisions

Authors: Zhexiong Liu, Diane Litman

Abstract Text revision has become an essential part of writing, as it often improves the text through iterative refinement. However, large language models (LLMs) have primarily regarded text revision as a holistic task that simultaneously revises grammar, logic, and content. This might be because LLMs are pre-trained only to generate final text rather than to learn the iterative revision process. Therefore, such LLMs largely overlook the dynamic nature of human revision, where writers progressively improve their writing through a sequence of intended edits. To bridge the gap, we investigate text revision generation from an iterativeness perspective and propose a chain of revisions that focuses on revision orders to iteratively improve text quality. Experimental results suggest that our proposed revision chains effectively capture the dynamics of text revision in multi-intent scenarios and outperform holistic revision baselines, which provide a valuable pilot study for iterative revision research.

Paper 327 Persona Non Grata: LLM Persona-Driven Generations in MCQA are Unstable in Distinct Dimensions

Authors: César Guerra-Solano, Xiang Lorraine Li

Abstract Persona-driven generations (PDGs) have seen prolific use in research and industry applications, where a large language model (LLM) takes on a “persona” while completing some task. While persona expressed through free-form text (like dialogue) has substantial work investigating stability or consistency, relatively, persona expressed in non-text-heavy outputs (like in multiple-choice question answering, or MCQA) is often overlooked. We work to address this gap, seeking to understand the instability of LLM PDGs in MCQA tasks. We develop three metrics investigating the performance, outcome, and question correctness stability, evaluating three distinct dimensions. Using these metrics, we find that instability varies consistently between model families and model size, and across question domains, with math/commonsense questions leading to greater instability. We also find task prompt format introduces more prediction instability than other hyperparameters, like temperature. Finally, we find that instability is related to task accuracy, and using our instability metrics, find different experimental settings that result in different best and worst personas for tasks, despite their similarity. This reveals the importance of checking hyperparameter instability in PDGs.

Paper 328 Large Language Models Approach Expert Pedagogical Quality in Math Tutoring but Differ in Instructional and Linguistic Profiles

Authors: Ramatu Oiza Abdulsalam, Segun Aroyehun

Abstract Recent work has explored the use of large language models (LLMs) to generate tutoring responses in mathematics, yet it remains unclear how closely their instructional behavior aligns with expert human practice. We analyze a dataset of math remediation dialogues in which expert tutors, novice tutors, and seven LLMs of varying sizes, comprising both open-weight and commercial models, respond to the same student errors. We examine instructional strategies and linguistic characteristics of

tutoring responses, including uptake (restating and revoicing), pressing for accuracy and reasoning, lexical diversity, readability, politeness, and agency. We find that expert tutors produce higher-quality responses than novices, and that larger LLMs generally receive higher pedagogical quality ratings than smaller models, approaching expert performance on average. However, LLMs exhibit systematic differences in their instructional profiles: they underuse discursive strategies characteristic of expert tutors while generating longer, more lexically diverse, and more polite responses. Regression analyses show that pressing for accuracy and reasoning, restating and revoicing, and lexical diversity, are positively associated with perceived pedagogical quality, whereas higher levels of agentic and polite language are negatively associated. These findings highlight the importance of analyzing instructional strategies and linguistic characteristics when evaluating tutoring responses across human tutors and intelligent tutoring systems.

Paper 329 Differential Transfer Lenses: What Behavioural Contrasts Miss in the Residual Stream

Authors: Junjie Liu, Yu Hui DENG, Guoxin Wang

Abstract Mechanistic interpretability often enters the model through a deceptively ordinary object called a behavioural contrast. Clean against corrupted prompts; conflict against control; one linguistic role against another. Yet this choice is not merely experimental bookkeeping. It helps define the estimand itself. In this paper, we identify a failure mode in which a contrast that is behaviourally well formed can still erase mechanism-relevant structure. When examples share the same behavioural label but differ in latent linguistic state, their case-specific residual-write transitions may average destructively in the pooled estimate. The damage is not cosmetic: the resulting profile can misplace an effect across layers and pathways. To expose this failure before intervention, we introduce Differential Transfer Lenses (DTL), a linear operator for pos-hoc auditing contrast-conditioned residual writes. DTL estimates task/control state-to-update maps in control-whitened coordinates, yielding layer- and pathway-resolved transition profiles together with singular-direction summaries for selected transitions. Subject--verb agreement (SVA) provides the controlled test case. There, stratifying by number case reveals residual-write structure that the pooled conflict/control contrast largely suppresses, and it changes the apparent localization of the agreement-related effect. Indirect Object Identification (IOI), with its established mechanistic structure, serves as a positive control; Factual Recall (FR), by contrast, supplies a geometry-driven comparison whose DTL profile closely follows mean/covariance baselines. Taken together, these results make contrast cancellation a concrete failure mode for mechanistic interpretability: before probing, patching, or causal localization, the behavioural contrast itself must be audited.

Paper 330 Text-Preserving Lossy Text Compression: A Study of Strategic Deletion and LLM Reconstruction

Authors: Yuchun Zou, Junhong Tong, Jun Li

Abstract Traditional lossless text compression preserves every byte, but its gains on natural language are often modest. We study lossy semantic text compression, where the encoder strategically deletes parts of the text and a large language model (LLM) reconstructs the original from the retained skeleton. We benchmark a progression of deletion strategies, from character-level and word-level rules driven by word length or corpus frequency to surprisal-based and hybrid rules that require a neural encoder, and reach three findings. First, deletion by corpus word frequency is a strong low-cost baseline: a static frequency lookup remains competitive with far more expensive semantic methods. Second, semantic and hybrid methods gain most at mild-to-moderate compression, whereas at the lowest retention rate frequency-based deletion matches them and keeps more named entities. Third, QLoRA fine-tuning yields a compact local decoder that, once adapted, is competitive with zero-shot Gemini 2.0 Flash. Additional English and Chinese experiments show that the framework transfers across domains, while the best deletion rule remains dataset-dependent.

Paper 331 INMS: Memory Sharing for Large Language Model based Agents

Authors: Hang Gao, Yongfeng Zhang

Abstract While Large Language Model (LLM) based agents excel at complex tasks, their performance in open-ended scenarios is often constrained by isolated operation and reliance on static databases, missing the dynamic knowledge exchange of human dialogue. To bridge this gap, we propose the INteractive Memory Sharing (INMS) framework, an asynchronous interaction paradigm for multi-agent systems. By integrating real-time memory filtering, storage, and retrieval, INMS establishes a shared conversational memory pool. This enables continuous, dialogue-like memory sharing among agents, promoting collective self-enhancement and dynamically refining the retrieval mediator based on interaction history. Extensive experiments across three datasets demonstrate that INMS improves agent performance by effectively modeling multi-agent interaction and collective knowledge sharing.

Paper 333 ATOD: An Evaluation Framework and Benchmark for Agentic Task-Oriented Dialogue Systems

Authors: Yifei Zhang, Hooshang Nayyeri, Rinat Khaziev, Emine Yilmaz, Gokhan Tur, Dilek Hakkani-Tür, Hari P Thadakamalla

Abstract Agentic task-oriented dialogue (TOD) requires systems to track concurrent goals, dependencies, and long-horizon state. We examine goal-lifecycle recovery from fixed dialogue trajectories. ATOD contains 1,000 synthetic dialogues annotated for six advanced-TOD properties, and ATOD-Eval defines metrics for dependency-sensitive completion, memory recall, and proactivity. We implement a symbolic–vector memory evaluator for ATOD-Eval. Among five prompt-only predictors and six matched-backbone memory baselines, our evaluator is the only configuration above 90% in both goal detection F1 and conditional status accuracy on medium dialogues. On complex dialogues, no baseline exceeds it on both metrics; it has the highest conditional status accuracy within the memory-based block and the lowest measured per-turn latency. These experiments assess lifecycle tracking rather than interactive agent task success. A cross-family judge swap and a manual audit with 94.0% agreement provide initial checks on measurement reliability. Code and data will be released at https://github.com/amazon-science/ATOD.

Paper 337 Embarrassingly Simple Access Control for Large Language Models

Authors: Veronica Rammouz, Đorđe Klisura, RAM KRISHNAN, Anthony Rios

Abstract Role-Based Access Control (RBAC) is fundamental in secure systems handling healthcare, financial, and government data. Prior work on LLM safety under role constraints has largely relied on fine-tuning or prompting-based enforcement. In contrast, we take a mechanistic view and probe internal representations to identify signals associated with roles, permissions, and refusal behavior. We evaluate this approach across three datasets, database- and department-separated splits, varying policy complexity levels, and adversarial prompts. On the structured Text-to-SQL datasets, simple probes achieve strong performance, reaching average macro-F1 scores of $.913$ on \textsc{BIRD} and $.926$ on \textsc{Spider}, while remaining more stable than prompting as policies grow longer. However, probing does not perform equally well in all settings: it underperforms prompting in most \textsc{OrgAccess} comparisons and can degrade substantially under adversarial instructions. Overall, our results show that access-control information can be strongly represented in LLM hidden states, but its usefulness depends on the model, domain, and prompt.

Paper 339 A Unified Ambiguity Signal for Calibrated Multimodal Emotion Recognition

Authors: Tahir Ahmad, Raza Mukhtar, Chanjun Park, Yeo Chan Yoon

Abstract Multimodal emotion recognition (MER) in conversation is inherently ambiguous, as linguistic, acoustic, and visual cues may support multiple plausible interpretations. Existing approaches often treat ambiguity as prediction uncertainty rather than modeling its structured sources. We propose UAF, an ambiguity-aware framework for MER built around a single unified ambiguity signal. UAF estimates a normalized per-sample Model-based Ambiguity Degree (MAD) from the predicted emotion distribution by measuring its proximity to the maximally ambiguous uniform distribution, and uses it as a shared control signal rather than a post-hoc uncertainty score. This single scalar coordinates three parts of learning: it aligns representation radius with ambiguity, emphasizes ambiguous samples during optimization, and scales OT-based perturbation consistency. The fused evidence is organized into context, speaker, temporal, cross-modal, and residual subspaces that serve as the substrate on which this signal operates. Experiments on IEMOCAP, CMU-MOSEI, and MELD show that UAF yields consistently better calibration and high-ambiguity reliability alongside competitive accuracy and weighted-F1, showing that a single ambiguity signal can be an effective shared training signal for reliable MER

Paper 341 Penalty-Framed No-Valid-Option MCQA: Analyzing LLM Abstention under Invalid Choices

Authors: Jinhyeok Kim, Hye-Young Jung

Abstract Multiple-choice question answering (MCQA) is commonly used to evaluate large language models under the assumption that one of the provided options is correct, typically using answer-selection accuracy. However, in real deployments, users or retrieval systems may provide invalid option sets in which none of the listed choices is correct, and selecting one of them may incur downstream cost. We study this setting as penalty-framed no-valid-option MCQA. Using the mathematics subset of MMLU-Pro, we remove the labeled correct option, allow models to either choose a remaining option or output ABSTAIN, and penalize invalid forced-choice responses. We further introduce correct-conditioned analysis, evaluating abstention only on instances that the model originally answered correctly. Experiments show that high MCQA accuracy does not fully guarantee abstention reliability: even under explicit no-valid-option-aware instructions and penalty-based scoring, models still produce invalid forced-choice responses for a subset of originally correct instances. These results show that penalty-framed no-valid-option MCQA reveals an aspect of model reliability not captured by standard answer-selection accuracy.

Paper 342 mSFT: Addressing Dataset Mixtures Overfitting Heterogeneously in Multi-task SFT

Authors: Woosung Koh, Jeyoung Jeon, Youngjin Song, Yujin Cheon, Soowon Oh, Jaehyeong Choi, Se-Young Yun

Abstract Current language model training commonly applies multi-task Supervised Fine-Tuning (SFT) using a homogeneous compute budget across all sub-datasets. This approach is fundamentally sub-optimal: heterogeneous learning dynamics cause faster-learning tasks to overfit early while slower ones remain under-fitted. To address this, we introduce mSFT, an iterative, overfitting-aware search algorithm for multi-task data mixtures. mSFT trains the model on an active mixture, identifies and excludes the earliest overfitting sub-dataset, and reverts to that specific optimal checkpoint before continuing. Extensive evaluations demonstrate that mSFT *consistently* outperforms **4** baselines across **10** benchmarks and **6** base models. Further analysis confirms mSFT maintains robust gains across diverse dataset sizes, task granularities, and is insensitive to its single new hyperparameter, compute budget. Notably, at a low compute budget, mSFT can improve performance *while* lowering training FLOPs. Ultimately, mSFT establishes a *practical* overfitting-aware algorithm for multi-task SFT that maximizes the potential of models across diverse data mixtures.

Paper 344 MisconceptDialogue: Evaluation Criteria Shape Measured Success in LLM Math Tutoring

Authors: Woojin Lee, Yoonkyu Woo, Jinah Kim, Hyeok-Min Gwon, Jin-Xia Huang

Abstract Large language models are increasingly used as math tutors, raising two questions: can a single LLM detect a student's error, diagnose the underlying misconception, and remediate it through multi-turn dialogue—and can we trust the numbers that report how well it does? We address both with MisconceptDialogue, a fully automated framework that runs the full pipeline on five open-weight LLM tutors (7B–14B) interacting with LLM-simulated students over 5,000+ dialogues—a deliberately simulated protocol (paired conditions, a wrong-diagnosis control unsafe with real learners, scale) whose results do not establish real-student learning. On measurement, our headline finding is uncomfortable for the field: benchmark-style exact-match scoring common in standard math NLP—a lower-bound stress test—reports 42.1% measured Remediation Success Rate on identical dialogues where a human-calibrated LLM-as-judge reports 81.3%, a 39.2 percentage-point spread larger than the combined effects of tutor model, teaching strategy, and supplied diagnostic information. This is not a single-judge artifact: four independent LLM judges and 200 human annotations converge to within a few points (κ up to 0.91). On capability, the verdict is equally striking: the score barely changes whether we tell the tutor the student's actual misconception or a deliberately swapped one—under the tested diagnosis-presentation format, these tutors show little evidence of reliably conditioning on the diagnostic information they are handed. For LLM tutoring evaluation to mature, multi-criterion scoring with multi-judge convergence and human anchoring should become standard practice—and the measurement question deserves as much scrutiny as the capability question it claims to answer.

Paper 345 YouRA: A Persistent-State Architecture for Evidence-Traceable Autonomous Research Agents

Authors: Yoonkyu Woo, Woojin Lee, Jin-Xia Huang

Abstract End-to-end research agents can now produce complete scientific papers, yet manuscript claims often diverge from executed experiments. This gap is structural: research state, failure histories, and claim–evidence alignment are not maintained as persistent, verifiable state across long-horizon pipelines. We present YouRA (Your Research Agent), an architecture for stateful, evidence-traceable autonomous research. YouRA preserves research state, execution evidence, and failure history across the research trajectory by integrating three components: a Verification State Architecture (VSA) that tracks hypotheses, gates, and evidence pointers; an Independent Controller that turns state and reflection records into lifecycle, recovery, and debate/review control while separating control from execution; and Stateful Reflection that logs failures as structured lessons and routes recovery through bounded repair, redesign, or reset. On MLR-Bench's predefined ten-task end-to-end subset, YouRA improves over both MLR-Agent and AI Scientist V2 on scalar Overall across all three matched backbones. An automated diagnostic using MLR-Bench's hallucination taxonomy reports intersection/union counts for four fact-based failure types, and data-provenance diagnostic shows more real-data-based outputs. Ablating each of the four components (the VSA, the Independent Controller, MCP tool access, and reflection-guided recovery) supports their separable contributions. Removing either core-state component drops YouRA below the full system. Code: https://github.com/PrayPrey/Your-Research-Agent.

Paper 347 TRIAGE: Severity-Ranked Multi-Agent Failure Attribution

Authors: Xizhi Wang

Abstract Failure attribution for multi-agent LLM systems is typically formulated as identifying a single decisive error in a failed execution trace. Yet failed traces often contain multiple interacting errors and admit several plausible intervention points. To examine whether an annotated decisive error is also the most promising point for repair, we introduce TriageBench, comprising LLM-council judgments over 1,184 Who&When intervention candidates. The council prioritizes a different step from the original

label in 59.3% of traces. We therefore reframe failure attribution as severity-ranked diagnostic triage and instantiate it as TRIAGE, a zero-shot, cost-aware cascade combining LLM-as-judge suspicion scoring, structural impact, and LLM-simulated replay over generated repair patches. Under top-K evaluation, TRIAGE reaches 73.9% top-5 on Who&When (vs. 42.4% top-1) and 93.2% top-10 on multi-error TRAIL (vs. 56.8% top-1). A separate evaluation of generated top-3 repair sets finds at least one plausible repair on 98.5% / 99% of 326 traces (human rater / 3-LLM panel) under lenient (good $+$ ok) rating. Our code (https://github.com/jimmywang585/triage-failure-attribution) and TriageBench (https://huggingface.co/datasets/jimmywang585/triage-bench) are publicly available.

Paper 348 Tracing the Genesis of Logic: Causal Data Attribution of Mechanistic Features via Counterfactual Poisoning

Authors: Xiaokun Wang, Siyu Song, Hongbo Zhao, Wentao Liu, Jiaqi Jiang, Qingqing Mao

Abstract How do interpretable internal features that implement a learned capability emerge during training? We present a mechanistic interpretability study of 3-digit addition in a 19M-parameter transformer trained from scratch, using the grokking transition as a natural microscope for feature genesis. Our framework, Causal Genesis Attribution (CGA), combines checkpoint-resolved training dynamics, Sparse Autoencoder (SAE) feature analysis, and counterfactual training interventions to trace arithmetic and carry features from emergence to causal function. We find layer-4 SAE features that are silent on non-arithmetic text, selective for the arithmetic probes, and co-emerge with validation accuracy at step 500 (near-zero to 85%). Zero-ablating the top arithmetic features drops exact-match accuracy from 100% to 0%, showing that these directions are necessary for the learned addition algorithm. A four-arm intervention study reveals what sustains these features: removing arithmetic data collapses feature activation by 81%; matched numeric-noise replacement also fails, ruling out surface token statistics; and poisoning answers by +137 hyperactivates features by 27% while making the model produce systematically shifted outputs. Across four seeds, the genesis detector fires consistently, with detection accuracy stable at 85.6% ± 1.0%. Together, these results provide a mechanistic data–feature–capability chain for an emergent arithmetic reasoning ability.

Paper 349 Three Numbers and a Precondition: Auditing Multi-Source LLM Benchmarks for Source-Identity Leakage

Authors: Tian Xia, Sarp Aykent

Abstract LLM benchmarks often pool rows from several sources (judges, retrieval configurations, training frameworks) and report one headline score. When label rates differ by source, a detector can earn that score by recognizing a row's source, not its quality, and such a score is no promise about a new source. Our audit reports three AUCs of one detector: given features plus source (AUC_full, the headline), features alone, and source alone. Their ratio $L=\mathrm{AUC}_\text{source}/\mathrm{AUC}_\text{full}$ shows how close the source alone comes to the headline. The precondition: L is read only above a default floor, $\mathrm{AUC}_\text{full}\geq 0.65$, because two near-chance AUCs give a ratio near 1 although nothing leaks. On four RAGBench configurations (adherence rates 40.2% to 94.2%) the headline is 0.843 and the configuration label alone reaches 0.811 ($L=0.96$), above every per-query feature. When five LLM judges grade the same hard-math answers, $L=0.84$ looks similar, but the headline (0.594) fails the floor: judge identity predicts nothing. The floor, not L, separates the two panels. Six further audits with the floor and a 0.90 band on L frozen add the other outcomes, including a floor-passing panel whose source carries nothing.

Paper 351 PatchHolmes: Agentic Patch Retrieval via Listwise Selection

Authors: Guanqun Yang, Yingming Zhou, Jiangrui Zheng, Shudong Hao, Xueqing Liu

Abstract Patch retrieval, the task of finding the commit that fixes a known vulnerability, is the foundation of vulnerability management workflows, yet 60% to 63% of CVEs in the major advisory databases lack a patch link. We present PatchHolmes, a two-phase patch retrieval system that pairs a hybrid first-stage retriever with an agentic second-stage inspection loop. Unlike pointwise prior work that scores each candidate independently, the Phase 2 agent reads the top-100 listwise: it sees the full candidate list at once and selectively reads 3 to 10 commits through four budgeted tools before submitting a single best commit. On GitHubAD, PatchHolmes beats the pointwise binary classifier Favia by 25.34% Recall@1 and the retrieve-and-CoT baseline IRCoT by 31.40%, at one agent conversation per CVE versus Favia's ten; with the candidate set held identical, the agent adds 27.32% Recall@1 over taking the retriever's top candidate, and the same agent, transferred unchanged to PatchFinder_top10, lifts Recall@1 from PatchFinder's own top-1 pick (24.28%) to 39.86%. Swapping the LLM backbone within the Qwen family changes Recall@1 by under 1%, and a second model family (gpt-oss) stays far above the no-agent floor, so the gain comes from the listwise agent loop; the entire system runs on a frozen open-weight model over a local Git repository, without fine-tuning or external search APIs.

Paper 357 Schema-Key Wording as an Instruction Channel in Structured Generation under Constrained Decoding

Authors: Yifan Le

Abstract Constrained decoding is widely used to make large language models produce structured outputs that satisfy schemas such as JSON. Existing work mainly treats schemas as structural constraints, overlooking that schema-key tokens also enter the autoregressive context and may guide generation. To the best of our knowledge, we present the first systematic study of schema keys as an implicit instruction channel under constrained decoding. We formulate structured generation as a multi-channel instruction problem, where task signals can be placed in prompts, schema keys, or both. We further provide a projection-aware analysis that gives a sufficient condition under which an unconstrained expected-score advantage of an instructional key is preserved after grammar projection. Experiments on GSM8K and Math500 across seven language models show that changing only schema-key wording can substantially affect accuracy, with both positive and negative effects across models. Prompt-level and schema-level instructions also interact non-additively. The evidence is substantially stronger on GSM8K than on Math500. Our findings show that schema design is not merely output formatting, but part of instruction specification in structured generation.

Paper 359 Anchored Turns: A Length-Driven Gradient in Chain-of-Thought Causal Necessity Across Multi-Turn Conversations

Authors: DHRUV GUPTA, Vasudev Majhi, Jagat Sesh Challa, Murari Mandal, Dhruv Kumar

Abstract Chain-of-thought (CoT) faithfulness is measured almost exclusively in single-turn settings, leaving open how the reasoning-trace-to-answer relationship evolves as conversations lengthen. We introduce turn-level CoT counterfactual deletion: at every assistant turn we regenerate the answer at five truncation levels of the current-turn thinking block under greedy decoding, and label a turn anchored—a descriptive label for causal redundancy of the trace, not a mechanistic claim—when all five regenerations agree on the same answer. Across 67 multi-turn sharded-GSM8K conversations from DeepSeek-R1-Distill-Qwen-7B, the anchored fraction grows 3.9× from short (< 10 turns, 7.2%) to long (> 20 turns, 28.4%). Strict early < mid < late monotone ordering holds in five of seven reasoning models from three base families and four post-training pipelines (binomial sign test, $p=0.002$). A supplementary pooled GEE across all seven models converges on this direction ($\beta=1.12$, $p=5.1\times10^{-8}$, n=2,838 turn-

level rows). A surface-feature predictor flags anchored turns at AUC= 0.71 (+0.07 over a length-only baseline) without additional generation, providing a cheap online signal for CoT monitors calibrated on short conversations. Single-turn CoT benchmarks, by construction, cannot detect this gradient.

Paper 361 LKValues: Aligning Large Language Models with Sri Lankan Societal Values

Authors: ~Nethmi_Muthugala1, ~Supryadi1, ~Surangika_Ranathunga1, ~Nisansa_de_Silva1, ~Ruijie_Tao2, ~Ovindu_Gunatunga2, ~Pengyun_Zhu2, ~Shaowei_Zhang5, ~Jingting_Zheng1, ~Deyi_Xiong2

Abstract Value alignment of Large Language Models (LLMs) has been shown to be culturally biased toward Western norms. This results in the mishandling of local values in multilingual societies such as Sri Lanka that have their unique cultural dynamics. Existing benchmarks overlook Sri Lankan-contextualized values in its official language Sinhala, hindering culturally sensitive evaluation and fine-tuning. To bridge this gap, we propose LKValues, the first survey-grounded resource suite for Sri Lankan value alignment. From a trilingual survey of 205 respondents, blending adapted global frameworks and LLM-elicited local constructs, we derive 40 majority-endorsed societal values. Using these values, we construct LKvaluesIT, a Sinhala-English news-derived instruction corpus containing 150K scenario-based instances, and LKvaluesBench, a value-sensitive evaluation benchmark of 1,000 instances. We evaluate a set of proprietary and open-weight LLMs with LKvaluesBench. We fine-tune three open-weight base models (Qwen3.5-4B-Base, Qwen3.5-9B-Base, and Aya-Expanse-8B-Base). Our experiments show that newer and larger LLMs still exhibit low-resource and cultural value-alignment gaps. LKValues fine-tuning improves Qwen-family models in English and Sinhala, reducing invalid outputs and cross-lingual disparities, though gains remain model-family dependent. These highlight LKValues' efficacy in embedding Sri Lankan values, offering a replicable pipeline for low-resource, country-specific pluralist value alignment. The datasets are publicly available at https://github.com/NextME14/LKValues.

Paper 363 EcoStatKG: A Domain-Specific Statistical Knowledge Graph and Benchmark for Evaluating Numerical Hallucination in LLMs

Authors: Ratnesh Kumar Joshi, Sagnik Sengupta, Asif Ekbal

Abstract Large language models routinely produce numerical hallucinations---plausible but incorrect quantities, percentages, and statistics---yet existing benchmarks focus almost exclusively on entity-level factuality. We introduce \textbf{EcoStatKG}, a knowledge graph (KG) of 50,746 statistical triples and 306,231 structural triples from 6,871 Wikipedia articles on environmental sustainability, and \textbf{NumHallu}, a 500-question numerical benchmark grounded in held-out KG topics. We use these resources to diagnose how knowledge-base coverage affects eight retrieval strategies across six LLMs. In the full-index, oracle-coverage condition, simple cosine retrieval reaches 0.989 Numerical Exact Match (NEM), statistically indistinguishable from Graph RAG (0.992; paired-bootstrap $p=0.14$). In the deliberately constructed held-out condition, which removes answer-bearing triples, retrieval methods fall below the zero-shot baseline (0.506 NEM), exposing the risk that topically related numerical distractors can override parametric knowledge. Similarity-threshold filtering restores KG Cosine to near-baseline performance (0.511), rather than establishing a new state-of-the-art. A bounded clinical-domain proof of concept reproduces the coverage and distractor patterns. The contribution is a diagnostic study and resource: its results clarify when numerical retrieval helps, when missing coverage can make it harmful, and why confidence gating is best viewed as damage control.

Paper 364 Automatic Evaluation of Mental Health Stigma in Online Communication

Authors: Naomi Baes, Jemima Kang, Nick Haslam, Chris Groot, Alsa Wu, Luc Raszewski, Yulia Otmakhova

Abstract Mental health stigma has profoundly harmful impacts but its complexity makes it difficult to evaluate. Stigma may involve explicit derogation, but also subtler forms of blame, fear, paternalistic pity, social distancing, structural exclusion, and discrimination. We introduce a theory-grounded benchmark for automatic evaluation of mental health stigma in online communication, consisting of naturally occurring online news and social media text annotated with a fine-grained taxonomy of stigma across multiple mental health conditions. Our annotation framework comprises a binary stigma-detection task and a multi-level taxonomy covering (i) stigma mode, (ii) domain, and (iii) specific components of certain forms of stigma. We apply this framework to texts mentioning six mental health conditions and evaluate large language models alongside stigma-related classifiers for detecting sentiment, toxicity, and hate speech. Results show that mental health stigma is not well captured by models trained to detect these neighboring constructs, and that LLMs often overpredict stigma unless given explicit operational rules - mirroring the importance of decision rules in human annotation

Paper 365 Domain Knowledge Annealing: A Structured Local-to-Global Paradigm for Constructing Expert LLMs

Authors: Yikang Zhang, Zuchao Li, Ping Wang, Lefei Zhang, hai zhao, Bo Du

Abstract Expertise in vertical domains requires more than exposing LLMs to vast corpora; it demands the systematic learning of complex logic. Standard fine-tuning often treats data as an unstructured collection, failing to capture the dependencies needed for high-level reasoning. We propose Domain Knowledge Annealing (DKA), a curriculum learning paradigm designed to improve domain proficiency. Inspired by physical annealing, DKA organizes training from local to global: it starts by injecting discrete concepts via entity-centric samples and progressively advances to synthesizing complex, inter-document relationships. This progression allows the model to integrate isolated facts into a coherent knowledge system. Experiments on Linguistics and Law benchmarks show that DKA outperforms standard strategies, achieving state-of-the-art results among comparably sized models in domain-specific reasoning.

Paper 366 The Geometry of Uncertainty in Language Model Representations

Authors: Zy Li

Abstract Hidden-state probes have been used to predict factual errors in language-model outputs. We study whether the same representations also distinguish four types of prompt-model relationship: known, unknown, ambiguous, and misleading (built on a false premise). On a controlled 200-prompt dataset constructed to balance six surface features, all six pairwise type contrasts are linearly recoverable from hidden states at the position where generation begins, at AUROC ≥ 0.986 on three open-weight instruction-tuned models spanning two families and a fourfold parameter range. The probe exceeds surface-feature, TF-IDF, and frozen sentence-embedding baselines on every pair. Under group-aware evaluation on paraphrased prompts with group-level shuffled controls, probe AUROC remains at least 0.949 on every pair while lexical baselines degrade. These findings frame hallucination as one downstream consequence of a four-type representational structure that also captures ambiguity and false premises.

Paper 367 Language Models Detokenize, But Only for Themselves

Authors: Qingyun Li, Ruoyun Li

Abstract Language models often appear to become insensitive to how text is tokenized. Models trained with different tokenizers can reach similar language-modeling loss, and prior work shows that they recover word-level information from the raw subword segmentations (a phenomenon referred to as "detokenization"). This raises a natural question: once that information has been recovered, do the resulting hidden states become independent of the tokenizer that produced them? We study this using *tokenizer twins*, pairs of models matched in architecture, training data, and random seed but trained with different tokenizers. We learn small translators between their hidden states at positions where the two tokenizers agree, and then test the same mapping where their tokenization paths differ. Transfer works well at agreement positions, adding less than one nat of loss, but breaks sharply at disagreement positions, where the loss increases by more than seven nats. This gap appears from the first hidden layer and persists through the network. At the same time, word-identity probes can still recover much of the underlying content from these representations. In other words, language models can abstract beyond their individual subword tokens while still carrying a lasting imprint of how the text was tokenized.

Paper 375 PAIRSAFE: Proxy-Resistant Paired Safety Evaluation for Refusal Boundaries

Authors: ~Wenzhang_Du1

Abstract Aggregate refusal scores on mixed safe and unsafe prompts can obscure local boundary failures and label-correlated metadata shortcuts. We present PAIRSAFE, Proxy-Resistant Paired Safety Evaluation, a score-agnostic diagnostic protocol built from 150 deterministic, matched XSTest-derived safe/unsafe pairs. PAIRSAFE distinguishes strict pair locks from audited residual fields, then asks whether nuisance proxies collapse, unsafe twins rank above safe twins, and one threshold yields pair-consistent decisions; generated behavior is evaluated separately. Strict source, normalized-type, and pair-ID controls collapse to PSA 0.50 and PCDA 0.00, as required by construction. Residual metadata remain informative: a leakage-safe combined probe obtains nested Cal-PCDA 0.397, yet residualizing model scores against measured metadata leaves substantial paired signal. Direct-logit margins for Qwen2.5-3B-Instruct and SmolLM2-1.7B-Instruct reach PSA/Cal-PCDA of 0.987/0.880 and 0.973/0.760, respectively. Qwen's secondary generation check is lower again (PCGA 0.700). These disagreements show why local ranking, shared-threshold decisions, and generated responses should not be collapsed into one score. PAIRSAFE is a bounded diagnostic protocol, not a new classifier or deployment certificate.

Paper 377 KaiToM: A Benchmark for Culturally Informed Theory-of-Mind in Thai Context

Authors: Kun Kerdthaisong, Pasin Buakhaw, Wayupuk Sommuang, Nutchanon Yongsatianchot

Abstract Large Language Models (LLMs) show impressive reasoning capabilities, yet their understanding of culturally specific social dynamics remains limited. We present KaiToM, a Theory of Mind (ToM) benchmark for Thai cultural contexts, focusing on Kreng-Jai, a uniquely Thai concept of social reluctance and consideration. Extending TactfulToM, we construct 882 conversations (39.2K questions) spanning three Kreng-Jai dimensions: White Lies, Avoidance of Confrontation, and Fear-Based Respect. To disentangle language proficiency from cultural knowledge, the dataset includes Thai scenarios at varying cultural depths, general English scenarios, and English translations of Thai content. We evaluate Thai-specialized (Typhoonv2.1, SEA-LION-v3), open-weight (Gemma-3, DeepSeek-V3.2, Kimi-K2, Qwen), and proprietary models (GPT-4o-mini, GPT-5.1, Gemini-3-Flash). Models achieve only moderate performance, struggle with Fear-Based Respect, and cannot reliably identify specific Kreng-Jai types, a failure shared by Thai-specialized models. Human evaluation shows that native speakers significantly outperform models on culturally loaded tasks. Performance also degrades in deep relative to surface-level cultural contexts, and

generalist models often perform better on English translations than on the original Thai, suggesting that models capture surface cues but lack deeper cultural alignment. KaiToM supports the development of culturally aware systems for low-resource languages. Code is available at https://github.com/Augustus2011/KAI-ToM.

Paper 381 A Reliability Audit of Aggregated MTEB Scores: Coverage Asymmetry and Language-Specific Rank Inversions Across 17 Languages

Authors: Archit Rastogi

Abstract Practitioners select multilingual embedding models from aggregated MTEB leaderboard scores, but the same global score can correspond to very different per-language behaviours. We audit MTEB(Multilingual) retrieval rankings against language-specific rankings on 17 languages across three nested model rosters (restricted $n{=}7$; deployment-relevant curated $n{=}25$; broad MMTEB-eligible $n{=}28$-57), using Kendall τ and pairwise inversion rate. At deployment-realistic roster size, median pairwise inversion rate is 33% (per-language Kendall τ reliable after Benjamini-Hochberg FDR in 11/17 languages); the extended tier shows higher agreement ($\tau{=}0.69$), reflecting its multilingual-first composition. Per-language task coverage is asymmetric, and inversion rate decreases with the number of MTEB retrieval tasks available for a language (Spearman $\rho{=}{-}0.59$ curated). A language-balanced aggregator (per-language mean first, then macro-average) drops the median inversion rate from 33% to 14.5% (leave-one-out, $\tau{=}0.71$), the lowest among the held-out aggregators we test and our most actionable mitigation. A cluster of hidden-failure models, headed by granite-311m (16/17 languages) and harrier-0.6b (10/17), ranks in the global top half but collapses to the local bottom quartile, with a distinctively shallow English-multilingual NDCG gap. We release the three-tier audit data, a multilingual RAG benchmark (IT/JA/HI $\times$ finance/law), and the analysis pipeline.

Paper 382 Diagnosing Prompt Sensitivity and Interaction Instability in LLM-Based Humor Judgment

Authors: ~Dohyeon_Kim2, ~Jihoon_Yang1

Abstract Humor judgment is subjective and often depends on ambiguity, expectation violation, and contextual reinterpretation, so the reliability of an LLM judge cannot be assessed from final detection performance alone. Existing LLM-based humor research and LLM-as-a-Judge studies largely focus on final performance, generation quality, or agreement with human evaluation, leaving prompt sensitivity and rebuttal robustness in subjective judgment tasks relatively underexplored. This paper analyzes how prompt design, structured interpretation, and user rebuttals shape LLM-based humor judgment. In zero-shot sentence-only judgments, models with similar F1 scores show substantially different FPR--FNR values, shifting between over-accepting and under-predicting humor depending on prompt formulation. Few-shot examples partially reduce these directional fluctuations, but their effects remain model-dependent. We further propose the Incongruity-Resolution Judgment framework, which compares a humor-supporting interpretation with a counter interpretation before producing a final judgment. Rather than uniformly improving performance, this framework serves as a diagnostic structure for examining how competing interpretations reshape model judgment behavior. Finally, explicit user rebuttals show that high initial performance does not guarantee robustness, highlighting the need to evaluate judgment persistence, directional error patterns, and interaction-level judgment trajectories.

Paper 384 WiC is Not WSD: A Study on LLMs and Lexical Ambiguity Resolution

Authors: ~Yi_Zhou14, ~Kiamehr_Rezaee1, ~Danushka_Bollegala1, ~Mohammad_Taher_Pilehvar2, ~Jose_Camacho-Collados1

Abstract Word-in-Context (WiC) remains challenging for language models, despite recent progress on lexical-semantic tasks. We hypothesise that this difficulty arises not only from comparing two contextual uses of a word, but also from the absence of an explicit sense inventory that specifies the relevant level of semantic granularity. We evaluate open LLMs on WiC and traditional Word Sense Disambiguation (WSD) under similar settings. We find that providing candidate senses, similar to what is done in traditional WSD, improves WiC performance in all settings. In general, explicit sense information helps models make more consistent and targeted judgements. Human evaluation further shows that many apparent WiC errors reflect label ambiguity or mismatches between model and annotator sense boundaries rather than simple failures of lexical understanding. In particular, results show that LLMs overthink the sense distinction often leading to errors based on overly fine-grained distinctions.

Paper 385 MM-NeuroOnco: A Multimodal Benchmark and Instruction Dataset for MRI-Based Brain Tumor Diagnosis

Authors: Feng Guo, Jiaxiang Liu, Yang Li, Qianqian Shi, Mingkun Xu

Abstract Brain tumor MRI diagnosis requires models not only to identify lesions but also to perform fine-grained medical reasoning based on imaging modality, tumor location, and key imaging findings. However, existing medical multimodal datasets often lack structured diagnostic semantics, and closed-ended evaluations may overestimate model ability due to option cues or elimination shortcuts. To address this issue, we introduce MM-NeuroOnco, a large-scale multimodal instruction dataset and benchmark for brain tumor MRI diagnostic reasoning, containing 24,726 MRI slices from 20 public data sources and approximately 200,000 multimodal instructions grounded in imaging evidence. We develop a multi-model collaborative pipeline for semantic completion and quality control, which supplements original tumor labels and segmentation masks with key diagnostic attributes, and further construct MM-NeuroOnco-Bench with a no-LLM-inference policy for ground-truth labels. The benchmark adds a “None of the above” option to closed-ended questions to reduce evaluation bias caused by option shortcuts. Experiments show that existing general-purpose and medical-specialist multimodal large models remain limited on this task, with the strongest baseline achieving only 41.88% accuracy on diagnosis-related questions. Fine-tuning on MM-NeuroOnco yields NeuroOnco-GPT, achieving a 27% absolute accuracy improvement and demonstrating the effectiveness of our dataset in enhancing fine-grained multimodal medical reasoning. Code and dataset documentation are available at https://github.com/gfnnnb/MM-NeuroOnco.

Paper 388 MGSM-Pro: A Simple Strategy for Robust Multilingual Mathematical Reasoning Evaluation

Authors: Tianyi Xu, Kosei Uemura, Alfred Malengo Kondoro, Catherine Nana Nyaah Essuman, Tadesse Destaw Belay, Ifeoma Okoh, Ganiyat Afolabi, Ayodele Awokoya, David Ifeoluwa Adelani

Abstract Large language models have made substantial progress in mathematical reasoning. However, benchmark development for multilingual evaluation has lagged behind English in both difficulty and recency. Recently, GSM-Symbolic showed a strong evidence of high variance when models are evaluated on different instantiations of the same question; however, the evaluation was conducted only in English. In this paper, we introduce MGSM-Pro, an extension of MGSM dataset with GSM-Symbolic approach. Our dataset provides five instantiations per MGSM question by varying names, digits and irrelevant context. Evaluations across nine languages reveal that many low-resource languages suffer large performance drops when tested on

digit instantiations different from those in the original test set. We further find that models robustness in HRL setting do not necessarily translate to LRL. Moreover, proprietary models, such as Gemini 2.5 Flash and GPT-4.1 are less robust to digit, whereas Gemini 3.0 Pro is more robust. Among open models, GPT-OSS 120B and DeepSeek v3 show stronger robustness. Based on these findings, we recommend evaluating each problem using at least five digit-varying instantiations to obtain a more robust and realistic assessment of math reasoning.

Paper 389 IndicIFEval: A Benchmark for Verifiable Instruction-Following Evaluation in 14 Indic Languages

Authors: Thanmay Jayakumar, Mohammed Safi Ur Rahman Khan, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan

Abstract Instruction-following benchmarks remain predominantly English-centric, leaving a critical evaluation gap for the hundreds of millions of Indic language speakers. We introduce IndicIFEval, a benchmark evaluating constrained generation of LLMs across 14 Indic languages using automatically verifiable, rule-based instructions. It comprises around 800 human-verified examples per language spread across two complementary subsets: IndicIFEval-Trans, translated prompts from IFEval carefully localized for Indic contexts, and IndicIFEval-Ground, synthetically generated instructions grounded in native Indic content. We conduct a comprehensive evaluation of major open-weight and proprietary models spanning both reasoning and non-reasoning models. While models maintain strong adherence to formatting constraints, they struggle significantly with lexical and cross-lingual tasks -- and despite progress in high-resource languages, instruction-following across the broader Indic family lags significantly behind English. We release IndicIFEval and its evaluation scripts to support progress on multilingual constrained generation.

Paper 394 Rethinking Cross-Modal Attention for Visual Token Pruning in Multimodal LLMs

Authors: Diego Weng Qiu, Lin Shang

Abstract Recent studies challenge the reliability of cross-modal attention mechanisms as an indicator for visual token pruning in Large Vision-Language Models (LVLMs). In this work, we demonstrate that such negative conclusions can be substantially influenced by an overlooked phenomenon termed task-bound coupling, wherein cross-modal attention patterns become entangled with task-specific prompts. Through controlled configurations that decouple visual inputs from downstream queries, we observe that attention distributions can be relatively consistent across the evaluated settings. Concurrently, we identify an empirical scaling phenomenon which we term Attention Fidelity Decay, where the downstream effectiveness of attention maps exhibits semantic dilution in higher-capacity models or advanced architectures. Motivated by these insights, we introduce TAP-LLaVA, a simple training-free framework that leverages localized task-agnostic attention for efficient, reusable visual token pruning across multi-turn sessions. Our approach achieves competitive downstream performance while substantially lowering computational overhead. This suggests that attention can provide a reliable structural prior at moderate model scales when isolated from textual confounders.

Paper 398 Corrective Feedback-Driven Humor Generation for Tiny Language Models

Authors: Ramakrishna Pinninti, Dibyanayan Bandyopadhyay, Lalit Chandra Routhu, Gaël Dias, Alexis Lechervy, Jeremie Pantin, Mohammed Hasanuzzaman, Asif Ekbal

Abstract Humor generation remains a challenging problem for language models due to its subjective, creative, and culturally sensitive nature. While large language models (LLMs) demonstrate strong generative capabilities, their deployment is often constrained by computational cost. We propose a lightweight, corrective feedback-driven framework in which a compact generator (GPT-2, 140M parameters) produces punchlines under the guidance of an LLM acting as a critic. Unlike traditional knowledge distillation, our method leverages the LLM as a feedback provider rather than a teacher to be mimicked, enabling the

student model to iteratively refine its outputs through corrective supervision. We systematically evaluate multiple LLMs (LLaMA3-8B, Qwen3-8B and Mistral-7B) as guides under different strategies, including frozen and learnable configurations, to study their impact on humor quality and diversity. To support responsible humor generation, we curate a high-quality setup–punchline dataset from Reddit and explicitly filter out harmful content. Experimental results show that corrective feedback-guided training substantially improves humor quality, fluency, and diversity over standard fine-tuning and knowledge distillation baselines, while preserving the efficiency of compact models. We observe that, among the evaluated guides, Mistral-7B consistently provides particularly strong guidance on the curated joke dataset. Our code and resources are available at: https://github.com/Ramakrishna176420/Corrective-Feedback-Driven-Humor-Generation-for-Tiny-Language-Models.

Paper 399 Training-Free Residual-Guided Retrieval for Multi-hop KGQA with Single-Step LLM Inference

Authors: Keito Sei, Sunil Kumar Maurya, Yohei Kobashi, Yutaka Matsuo

Abstract Graph-based retrieval is effective for knowledge graph question answering (KGQA), but retrieving the correct reasoning path remains challenging in multi-hop settings because the retrieval objective changes as reasoning progresses. Existing methods often address this issue with repeated LLM-based planning, which improves adaptivity but increases inference cost. We present a training-free residual-guided retrieval framework that updates the retrieval signal according to the current reasoning path in embedding space. The method performs multi-hop graph retrieval without repeated hop-wise LLM calls and invokes the LLM only once for final answer generation. Experiments on WebQSP and CWQ show that our method achieves superior Hit rates with only a single LLM call. Additional comparisons with supervised graph-based KGQA methods show that the proposed method remains effective without task-specific training data. These results suggest that residual-guided retrieval is an efficient alternative for multi-hop graph-based question answering.

Paper 401 REMIND: Enhancing CoT Reliability with Self-Supervised Introspection

Authors: Souvik Maji, Dibyanayan Bandyopadhyay, Asif Ekbal

Abstract Chain-of-Thought (CoT) prompting has significantly enhanced the reasoning capabilities of Large Language Models (LLMs). However, CoT can also produce fluent but logically flawed reasoning chains that make incorrect answers appear more reliable. In this work, we introduce REMIND (Reliability Enhancement via Metacognitive INtrospection and Debiasing), a framework that enables LLMs to learn from their own failures via episodic memory. Unlike standard error-correction methods that rely on external feedback or extensive fine-tuning, REMIND leverages a self-supervised introspection cycle. We first isolate epistemic failure—instances where CoT decreases model confidence despite yielding incorrect results. The model then performs structured introspection to diagnose the root cause and synthesizes a reusable natural language lesson. These lessons are stored in a vector database and retrieved during future inference steps to correct similar future errors. Experiments on eight diverse reasoning benchmarks spanning mathematical, scientific, multi-hop, truthfulness, and commonsense reasoning demonstrate that REMIND significantly improves reliability by effectively converting past CoT failures into actionable insights for unseen problems. Codes are available in https://github.com/newcodevelop/REMIND.

Paper 402 DeepVision-103K: A Visually Diverse, Broad-Coverage, and Verifiable Mathematical Dataset for Multimodal Reasoning

Authors: ~Haoxiang_Sun3, ~Lizhen_Xu1, ~Bing_Zhao7, ~Wotao_Yin1, ~Wei_Wang162, ~Boyu_Yang5, ~Rui_Wang10, ~HU_WEI4

Abstract Reinforcement Learning with Verifiable Rewards (RLVR) has been shown effective in enhancing the visual reflection and reasoning capabilities of Large Multimodal Models (LMMs). However, existing datasets are predominantly derived from either small-scale manual construction or recombination of prior resources, which limits data diversity and coverage, thereby constraining further gains in model performance. To this end, we introduce \textbf{DeepVision-103K}, a comprehensive dataset for RLVR training that covers diverse K12 mathematical topics, extensive knowledge points, and rich visual elements. Models trained on DeepVision achieve strong performance on multimodal mathematical benchmarks, and generalize effectively to general multimodal reasoning tasks. Further analysis reveals enhanced visual perception, reflection and reasoning capabilities in trained models, validating DeepVision's effectiveness for advancing multimodal reasoning.

Paper 405 SurgAgents: A Multi-agent Framework for Embodied Cataract Surgical Intent Prediction

Authors: Shengyuan Wang, Yanzhe Ji, Yue Yao, Zhenyue Qin, Yang Liu

Abstract Predicting next surgical phase, viewed as a form of surgical procedural intent, with high accuracy is essential for robotic real-time surgical systems. Previous approaches struggle with short-term dynamics, occlusions, and phase persistence; end-to-end large vision-language models (LVLMs) can reason from visual inputs but often fail to exploit structured temporal observations. Thus, we introduce SurgAgents, a multi-agent framework for embodied surgical intent prediction. SurgAgents decomposes prediction into specialized agents: (1) a ResNet-based single-frame classifier agent extracts per-frame phase information; (2) a YOLO-based instrument specialist agent extracts per-frame instrument sets; (3) a history descriptor agent composes these facts into frame-level descriptions; (4) a history summarizer agent compresses long histories into a compact temporal narrative; (5) a phase predictor agent assembles the narrative and the current frame into a formatted LVLM prompt; (6) a retrieval agent with a confidence-gated Retrieval-augmented generation (RAG) module selectively retrieves similar case patterns when prediction confidence is low. On Cataract-1K, SurgAgents attains 46.54% macro-F1 with GPT-4o, outperforming the baseline, with systematic ablation studies showing that each agent materially contributes to performance. Latency measurements demonstrate feasibility for robotic real-time surgical support during surgery procedures.

Paper 406 GROWTHEVAL: Evaluating Stage-Conditioned Developmental Realism in Role-Playing LLMs

Authors: Danping Yan

Abstract Large language models (LLMs) are increasingly used as role-playing dialogue agents in long-horizon applications such as tutoring, companionship, and narrative interaction, where users may expect stable persona consistency as well as plausible stage-conditioned change. Existing evaluations emphasize static persona fidelity, temporal factuality, or isolated social skills, but rarely test whether role-playing models produce coherent developmental differentiation across ordered stages. We introduce GROWTHEVAL, a psychology-inspired diagnostic framework for assessing stage-conditioned developmental realism in single-turn, stage-isolated role-playing responses. Using stage-conditioned scenarios and multi-axis scoring, \textsc{GrowthEval} evaluates cross-stage patterns in developmental appropriateness, role-play coherence, and canon fidelity across five socio-emotional dimensions. Across the evaluated LLMs, we observe a model-tiered empirical pattern: higher-capacity models respond more consistently to stage conditioning but can over-mature at early stages, whereas lower-capacity models exhibit flatter and weakly differentiated profiles. Reliability and robustness analyses provide evidence for the stability of these findings.

Overall, the evaluated LLMs only partially realize coherent, stage-appropriate developmental change. GROWTHEVAL provides an extensible diagnostic framework for analyzing stage-conditioned behavioral differentiation and associated failure modes in single-turn role-playing responses.

Paper 410 S^{4}R: Selective Sampling, Subspaces, and Sparse Reconstruction for Compressed Long-Context KV Caching

Authors: Jialong Han, You Wu, Kewei Tu

Abstract The growth of context window lengths in Large Language Models (LLMs) significantly enhances their long-context capabilities but incurs prohibitive memory costs due to the Key-Value (KV) cache. Although low-rank compression of KV cache is a promising remedy, existing methods face a dilemma: offline approaches depend on external calibration data, whereas online approaches incur substantial compute for full-prompt decomposition and reconstruction. In this paper, we propose **S^{4}R**, which builds low-rank subspaces from selectively sampled tokens and computes attention over a sparsely reconstructed KV representation. S^{4}R uses prompt-aware initialization to build initial key/value bases from a representative prompt subset, trading off calibration-data dependence against prefilling cost. Because fully reconstructing the cache at every decoding step is prohibitively expensive and hurts throughput, we further adopt sparse reconstruction to retain only informative positions during decoding. Extensive experiments on LongBench and RULER with Llama and Qwen model families show that S^{4}R achieves up to 5x KV compression with near full-cache accuracy, combining the efficiency of fixed compression with the adaptability of prompt-dependent methods. Our code is publicly available at https://github.com/Draconis98/S4R

Paper 411 SEAL-PR: Scientific Evidence-Audited and Locally Isolated Peer Review under Untrusted Manuscripts

Authors: Zida Yang, Guquan You, Weiqing Zeng, Yu Zheng

Abstract Large language models (LLMs) are increasingly used to support scientific peer review, yet most review systems treat the manuscript as a trustworthy input channel. This assumption fails when manuscripts contain hidden instructions, rubric-replacing text, citation pollution, fabricated evidence, or formatting attacks aimed at generative reviewers. We formalize Untrusted-Manuscript Peer Review, where scientific judgment should remain invariant to non-scientific perturbations and every certified critique must be bound to trust-screened evidence. We introduce SEAL-PR, a five-stage architecture combining a forensic span parser, a five-label trust lattice, a typed evidence-audit graph, sealed/shadow dual-view reviewing, and evidence-constrained decoding. We separate two classes of claim throughout. Structurally, no manuscript span can enter the trusted-instruction channel and no certified critique can cite an inadmissible span, so evidence admissibility is 1.000 by construction rather than by measurement. Empirically, on SEAL-PR-BENCH—a paired benchmark of 5129 papers across 7 venues, 129 independent evaluation manuscripts, 1290 clean+attacked records and 9 attack families—SEAL-PR attains 0.994 clean expert weakness recall, preserves 0.993 under attack, and cuts the text-acceptance attack rate from 0.866 to 0.000, without reviewer-LLM fine-tuning. Because the corpus predates our backbones, absolute EWR is an in-corpus ceiling; the paired clean–attacked contrast is the quantity we treat as informative. SEAL-PR reframes AI peer review from fluent review generation to locally isolated, evidence-audited scientific assistance.

Paper 412 QUILT: Query–User Interaction Linguistic Taxonomy

Authors: ~Yuyang_Dai4

Abstract Large language models are used by people with very different expertise levels, but existing work studies what users ask more than who is asking. We introduce QUILT (Query User Interaction Linguistic Taxonomy) , a two-dimensional framework that jointly characterizes LLM conversations by task type and user expertise level. Applying QUILT to 3,000 WildChat dialogues, we construct the first large-scale corpus linking expertise level to task-level interaction patterns in open-ended LLM conversations. Corpus analysis reveals clear expertise-dependent asymmetries: debugging queries are strongly associated with expert users, novice users rarely engage in reasoning-intensive tasks, and expert queries are substantially longer across all task types. These patterns are statistically significant and recoverable by automatic classifiers, indicating that expertise differences are systematically reflected in conversational language. Together, QUILT and the QUILT Corpus provide an empirical foundation for studying user heterogeneity in LLM interaction and for developing expertise-aware response strategies.

Paper 413 You Don’t Understand: Benchmarking Experiential Comprehension Gaps in LLMs

Authors: Yuyang Dai

Abstract Current LLM evaluation frameworks primarily assess understanding through task performance, focusing on factual accuracy, reasoning, and instruction following, while overlooking the experiential dimensions of human comprehension. Models that produce fluent, empathetic responses score highly even when they claim experiential access they cannot possess. We introduce ComprehensionGap, a benchmark of 500 scenarios spanning five empirically grounded categories of experiential incommensurability: Embodied Qualia, Episodic Incommensurability, Cultural Tacit Knowledge, Relational Situatedness, and Existential Finitude. Using a human-judged protocol measuring Acknowledgement of Limit, Absence of False Empathy, and Situational Relevance, we evaluate six frontier and open-weight models. Our experiments reveal a consistent Grounding Gap: all models maintain high fluency while systematically overclaiming experiential understanding, with larger models producing more convincing yet less honest mimicry. Models can calibrate for information gaps but not experiential ones. These findings argue for integrating experiential calibration into future alignment and evaluation frameworks.

Paper 416 VERGE: Verification-Enhanced Generation of Multi-Hop Evaluation Datasets for Task-Specific RAG

Authors: Kyosuke Morita

Abstract Evaluating retrieval-augmented generation (RAG) on custom knowledge bases requires task-specific benchmarks of multi-chunk evidence synthesis, yet most existing generation pipelines apply quality control only after the full question set has been solved. We present VERGE, which moves quality control inside the generation loop: a verification agent targets, per item and without an external solver pass, an operational multi-chunk criterion and a QA-integrity criterion, refining items that fail and discarding those it cannot repair. From 11,052 candidates across six domains, VERGE yields 2,854 verified multiple-choice items targeting aggregation- and comparison-style synthesis over topically coherent evidence. We contribute a method-agnostic validation protocol — exhaustive strict-subset ablation, chunk-removal ablation, format-unconfounded open-ended re-evaluation, and inter-chunk-similarity stratification — and use it to localise where the multi-chunk dependence holds. On a stratified sample (n = 250), 69.8\% of items not answerable closed-book cannot be answered from any strict subset of their supporting chunks (strict-necessity violation rate 15.2\%). A human annotation study finds the verifier reliable (Cohen's κ = 0.903) and VERGE items higher in QA integrity and distractor quality than the closest prior method (p $<$ 0.001).

Paper 417 Encoded but Not Decoded: Layer-Localized Evidence for a Three-Level Gap in LLM Syntax

Authors: ~Zhenyan_Lu2, ~He_Wang37, ~Xiaohui_Huang4

Abstract A language model can fail a syntactic test in two distinct ways: by not encoding the relevant structure, or by encoding it but failing to use it at the output. Behavioral evaluation alone cannot tell these apart. We propose a three-level evaluation framework (behavioral deployment, LM-head readout, and probe recoverability) measured on the same items under the same binary decision. Using a compact trilingual (English, Chinese, German) control-dependency benchmark, we find that probe recoverability exceeds or equals LM-head readout, which in turn exceeds or equals behavioral deployment, across seven models and all three languages in the aggregate. The recoverability surplus is never negative across all 14 (model, task) conditions. The disconnect concentrates in subject-control, where a nearest-noun heuristic gives the wrong answer. The single largest gap (0.653) appears on Qwen3-0.6B Instruct in question answering. The gap persists at Qwen3-14B Instruct. Instruction tuning degrades deployment more than encoding in percentage terms. We rule out option-position bias, late-layer erasure, output-formatting artifacts, and probe-training variance. The pattern is consistent with decoding that favors surface shortcuts, and the behavior-probe gap measures the strength of that preference. Activation patching shows the gap is layer-localized. Under instruction tuning, the LM-head-decoded layer shifts approximately ten layers later than the probe-decoded layer. These findings argue that behavioral evaluation understates what models encode, while probing alone overstates what they deploy.

Paper 421 Auditing PEFT Claims: A Stability Probe Unifying Three LoRA-Variant Fragility Patterns

Authors: Suan Lee, JaeSeong Kim

Abstract Most published PEFT improvements over LoRA do not survive paired multi-seed evaluation. We audit four methods — DoRA, PiSSA, AdaLoRA, and rsLoRA — against LoRA across 16 claim cells spanning 125M–7B, using paired seeds, bootstrap confidence intervals, sign-flip permutation tests, and Holm–Bonferroni correction, with adjudication rules committed before the confirmatory runs. Only 2 of 15 confirmatory cells remain supported after correction, and 4 reverse outright. Three findings stand out. (i) DoRA shows no supported gain in any of its seven audited cells. (ii) PiSSA's +2.6% GSM8K advantage under a common budget reverses under paper-faithful hyperparameters, including in the cleanest claim-matched 7B setting (Gemma-7B, MetaMathQA → GSM8K: $\Delta = -1.58$pp, $p \approx 0.004$, 1W/9L). (iii) PiSSA's in-distribution NLI gain reverses on adversarial OOD yet improves on counterfactual sentiment, so in-distribution audits alone can mislead. A 303-run stability landscape then supplies a mechanism for three of the four variants. The effective adapter learning rate, $\mathrm{lr} \times \alpha/\sqrt{r}$, marks a risk boundary near 5×10^{-4} on BERT-base/MNLI/AdamW: nothing below it collapsed (40/40 stable), while 11 of 32 configurations above it did. The boundary transfers to 7B (Mistral rsLoRA: 3/10 seeds degrade at $\alpha/\sqrt{r} = 4$) and predicts DoRA's α-driven failures at high rank, but it is calibrated rather than universal: a second 7B cell trains stably well above it once dtype is controlled, and we withdraw that cell as supporting evidence. AdaLoRA is excluded from this account: its failure is schedule-driven rather than driven by $\mathrm{lr} \times \alpha/\sqrt{r}$. We release all configs, seeds, and scripts, and propose a minimal reporting standard: ≥10 paired seeds, bootstrap CIs, a regime ablation, and disclosure of $\alpha/\sqrt{r}$.

Paper 422 Memorization or Extraction? Auditing LLM-Based Recovery of Scientific Assumptions

Authors: Suan Lee, JaeSeong Kim

Abstract We audit whether benchmarks for LLM-based recovery of scientific assumptions measure corpus-grounded extraction or pretraining memorization. Using author-defined targets for 10 historical AI breakthroughs (frozen 4-case protocol plus 6 post-hoc stress cases) and a "field-wide paradigmatic assumption" prompt, GPT-4o's top-10 extractions cover the target in 6/10 cases under a generous any-of-top-K criterion---an upper bound that a 10-rater study corroborates for only ${\sim}0.19$ of sampled

top-ranked pairs (human--rule $\kappa{=}0.70$, $\alpha{=}0.634$). Controlled stress tests structure the diagnosis: a no-corpus control (field name only) still scores 4/4 on the primary shifts (memorization dominates), while a wrong-corpus control collapses to 0/12 across all $4{\times}3$ cross-pairings (corpus can be causally read); a five-LLM canonical sweep (GPT-4o 4/4, Qwen 2/4, Claude 1/4, Llama-3.1 1/4, Mistral 0/4) shows the gap is model-specific rather than closed-vs-open, with format compliance a major but not fully isolable contributor; a Min-K\% Prob cross-check actually mis-ranks synthetic non-members as more contaminated than real pre-shift papers, motivating the behavioral framing; and synthetic/niche controls are mixed (causal in 1 of 3 synthetic scenarios, one shows corpus harm). We release an audit kit --- five factorable controls (no-corpus, wrong-corpus, pair-level precision, ${\geq}2$-LLM, multi-seed; ${\sim}\$15$) --- as a reusable benchmark-validity probe, with all artifacts released.

Paper 424 CAFE: Caption-Anchored Feature Enrichment and Evidence Routing for Multimodal Sarcasm Detection

Authors: Puchu Zeng, Yun Luo, qianqian li, Yanzhou Ma, Liu Junyi, Song Fengxuan, chonghuifang

Abstract Multimodal sarcasm detection aims to identify sarcasm in image-text posts by capturing how visual content changes the interpretation of short informal text. Existing methods have advanced the task through multimodal fusion, caption-based cues, and fact/sentiment incongruity modeling, but their evidence use is often controlled inside latent feature channels or reduced discrepancy scores. This makes it difficult to decide which visual information should be selectively weighted for a particular post. We propose CAFE, a caption-anchored feature enrichment and evidence routing framework that decides how to use visual evidence for each sample under ambiguity. CAFE enriches each image with two generated descriptions, one for visible facts and one for the image-text relation, then estimates how useful each description is for the current post and fuses caption-calibrated contrast evidence with direct cross-modal evidence for prediction. Experiments show that CAFE achieves state-of-the-art performance on the primary MMSD2.0 benchmark with 87.81+/-0.33% F1 (mean over 5 seeds). Component ablations on the original MMSD provide an additional evaluation.

Paper 427 ILRR: Inference-Time Steering Method for Masked Diffusion Language Models

Authors: Eden Avrahami, Eliya Nachmani

Abstract Discrete Diffusion Language Models (DLMs) offer a promising non-autoregressive alternative for text generation, yet effective mechanisms for inference-time control remain relatively underexplored. Existing approaches include sampling-level guidance or trajectory optimization mechanisms. In this work, we study the paradigm of reference-based latent steering for DLMs. We introduce Iterative Latent Representation Refinement (ILRR), an efficient framework for steering DLMs using a reference text as a high-level semantic blueprint. ILRR extracts and injects reference-derived semantic signals into the evolving activations of the generated sequence, enabling tunable transfer of coarse properties such as sentiment. We further introduce Spatially Modulated Steering, an extension that enables long-form generation to be guided by shorter references by regulating intensity across the sequence. Empirically, we demonstrate that ILRR achieves effective control on LLaDA and MDLM architectures with low computational overhead, requiring only one additional parallel forward pass per denoising step. Under comparable compute budgets, ILRR improves attribute accuracy over baselines by 10% to 60% points. Our results suggest that the iterative, global denoising process makes DLMs a natural substrate for effective sequence-wide activation-level control.

Paper 428 How Far Do Auto-Interpretation Labels Generalize: A Controlled Study Across Languages, Scripts, and Rewordings

Authors: Sripad Karne

Abstract Sparse autoencoder (SAE) features are increasingly used to interpret language models, with auto-generated natural-language labels serving as the primary interface for understanding what each feature represents. We ask whether these labels generalize: does a feature labeled for a concept actually track that concept across languages and scripts? Using Serbian digraphia as a controlled testbed---the same language written in both Latin and Cyrillic via deterministic transliteration---we first find that SAE feature sets activated by the same content in different languages, scripts, and wordings share substantial overlap (mean Jaccard 0.39 vs. 0.13 random baseline, peaking at 0.57), suggesting genuine cross-lingual semantic features. We then test whether auto-interpretation labels keep pace. They often do not: features whose labels describe semantic content miss the same meaning in Serbian up to $4\times$ more often than within English, and miss Serbian Cyrillic more than Serbian Latin---two scripts that are deterministic transliterations of each other. The gap grows with network depth, yet the labels give no indication that they fail. These results suggest that auto-interpretation labels reflect a feature's behavior on the languages and scripts a model has seen most in training, rather than the concept itself.

Paper 429 The Missing "I Don't Know": Why Three Reasoning-Reliability Findings Converge on Calibrated Abstention

Authors: Srijith Ravikumar

Abstract Three recent results describe what look like unrelated LLM reliability problems. Yin et al. (2026) show reasoning RL collapses tool-reliability representations. Suleymanov et al. (2026) show that under safety-constrained generation, large models rewrite flagged spans while small models truncate. Bastounis et al. (2024) prove any consistent-reasoning system without an implicit "I don't know" function must hallucinate infinitely often on broad problem classes. We argue these findings converge on a single intervention: calibrated abstention is what each independently identifies as the missing capability, even though the unavailability they document, a capability gap, a policy gap, and a recursion-theoretic gap, has a different source in each case. Honesty post-training has narrowed the gap in deployed models, but principled closure of the class Bastounis identifies requires a calibrated abstention function whose training signal at the leaderboard level is absent: dominant benchmarks assign zero reward to decline, so the leaderboard gradient that would select for the function does not exist. We propose four changes to evaluation: triple-scoring, abstention-rate reporting, capability-stratified evaluation, and mandatory calibration metrics. Benchmark reform is necessary, not sufficient, for closing the gap the theorem identifies.

Paper 432 The Limits of Simulated Societies: How LLM Alignment Erases Cross-Cultural Variance

Authors: Rojin Ziaei

Abstract Using large language models (LLMs) to simulate diverse human populations promises to transform computational social science. However, the accuracy of these simulations remains unclear because current evaluations prioritize predicting an average response over capturing the rich diversity of opinions within real human groups. Here, we develop a diagnostic framework measuring both accuracy and viewpoint retention across 10,000 respondent-question pairs from the World Values Survey, spanning twelve countries and six continents. By evaluating nine LLMs across four alignment paradigms (zero-shot, SFT, DPO, GRPO) against ground-truth human responses, we identify a failure mode we term consensus collapse. While training alignment improves point accuracy, we find that it compresses the range of outputs toward a single demographic stereotype. The highest-performing model (Tulu 3 70B DPO, 57.9% accuracy) retains only 50% of real human variance (σr =

0.50). We demonstrate this collapse is within the models' underlying parameters rather than decoding procedures; temperature scaling up to T = 1.0 alters fewer than 3.7% of predictions. Furthermore, this diversity loss is inequitable, erasing significantly more variation for Global South populations like Nigeria (σr = 0.11) and Ethiopia (σr = 0.16) compared to Western democracies (σr = 0.70–0.87), with accuracy-rewarded GRPO unable to reverse this effect. While mixing an aligned model with an unaligned prior partially mitigates the collapse (reducing W1 from 0.55 to 0.52), substantial gaps remain (Nigeria recovering from σr = 0.14 to 0.47). Ultimately, we find that modern AI alignment methods trade distributional diversity for point accuracy, challenging existing assumptions for using LLMs as substitutes for human populations. Yet, by exposing the mechanics of this collapse, our work provides insight for future methods designed to value and preserve diverse human viewpoints.

Paper 434 LLM-Ideoplasticity: Measuring Ideological Plasticity in the Political Behavior of LLMs as a Context-Conditioned Distribution

Authors: Adib Sakhawat, Syed Rifat Raiyan, Tahsin Islam, Takia Farhin, Hasan Mahmud, Md Kamrul Hasan

Abstract We argue, with systematic empirical evidence, that a large language model's political ideology is not a fixed point, but a *conditional distribution* $\mathbb{P}($position$\mid$context$)$ over a real political space. We evaluate nine current LLMs using a unified measurement framework anchored by VAA–CHES projection models, which map responses onto three validated dimensions (*lrgen*, *lrecon*, *galtan*) across six contextual axes. Our findings reveal high sensitivity to context: persuasive framing and under-represented languages displace coordinates by up to 0.57 and 0.52 units, respectively, while chain-of-thought reasoning often amplifies rather than dampens paraphrase instability. Despite this local plasticity, the model cohort occupies a remarkably narrow Overton envelope overall, occupying roughly one-third the spread of major European parties. Supported by a multi-trait multi-method (MTMM) analysis, we conclude that a single point cannot summarize LLM political behavior; it must be characterized as a shape. Our code and data are publicly available.

Paper 435 ProMoral-Bench: Comparing Prompting Strategies for Moral Judgment and Jailbreak Robustness in LLMs

Authors: ~Rohan_Subramanian_Thomas1, ~Shikhar_Shiromani1, ~Abdullah_Chaudhry1, ~Ruizhe_Li2, ~Vasu_Sharma1, ~Kevin_Zhu3

Abstract Prompt design affects the moral judgments and refusal behavior produced by large language models (LLMs), yet controlled comparisons remain fragmented across datasets and models. We introduce ProMoral-Bench, a benchmark-style comparison of 10 prompting paradigms across four LLM families. Using ETHICS, Scruples, WildJailbreak, and our 200-pair ETHICS-Contrast set, we report task-level metrics together with Moral Competence Score (MCS), jailbreak/over-refusal Safety Robustness Score (SRS), and Unified Moral Safety Score (UMSS). UMSS is a study-specific relative summary of MCS and SRS, not an absolute measure of model safety or moral grounding. In this fixed-template evaluation, compact exemplar-guided scaffolds attain stronger competence-robustness trade-offs at lower token cost than the particular multi-stage variants tested, although multi-stage prompting occasionally improves jailbreak refusal. The few-shot demonstrations are non-overlapping but drawn from the evaluation distributions, so their advantage should not be read as out-of-distribution robustness. Code and data are available at https://github.com/rohanthomas334-sketch/ProMoral_Bench.

Paper 439 Probing for Negation: Extracting Compositional Negation Signals from Pretrained Language Representations for Image Retrieval

Authors: Animesh Singh, Brejesh Lall

Abstract Vision-language models such as CLIP exhibit strong zero-shot performance on image-text matching yet consistently fail to handle negation compositionally. Queries like "a dog that is not running" are collapsed into co-occurrence statistics, with the negation cue ignored. Critically, this failure is not random: because negation removes the most discriminative concept from a query, negation-blind models are systematically misled, they confidently rank the wrong images highest rather than producing an uninformative ranking. Prior work addresses this by retraining with negation-augmented data or learning embedding-level transformations. We propose vpCLIP (verypositiveCLIP), a fundamentally different approach: we extract the negation target at the query level using a frozen DistilBERT model (66M parameters). vpCLIP probes intermediate attention layers for negation-conditioned attention flow, removes candidate concept directions from hidden representations, and scores the resulting semantic shift to identify the negated concept. The extracted signal drives a retrieve-and-rerank pipeline on top of unmodified CLIP. On three curated negation benchmarks (Visual Genome, MS-COCO, Flickr30k; 800 queries total) where random retrieval averages 47-50%, vpCLIP achieves 68-79% Top-1 accuracy with CLIP ViT-B/32, outperforming CLIP (27-41%), NegCLIP (24-37%), ConCLIP (41-58%), and BLIP-2 (26-34%), all of which score below random on average, while requiring no task-specific training. Because the extraction stage is backbone-agnostic, vpCLIP generalizes without modification: paired with SigLIP2 SO400M it reaches 87.5% on VG, 93.2% on COCO, and 85.6% on F30k (88.8% average). We further demonstrate zero-shot transfer to an affirmation-biased medical encoder (PLIP), achieving 96% Top-1 on histology images. Our results show that negation signals are latent in the intermediate representations of an auxiliary pretrained language model and require structured probing to extract, challenging the assumption that compositional negation necessitates large-scale retraining of the vision-language backbone.

Paper 440 Enhancing Safety Alignment by Universal Adversarial Prompts

Authors: ~Yuan_Zhou14, ~Xiangyu_Zhang3

Abstract Large Language Models (LLMs) exhibit remarkable capabilities but remain susceptible to adversarial attacks aimed at eliciting harmful responses. Current defense mechanisms, while partially effective, often compromise the models' performance on benign tasks. To address this issue, we propose a novel two-stage framework to enhance LLM robustness against adversarial attacks. Initially, we train a universal adversarial prompt designed to align the hidden representations of harmful inputs with benign ones, effectively emulating adversarial attack patterns. Subsequently, we utilize this soft prompt to augment training data, strengthening the model's capability to reject harmful content. We conduct extensive evaluations to empirically demonstrate that our method outperforms existing methods such as Circuit Breaker (CB) and Deliberative Alignment SFT (DSFT), achieving average defense rate of 96.4% on llama3-8b-instruct model, compared to CB (90.6%) and DSFT (87.9%).

Paper 442 RL-MR: Reinforcement Learning based Budget-Aware Sequential Pretrained Model Ranking

Authors: ~Saddam_Mukta1, ~Md._Ahashan_Habib1, ~Mir_Tafseer_Nayeem1, ~Md_Adnanul_Islam1

Abstract The rapid growth of pre-trained models has made selecting suitable models for downstream tasks challenging. Exhaustively fine-tuning or evaluating every candidate model is computa- tionally expensive, while existing proxy-based metrics such as LEEP, NCE, and LogME pro- duce one-shot static rankings that do not use feedback from earlier evaluations. We propose Reinforcement Learning–based Model Rank- ing (RL-MR), a budget-aware framework that reformulates pretrained model selection as a sequential decision-making problem. RL-MR learns a ranking policy that uses proxy-based transferability signals as rewards to guide the order in which candidate models are evaluated. In controlled experiments on SST-2, Yelp Po-

larity, Amazon Polarity, and DynaSent, RL- MR consistently improves ranking correlation over static proxy baselines, achieving up to Kendall's τ = 0.8695 while maintaining Top-1 selection accuracy. These results suggest that sequential proxy feedback can provide a use- ful alternative to one-shot model ranking under limited evaluation budgets.

Paper 447 Effects of Theory of Mind and Prosocial Beliefs on Steering Human-Aligned Behaviors of LLMs in Ultimatum Games

Authors: Neemesh Yadav, Yihuai Lan, Shan Dong, Mai Hieu Hien, Palakorn Achananuparp, Jing Jiang, Ee-Peng Lim

Abstract Large Language Models (LLMs) have shown potential in simulating human behaviors and performing theory-of-mind (ToM) reasoning, crucial for complex social interactions. We investigate ToM reasoning's role in aligning agentic behaviors with human norms in negotiation tasks, using the ultimatum game as our referenced task. We initialized LLM agents with different prosocial beliefs (Greedy, Fair, Selfless) and reasoning methods (chain of thought and ToM reasoning of varying levels), examining their decision-making processes and outcomes across multiple LLMs, including reasoning models like o3-mini and DeepSeek-R1 Distilled Qwen 32B. We perform 2,700 simulations to show that ToM reasoning enhances behavioral alignment with human norms, decision-making consistency, and negotiation outcomes. Consistent with prior findings, reasoning LLMs exhibit limited capability compared to ToM-enhanced LLMs, with different game roles benefiting from different ToM orders.Fair proposers and responders accepting offers were the most consistent with their strategic reasonings, whereas all agents showed strong consistencies with human beliefs when rejecting offers, except when the offer was fair. Human verification further revealed that Llama 3.3 70B produces reasoning most consistent with its actions and beliefs. Our findings advance understanding of ToM's role in human-AI interaction and cooperative decision-making. The code used for our experiments can be found at https://anonymous.4open.science/r/UltimatumToM.

Paper 450 From Relation-Specific Heads to Shared Interventions: Contextual Entrainment in Language Models

Authors: Jinghan Zheng, Jiajun Shi

Abstract Large language models rely on prompt context to perform in-context learning, follow formats, and adapt to local evidence. This ability, however, also raises a robustness concern: models may reuse irrelevant tokens from earlier context when answering a later query. Prior work identifies this behavior as contextual entrainment and finds attention heads whose ablation reduces it, arguing that these heads are largely task- or relation-specific. But are the mechanisms behind contextual entrainment truly isolated by relation, or do relation-conditioned searches recover reusable components shared across relations? Using factual-distractor relation-completion prompts across 19 lexical-semantic subtypes, we identify attention heads whose ablation reduces entrainment and test how the resulting head sets transfer recovery ability across exact relations, semantic families, and mixed relation distributions. Same-family transfer exceeds different-family transfer, family residuals weaken once globally shared heads are removed, and subtype-specific searches repeatedly rediscover many global heads. Clean-completion and context-use probes further suggest that this shared intervention can reduce harmful copying while preserving ordinary relation completion and simple in-context abilities. Our findings provide a more precise perspective on contextual entrainment and its mitigation: harmful context following is not best addressed by searching for isolated relation-specific circuits, but by using heterogeneous attack prompts to reveal reusable intervention targets whose effects are shaped by semantic relation structure.

Paper 451 Edit Policy-Aligned Text Simplification

Authors: Xuanxin Wu, Yuki Arase, Masaaki Nagata

Abstract Text simplification aims to modify documents and sentences to make them easier to read and understand while preserving their meanings. Different applications require distinct simplification policies, e.g., lexical-oriented simplification and overall rewriting while trading off detail for simplicity. However, achieving such policy-driven control remains an open challenge. In this work, we introduce a simple yet powerful approach that leverages Large Language Model-as-a-Judge (LLM-Judge) to automatically construct policy-aligned training data, removing the need for costly human annotation or parallel corpora. Our method enables the construction of simplification models that adapt to diverse simplification policies at both the sentence and document levels. Remarkably, small-scale open-source LLMs match or surpass GPT-5, as verified by both automatic metrics and edit analysis. The consistent improvements across model families and sizes demonstrate the robustness of our approach.

Paper 453 Do LLM Embedding Spaces Recover Expert Structure? Evidence from Mental-Health Language

Authors: ~Yixuan_Zhu6, ~Zhenke_Duan1, ~Fanghen_Li1

Abstract Pretrained text embeddings are increasingly used as representational maps, yet high category separability does not imply that their geometry recovers expert-defined structure. We study this problem in mental-health-related language, where symptom relations provide an external reference and online communities introduce strong domain, affective, stylistic, and discourse confounds. Using 28 Reddit communities, we compare pretrained and supervised fine-tuned Qwen3 embedding spaces at two scales (0.6B and 4B). We construct category prototypes, evaluate their representational dissimilarity matrices against an expert symptom matrix with representational similarity analysis, and complement this global test with prototype-based typicality and multi-baseline confound controls. Pretrained embeddings show measurable alignment with expert structure within the mental-health subset; fine-tuning strengthens this alignment most at the finest category level; and larger scale improves both zero-shot alignment and supervision-induced gains. Residual alignment remains substantial after controlling for VAD, LIWC, lexical style, and topic-distribution structure. These results suggest that LLM embeddings can recover expert-relevant category geometry, but this recovery is level-dependent and should be tested against explicit confounds rather than inferred from classification alone.

Paper 455 The Model Knows Where to Look: Visual-Grounding-Guided Head Suppression for Mitigating Hallucination in LVLMs

Authors: Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, Byron Fang

Abstract Large Vision-Language Models (LVLMs) achieve strong multimodal reasoning capabilities, yet are still prone to object hallucination - often generating content that is not grounded by the visual input. Existing attention-based training-free mitigating methods rely on visual attention heads identified by high attention weights or sensitivity to visual token perturbations. However, we show that these identified visual attention heads are not necessarily responsible for visual grounding, and using this as guidance is suboptimal. Motivated by this observation, we propose a novel training-free decoding method that identifies visually sensitive heads by counterfactual region masking. We find that these heads provide a clear signal when visual grounding is required and can therefore be used to mitigate hallucination by guiding the suppression of hallucination heads. Across multiple LVLMs and benchmark image captioning and VQA datasets, our method achieves best or near-best performance in reducing hallucination while preserving generation quality.

Paper 457 Prompt-Based Abstention Fails Under Misleading Context: A Controlled Study of Small Frozen RAG Models

Authors: Yohanes Andre Setiawan

Abstract Missing and misleading evidence are not the same problem in retrieval-augmented generation (RAG), but prompt-based abstention treats them alike. Small frozen models abstain when context is absent, not when it is misleading. We introduce GRAB-RAG (Graded Abstention Benchmark for Retrieval-Augmented Generation), a paired benchmark that tests the same questions across four context conditions (supportive, degraded, missing, and misleading) in Natural Questions and HotpotQA. In the misleading condition, we edit a gold passage to support a wrong answer and place it among other retrieved passages. We test five abstention policies on three small frozen models (3.8B–8B) across two English QA benchmarks. Models abstain reliably when evidence is missing, but under explicit abstention prompting still answer 41.6% of misleading questions, with 63% of those answers echoing the planted wrong entity. Chain-of-thought provides little additional benefit. A generator-side conflict check cuts the rate to 13.3% but discards many correct answers, while a natural language inference (NLI) verifier recovers that coverage but fails when parametric memory and the misleading passage agree on the same wrong answer. Prompt-based abstention asks whether context is sufficient, not whether it is correct. The two-pass checks are diagnostic baselines, not deployable mitigations.

Paper 459 PersonaPath: Towards Knowledge-Centric Personalized Learning Path Planning

Authors: Yu Liu, Zeming Liu, Tianle Zhang, Zihao Cheng, Yuhang Guo, Kehai Chen, Min Zhang, Yunhong Wang, Haifeng Wang

Abstract Adaptive learning systems commonly formulate learning path planning as Exercise-Centric (EC) recommendation, where the next step is inferred from item-level interaction logs. Evaluating goal-oriented guidance additionally requires explicit learner goals and curriculum-scale prerequisites: learners with similar exercise records may need different paths toward their targets. We therefore study Knowledge-Centric (KC) personalized learning path planning, where a planner must reason over learner profiles, mastery states, and prerequisite knowledge structures to decide which textbook, unit, and concept should be studied next. To support this setting, we introduce PersonaPath, a benchmark that pairs 2,000 fine-grained learner personas with a hierarchical knowledge graph of 347 textbooks, 1,751 units, and 4,092 concepts across 77 subjects. We evaluate representative LLMs on PersonaPath. Results show that even the strongest LLM reaches only a 29.5% final pass rate in Basic Education, and that the main bottleneck lies in adaptivity, where no model exceeds 44.7% in tailoring paths to individual learners.

Paper 460 CogMem: A Cognitive Memory Architecture for Sustained Multi-Turn Reasoning in Large Language Models

Authors: Yiran Zhang, Jincheng Hu, Mark Dras, Usman Naseem

Abstract Large language models (LLMs) excel at single-turn reasoning but often lose accuracy and coherence over extended, multi-turn interactions. Recent evaluations such as TurnBench highlight recurring failure modes, including reasoning bias, task drift, hallucination, overconfidence, and memory decay. Current approaches typically append full conversational histories, causing unbounded context growth, higher computational costs, and degraded reasoning efficiency. We introduce CogMem, a cognitively inspired, memory-augmented LLM architecture that supports sustained iterative reasoning through structured, persistent memory. CogMem incorporates three layers: a Long-Term Memory (LTM) that consolidates cross-session reasoning strategies; a Direct Access (DA) memory that maintains session-level notes and retrieves relevant long-term memories; and a Focus of Attention (FoA) mechanism that dynamically reconstructs concise, task-relevant context at each turn. Experiments on TurnBench show that this layered design mitigates reasoning failures, controls context growth, and improves consistency across

extended reasoning chains, suggesting that structured hierarchical memory is a promising direction for improving long-horizon multi-turn reasoning in large language models.

Paper 461 Multicultural Spyfall: Assessing LLMs through a Dynamic Multilingual Social Deduction Game

Authors: Haryo Akbarianto Wibowo, Alaa Elsetohy, Qinrong Cui, Alham Fikri Aji

Abstract The rapid advancement of Large Language Models (LLMs) has necessitated more robust evaluation methods that go beyond static benchmarks, which are increasingly prone to data saturation and leakage. In this paper, we propose a dynamic benchmarking framework for evaluating multilingual and multicultural capabilities through the social deduction game Spyfall. In our setup, models must engage in strategic dialogue to either identify a secret agent or avoid detection, utilizing culturally relevant locations or local foods. Our results show that our game-based rankings are consistent with Chatbot Arena rankings on the overlapping model. However, we find a significant performance gap in non-English contexts: models are generally less proficient when handling locally specific entities and often struggle with rule-following or strategic integrity in non-English languages. We demonstrate that this game-based approach provides a scalable, leakage-resistant, and culturally nuanced alternative to traditional NLP benchmarks. The dataset (game history) and source code can be accessed here.

Paper 462 Counting Clues: A Lightweight Probabilistic Baseline Can Match an LLM

Authors: Furong Jia, Yuan Pu, Finn Guo, Monica Agrawal

Abstract Large language models (LLMs) excel on multiple-choice clinical diagnosis benchmarks, yet it is unclear how much of this performance reflects underlying probabilistic reasoning. We study this through questions from MedQA that ask for the most likely diagnosis. We introduce Frequency-Based Probabilistic Ranker (FBPR), a lightweight method that scores options with a smoothed Naïve Bayes over concept-diagnosis co-occurrence statistics from a large corpus. When co-occurrence statistics are sourced from Dolma, the pretraining corpus for OLMo, and RedPajama, an open reproduction of the LLaMA pretraining-data recipe, FBPR achieves comparable performance to the corresponding LLMs. Furthermore, direct LLM inference and FBPR exhibit complementary strengths, correctly answering largely different subsets of questions with overlap barely above chance. Combining them with a simple hybrid method yields a notable gain over either method alone. These findings show that explicit probabilistic baselines can account for a substantial portion of the benchmark performance on diagnosis questions, while serving as historically grounded reference points and complementary signals for hybridization.

Paper 464 Benchmarking Fine-Grained Readability Explanations for Arabic

Authors: Nour Rabih, Khalid N. Elmadani, Hanada Taha, Nizar Habash

Abstract In this work, we introduce BAREC-X, the first explanation-annotated Arabic readability dataset, comprising 1,793 sentences drawn from the Balanced Arabic Readability Evaluation Corpus (BAREC) and labeled by expert annotators with linguistic justifications aligned with the BAREC guidelines. Annotators reached an average inter-annotator agreement of 93.3\% under the At-Least-1 metric (73.9\% Exact Match at the Category+Reason level, and 81.9\% at the Category level). We frame readability explanation as a gold-level-conditioned classification task: given a sentence and its gold readability level, a system must return the linguistic reasons that justify the label. We present the first systematic benchmark for this task on Arabic. We evaluate two contrasting paradigms: (i) eight LLMs under zero- and few-shot prompting, and (ii) a fully interpretable, training-free rule-based system grounded in 67 linguistic features spanning morphology, syntax, vocabulary, and syllabic structure. Gemini~3 and GPT-5.5 achieve the highest exact-match agreement on fine-grained reason prediction (55.8\% and 56.4\% Category+Reason), while the rule-based system achieves the strongest exact-match performance at the Category level (65.9\%). Together, BAREC-X and our benchmark establish a transparent foundation for interpretable Arabic readability explanation.

Paper 468 Dating the Model: Hidden Dates in System Prompts Affect LLM Evaluation

Authors: Mario Sanz-Guerrero, Minh Duc Bui, Manuel Mager, Katharina von der Wense

Abstract Reproducibility is essential for scientific research, yet prior work shows that LLM outputs vary with hardware and batching. We identify an overlooked factor: the hidden injection of the current date into system prompts, which users cannot control and which changes every day. Across 9 recent LLMs and 6 datasets spanning multiple-choice QA (MCQA), math reasoning, code generation, and machine translation, performance varies *solely* with the current date, with deltas of up to 6% on MCQA, 14% on math reasoning, 7% on code generation, and 2.84 BLEU on machine translation. Model rankings also shift, affecting leaderboards. This date effect exceeds other sources of non-determinism, such as batch size and numerical precision. Standard prompting techniques – chain-of-thought and few-shot prompting – do not reduce the sensitivity; chain-of-thought even amplifies it. Our findings underscore the need for careful evaluation protocols to ensure reproducibility and fair comparisons in LLM research.

Paper 470 Multi-Task Instruction Tuning via Data Scheduling for Low-Resource Arabic SpeechLLMs

Authors: Hunzalah Hassan Bhatti, Firoj Alam, Shammur Absar Chowdhury

Abstract Audio large language models (LLMs) enable unified speech understanding and generation, but adapting them to linguistically complex and dialect-rich settings such as Arabic-English remains challenging. We present a controlled study of multi-task instruction tuning for an Arabic-centric audio LLM across generative tasks including ASR and speech and text summarization, and discriminative tasks including dialect and emotion recognition, in a resource-constrained setting. To support end-to-end Arabic speech summarization, we introduce AraMega-SSum, a first speech summarization resource for training and benchmarking Arabic-centric Audio-LLMs. We compare four training strategies (i) Uniform Mixing (UM), (ii) Task-Progressive Curriculum (TPC), (iii) Aligner-Based Diverse Sampling (ADS) for training-time batch construction, and (iv) A two-stage TPC→ADS strategy. Our results show a clear task-dependent trade-off: TPC and uniform mixing favor generative tasks, while ADS is strongest for SER. The two-stage TPC→ADS strategy achieves the best DID performance across both backbones while preserving competitive generative quality. The adapted models also outperform zero-shot proprietary baselines, including Gemini-2.5-Pro, on discriminative tasks. We made AraMega-SSum and all experimental resources available to the community (https://huggingface.co/datasets/QCRI/AraMega-SSum).

Paper 471 ROCOB: Dual-Manifold Alignment with Optimal Budgeting for Training-Free PTQ Recovery

Authors: Haoqi Yang, Teng Xiao, Zuchao Li, Hao Huang, Lefei Zhang, hai zhao

Abstract Post-training quantization facilitates efficient LLM deployment, yet it disproportionately erodes reasoning and mathematical performance. To recover fidelity without costly fine-tuning, we identify a critical vulnerability in existing output-space methods: the tendency to overfit the limited calibration manifold, which necessitates structural regularization for unseen data. We propose **ROCOB** (**R**esilient **O**utput **C**ompensation with **O**ptimal **B**udgeting), a training-free framework based on *Dual-Manifold Alignment*. ROCOB couples a Principal Activation Component (PAC), which minimizes error on the calibration manifold, with an Orthogonal Geometric Component (OGC), which recovers residual weight-space structure without fitting a second calibration objective. To compare methods fairly under the same added-parameter budget, ROCOB uses Global Spectral Energy Maximization (GSEM) to allocate ranks according to recovered spectral energy and module cost. Across Qwen3 and Llama-3 models, quantizers, and reasoning benchmarks, ROCOB improves average post-quantization recovery in most evaluated model–quantizer settings while preserving a training-free construction pipeline.

Paper 472 Synergizing Large Language Models and Theory for Human-like Causal Reasoning

Authors: Yanxi Zhang, Xin Cong, Zhong Zhang, Xiao Liu, Dongyan Zhao, Yesai Wu

Abstract Performing causal reasoning akin to humans is fundamental to achieving strong artificial intelligence. However, existing studies on causal reasoning and large language models (LLMs) often underestimate the role of human causal cognition. Human causal reasoning is far more than recognizing causes and effects. Humans often factor in many other considerations when establishing a causal link between two events, such as norm violations, intentionality, morality, counterfactual scenarios, etc. To simulate such reasoning, it is intuitive to identify cause-effect pairs using token causality formalisms while modeling the factors that influence human causal judgment using findings from cognitive science experiments. To support the evaluation and modeling of human-like causal reasoning in LLMs, we introduce HCR-Bench, a challenging benchmark with 1,093 instances and detailed human causal reasoning step annotations, and HCR-Reasoner, an approach that systematically integrates theories of token causality and cognitive science into LLMs for human-like causal reasoning. Results suggest that HCR-Reasoner consistently improves LLMs' causal alignment with humans, and that theory-guided reasoning is highly effective.

Paper 476 Beyond Isolated Dots: Benchmarking Structured Table Construction as Deep Knowledge Extraction

Authors: Tianyun Zhong, Guozhao Mo, Yanjiang Liu, Yihan Chen, Lingdi Kong, Xuanang Chen, Yaojie Lu, Hongyu Lin, Shiwei ye, Shanshan Jiang, Bin Dong, Le Sun, Ben He, Xianpei Han

Abstract With the emergence of large language models (LLMs), there is an expectation that they can effectively extract structured information from complex, real-world documents. However, most LLMs generate paragraph-style answers that are chaotic, disorganized, and untraceable. To bridge this gap, we introduce the Arranged and Organized Extraction Benchmark (AOE), a new bilingual benchmark featuring long-form documents (171k tokens on average) designed to systematically evaluate the ability of LLMs to comprehend fragmented information and reconstruct it into an organized table. Unlike conventional text-to-table tasks, which rely on fixed schema and narrow domains, AOE includes 12 carefully designed tasks across four diverse domains, requiring models to generate context-specific schema tailored to varied input queries. In our experiments, we evaluated 12 open-source and proprietary LLMs. Even the most advanced models struggle significantly.

Paper 478 Explainable End-to-End Argument Mining with Toulmin-Structured Explanations

Authors: Zhiqiang Shi, Oana Cocarascu

Abstract Large language models (LLMs) have advanced argument mining (AM), yet most approaches treat component classification, relation identification, and explanation generation as separate tasks, limiting coherence and explainability. We propose the first explainable end-to-end AM framework grounded in Toulmin's theory, jointly predicting argument components, argumentative relations, and structured explanations in a single generation step. This unified formulation encourages LLMs to produce internally consistent argument graphs and reasoning traces aligned with their predictions. Further, we introduce five explainability criteria tailored to AM and grounded in Toulmin's theory: completeness, correctness, faithfulness, soundness, usefulness, and develop a unified evaluation protocol combining structural metrics, neural similarity, LLM-based judgment, human evaluation and qualitative analysis. Extensive evaluation on several LLMs shows that our approach achieves competitive performance while substantially improving explainability, generating Toulmin-based explanations that are complete, grounded and faithful, uncovering reasoning errors not captured by standard AM metrics, and being consistently preferred in usefulness evaluations.

Paper 482 WebWorld: Web-Simulated World Model

Authors: Ting Huang, Zeyu Zhang, Hao Tang

Abstract Large language models (LLMs) have recently shown promise as planners across web, spatial, and embodied domains. However, existing agents often handle information seeking, spatial grounding, and action planning as separate capabilities, limiting their ability to solve long-horizon compositional tasks. To address these challenges, we present **WebWorld**, a web-simulated world model for long-horizon reasoning over web evidence, scene observations, and action plans. WebWorld maintains structured intermediate states across heterogeneous modalities and verifies them through schema-guided hierarchical planning. The framework introduces a Mixture-of-Transformers architecture that coordinates website, scene, and motion experts via multi-modal self-attention. To support training and evaluation, we develop WebWorld-CoT-55K, a large-scale dataset that integrates heterogeneous modalities into schema-guided reasoning traces, and **WebWorldBench**, a curated benchmark for evaluating long-horizon compositional tasks. Training combines cold-start supervised fine-tuning with offline reinforcement learning based on structural, semantic, and planning rewards; inference uses test-time scaling to aggregate multiple reasoning trajectories. Experiments on WebWorldBench show that WebWorld improves the overall score by up to 8.4\% over strong LLM-agent baselines, with gains in outdoor reasoning, geolocation, and travel planning, while still leaving substantial room for progress toward human-level performance.

Paper 483 DramaAgent: Agentic Storytelling Video Generation

Authors: Ting Huang, Biao Wu, Ronghao Chen, Zeyu Zhang, Tengfei Cheng, Qizhen Lan, Huacan Wang, Hao Tang

Abstract Recent diffusion and autoregressive video generators have substantially improved short-form synthesis, yet composing multiple independently generated clips into a coherent audiovisual story remains challenging. Character identities may drift across scenes, narrative events can become inconsistent, and dialogue or sound may not align with the evolving visual context. We present **DramaAgent**, a hierarchical, agentic, and model-agnostic framework for multi-scene long-horizon audiovisual story generation. Rather than modifying the underlying video backbone, DramaAgent introduces a stateful control layer that combines hierarchical story planning, persistent character conditioning, scene-wise generation, and reflection-guided targeted repair. The framework maintains reusable story and character states across scenes, diagnoses failures along identity, semantic, temporal, and audio-visual dimensions, and maps each diagnosed failure to a corresponding localized repair operation. Experiments on 50 narrative scripts across five video-generation backbones show consistent improvements in semantic alignment, character consistency, narrative fidelity, and scene-level audio-visual compatibility. Controlled same-backbone comparisons further show that these gains cannot be explained solely by stronger generators or additional sampling. Our results suggest that persistent state, failure diagnosis, and targeted repair provide a practical control paradigm for coherent multi-scene audiovisual generation.

Paper 486 You Can Create Words Rather Than Just Detect Them: Towards A New Paradigm of Chinese Neologisms

Authors: Qiliang Liang, Hansi Wang, Yue Wang, Yang Liu

Abstract Lexical evolution is universal across languages, yet the mechanisms driving neologism formation vary substantially from one language to another. In Chinese, this variation is particularly reflected in the productivity of compounding, where morphemes are combined through word-formation relations to create interpretable lexical forms. This morphological productivity implies that the Chinese lexicon can be viewed not only as a set of attested words, but also as a dynamically latent space with potential neologisms to be realized. However, this perspective still remains underexplored in computation, as current approaches

are largely reactive, focusing on detecting attested words from corpora. This raises a question about word-formation itself: which morpheme combinations do the rules license beyond those that corpora already attest? Motivated by this, we propose a proactive creation paradigm for Chinese neologism research. We introduce a three-stage pipeline: KG-Based Candidate Generation, Morphological Discarding, and LLM-Based Filtering. We then probe the candidates with three complementary signals: expert judgment, model perplexity, and corpus attestation. The three signals agree: accepted candidates sit close to real words and clearly apart from rejected combinations. The pipeline yields 18,662 candidates with definitions, examples, and formation metadata, released as material for follow-up work, such as lexical gaps and lexical change. Overall, this study substantiates a proactive creation paradigm for Chinese neologism research, and uses it to characterize what Chinese word-formation rules license beyond what dictionaries have already recorded.

Paper 488 From Construction to Injection: Edit-Based Fingerprints for Large Language Models

Authors: ~Yue_Li31, ~Xin_Yi3, ~Dongsheng_Shi1, ~Yongyi_Cui1, ~Gerard_de_Melo3, ~Linlin_Wang1

Abstract Reliable model fingerprints are essential for protecting large language models (LLMs) against unauthorized redistribution and commercial misuse. In black-box deployment, verification is hindered by defensive filtering of suspected fingerprint queries, as well as by downstream model modifications that may weaken embedded ownership evidence. These risks require fingerprints to be robust in both construction and injection. For construction, prior paradigms face an imperceptibility trade-off: natural-language fingerprints may be accidentally activated, whereas garbled fingerprints are statistically exposed and easier to filter. For injection, existing methods struggle to preserve persistent trigger--target behaviors under model modification. We propose an end-to-end injected fingerprinting framework to address these challenges. Code-mixing Fingerprints (CF) use lowest-perplexity code-mixing under a high-complexity constraint to mitigate this two-sided imperceptibility trade-off. Multi-Candidate Editing (MCEdit) constructs structurally redundant, margin-separated trigger--target mappings to enable graceful degradation under model modification. Extensive evaluations on imperceptibility, detectability, and harmlessness demonstrate robust ownership verification with negligible impact on utility.

Paper 491 Is a Picture Worth a Thousand Words? Adaptive Multimodal Fact-Checking with Visual Evidence Necessity

Authors: Jaeyoon Jung, Yejun Yoon, Kunwoo Park

Abstract Automated fact-checking is a crucial task that supports a responsible information ecosystem. While recent research has progressed from text-only to multimodal fact-checking, a prevailing assumption is that incorporating visual evidence universally improves verification accuracy. In this work, we challenge this assumption and show that the indiscriminate use of visual evidence can reduce accuracy. Building on this finding, we propose AMuFC, a modular fact-checking framework that employs two collaborative vision-language models with distinct roles to enable the adaptive use of visual evidence. Experimental results on three datasets, including WebFC, introduced in this study, demonstrate the effectiveness of adaptive visual evidence use in fact-checking.

Paper 494 Local vs. Global Reasoning Failures in LLM-as-a-Judge Evaluation of Structured Physics Reasoning Decompositions

Authors: Ansh Goel, Atin Aggarwal, Aryan Vashishtha, Rudali Huidrom, Vikram Goyal, Mukesh Mohania

Abstract Rubric-based LLM evaluation is increasingly used for structured reasoning, yet the reliability of such evaluators across multi-dimensional criteria is poorly understood. Using structured pedagogical sequences in STEM problem-solving, we study a

five-dimensional framework that separates localized properties (redundancy, relevance) from globally entangled ones (conceptual completeness, diagnostic depth). Benchmarking open-source evaluators against human annotations across decomposed prompting, self-correction, and multi-agent pipelines, we find a consistent asymmetry: localized dimensions benefit from decomposition and pairwise scoring, while globally entangled dimensions stay unstable across every strategy, and on the most global dimension decomposition often degrades alignment. The open-closed alignment gap ranges from $1.0\times$ on the most local dimension to $6\times$ on global ones. For small open-source evaluators, dimensional locality is thus a more fundamental bottleneck than prompting strategy, and decomposition should follow locality rather than be applied uniformly.

Paper 497 When Token Supervision Fails: Segment-Level On-Policy Distillation for Long-Horizon Agents

Authors: Jialong Liu, Zuchao Li, Yuling Shi, Lefei Zhang, Baoyuan Qi, hai zhao

Abstract On-policy distillation (OPD) for long-horizon language agents fails even when the teacher remains sequence-informative: in a failing Qwen3 configuration, the teacher still separates good from bad rollouts (sequence AUROC $\approx$ 0.75), yet the local token-gradient alignment α_{diag} decays from 0.71 to 0.09 and standard OPD recovers only 1.3\% of the teacher--student gap. We restrict our claims to two coupled mechanisms, jointly termed CHOP (\emph{Counterfactual Hierarchical On-Policy distillation}): (i) a per-state reliability gate routing between per-token reverse Kullback--Leibler matching and per-segment sliced-Wasserstein matching, and (ii) a counterfactual total-variation gate activating privileged information (PI) only when the PI changes the teacher's distribution. The scope is explicit: CHOP targets \emph{low-reliability, long-horizon traces where token-local support matching plateaus}; in short-horizon or high-reliability regimes Top-K local-support matching (LSM) is already sufficient and CHOP's segment branch is rarely activated. CHOP recovers 31.7\% of the anisotropic-Qwen3 gap (vs.\ 1.3\%/14.5\%/24.7\% for OPD/Top-K LSM/Li et al.\ 2026); \emph{on SWE-bench Verified under a fixed OpenHands harness, CHOP-deployable prevents the OPD-induced regression} ($6.5\%\!\to\!13.2\%$, matched to off-the-shelf 11.8%, not a significant gain over off-the-shelf) and \emph{oracle PI reveals further headroom} (14.2\%, 95% CI $[11.4,17.3]$); on AIME 2024 at a 60K-token \emph{stress cap} (a diagnostic horizon, not a deployment target), CHOP holds 44.9 vs.\ 42.8 for Top-K LSM ($p=0.032$), while standard OPD collapses to 28.4. Compute-matched, leakage, and PI-removed-teacher controls attribute the gain to supervision-object routing rather than extra compute or PI contamination.

Paper 498 Learning Verifiable Reasoning Programs through Cohort Consistency

Authors: Xiao Ye, Zhaonan Li, Jacob Dineen, Zhikun Xu, Shijie Lu, Ming Shen, Shaswat Shrivastava, Avneet Ahuja, Ben Zhou

Abstract In high-stakes settings, a correct answer from a large language model (LLM) is not enough: the decision procedure must also be auditable. Free-form chain-of-thought is poorly suited to this: it is a natural-language rationale rather than an explicit rule, and may not faithfully reflect the computation that produced the answer. We instead have the model emit a short executable program whose execution defines the answer, making the decision logic explicit and checkable. The central challenge is not writing runnable code but producing logic that is complete and reusable rather than an instance-specific shortcut. Our hypothesis is that a genuine reasoning rule survives factual substitution while a shortcut does not; hence consistency across a cohort of factually varied questions is a verifiable signal of reuse. We therefore train a single program to run unchanged across each cohort, using consistency as the main reinforcement-learning signal. Across five in-domain and three out-of-domain benchmarks at two model scales, our method delivers 10–20 absolute-point gains over strong vanilla, SFT, and per-instance RL baselines. Together, these results show that cohort consistency is an effective, checkable signal for learning complete and reusable reasoning.

Paper 499 Private Evaluation Protocol: Contamination-Resistant Black-Box Evaluation via Public-Slice Surrogates

Authors: Md. Mahadi Hassan, Sazzadur Rahaman, Yusuf Sarwar Uddin, Santu Karmaker

Abstract Existing paradigms for black-box LLM evaluation by independent third parties share a structural weakness: the testing benchmark must reach the model being evaluated, enabling model developers to game future evaluations through logging or retraining. Ideally, rigorous evaluation would grant evaluators direct access to the model so testing can occur without exposing benchmark contents. However, proprietary black-box models cannot be released to external evaluators, creating a fundamental barrier to independent evaluation. To address this gap, we propose the Private Evaluation Protocol (PEP), a surrogate-based framework for independent evaluation of black-box models without sending the private benchmark to the model developer. PEP releases a small public slice of a held-out benchmark, queries the target model only on that slice, and trains a surrogate from the resulting responses, balancing fidelity to the target model against leakage of the private remainder over a tunable revelation budget. Across six benchmarks, composite leakage scores reliably predict behavioral exposure (r=0.788), and fine-tuned surrogates reach 0.85--0.90 agreement on template-heavy datasets while plateauing below 0.73 on reasoning-heavy datasets. The optimal revelation budget is dataset-dependent, and PEP estimates black-box performance with 5.88 percentage-point MAE, showing the feasibility of private surrogate-based evaluation for structured tasks while protecting the privacy of the held-out benchmark.

Paper 500 DeceptGuard: A Constitutional Oversight Framework For Detecting Deception in LLM Agents

Authors: Snehasis Mukhopadhyay

Abstract Reliable detection of deceptive behavior in Large Language Model (LLM) agents is an essential prerequisite for safe deployment in high-stakes agentic contexts. Prior work on scheming detection has focused exclusively on black-box monitors that observe only externally visible tool calls and outputs, discarding potentially rich internal reasoning signals. We introduce DeceptGuard, a unified framework that systematically compares three monitoring regimes: black-box monitors (actions and outputs only), CoT-aware monitors (additionally observing the agent's chain-of-thought reasoning trace), and activation-probe monitors (additionally reading hidden-state representations from a frozen open-weights encoder). We introduce DeceptSynth, a scalable synthetic pipeline for generating deception-positive and deception-negative agent trajectories across a novel 12-category taxonomy spanning verbal, behavioral, and structural deception. Our monitors are optimized on 4,800 synthetic trajectories and evaluated on 9,200 held-out samples from DeceptArena, a benchmark of realistic sandboxed agent environments with execution-verified labels. Across all evaluation settings, CoT-aware and activation-probe monitors substantially outperform their black-box counterparts (mean AUROC improvement of +0.097), with the largest gains on subtle, long-horizon deception that leaves minimal behavioral footprints. We empirically characterize a transparency–detectability trade-off: as agents learn to suppress overt behavioral signals, chain-of-thought becomes the primary detection surface—but is itself increasingly unreliable due to post-training faithfulness degradation. We propose HybridConstitutional ensembles as a robust defense-in-depth approach, achieving a AUROC of 0.934 on the held-out test set, representing a substantial advance over the prior state of the art.

Paper 501 Llm agent memory: a survey from a unified representation–management perspective

Authors: Zhenheng Tang, Xin He, Tiancheng Zhao, Fanjunduo Wei, Xiang Liu, Peijie Dong, Qian Wang, Qi Li, Huacan Wang, Ronghao Chen, Sen Hu, Weidong Guo, Yu Xu, Haolan Chen, Kunfeng Lai, Kaiyong Zhao, Keyan Ding, Ivor Tsang, Yew-Soon Ong, Bo Li, Xiaowen Chu

Abstract Large language models (LLMs) face significant challenges in sustaining long-term memory for agentic applications due to limited context windows. To address this limitation, many work has proposed diverse memory mechanisms to support long-term, multi-turn interactions, leveraging different approaches tailored to distinct memory storage objects, such as KV caches. In this survey, we present a unified taxonomy that organizes memory systems for long-context scenarios by decoupling memory abstractions from model-specific inference and training methods. We categorize LLM memory into three primary paradigms: natural language tokens, intermediate representations and parameters. For each paradigm, we organize existing methods by three management stages, including memory construction, update, and query, so that long-context memory mechanisms can be described in a consistent way across system designs, with their implementation choices and constraints made explicit. Finally, we outline key research directions for long-context memory system design.

Paper 502 Emotion Is Not External Knowledge: Internal Attention Signals for Affective State Reasoning

Authors: Wenqiu Tang, Zhen Wan, Yihui Wang, Takahiro Komamizu, Ichiro Ide

Abstract Large Language Models (LLMs) increasingly operate in multi-turn interactions that require maintaining latent conversational states, not only explicit facts. We study affective state as an implicit state inferred from sparse contextual cues that remain relevant after topic shifts, and ask whether LLMs expose internal attention signals for routing such evidence at inference time. Prior attention interventions rely on labeled in-domain head profiles and externally specified target spans, limiting transfer across affective-evidence distributions. We propose Self-targeted Internal Attention Boosting (SIAB), which localizes sparse affective heads via causal localization (in detail activation patching) on synthetic minimal pairs and boosts, at the answer position, the content tokens these heads already attend to; both edited heads and target tokens are derived from the model's own behavior, enabling the same head set to be reused across tasks.We evaluate on ECC and GoEmotions across three LLMs and multiple prompting settings. Across settings, SIAB improves prompt-matched baselines by up to +10.3 macro-F1 on ECC and +5.5 on GoEmotions. Combined with PASTA-style external supervision, it achieves up to +13.84 accuracy points and +12.51 macro-F1 points, while mechanistic analyses trace the gains to sparse, emotion-enriched heads and show that layer proximity affects additivity.

Paper 504 IndicRipple: Measuring Cross-Lingual Ripple Effects of Knowledge Editing in Indic Languages

Authors: ~Gunjan_Rawat1, ~Sneha_Venkatesh1, ~Shashidhar_Reddy_Javaji1, ~Zining_Zhu1

Abstract Large language models can be updated after training by modifying parameters, conditioning on edit demonstrations, or supplying retrieved evidence. Yet successful factual updates should do more than change one prompt: they should generalize to related questions, transfer across languages, suppress obsolete facts, and preserve unrelated knowledge. This is the problem of measuring cross-lingual ripple effects of factual updates. The existing evaluations on multilingual knowledge injection, however, focus on high-resource languages like English, and it is unknown whether the findings can transfer to Indic languages. In this paper, we present IndicRipple, a controlled evaluation of multilingual factual updating over 500 factual updates across five domains and five languages: English, Hindi, Marathi, Telugu, and Kannada. We compare parametric editing, in-context editing, and retrieval-based correction using ripple-oriented probes for logical generalization, compositionality, subject aliasing, relation specificity, and forgetfulness. Our results show that current factual update mechanisms show limited cross-lingual ripple

effects. IKE performs best overall but reaches only 21.07% strict semantic-plus-language accuracy; RAG often recovers the edited fact but fails to answer in the requested language; and ROME/MEMIT are frequently language-compliant without semantically propagating the update. Strict accuracy is consistently higher for English evaluation than for the Indic-language evaluations. These findings show that direct update success overstates progress: multilingual factual updating must be evaluated as stable cross-lingual propagation, not isolated prompt-level overwrite.

Paper 509 A Benchmark for Controllable Speaking-Style Captioning in Audio-Language Models

Authors: Shreeram Suresh Chandra, Ashwyn Sharma, Liang Lu, Xiutian Zhao, Leda Sari, Björn Schuller, Berrak Sisman

Abstract Speaking-style captioning (SSC) aims to generate natural language descriptions of how speech is delivered, capturing paralinguistic attributes such as vocal timbre, prosody, and expressivity. Many downstream applications, including conversational AI agents, controllable speech generation, and large-scale audio annotation, require dimension-specific style captions that describe targeted aspects of speech (e.g., speaker traits, emotion, or delivery style), rather than a single undifferentiated description. However, existing work lacks a unified formulation for controllable speaking-style captioning. We address this gap by formulating SSC as an instruction-following audio-language task, where explicit instructions specify the speaking-style dimensions to be described. We introduce StyleInstructCaps, the first standardized benchmark for controllable speaking-style captioning in audio-language models. StyleInstructCaps provides a dataset and evaluation framework that measures metadata groundedness, hallucination, instruction-following ability, generalization to unseen instructions and audio datasets, and speaker-style consistency. Across nine state-of-the-art audio-language models, we find that strong instruction-following ability does not necessarily imply grounded speaking-style understanding, and that SSC-specific instruction-diverse training substantially improves groundedness, controllability, and speaker consistency.

Paper 511 Hallucinate at the Last in Long Response Generation: A Case Study on Long Document Summarization

Authors: Joonho Yang, Seunghyun Yoon, Hwan Chang, Byeongjeong Kim, Hwanhee Lee

Abstract Large Language Models (LLMs) have significantly advanced text generation capabilities, including tasks like summarization, often producing coherent and fluent outputs. Despite these advances, hallucinations remain a critical challenge. While prior studies primarily focus on detecting or mitigating hallucinations, little attention has been paid to where hallucinations occur within long generated outputs. In this work, we investigate where hallucinations occur in LLM-based long response generation, using long document summarization as a key case study. Focusing on the challenging setting of long context-aware long response generation, we find a consistent and concerning phenomenon: hallucinations tend to concentrate disproportionately in the latter parts of the generated long response. To understand this bias, we explore potential contributing factors related to the dynamics of attention and decoding over long sequences. Furthermore, we investigate methods to mitigate this positional hallucination, aiming to improve faithfulness specifically in the concluding segments of long outputs.

Paper 514 What's Worth Asking? Evaluating Preference Elicitation in Transactional Agents

Authors: Kandarp Joshi, Tushar Mittal, Chirag Soni, Lillian Zhou, Petar Aleksic

Abstract Transactional agents must act on underspecified requests. "Book me a flight from Chicago to San Francisco" omits budget, timing, stop tolerance, and airline loyalty, any of which can change the best option by hundreds of dollars. The challenge is identifying, without a user-specific prior, which unstated attribute preference is most likely to change the recommendation. We introduce ELICITBENCH, a benchmark of 1,387 travel and housing scenarios pairing real search-result pages (SERPs) with

transaction-grounded preference profiles. Each scenario partitions attributes into stated and hidden preferences. An ex-post analytical oracle assigns question-value labels based on the normalized reduction in market-conditioned recommendation regret. Agents are evaluated by task success, question count, and net utility under a super-linear patience cost. ELICITBENCH reveals that plausible questions are often not decision-relevant: on Gemini 2.5 Flash and Claude Sonnet 4.5, unguided agents ask natural follow-up questions with little downstream value. SERP-aware prompting raises net utility from 0.073 to 0.306 and task success from 21.7% to 57.6% on Gemini, and net utility from 0.221 to 0.346 on Sonnet while asking fewer questions. The selection of items is a second bottleneck: with every preference stated, agents still pick the wrong item in about 11% of the scenarios.

Paper 515 The Same Answer Every Time: Auditing Small Medical LLM Agents

Authors: Abhineet Rajendra Kulkarni, Pranav Madhav Kuber, Anurag Bihani

Abstract We investigate how small open-weight language models make decisions in medical tasks and whether their failures also occur in other settings. We evaluated nine models on wearable fatigue and intensive-care tasks, where they chose among six recommendations, including continued monitoring. We then extended testing to a non-medical tool router and subsets of the Berkeley Function-Calling Leaderboard and JSONSchemaBench. We introduce the Minimum Viable Audit (MVA) to examine which actions models choose, whether their responses contain usable decisions, and whether they copy prompt content. Comparisons across tasks help establish where these behaviors recur. Under the baseline intensive-care prompt, MedGemma chose continued monitoring in all 450 attempts, achieving 33.3% accuracy. On the router, it used all six tools and answered 144 of 150 requests correctly. Meditron-7B produced unusable responses and copied prompt content across intensive-care monitoring and the non-medical evaluations at both four-bit and eight-bit quantization. MedGemma's results motivate checking the medical instructions. Meditron's recurring failures warrant checking the model configuration and how outputs are generated. Differences between tasks and mismatches between medical prompts and scoring rules limit conclusions about the causes of these failures. We recommend reporting selected actions, invalid and copied responses, and results for each prompt condition alongside accuracy.

Paper 516 CRAFT: A Unified Counterfactual Reasoning Framework for Tabular Question Answering and Fact Verification

Authors: Chenshuo Pan, Yu Zhao, Jie Zhang, Changzai Pan, Jiayi Liang, Yujie Mao, Zhenhe Wu, Shuangyong Song, Yongxiang Li, Zhongjiang He

Abstract Table reasoning remains challenging for large language models (LLMs), particularly in tasks that require multi-step inference over long and structured tables. Existing approaches predominantly rely on single-direction reasoning, which limits their ability to explore alternative hypotheses across tasks. In this work, we propose CRAFT, a unified Counterfactual Reasoning Framework that reformulates Tabular question answering and fact verification into a general bidirectional verification process. Our method explicitly constructs both declarative statements and their counterfactual variants. Evidence is then extracted from reasoning along both the original and counterfactual paths, and integrated via a weighted mechanism to arrive at the final answer. Experimental results show that our approach consistently surpasses representative baselines on table reasoning datasets such as WikiTQ and TabFact, achieving especially large improvements on complex question answering. Our framework also significantly mitigates performance gaps between different backbone LLMs. This indicates that counterfactual reasoning effectively overcomes the limitations of single-direction inference, guiding LLMs toward more discerning reasoning and establishing a more principled paradigm for structured reasoning tasks.

Paper 517 Multidimensional Consistency Improves Reasoning in Language Models

Authors: Huiyuan Lai, Xiao Zhang, Malvina Nissim

Abstract While Large language models (LLMs) have proved able to address some complex reasoning tasks, we also know that they are highly sensitive to input variation, which can lead to different solution paths and final answers. Answer consistency across input variations can thus be taken as a sign of stronger confidence. Leveraging this aspect, we introduce a framework where models are systematically pushed to diversify solution paths towards a final answer, thereby testing them for answer consistency across multiple input variations. Focusing on math problems, we induce variations in (i) order of shots in prompt, (ii) problem phrasing, and (iii) languages used. Experiments on a wide range of open-source LLMs of various sizes show that reasoning consistency differs by variation dimension, and that by aggregating consistency across dimensions, our framework enhances mathematical reasoning performance on monolingual datasets GSM8K and MATH500, and the multilingual dataset MGSM. We also discuss the different efficacy of the types of variations induced and how they could be further leveraged in future work.

Paper 519 CriticGen: Generation-Aware Evaluation as Actionable Feedback

Authors: Huifang Du, Zecheng Zuo, Sen Wang, Haofen Wang, Yehui Yang

Abstract Current evaluation methods for large language models are coarse-grained and decoupled from generation, producing generic explanations that fail to provide actionable feedback for model improvement. We propose CriticGen, a fine-grained, generation-aware evaluation framework that turns evaluation into actionable control for answer improvement. CriticGen first generates sample-specific evaluation dimensions and scoring criteria, then uses the induced rubric to jointly produce a score, a reason, an executable refinement suggestion, and a refined answer. Experiments show that fine-grained evaluation should be both instance-specific and actionable. CriticGen improves human-rated rubric relevance/coverage from 3.33/4.03 to 3.97/4.24. On the score-consensus-filtered \textsc{RefineData} test set, the best CriticGen variant reaches 0.9556/0.9560 Pearson/Spearman agreement with the reference scores and improves reason/suggestion F1 to 0.7554/0.7900. Independent validation provides additional support: on 180 human-scored instances, CriticGen-Qwen2.5-7B obtains higher observed score correlations than GPT-5, with 0.8868/0.8969 Pearson/Spearman versus 0.8542/0.8615. Crucially, its feedback translates into substantial answer improvement, raising accuracy from 45.36\% to 91.26\% on 5,000 examples with reliable gold answers, with 47.30\% of answers corrected from Wrong$\rightarrow$Right and only 1.40\% degraded from Right$\rightarrow$Wrong.

Paper 521 LLM-Generated Social-Desirability Ratings for Forced-Choice Questionnaire Construction: A Preregistered Downstream Validation

Authors: Yui Furukawa, Kyosuke Bunji, Kensuke Okada

Abstract Large language models (LLMs) increasingly produce evaluative judgments, but rating-level human--model agreement does not establish whether they can replace human inputs in a downstream task. Using HEXACO personality items as a case study, we conduct a preregistered three-stage validation of whether LLM-generated social-desirability ratings can serve as inputs to forced-choice scale construction, a psychometric task that requires costly human ratings. In Stage I, three LLMs served as raters of item-level social desirability for sentence- and adjective-format items, with human ratings serving as references. In Stage II, human- and LLM-based ratings were entered into the same constrained optimization procedure. In Stage III, 400 participants completed the resulting forms, and latent trait estimates were compared using graded Thurstonian item response theory. LLM ratings showed strong item-level correspondence with human ratings (sentence: $r = 0.942$--0.963; adjective: $r = 0.979$--0.980), but optimized forms showed partial item overlap and sparse exact pair overlap. Despite these design

differences, LLM-based forms produced latent trait estimates that closely matched those from human-based forms (mean $r = 0.815$ for sentence items and 0.768 for adjective items). These results support the conditional use of LLM-generated ratings as scalable construction inputs while showing why downstream validation is necessary.

Paper 523 What Changes in Hidden States When Model Merging Hurts Task Performance?

Authors: Canasai Kruengkrai, Koichiro Yoshino

Abstract Model merging is a practical way to combine several fine-tuned models, but merged models can lose performance relative to their task-specific counterparts. Task accuracy shows the degradation, but not how merging changes the model's internal representations. We study this question through hidden-state distortion, which measures the difference between a merged model's hidden states and those of the corresponding task-specific reference model on the same examples. Across two model families, six tasks, and three merge methods, we find that the distortion is highly anisotropic: much of the merged-reference change is concentrated in a small number of directions. Removing distortion components along dominant directions often improves accuracy more than removing those along unrelated directions. When we jointly fit a single distortion basis across tasks for each model family, this removal still improves accuracy, suggesting that the tasks share some useful distortion structure. In addition, the distortion score provides a label-free criterion for choosing among candidate task-arithmetic merge scales, matching supervised selection for both model families.

Paper 528 STRIDE: Structural Reasoning for In-context Demonstration Extraction in Multi-hop Question Answering

Authors: ~Shiwei_Chen1, ~Bin_Liang6, ~Yue_Yu8, ~Kam-Fai_Wong2, ~Hui_Wang13, ~Ruifeng_Xu1

Abstract In-context learning (ICL) for multi-hop question answering (MHQA) depends critically on the choice of demonstrations. Yet most existing methods prioritize semantic similarity, even though semantically similar examples may follow different reasoning paths and therefore be poor demonstrations for the target query. We present \textbf{STRIDE} (\textbf{ST}ructural \textbf{R}easoning for \textbf{I}n-context \textbf{D}emonstration \textbf{E}xtraction), a structure-centric framework for multi-hop QA. STRIDE first infers the latent reasoning structure of a query. It then performs positive–contrastive structural routing, selecting structurally compatible demonstrations as positives and semantically plausible but structurally mismatched examples as contrastives. Finally, it constructs structural prompts that expose compact, role-consistent reasoning traces over supporting evidence. Together, it turns demonstration retrieval from semantic matching into structure-aware problem-solving guidance. Empirical results show that STRIDE consistently improves downstream QA performance across settings.

Paper 530 Efficient Task Adaptation in Large Language Models: A Survey of Weight-Based, Prompt-Based, and Embedding-Based Adaptations

Authors: Jungwon Park, Changin Choi, Jimyeong Kim, Nojun Kwak, Wonjong Rhee

Abstract As large language models are increasingly deployed across diverse downstream tasks, efficient task adaptation has emerged as a central challenge. In response, a wide range of task adaptation methods have been proposed, spanning parameter-efficient fine-tuning, in-context learning, and embedding-injection approaches. However, these lines of work have largely evolved within individual paradigms, leaving their cross-paradigm relationships and trade-offs underexplored, especially for recently emerging embedding-based adaptations. This survey presents a unified framework that categorizes task adaptation methods by where and how task information is encoded: model weights, input prompts, or injected task embeddings. We provide

a comprehensive taxonomy that integrates these paradigms, analyze their key strengths and limitations to explain how different adaptation paradigms have evolved, clarify relationships across paradigms, and highlight open problems for future research.

Paper 531 ArtHopQA: Diagnosing Hidden-Chain Recovery in Artwork Question Answering

Authors: ~Xuefeng_Wei1, ~Xuan_Zhou8, ~Yuta_Kato1, ~Shintaro_Ozaki1, ~Kazuki_Hayashi1, ~Yusuke_Sakai1, ~Hidetaka_Kamigaito2, ~Katsuhiko_Hayashi2, ~Taro_Watanabe1

Abstract Many artwork questions begin with vision but cannot be answered by visual recognition alone. A model may need to connect an artwork to its creator, depicted subject, collection, or location, and then follow multiple facts to reach the answer. We introduce ArtHopQA, a multimodal multi-hop question answering benchmark for artwork understanding, containing 4,885 image-grounded questions paired with supporting fact chains. Experiments with open and API large vision-language models show a large gap: even the strongest open model reaches only 36.87% accuracy when given the image and title, while gold-path performance reaches 97.77% to 99.80% when the supporting fact chain is provided. Text-only language models exhibit a similar gap, showing that the difficulty is not confined to visual input. Further analysis reveals high-prior geographic guesses, inconsistent answers under image-title conflict, and limited recovery of intermediate entities even when final answers are correct.

Paper 532 Role-aware Heuristic Episodic Attention for Conversational LLMs

Authors: hongwanyang, Zhaoning Zhang, Yi Chen, Libo Zhang, Baihui Liu, Linbo Qiao, Zhiliang Tian, Dongsheng Li

Abstract Large language models often lose track of persistent instructions and relevant information as multi-turn conversations grow. We study this cumulative contextual decay through three related failure modes: attention pollution, dilution, and drift. We propose **REA** (Role-aware Heuristic Episodic Attention), a context-management framework that assigns different persistence and representation policies to instructions and episodic interactions. Instructional Memory retains identified global constraints in a dedicated prefix. Episodic Memory preserves user inputs and compresses model replies, while heuristic retrieval selects raw text, compressed representations, or omission for each historical turn. On Long-MT-Bench+, REA improves the judge score from 6.32 to 7.36 on a 10-point scale, a 16.5% relative gain over the Vanilla baseline, and reduces average latency by 2.91×. Additional evaluations show aggregate gains on three backbones spanning 1.7B–7B parameters and on Chinese and English role-playing tasks. These results support role-aware context management as a practical approach to maintaining conversational continuity and instruction adherence.

Paper 536 AtomicCheck: Learning Structured Reasoning for Fact Verification via Protocol-Guided Reinforcement Learning

Authors: ~Xingquan_Zhang1, ~Jianghao_Chen1, ~Yang_Zhao26, ~Pu_Jian1, ~Haoran_Gao1, ~Fanyu_Meng5, ~Junlan_Feng3, ~Chengqing_Zong1

Abstract Fact verification is a critical primitive for ensuring the reliability of Large Language Models (LLMs). While frontier LLMs are effective verifiers, their repeated long-context inference incurs substantial costs; meanwhile, existing specialized verifiers face a fundamental dilemma: label-only models suffer from black-box opacity, while free-form rationale generators, whether trained by distillation or explanation-oriented RL objectives, can produce fluent but not structurally checkable explanations. In this work, we propose AtomicCheck a Protocol-Guided Policy Optimization framework that trains specialized verifiers to produce protocol-guided, locally checkable traces, moving beyond opaque label-only predictions and free-form rationales. Rather than cloning teacher traces or optimizing unstructured explanations, AtomicCheck uses RL to make the verification protocol itself part of the policy objective. To preserve semantic integrity during atomic decomposition, we introduce a Contrastive Decomposition

Reward that compares decomposed reasoning with a same-backbone holistic counterfactual and penalizes decompositions that reduce verification accuracy. Extensive experiments demonstrate that AtomicCheck outperforms existing specialized verifiers and approaches strong frontier LLMs while producing structured traces that expose local checks behind the final decision.

Paper 547 K-Prism: A Korean Diagnostic Benchmark for Multi-faceted Hallucination in Vision-Language Models

Authors: Min Hyuk Kim, DONG HYOK SHIN, Jimi Lee, Jin Yea Jang

Abstract Vision-language models (VLMs) often produce hallucinated outputs that are not supported by visual or textual evidence. While existing benchmarks have advanced hallucination evaluation, they primarily target object-level or task-specific errors and often assess different failure types on separate samples. This design limits direct comparison of hallucination patterns under shared evidence, and most prior evaluations remain English-centric. We introduce K-Prism, a Korean diagnostic benchmark for multi-faceted hallucination analysis in VLMs. K-Prism contains 1,002 multiple-choice questions across text and image tracks, organized into shared-evidence topic groups. Within each group, questions are designed to probe three diagnostic dimensions: factual, reasoning, and misleading. K-Prism uses a controlled dual-mode protocol comparing intact inputs with visually perturbed inputs to examine hallucination under reduced visual evidence. Experiments with closed- and open-source VLMs show dimension-specific degradation rather than a uniform accuracy drop. These findings demonstrate that hallucination in Korean VLMs cannot be captured by a single aggregate score and requires fine-grained diagnostic evaluation across evidence conditions and failure dimensions. Our dataset is publicly available at https://huggingface.co/spaces/KETI-NLP/K-Prism_Benchmark.

Paper 548 KoMMERCE: A Korean Multimodal and Multitask Evaluation Benchmark for Reliable Conversational E-commerce Agents

Authors: Byunggill Joe, San Kim, Minyoung Jung, DONG HYOK SHIN, Yu-Jeong Sung, Saim Shin

Abstract E-commerce conversations are increasingly multimodal, yet end-to-end response scores cannot localize which dialogue competence failed, and no public benchmark evaluates finegrained multimodal dialogue understanding for Korean commercial agents. We present KoMMERCE, a Korean multimodal fashion-domain benchmark that coordinates nine fine-grained dialogue understanding tasks on the same conversations: sentiment, intent, and strategy classification;image and text knowledge retrieval;dialogue state tracking; span detection; coreference resolution; and the first formulation of cross-modal dialogue relation extraction. The corpus contains 15,293 dyadic role-play dialogues under two-stage quality control. Baselines on six frontier multimodal LLMs reveal task-conditional patterns that should inform deployment choices: no model wins all nine tasks, and context length affects accuracy differently from one task to another.

Paper 549 KoEVD: A Korean Benchmark for Metadata-Conditioned Multi-Stage LLM Safety Evaluation

Authors: San Kim, Wonjung kim, Heesoo Kwon, Jae Hee Yang

Abstract Large language model (LLM) safety benchmarks often evaluate input risk, response selection, generation, and tool use separately, making comparisons across tasks difficult. We introduce KoEVD (Korean Ethical Violation Dataset), which links five evaluation and analysis tasks through annotated Korean source utterances: input-risk judgment, candidate-response selection, generation safety, response-strategy analysis, and non-executable mock-tool selection. The tasks share source IDs, with task-specific subsets and candidate sampling where required. Built through an approximately 3.5-year human-governed, AI-assisted process, KoEVD contains 13,552 curated user utterances and 71,395 linked response candidates. Across 12 open-

weight and commercial models, observed task profiles differ, and aggregate scores obscure variation across severity, explicitness, utterance type, and topic. In matched generation and mock-tool-selection cases, responses classified as safe coexist with unsafe mock-tool choices; pooled unsafe-call rates are 9.73% and 9.31% for the classifier-safe and classifier-unsafe groups, respectively, with framing-specific differences. These diagnostic measurements do not establish real-world safety. KoEVD supports detailed Korean safety analysis through contextual annotations and source-linked comparisons.

Paper 551 ConceptWeave: Concept-Driven Prompting through Adaptive Semantic Infilling

Authors: Atul Kumar Singh, Rahul Mishra

Abstract Conventional prompting and prompt-based finetuning offer only weak semantic structure, leaving models to derive task-specific reasoning from limited cues. Existing techniques, such as using demonstrations, chain-of-thought prompting, or retrieval augmentation, can improve performance, but often at the cost of longer, more redundant prompts, higher inference overhead, and reliance on external knowledge sources. We present ConceptWeave, a novel concept-centric prompt-based finetuning approach built on adaptive semantic infilling. Instead of depending on external resources, it extracts concise, task-relevant concepts directly from a small portion of downstream data and organizes them into a dynamically evolving hierarchical memory. During fine-tuning, ConceptWeave identifies informative spans in the input through semantic alignment and augments them with corresponding concept representations, enabling fine-grained, localized conditioning of the model. As training progresses, the concepts are continuously refined and structured into a hierarchy that supports efficient retrieval while preserving semantic richness and scalability. We validate ConceptWeave across a range of NLP tasks, including misinformation detection, question answering, sentiment analysis, and hate speech classification, where it consistently outperforms strong baselines and recent state-of-the-art methods.

Paper 554 CulturALL: Benchmarking Multilingual and Multicultural Competence of LLMs on Grounded Tasks

Authors: Peiqin Lin, Chenyang Lyu, Wenjiang Luo, Haotian Ye, Md Mehrab Hossain, Chunlan Ma, Shaoxiong Ji, Younes Samih, Bo Zeng, Fan Jiang, Yuanbin Cao, DILDA DUISENBEK, Adrian Neo Sau Xun, POZDNIAKOVA DARIA, MISEVICH LIUBOU, Marinković Nevena, Ngoc Gia Linh Nguyen, THI KHANH LINH DO, SARAKMATAK SOPHY, Baotian Hu, Guanhua Chen, Gongbo Tang, Alham Fikri Aji, Longyue Wang, Weihua Luo

Abstract Large language models (LLMs) are now deployed worldwide, inspiring a surge of benchmarks that measure their multilingual and multicultural abilities. However, these benchmarks prioritize generic language understanding or superficial cultural trivia, leaving the evaluation of grounded tasks—where models must reason within real-world, context-rich scenarios—largely unaddressed. To fill this gap, we present CulturALL, a comprehensive and challenging benchmark to assess LLMs' multilingual and multicultural competence on grounded tasks. CulturALL is built via a human–AI collaborative framework: expert annotators ensure appropriate difficulty and factual accuracy, while LLMs lighten the manual workload. By incorporating diverse sources, CulturALL ensures comprehensive scenario coverage. Each item is carefully designed to present a high level of difficulty, making CulturALL challenging. CulturALL contains 2,610 samples in 14 languages of 51 regions, distributed across 16 topics to capture the full breadth of grounded tasks. The experiments show that the best LLM achieve 44.48% accuracy on CulturALL, underscoring substantial room for improvement.

Paper 557 Mitigating Cross-Image Information Leakage in Multi-Image Understanding with Large Vision-Language Models

Authors: Yeji Park, Minyoung Lee, Sanghyuk Chun, Junsuk Choe

Abstract Large Vision-Language Models (LVLMs) exhibit strong performance on single-image tasks. However, their performance degrades significantly when handling multi-image inputs. While this degradation has been observed in prior work, its nature remains poorly understood. We empirically observe visual elements from different images become entangled in the model's representations and responses. We refer to this phenomenon as $\textit{cross-image information leakage}$. To address this issue, we propose $\textbf{FOCUS}$, a training-free and architecture-agnostic method. ${FOCUS}$ masks all but one image with random noise, guiding the model to focus on the single clean image. This process is applied across the target images to obtain logits under partially masked contexts. These logits are aggregated and then refined using a noise-only reference input, which suppresses the leakage and yields more accurate outputs. ${FOCUS}$ consistently improves performance on diverse multi-image benchmarks. We further show that generalizes to video understanding, extending its applicability beyond static multi-image inputs. This demonstrates that ${FOCUS}$ offers a general solution for enhancing multi-image reasoning without additional training or architectural modifications.

Paper 558 Self-Distillation for Multi-Token Prediction

Authors: Guoliang Zhao, Ruobing Xie, An Wang, Kai Zhang, Shuaipeng Li, Huaibing Xie, Samm Sun

Abstract As Large Language Models (LLMs) continue to scale, inference efficiency has emerged as a critical bottleneck. Multi-Token Prediction (MTP) offers a promising avenue to accelerate LLM inference by predicting multiple future tokens simultaneously. Wherein, DeepSeek MTP has gradually become an indispensable component in industrial LLMs. However, it still faces two challenges: limited accept rates of MTP heads, and difficulties in jointly training multiple MTP heads. To address these challenges, we propose \textbf{MTP-D}, a simple yet effective self-distillation method that substantially boosts MTP head accept rates while maximally preserving main-head performance. We further introduce a \textbf{looped extension strategy} for MTP-D, enabling effective and economical extension to multiple MTP heads and yielding significant additional inference speedup (up to $4.55\times$). Beyond these contributions, we conduct systematic experiments across seven benchmarks to explore and validate key insights into distillation strategies and the scalability potential of DeepSeek MTP. Results demonstrate that MTP-D and the grouped extension strategy consistently enhance MTP head performance and inference efficiency, facilitating the practical usage of MTP in LLMs.

Paper 559 CLEF: Closed-Loop Error Feedback for Multi-Bit Watermarking in Short LLM-Generated Text

Authors: Jing Wang, Kejun Qian, Pengyu Xiong, Qiankun Zhang, Yi Yang, Jing Fan

Abstract Multi-bit watermarking for large language models (LLMs) enables richer provenance information, but existing schemes remain unreliable for short texts because they rely on local or pre-specified allocation strategies. To address this limitation, we propose a Closed-Loop Error Feedback (CLEF) Multi-Bit Watermarking framework. By treating the LLM as a noisy channel, CLEF tracks the global feature state of generated tokens and dynamically measures its deviation from an ideal multidimensional target trajectory. The resulting dynamic error vector is then used to modulate the logits of candidate tokens via an inner-product scoring mechanism, biasing generation toward tokens that compensate for underrepresented payload dimensions. This closed-loop control enables robust multi-bit embedding with a simple, fully linear decoding process. Experiments on short-text generation show that CLEF improves bit extraction accuracy and complete message recovery rate over the evaluated multi-bit

watermarking baselines. Furthermore, CLEF achieves higher bit extraction accuracy under different text perturbations while maintaining acceptable generation quality across downstream tasks such as text summarization and machine translation.

Paper 560 Closing the Bilingual Gap: Cross-lingual Semantic Scaffolding for Fine-Grained Machine Translation Evaluation in Taiwanese Hokkien

Authors: You-Wei Wu, En-Shiun Annie Lee, Richard Tzong-Han Tsai

Abstract Despite the growing adoption of large language models (LLMs) for machine translation (MT) evaluation, reliably detecting fine-grained error spans in low-resource languages remains challenging. A key limitation is that LLMs often lack sufficient cross-lingual knowledge, making it difficult for them to accurately localize errors at the span level. In this work, we introduce Cross-lingual Semantic Scaffolding, a two-part prompting framework that incorporates sentence-level parallel examples and word-level lexical mappings at inference time. We evaluate this approach on five LLMs using controlled ablation experiments on Taiwanese Hokkien. The proposed framework consistently improves performance across all tested models, increasing average Span F1 by +0.175 (+33.6%) and average Span Recall by +0.351 (+75.9%), and shows that the gains are mainly driven by improved error coverage. It also supports AI-assisted human annotation, reducing expert annotation time from 450 seconds to 90.92 seconds per sentence, with 98.2% of AI-proposed spans accepted without modification. These results demonstrate that Cross-lingual Semantic Scaffolding is both an effective detection method and a practical tool for reducing expert annotation costs in low-resource MT evaluation.

Paper 563 “Do You Truly Love Me?” Benchmarking LLM Capability on Hierarchical Pragmatic Tactic Conversations

Authors: Peter Q. Chen, Pangjing Wu, Zhuohang Jiang, Xiaodong Li, Peter H. F. Ng

Abstract Pragmatic reasoning, including inferring intent, manipulation, and hidden meaning in dialogue, is essential for trustworthy language understanding yet remains underexplored in current large language models (LLMs). We introduce \textbf{HIPO}, a new benchmark designed to assess \textbf{\underline{Hi}}erarchical \textbf{\underline{P}}ragmatic Tactic in multi-turn C\textbf{\underline{o}}nversations. Inspired by linguistic pragmatics, HIPO operationalizes a hierarchical annotation schema for multi-turn conversations and labels each utterance along one dialogue-level goal and three utterance-level dimensions: communicative intention (\emph{why}), veracity strategy (\emph{how}), and illocutionary act (\emph{what}). To ensure reliable supervision, we design a structured generation pipeline to generate high-quality synthetic dialogues. The benchmark comprises over 4,088 benchmarking utterances with 6,350 contextual utterances across 1,131 dialogues drawn from 31 real-world-inspired scenarios. Benchmarking 22 state-of-the-art LLMs reveals a striking gap: while models achieve higher performance at recognizing surface speech acts(83.3% accuracy), they perform poorly on detecting veracity strategies(32.2%) and speaker intentions(45.1%). These results suggest current LLMs remain challenged by the pragmatic inferences operationalized in HIPO, particularly those involving information manipulation and intended effects beyond literal content. We further show that supervised fine-tuning on additional synthetic data generated by our pipeline can improve performance on HIPO, suggesting a practical path for targeted capability improvement within this benchmark setting.

Paper 564 Contextual Cue Sensitivity in LLM Evaluation of Grant Proposals: A Controlled ERC-Style Study

Authors: Marco Rospocher

Abstract Large language models (LLMs) are increasingly used as evaluators, but in high-stakes settings reliability requires more than consistency: scores should also be robust to contextual information that is not intended to alter the criterion being evaluated.

We study this question in a controlled English-language ERC-style proposal-scoring setup. Starting from 40 Horizon Europe ERC-related project summaries, we construct synthetic Part B1-like proposal texts and evaluate them with eight LLMs under manipulations of principal-investigator (PI) profile information—name, institution, and bibliometric profile—and AI-use disclosure. Across six ERC-style scoring questions, we observe systematic score shifts relative to the blind condition when contextual information is introduced while proposal content is held fixed. The strongest penalties are associated with lower bibliometric profiles, followed by AI-use disclosure and lower-prestige institutions; name-based effects are smaller. These effects are consistent across prompts and evaluator models and are large enough to alter counterfactual rankings and top-K selection outcomes. The results show that current LLM evaluators can be sensitive to contextual information attached to otherwise identical proposal content in ways that are fairness-relevant for research assessment.

Paper 566 What Makes User Reactions Useful for Fake News Detection?

Authors: Yassine Toughrai, David Langlois, Smaili

Abstract Fake news detection is often treated as a content-centered task, yet social media posts are embedded in user reactions that express skepticism, endorsement, correction, ridicule, and affective responses. This paper asks what makes user reactions useful for fake news detection. We study performance according to different dimensions: text only, emotions in comments, and number of available comments. We study this question across Fakeddit and Weibo20 under both standard and temporal splits (in a temporal split, most recent posts are included in the testing split). Robustness of results is measured across two encoder families (BERT-base, RoBERTa-base) and two English emotion taxonomies (dair-ai/emotion, GoEmotions). Across these configurations, comment-only representations outperform post-only baselines in most evaluation regimes, though this gain depends on dataset, split, and number of available comments. Emotion-fine-tuned comment encoders produce a temporal-shift-specific significant benefit compared to non-temporal splits. We compare post-only representations, generic comment representations, emotion-fine-tuned comment representations, and post--reaction fusion models, while varying comment availability, showing that a minimum number of comments is necessary to improve detection of fakes. In addition, we show that fine-tuning comment or reaction-encoders on relevant auxiliary tasks such as emotions has a significant impact on the distribution of embeddings of fake and real posts in their representation space. In light of these findings, we believe that user responses—and their emotional structure for example—should be taken into account when detecting fake news or compiling datasets for this purpose.

Paper 567 Safety in Batches? Understanding and Mitigating Safety Failures in Batch Prompting

Authors: Kihyun Kim, Hee-Seon Kim, Wonjun Lee, Changick Kim

Abstract Batch prompting is a practical inference strategy for large language models, but its safety implications remain underexplored. We show that the success of batch prompting for utility does not extend to safety: a harmful question that is reliably refused in isolation can elicit a harmful response when embedded in a batch of benign questions. We identify this as a distinct safety failure mode, not reducible to known vulnerabilities such as in-context learning or long-context effects—and analyze its causes from two complementary perspectives: alignment signal weakening and refusal signal dilution. Across widely used open-source and frontier commercial models, batch prompting consistently achieves high attack success rates as a simple black-box attack. We further show that batch-aware preference optimization effectively mitigates the vulnerability. These findings highlight a blind spot in current safety alignment and point to batch-aware alignment as a necessary step toward robust deployment.

Paper 568 Positive-Unlabeled Contrastive Retrieval for Graduate Program Recommendation

Authors: Shiyan Liu, Rui Qu

Abstract Graduate program selection is a high-stakes decision where applicants struggle to identify programs matching their latent academic profiles. We address this as a retrieval problem: given a student profile, rank 531 graduate programs by estimated admission compatibility, i.e., how closely the student's profile resembles each program's historical admit pool. Because applicants rarely report rejections, training data consists only of positive labels, making negative supervision unavailable. We propose PURE (Positive-Unlabeled Retrieval Encoder), a positive-unlabeled biencoder with InfoNCE contrastive loss and a gated structural residual that fuses GPA and experience counts with free-text descriptions. With a novel multi-positive MRR protocol that corrects for false negatives, PURE achieves MRR=0.2696, outperforming BM25 (+4.9 points) and GPT-4o-mini (+8.3 points), while remaining statistically indistinguishable from Biencoder + GPT-4o-mini reranking at negligible cost.

Paper 571 Does Translation Preserve Character Personality? Quantifying Perceived Character-Personality Consistency Across Languages

Authors: Xinger Fu, Masaaki Nagata, Chenhui Chu

Abstract Translation conveys meaning across languages, but does it preserve the personality a character conveys? We introduce an evaluation framework for quantifying how well translation systems maintain perceived character personality, the personality impression a reader forms from a character’s dialogue, described along the five Big Five dimensions; we use these dimensions as descriptive labels for the impression, not as a psychometric measurement of the character. Our framework applies personality predictors to both source and translated utterances, measuring cross-lingual consistency via Spearman rank correlation at the speaker level. We study entertainment subtitle translation, where every utterance is produced by a named character, and compare evaluation conditions that include, shuffle, or withhold the character names. We evaluate five MT systems, including dedicated MT APIs (Google Translate, DeepL) and LLMs (GPT-4o, Qwen3- 32B, Aya-expanse-32B), and additionally report human translation as a reference condition on GAME OF THRONES (English→12 target languages, 11 characters) and ONE PIECE (Japanese→English, 15 characters). Under our evaluation protocol, measured personality consistency varies across systems and language pairs, and high BLEU/COMET translation quality does not reliably imply high personality consistency.

Paper 575 Cross-Lingual Semantic Fusion: Towards Universal Sentence Representations via LLMs

Authors: Zhen Hu, Zhu Liu, Ying Liu

Abstract Extracting sentence representations from pretrained large language models (LLMs) is a training-free approach that leverages their inherent semantic understanding. Existing methods primarily rely on prompt engineering to distill core semantics, but the resulting representations often retain language-specific features that hinder semantic tasks. Moreover, these approaches fail to exploit the cross-lingual capabilities of LLMs, remaining confined to monolingual settings. To address these limitations, we propose CLSF, a novel method that actively introduces diverse language perturbations when encoding the same sentence and then fuses the perturbed embeddings, effectively reducing language-specific biases and producing a more universal representation. CLSF is plug-and-play and can be integrated with existing prompt-based methods. Experiments on Semantic Textual Similarity (STS), multilingual alignment, and various downstream tasks show that CLSF consistently improves performance across different LLMs.

Paper 576 EviMem: Evidence-Gap-Driven Iterative Retrieval for Long-Term Conversational Memory

Authors: Yuyang Li, Yimu He, Zeyu Zhang, Dong Gong

Abstract Long-term conversational memory requires retrieving evidence scattered across multiple sessions, yet single-pass retrieval fails on temporal and multi-hop questions. Existing iterative methods refine queries via generated content or document-level signals, but none explicitly diagnoses the evidence gap-what is missing from the accumulated retrieval set—leaving query refinement untargeted. We present EviMem, combining IRIS (Iterative Retrieval via Insufficiency Signals), a closed-loop framework that detects evidence gaps through sufficiency evaluation, diagnoses what is missing, and drives targeted query refinement, with LaceMem (Layered Architecture for Conversational Evidence Memory), a coarse-to-fine memory hierarchy supporting fine-grained gap diagnosis. On LoCoMo, EviMem improves Judge Accuracy over MIRIX on temporal (73.3%→81.6%) and multi-hop (65.9%→85.2%) questions at 4.5× lower latency.

Paper 579 Bebras-SK: Evaluating Multimodal Large Language Models on Slovak Computational Thinking Tasks

Authors: Patrícia Vnenčáková, Marek Šuppa

Abstract Multimodal large language models are increasingly used for visual reasoning, yet their ability to solve computational thinking problems posed in less-resourced languages is rarely measured. We introduce Bebras-SK, a benchmark of 822 multiple-choice tasks from iBobor, the Slovak edition of the international Bebras challenge, spanning 16 competition years and 11 categories. The tasks are educator-authored and native Slovak. Of them, 629 include images and 321 include several, while text-only tasks, some written for visually impaired students, provide a built-in control. We evaluate six hosted models. The strongest, GPT-5.2, matches or exceeds the published average of Slovak students in most categories, yet in our original run its accuracy falls from 0.974 on text-only tasks to 0.632 on multi-image tasks. Across models, accuracy depends far more on input modality than on the age group a task was written for. Rendering these short tasks as screenshots increases token usage, often several-fold, without consistent accuracy gains. The model ranking is stable across three runs per model, and a guided-completion probe finds no evidence of verbatim memorisation in the five models it can test (GPT-5.2 declines it). We release the tasks and the evaluation code at https://github.com/NaiveNeuron/SlovakBebrasChallenge.

Paper 583 CGVR: Contradiction-Guided Visual Reasoning for Mitigating Hallucinations in Multimodal Large Language Models

Authors: Haoran Ding, Wanqi Zhou, Lei Xing, Lujuan Dang, Kaizhong Zheng, Badong Chen

Abstract Multimodal Large Language Models (MLLMs) frequently suffer from object hallucination, producing descriptions or answers unsupported by visual evidence, which undermines their reliability in real-world applications. This paper begins with an intriguing phenomenon: prefacing queries with statements explicitly contradicting image content often yields more accurate responses than direct questioning. We term this the Contradictory Prompting Effect (CPE). Priming inputs with assertions opposing reality triggers an intrinsic self-verification mechanism within MLLMs, where conflicting claims compel the model to prioritize visual evidence over linguistic priors. Inspired by this discovery, we propose Contradiction-Guided Visual Reasoning (CGVR), a novel training-free framework for hallucination mitigation. CGVR systematically generates deliberate contradictory captions alongside standard ones. These captions drive a dual-channel inference process, in which conflicting signals enforce rigorous cross-checking against visual evidence. Our work reveals MLLMs' latent self-verification capability triggered by contradictions and establishes a paradigm for lightweight, reliable multimodal reasoning. Extensive evaluations on POPE and

MME benchmarks demonstrate that our proposed CGVR significantly mitigates hallucinations across diverse MLLMs. Our code is available at https://github.com/dhr1698/CGVR.

Paper 584 KBS: A Korean Benchmark Suite for Foundation Model Evaluation

Authors: Yeonkyoung So, Jongmin Kim, Sungmok Jung, Gyuseong Lee, Sangho Kim, Jongyeon Park, Joonhak Lee, Seho Pyo, Gyeongje Cho, Seorin Kim, Jisoo Kim, Suyoung Park, Hyunji M. Park, Yelim Ahn, Yeongho Seo, Jaejin Lee

Abstract We present a Korean Benchmark Suite (KBS) comprising Ko-ARC, Ko-GSM8K, Ko-EQ-Bench, Ko-WinoGrande, Ko-LAMBADA, and Ko-IFEval, covering six capabilities across 9,396 items. The suite combines review, localization, direct Korean construction, and redesign for Korean, avoiding a translation-only approach to benchmark adaptation. We evaluate Korean-developed models, multilingual open models, and closed API models under protocols that account for model access, prompt format, and scoring differences. Results show that Korean benchmark scores depend on task construction, scoring rules, prompting, and model post-training. These findings suggest that Korean evaluation should report both final scores and the settings used to obtain them.

Paper 586 Constructing Disambiguated Knowledge Bases from Large Language Models at Scale

Authors: Yujia Hu, Tuan-Phong Nguyen, Simon Razniewski

Abstract Automated Knowledge Base Construction (AKBC) is a core NLP task, and recent work proposes generating knowledge bases directly from large language models (LLMs), treating the model itself as the knowledge source. However, LLMs natively possess no representation of entities, leading to duplicate entries as well as conflations. We propose DisamKB, a methodology for constructing disambiguated KBs directly from LLMs. DisamKB incorporates on-the-fly disambiguation of entities, relations and classes, and is meticulously designed to satisfy both scalability and disambiguation accuracy. We analyze the central design decisions and characterize the trade-offs between accuracy, scale, and cost. We execute DisamKB at scale, obtaining a materialized KB containing over 1M disambiguated entities and 38.4M triples. This represents the first million-scale LLM-native KB with explicit internal canonicalization of entities, relations, and classes, a significant departure from prior Wikimedia-centric works. DisamKB is anonymously available at https://gptkb.org/.

Paper 589 JanusPO: Asymmetric Clip and KL for Entropy-Stable Group Policy Optimization in Tool-Augmented LLM Reasoning

Authors: Hao Li, Man Fung Zhuo, Guangxin Wu, Yusen Wu, Jiayu Wu

Abstract Group-relative policy optimization (GRPO, DAPO, ARPO) has become the dominant recipe for RL fine-tuning of reasoning LLMs, yet in the resource-constrained regime most practitioners train in (a single GPU, small batches) it consistently suffers from rapid entropy collapse that kills exploration, together with sparse credit assignment in long rollouts, which is most extreme under multi-turn tool use but already present in long single-turn chain-of-thought. We address both failure modes with JANUSPO, a drop-in GRPO modification that introduces asymmetric PPO clipping \(\epsilon_{+}<\epsilon_{-}\) to tolerate larger downward than upward updates, pairs it with an asymmetric KL schedule \(\beta_{pos}<\beta_{neg}\) that doubles as a free entropy regulator without any explicit upper bound, and redistributes the terminal reward through a turn-weighted decomposition \(w_{t}(n)\) that generalises naturally from single-turn CoT to multi-turn rollouts. On Qwen2.5-3B-Instruct / GSM8K, JANUSPO improves from 74.8\% to 84.3$\pm$0.2\% (mean$\pm$std over 2 seeds) across 14 controlled ablations, and the same recipe lifts 7B from 85.3\% to 93.6\% with no hyper-parameter retuning. On multi-turn DeepSearch (103 GAIA + 100 HLE), SFT-warmup-then-JANUSPO is the only RL recipe whose paired comparison against SFT survives best@8 outlier verification,

achieving strict dominance on all four reward and F1 columns while a matched-init GRPO baseline's apparent F1 wins collapse to 0\%. Beyond this main result, we document a systematic negative finding: SFT warm-up is a necessary precondition for GRPO-style multi-turn RL on format-unaware base models.

Paper 590 Explainable Multimodal Aspect-Based Sentiment Analysis with Dependency-guided Large Language Model

Authors: Zhongzheng Wang, Yuanhe Tian, Hongzhi Wang, Yan Song

Abstract Multimodal aspect-based sentiment analysis (MABSA) aims to identify aspect-level sentiments by jointly modeling textual and visual information, which is essential for fine-grained opinion understanding in social media. Existing approaches mainly rely on discriminative classification with complex multimodal fusion, yet lacking explicit sentiment explainability. In this paper, we reformulate MABSA as a generative and explainable task, proposing a unified framework that simultaneously predicts aspect-level sentiment and generates natural language explanations. Based on multimodal large language models (MLLMs), our approach employs a prompt-based generative paradigm, jointly producing sentiment and explanation. To further enhance aspect-oriented reasoning capabilities, we propose a dependency-syntax-guided sentiment cue strategy. This strategy prunes and textualizes the aspect-centered dependency syntax tree, guiding the model to distinguish different sentiment aspects and enhancing its explainability. To enable explainability, we use MLLMs to construct new datasets with sentiment explanations to fine-tune. Experiments show that our approach not only achieves consistent gains in sentiment classification accuracy, but also produces faithful, aspect-grounded explanations.

Paper 591 HorizonBench: Evaluating Ideation Agents on Future Scientific Discoveries

Authors: ~Markus_Flicke1, ~Haoyu_He4, ~Kashyap_Chitta1, ~Yong_Cao3, ~Andreas_Geiger3

Abstract As ideation agents become more capable, evaluating them poses an increasing challenge. LLM-as-a-judge approaches are brittle and gameable, while human evaluation is expensive and hard to scale. To address this, we introduce HorizonBench, an evaluation methodology leveraging future scientific discoveries. HorizonBench uses published papers as ground truth for ideation quality. Specifically, ideation agents pretraining and context retrieval data are strictly constrained to pre-publication knowledge, and scored on how well their predicted ideas align with subsequently published work. Using HorizonBench, we conduct the first unified comparison of five published ideation agents across multiple frontier model backbones. We verify that our soft recall metric captures meaningful ideation quality aligned to human judgement via a user study with 51 domain experts. We then uncover a surprising finding: a naïve baseline that receives only the research goal as an input consistently matches or outperforms all existing agents that require retrieval or multi-step reasoning, highlighting the importance of the research goal for ideation. Based on this result, we develop OSCAR, an ideation agent that distils retrieved papers into insights and explicitly reinforces the research goal in each step of the pipeline. OSCAR outperforms all evaluated agents on HorizonBench. We publish the code and an online version of HorizonBench.

Paper 595 Language Triggers Hijack Language Circuits: A Mechanistic Analysis of Backdoor Behaviors in Large Language Models

Authors: ~Théo_Lasnier1, ~Wissam_Antoun1, ~Francis_Kulumba1, ~Benoît_Sagot1, ~Djamé_Seddah1

Abstract Backdoor attacks pose significant security risks for Large Language Models (LLMs), yet the internal mechanisms by which triggers operate remain poorly understood. We present the first mechanistic analysis of trigger-induced language-switching backdoors injected during pre-training, studying the Gaperon model family (1B, 8B and 24B). Using activation patching,

we localize trigger formation and identify which attention heads process trigger and natural language information. Our central finding is that trigger heads substantially overlap with heads naturally encoding output language across model scales, with Jaccard indices between 0.18 and 0.43 over the top 10 heads identified. This suggests that backdoor triggers do not form new circuits but instead co-opt the model's existing language components and representations. These findings have implications for backdoor defense as detection methods and mitigation strategies could leverage this entanglement between triggers and natural behaviors. More broadly, our work represents a first step toward a more realistic mechanistic understanding of pre-training-injected backdoors in LLMs, paving the way for principled, interpretability-driven defenses.

Paper 596 A Framework for Evaluating AI-Generated Creativity: Human–LLM Agreement in Evaluating AI-Generated Stories

Authors: Pharath Sathya, Yin Jou Huang, Fei Cheng

Abstract Creative text is inherently subjective: any opinion can be equally valid, and there is no objective ground truth. The only way to quantify such subjectivity is to model the crowd of opinions that produces it. Large language model (LLM)-as-a-judge has emerged as a scalable substitute for human raters, but whether an LLM can stand in for the human crowd on AI creativity remains unexplored. We develop a decomposition framework (four main components, 11 sub-components) and a three-step rating procedure, applied identically to human raters and to different LLM judging conditions on AI-generated stories. Humans rate AI creativity with the same subjectivity that creativity research documents for human-authored creative work. Across three structural comparisons (disagreement on the same story including ranking; revision from impressionistic to reflective; sub-component weighting of the final rating), no LLM reproduces all three human patterns. We frame the contribution as a methodological framework plus a quantification of where LLM judges currently fall short of the human ratings in our study. We release all stories, ratings, and code at https://github.com/ku-nlp/creativity-evaluation-artifact.

Paper 599 Evaluating Large Reasoning Models on Belief Revision

Authors: Nurul Fajrin Ariyani, Zied Bouraoui, Richard Booth, Steven Schockaert

Abstract Logic offers a natural framework for benchmarking LLM-based reasoning models. While frontier models perform well on standard logical reasoning tasks, it remains unclear to what extent these capabilities generalize to other forms of reasoning. In this paper, we use belief revision problems to stress-test the reasoning abilities of these models on a less familiar task, which builds on satisfiability and entailment checking but requires a careful strategy to manage complexity. We show that belief revision is indeed challenging for existing models, even when only small and relatively simple problem instances are considered.

Paper 601 GLM-RAG: Graph Language Models for Graph-Based Retrieval-Augmented Generation

Authors: Maya Arseven, Anette Frank, Beni Egressy, Johann Higl, Moritz Plenz

Abstract Retrieval-augmented generation (RAG) over knowledge graphs requires retrievers that can effectively capture both graph structure and semantic information. Recent approaches have explored graph neural network (GNN)-based retrievers to model graph topology in multi-hop reasoning tasks. In parallel, graph language models (GLMs) have emerged as a promising paradigm that integrates graph reasoning and the semantic capabilities of language models. In this work, we introduce a GLM-based retriever and investigate the comparative strengths of GLM-based, GNN-based, and traditional vector-search-based retrievers in single- and multi-hop RAG settings, and with a particular focus on transferability to unseen domains. Our findings suggest that finetuned GLM retrievers generalize better out of domain, achieving SOTA on two multi-hop benchmarks. On in-domain multi-hop QA datasets they remain comparable to prior work, with promising scaling as parameters and subgraph

coverage increase. GNN-based retrievers achieve higher graph coverage with an efficient training setup, whereas the vector-search baseline excels at single-hop datasets.

Paper 602 Structuring the Unstructured: A Systematic Review of Text-to-Structure Generation with a Universal Evaluation Framework

Authors: Zheye Deng, Chunkit Chan, Tianshi Zheng, Wei Fan, Weiqi Wang, Yangqiu Song

Abstract The evolution of AI systems toward agentic operation and context-aware retrieval necessitates transforming unstructured text into structured formats like tables, knowledge graphs, and charts. While such conversions enable critical applications from summarization to data mining, current research lacks a comprehensive synthesis of methodologies, datasets, and metrics. This systematic review examines text-to-structure techniques and the encountered challenges, evaluates current datasets and assessment criteria, and outlines potential directions for future research. We also introduce a universal evaluation framework for structured outputs, establishing text-to-structure as foundational infrastructure for next-generation AI systems.

Paper 604 A2M: Trace-Optimized Agent Hijacking in the MCP Ecosystem

Authors: Laizhen Li, Xuan Wang, Peicheng Zhao, Juanjuan Zhao, Kejiang Ye, Cheng-zhong Xu, Xitong Gao

Abstract Agents using the Model Context Protocol (MCP) rely on semantic matching to select tools from third-party servers, exposing a semantic supply-chain risk through attacker-controlled metadata and outputs. We introduce A2M (Attraction-to-Manipulation), a two-stage black-box framework for hijacking MCP agents. The Attraction phase optimizes tool metadata to increase invocation probability; the Manipulation phase uses execution traces to refine adversarial tool returns that steer agents toward attacker-desired outcomes. On LiveMCPBench, direct attacks optimized and evaluated on GLM-4.6 achieve a macro-average malicious tool invocation rate of 93.6% across four scenarios, increase weighted token costs to $32.4\times$ the benign baseline under Cognitive Denial of Service, and attain a mean attack success rate of 74.4% across Information Exfiltration, Environment Integrity Compromise, and Reasoning Derailment. Transfer to four other models without re-optimization yields corresponding macro-averages of 63.6%, $2.7\times$, and 24.5%. These findings motivate stronger tool vetting and runtime isolation in MCP ecosystems. Code is publicly available at https://github.com/Lilaizhen/A2M.

Paper 605 All Verdicts are Not Equal: Rethinking LLM Judge Reliability

Authors: Vineet Kumar, Darshita Rathore, ANINDYA MOITRA

Abstract LLM-as-a-Judge is the standard paradigm for NLP evaluation, yet its systemic reliability remains poorly understood despite being widely treated as a deterministic ground truth. We present a comprehensive reliability audit, stress-testing six frontier models across four benchmarks, five prompt formats, two presentation orders, three sampling temperatures, and ten repetitions per condition. Our empirical analysis reveals severe vulnerabilities: verdicts change across identical replications at temperature zero, position-order swaps flip the majority of verdicts on challenging tasks, and the most deterministic judge achieves perfect consistency by trivially repeating incorrect verdicts, agreeing with ground truth only 51% of the time. To formalize these multi-faceted failure modes, we introduce the trustworthy verdict rate (T), a unified metric capturing the joint probability that an evaluation is reproducible, order-invariant, and accurate. Using T , we derive a theoretical upper bound on accuracy imposed by position bias and show that reliability is item-specific rather than model-level. Finally, we demonstrate that shifting from pairwise win-rate to holistic rubric scoring improves trustworthiness more than any single-format prompting intervention, offering an actionable framework for robust NLP evaluation.

Paper 608 Target-Conditioned Gating for Many-to-Many Argument Assignment in Document-Level Event Extraction

Authors: Junru Liao, Haoran Wang, Xuanjing Huang, Yatian Shen

Abstract Document-level event extraction often requires the same argument candidate to fill multiple valid event-slot targets, with each assignment grounded in different evidence. Many existing methods introduce target-specific signals during event construction, matching, or role scoring, while explicit per-target argument views jointly conditioned on event-slot semantics and instance-specific information before role scoring remain less explored. This late arrival of target-specificity is especially problematic for many-to-many (M2M) assignment, where the same candidate must preserve distinguishable evidence for several valid targets before decoding. We propose Target-conditioned Gating for Inter-target Decoupling (TGID), an encoding-stage MoE framework that produces target-conditioned argument views before decoding. TGID combines event-slot semantics with instance probes to form target queries and uses independent sigmoid gates to remove direct cross-gate coupling induced by softmax normalization, while schema-structured routing initialization and a multi-positive target-view separation objective stabilize the resulting views. By incorporating instance probes, TGID can distinguish Cross-Type, Cross-Role, and Same-Type Multi-Instance targets before decoding. Across ChFinAnn, DuEE-Fin, and FNDEE, TGID achieves consistent gains over strong trigger-free DEE baselines, especially on M2M subsets. Parameter-matched softmax controls and routing diagnostics indicate that target-sensitive routing contributes beyond additional model capacity. Code is available at https://github.com/LeonaKTH/TGID.

Paper 609 The Guardrail Trinity: Procedural Safety Schemas, Reasoning Chain Decomposition, and Dual-Mode Inference for LLM Guardrails

Authors: ~Huy_H._Nguyen1, ~Pride_Kavumba1, ~Jiaxuan_Li3, ~MinhDuc_Vo1, ~Koki_Wataoka1

Abstract Effective LLM safety guardrails must balance adversarial robustness, custom-policy adaptability, and precision on benign prompts with sensitive vocabulary. We introduce the Guardrail Trinity, a data-curation recipe combining procedural safety schemas, reasoning chain decomposition with hard-negative supervision, and a hybrid data blend for fast and slow inference. Experiments across Llama-3.1 and Qwen3 models yield three practical insights. First, strong standard moderation scores can conceal a custom-policy compliance gap; procedural schema variation improves adaptation on our held-out policy-shift probe. Second, augmentation must be matched to model scale and training recipe: smaller models can exhibit training instability, while hard negatives at larger scales can reduce false alarms at a cost to harmful-input recall. Third, hybrid training supports both direct classification and reasoning, with the fast mode retaining strong custom-policy compliance without generating a chain. Together, these findings motivate a deployment playbook that selects data composition, inference mode, and scale jointly, and validates both policy adherence and the trade-off between missed harms and false alarms.

Paper 610 When Ethics and Payoffs Diverge: LLM Agents in Morally Charged Social Dilemmas

Authors: Steffen Backmann, David Guzman Piedrahita, Terry Jingchen Zhang, Emanuel Tewolde, Rada Mihalcea, Bernhard Schölkopf, Zhijing Jin

Abstract Recent advances in LLMs have enabled their use in complex agentic roles, involving decision-making with humans or other agents, making ethical alignment a critical concern. While prior work has examined LLMs' moral judgment and strategic behavior separately, there is limited understanding of how they act when moral imperatives directly conflict with profit incentives. We introduce MoralSim to evaluate how LLMs behave in the prisoner's dilemma and public goods game embedded in morally

charged contexts, varying moral framing, opponent behavior, and survival pressure across nine models. Beyond measuring behavior, we estimate the causal effect of each factor via average treatment effects (ATEs) and analyze agents' own reasoning traces to characterize the motives behind their choices. We find that no model remains consistently moral, with cooperation rates ranging from 7.9% to 76.3%. Game structure and moral framing are the strongest causal drivers of moral behavior, while reasoning-trace analysis reveals distinct motive profiles across models, ranging from predominantly payoff-maximizing to moral- and reputation-oriented. Together, these results expose the situational brittleness of current LLMs' moral behavior and the risk of deploying them where profit incentives conflict with ethical guidelines.

Paper 613 MedHal: A Synthetic Dataset for Medical Hallucination Detection

Authors: Fabrice Lamarche, Gaya Mehenni, Neshat Elhami Fard, Odette Rios-Ibacache, Li Ming Wang, John Kildea, Amal Zouaq

Abstract Hallucination, the generation of non factual content by AI systems, poses a serious risk in medical contexts, where errors can directly affect patient outcomes. We present MedHal, a novel large-scale dataset specifically designed to assess capabilities and train models on the task of hallucination detection in medical texts. Current hallucination detection methods face significant limitations when applied to specialized domains like medicine, where they can have disastrous consequences. MedHal addresses this issue by incorporating diverse medical text sources and tasks covering both intrinsic and extrinsic hallucinations, and by providing a substantial volume of data samples suitable for training medical hallucination detection models. We demonstrate MedHal’s utility by training and evaluating a baseline medical hallucination detection model, showing improvements over general-purpose hallucination detection approaches. This resource enables more efficient evaluation and training of medical text generation systems while reducing reliance on costly expert review, potentially accelerating the development of medical AI research.

Paper 620 Structure Tax: How Structured Output affects LLMs Performance

Authors: Vineet Kumar, Kanishka, Bhuvanesh Mandora

Abstract Deploying large language models in production often requires constraining outputs to structured formats such as JSON or XML, and prior work treats the resulting accuracy loss as an inherent "structure tax". We re-examine this claim by evaluating a battery of models, datasets and schemas, measuring task accuracy, confidence calibration, and hidden-state geometry. The tax turns out to depend on schema design rather than on structure per se: reasoning-first field ordering matches or exceeds free-form accuracy, while answer-first ordering causes steep drops, particularly in smaller models. Format sensitivity scales inversely with a task's own structural constraints, and schemas that preserve reasoning order also improve calibration with CKA showing greater separability between correct and incorrect representations in middle transformer layers. Our findings indicate that properly designed structured formats can match or exceed free-form performance, reframing the critical question from 'whether to structure' to 'how to structure' for optimal reasoning preservation.

Paper 622 Grounded or Guessing? LVLM Confidence Estimation via Blind-Image Contrastive Ranking

Authors: Reza Khanmohammadi, Erfan Miahi, Simerjot Kaur, Charese Smiley, Ivan Brugere, Kundan S Thind, Mohammad M. Ghassemi

Abstract Large vision-language models (LVLMs) suffer from visual ungroundedness: they can produce a fluent, confident, and even correct response driven entirely by language priors, with the image contributing nothing to the prediction. Existing confidence estimation methods cannot detect this, as they observe model behavior under normal inference with no mechanism

to determine whether a prediction was shaped by the image or by text alone. We introduce BICR (Blind-Image Contrastive Ranking), a model-agnostic confidence estimation framework. The LVLM stays entirely frozen; instead, we extract its hidden states twice per sample, once with the real image-question pair and once with the image blacked out while the question is held fixed. A small, separate probe is then trained on the real-image hidden state, regularized by a ranking loss that penalizes higher confidence on the blacked-out view, teaching it to treat visual grounding as a signal of reliability. At inference only the single real-image pass is used, so BICR adds zero cost beyond a standard single-view probe. Evaluated across five modern LVLMs and seven baselines on a benchmark covering visual question answering, object hallucination detection, medical imaging, and financial document understanding, BICR achieves the best cross-LVLM average on both calibration and discrimination simultaneously, with statistically significant discrimination gains robust to cluster-aware analysis, at 4–18× fewer parameters than the strongest probing baseline.

Paper 623 Exploring Cross-Scenario Generality of Agentic Memory Systems: Diagnostics and a Strong Baseline

Authors: Zhikai Chen, Jialiang Gu, Junyu Yin, Xianxuan Long, Shenglai Zeng, Xiaoze Liu, Kai Guo, Keren Zhou, Jiliang Tang

Abstract LLM agents accumulate interaction histories that exceed their context windows, motivating extensive research on memory systems. Yet most existing designs are tailored to a single scenario (such as multi-session chat or a specific trajectory format), offering little evidence that they generalize across the heterogeneous trajectories encountered in deployment. We evaluate twelve baselines spanning specialized memory systems, long-context models, and an unindexed search agent across five scenarios: single-turn QA, multi-session chat, agentic-trajectory QA, memory stress tests, and long-horizon agentic tasks. Notably, DCI-Lite, an unindexed baseline that searches raw text via grep and bounded reads, is among the strongest existing baselines. This suggests that memory performance depends partly on granting the agent active control over retrieval rather than coupling a static store to a single retrieval route. We instantiate this diagnostic insight in **AutoMEM**, a strong baseline that exposes both raw-text and indexed memory views through an agent-controlled interface. Across the evaluated systems, AutoMEM achieves the best mean cross-scenario rank.

Paper 626 Forgetting Only What Matters: Layer-Selective Unlearning toward Robust LLMs

Authors: Ravi Ranjan, Olivera Kotevska, Agoritsa Polyzou

Abstract Large Language Models (LLMs) can memorize and reproduce sensitive, copyrighted, or otherwise undesirable training content, creating privacy, safety, and regulatory concerns. Machine unlearning offers a practical alternative to full retraining, but many existing methods apply broad or fixed parameter updates that can degrade utility and remain brittle under deployment changes such as post-training quantization, where forgotten knowledge may partially re-emerge. We propose Forgetting Only What Matters via Unlearning Layers (FOM-UL), a layer-level unlearning framework that selects transformer layers using a forget-to-retain significance score. This score identifies layers with high influence on the forget set and low sensitivity to the retain set, allowing FOM-UL to concentrate updates where they are most effective while leaving most of the model unchanged. This targeted update strategy improves the forgetting-utility trade-off and provides an empirical path toward quantization-resilient unlearning by reducing the chance that small, diffuse updates are erased by low-bit rounding. Across TOFU, KnowUnDo, and MUSE-style evaluations, FOM-UL reduces residual memorization compared with strong GA, NPO, KLD, SURE, ReLearn, and LUNAR-based baselines while preserving retain-set utility close to the vanilla model. Under 8-bit and 4-bit post-training quantization, FOM-UL maintains a favorable forgetting-utility trade-off, combining competitive memorization suppression with stronger retained utility. Overall, FOM-UL provides an efficient unlearning strategy that improves targeted forgetting, utility preservation, and deployment robustness without claiming formal guarantees of erasure.

Paper 628 A Pre-Training Analogue of Grokking in Language Models: Tracing Delayed Grammatical Generalization

Authors: Sherin Muckatira, Namrata Shivagunde, Vijeta Deshpande, Anna Rumshisky

Abstract Grokking, the phenomenon in which neural networks generalize long after fitting their training data, has been studied in supervised settings on many epochs. LLM pre-training instead involves next-token prediction over an unlabeled corpus, with limited data repetition and no explicit train/validation split. To address this, we propose an exposure-based framework that enables the study of grokking-like dynamics during LLM pre-training. We ground our evaluation in BLiMP minimal pairs, which provide controlled grammatical contrasts. For every BLiMP minimal pair, we identify a critical phrase, the smallest continuous span that captures the grammatical contrast and the phenomenon-relevant context. Examples whose critical phrase appears in the pre-training window are assigned to the proxy-train split; the remaining examples are assigned to the proxy-validation split. Across five grammatical phenomena, we observe delayed generalization. Analyzing pre-training checkpoints before and after generalization shows that grammatical concept vectors become more predictive of grammatical acceptability and occupy a higher-dimensional subspace after generalization. We also find that attention from the critical token to the relevant context token is concentrated in a small number of heads.

Paper 633 BenHalluEval: A Multi-Task Hallucination Evaluation Framework for Large Language Models on Bengali

Authors: Shefayat E Shams Adib, Ahmed Alfey Sani, Ekramul Alam Esham, Ajwad Abrar, Ishmam Tashdeed, Md Taukir Azam Chowdhury

Abstract Despite Bengali being the sixth most spoken language in the world, no prior work has systematically evaluated hallucination in large language models (LLMs) for Bengali. We introduce \textbf{BenHalluEval}, a fine-grained hallucination evaluation framework for Bengali covering four tasks: Generative Question Answering (GQA), Bangla-English Code-Mixed QA, Summarization, and Reasoning. We construct 12,000 hallucinated candidates using GPT-5.4 across twelve task-specific hallucination types, drawn from three existing Bengali datasets. We evaluate seven LLMs spanning reasoning-oriented, multilingual, and Bengali-centric categories under a dual-track protocol that independently measures false-positive rate on ground-truth instances (Track A) and hallucination detection rate on hallucinated candidates (Track B). To jointly penalise both failure modes and prevent inflated scores from uniform response bias, we propose \textbf{BenHalluScore}, a dual-track calibration metric. BenHalluScore ranges from 7.72% to 55.42% across models and tasks, revealing substantial variation in hallucination calibration. Chain-of-thought prompting, applied as a mitigation strategy, shifts response distributions without consistently improving hallucination discrimination. BenHalluEval establishes the first dedicated hallucination benchmark for Bengali and highlights the inadequacy of single-track and prompting-only evaluation approaches for low-resource language settings. The dataset and code are available at \url{https://anonymous.4open.science/r/BanglaHalluEval-EB77}.

Paper 634 GEAR: Training-Free Rule Distillation for Advanced and Efficient Tool-Augmented Reasoning

Authors: ~Bhiman_Kumar_Baghel1, ~Fanyou_Wu1, ~Suraj_Maharjan2, ~Peiran_Yao1, ~Srinivasan_H._Sengamedu1

Abstract Improving tool use in large reasoning models (LRMs) typically demands expensive test-time compute, training resources unavailable to API-only practitioners, or expensive iterative rollouts with execution feedback required by training-free context-engineering methods, all incompatible with resource-constrained deployment. We introduce **GEAR**, a training-free framework that extracts compositional rules from tool-use failures via single-pass meta-reasoning and distills them into inference

prompts. We hypothesize that combining explicit rules with examples outperforms either alone. Across 78 API domains, **GEAR**+2-shot outperforms 3-shot by 24.9\% on Qwen3-30B, validating our complementarity hypothesis. On Qwen3-4B, **GEAR** exceeds LoRA SFT by 24.3\% and RAG-3-shot by 38.5\%, while **GEAR**+2-shot beats the state-of-the-art context-engineering method ACE by 38.8\% at 4.1$\times$ lower cost. Remarkably, Qwen3-30B with **GEAR**+2-shot nearly doubles zero-shot accuracy while remaining under 1\,s latency, demonstrating rollout-free adaptation for resource-constrained deployment.

Paper 636 HARPA: A Testability-Driven, Literature-Grounded Framework for Research Ideation

Authors: Rosni Vasu, Peter Jansen, Pao Siangliulue, Cristina Sarasua, Abraham Bernstein, Peter Clark, Bhavana Dalvi Mishra

Abstract Recent advances in large language models (LLMs) have accelerated automated scientific discovery (ASD), but existing ideation tools still struggle to generate hypotheses that are both testable, literature-grounded, and adaptive to prior experimental outcomes. We developed HARPA to address these challenges by following a human-inspired ideation workflow that mines literature trends, explores hypothesis design spaces, and converges on precise, testable hypotheses by pinpointing research gaps and justifying design choices. Human evaluations show that HARPA-generated hypothesis-driven research proposals performs comparably to a strong baseline system, AI-researcher, in novelty and overall quality, while showing a positive trend in feasibility (+0.78, $p<0.05$, bootstrap) and a significant gain in groundedness (+0.85, $p<0.01$, bootstrap) on a 10-point Likert scale. When evaluated with the ASD agent CodeScientist, HARPA produced more successful executions (20 vs. 11 out of 40) and fewer failures (16 vs. 21 out of 40) indicating that expert feasibility judgments track execution success. HARPA also learns a reward model that scores hypotheses based on prior experimental outcomes, achieving approx. a 28% absolute gain over HARPA's untrained baseline scorer.

Paper 637 Retrieve the Logic, Not the Topic: Taxonomic Strategy RAG for Cross-Domain Persuasion Agents

Authors: Sana Ayromlou, Purvi Sehgal, Pradyumna Narayana

Abstract Foundation-model agents in multi-step, open-ended environments frequently suffer from compounding errors, where early mistakes contaminate long-horizon trajectories. While Multi-Agent Debate (MAD) succeeds in deterministic domains with external validators~\cite{du2024improving}, agents in subjective, non-deterministic tasks such as persuasion experience problem drift and sycophantic conformity. Focusing on subjective argumentation, we identify semantic leakage in standard Retrieval-Augmented Generation (RAG) as a reproducible trigger for these failures: standard RAG prioritizes vocabulary overlap over logical structure. To mitigate this leakage, we introduce Taxonomic Strategy RAG (TS-RAG), a systems intervention that routes strategy retrieval through a discrete categorical bottleneck to decouple argumentative structure from topical content. Zero-shot, cross-domain evaluations on r/ChangeMyView show that TS-RAG improves the transfer of abstract argumentative logic where standard semantic retrieval degrades ($+14.5$ win-rate points over an unaugmented persuader for Gemini Flash 3.0, $p<0.05$). In asymmetric deployments that mix lightweight and frontier models, TS-RAG raises the win rate of a lightweight persuader against a stronger opponent (70.5\% $\rightarrow$ 78.5\%, $p<0.05$) and reduces rounds to consensus when the persuader is already dominant. Finally, we introduce trace-level diagnostics via a turn-by-turn Debate State Representation (DSR), and show that strict stance-holding constraints on the opponent are necessary to prevent evaluation collapse caused by default agentic sycophancy.

Paper 638 Fast Listwise Passage Reranking via Semantic Tokens

Authors: Omkar Prabhune, Xijun Wang, Nan Bi, Haiyan Wang, Shima Alizadeh, Anoop Jain, Yu Lou, Junbang Liang

Abstract Listwise passage reranking often achieves strong ranking performance, but it incurs substantial computational cost. We propose FLASH, an efficient listwise reranking framework based on semantic token compression and single-step listwise scoring. FLASH represents each query--passage pair with a small set of query-aware semantic tokens and jointly scores all candidates in a single decoder step, enabling query-conditioned joint scoring without expensive token-level listwise processing. We further introduce multi-cardinality training, which allows a single model to support different semantic-token budgets at inference time. We also propose FLASH-LITE, a lightweight variant with offline passage encoding for low-cost online inference. Across TREC and BEIR benchmarks, FLASH improves average ranking effectiveness at comparable online latency and slightly lower FLOPs, while FLASH-Lite substantially reduces query-time cost with a modest effectiveness trade-off. FLASH-LITE enables offline passage encoding, substantially reducing online query-time cost while maintaining competitive reranking performance. Code is provided in the supplementary material.

Paper 640 Influcoder: Distilling Decoders' Gradient Influence Rankings into an Encoder for Data Attribution

Authors: ~Dimitri_Kachler1, ~Damien_Sileo2, ~Pascal_Denis1

Abstract Data attribution traces model behavior back to individual training examples, with applications such as identifying training data that induces toxic responses or tracing a model's factual claims back to the examples that support them. Gradient-based influence methods provide reliable attribution performance, yet are impractical at the scale of LLM training datasets due to their high computational costs. We introduce Influcoder, which distills the LoRA-gradient cosine influence rankings of a target decoder into a compact encoder, reducing attribution at inference time to an embedding cosine similarity. It scores training examples 35x faster than gradient-based influence computed on the target model and 18x faster than influence computed on the smallest comparable proxy model. Our method remains competitive with gradient-based methods even when considering the setup cost, as evaluated on two downstream tasks of toxicity filtering and factual attribution.

Paper 646 TeleSWEBench: A Commit-Driven Benchmark for Evaluating LLM-Powered Software Engineering in Telecommunications

Authors: Pranshav Gajjar, Ali Mamaghani, Dinesh Bharadia, Vijay K Shah

Abstract With the telecommunications field embracing zero-touch management alongside novel O-RAN and AI-RAN frameworks, contemporary telecom networks now function as immensely intricate and heavily softwarized codebases. While automated software engineering (ASE) tools and Software Engineering (SWE) Agents hold the potential to alleviate the critical code generation bottleneck in this domain, their ability to navigate and modify specialized, mathematically rigorous wireless stacks like \texttt{srsRAN 5G} remains unverified. General-purpose coding benchmarks fail to capture the stateful logic and strict requirements of telecommunications, leaving a critical evaluation gap. In this paper, we introduce TeleSWEBench, the first commit-driven benchmark specifically designed to measure an agent's performance in the telecom domain. We mine real developer commits from the srsRAN 5G repository and distill them into structured test cases across three difficulty tiers (Easy, Medium, and Difficult). Our benchmark consists of 734 questions that are accompanied by executable unit tests. To avoid the rigidity of test cases, we further propose a hierarchical LLM-as-a-Judge framework called TeleJudge that scores agent outputs at the file level and aggregates verdicts holistically. This follows an evaluation based on context and semantic similarity in parallel to a standard unit test-based evaluation. Using this benchmark, we evaluate AIDER, OpenHands, and the Claude Code frameworks, powered by state-of-the-art reasoning LLMs, including Qwen, GPT-OSS, Gemma 4, Kimi, GLM, and Qwencoder

2.5. Our two-stage evaluation reveals that models suffer from a lack of both localization accuracy and functional correctness, with the strongest ASE tools achieving up to 25% of shippable changes.

Paper 648 Reasoning-Based Personalized Generation for Sparse Users

Authors: Bo Ni, Branislav Kveton, Samyadeep Basu, Subhojyoti Mukherjee, Leyao Wang, Franck Dernoncourt, Sungchul Kim, Seunghyun Yoon, Zichao Wang, Ruiyi Zhang, Puneet Mathur, Jihyung Kil, Jiuxiang Gu, Nedim Lipka, Yu Wang, Ryan A. Rossi, Tyler Derr

Abstract Large Language Model (LLM) personalization holds great promise for tailoring responses by leveraging personal context and history. However, real-world users usually possess sparse interaction histories with limited personal context, such as cold-start users in social platforms and newly registered customers in online E-commerce platforms, compromising the LLM-based personalized generation. To address this challenge, we introduce GraSPer (Graph-based Sparse Personalized Reasoning), a novel framework for enhancing personalized text generation under sparse context. GraSPer first augments user context by predicting items that the user would likely interact with in the future. With reasoning alignment, it then generates texts for these interactions to enrich the augmented context. In the end, it generates personalized outputs conditioned on both the real and synthetic histories, ensuring alignment with user style and preferences. Extensive experiments on three benchmark personalized generation datasets show that GraSPer achieves significant performance gain, substantially improving personalization in sparse user context settings.

Paper 650 Identifying User-Inference-Dependent Out-of-Character Behavior in LLM-Based Role-Playing Agents

Authors: ~Shinji_Iwata1, ~Shiki_Sato1, ~Jun_Baba1, ~Koya_Ihara1, ~Michimasa_Inaba2

Abstract For the practical deployment of the Role-Playing Agents with Large Language Models, prior work has mainly focused on suppressing out-of-character behavior relative to explicitly provided character information such as personas. However, as shown in psychological research, out-of-character behavior can arise not only from contradictions with explicit information but also from mismatches with character images implicitly inferred by users. Modeling user-inference-dependent out-of-character behavior is a challenging task, because it is unclear what kinds of inferences users make from existing information. To tackle this problem, we construct a dataset of 480 dialogue scenarios and collect 2,152 descriptions from users familiar with those titles. Using cases that annotators judged inappropriate despite the absence of contradiction with existing settings, we derive a 12-category typology, organized into general, title-specific, and individual-specific groups. We also conduct an exploratory evaluation of category-guided feedback, finding no consistent overall advantage over standard self-refinement but possible benefits for some characters. Our shareable code and dataset are at https://github.com/CyberAgentAILab/User-Inference-OOC.

Paper 655 The Error Is the Signal: Misconception Discovery via Step-Error Annotation Clustering

Authors: Robinson Vasquez Ferrer, Mrinmaya Sachan, Shashank Sonkar

Abstract Discovering student misconceptions is a key challenge in intelligent tutoring systems, where identifying recurring error-producing patterns can support targeted feedback and adaptive practice. In settings without a predefined misconception taxonomy, clustering is a natural approach for discovering such patterns at scale. Yet clustering is only as good as its representation: full student records contain problem wording, numerical context, and solution-specific details that embeddings can conflate with the student's error, obscuring the misconception behind the wrong answer. We construct an LLM step-error-annotation-first pipeline that generates step-error annotations, embeds them with or without problem and student response text,

and clusters the resulting representations. On MalruleLib, a 101-class misconception benchmark, clustering step-error annotations alone improves clustering accuracy by 0.22 and normalized mutual information by 0.05 on average over clustering the problem and student response, consistent with annotations capturing recurring error patterns more effectively. On Eedi and unlabeled MalAlgoQA, annotation-based representations show higher average held-out clustering accuracy and internal cluster separation, respectively, than their unannotated baselines. To help experts interpret these clusters, the pipeline then generates candidate misconception descriptors and coherence scores for human triage. Overall, these findings show that input representation substantially affects clustering results and that step-error annotations often improve clustering quality.

Paper 657 LLM-as-RNN: Feedback-Driven Memory Rewriting for Online Learning and Sequential Prediction

Authors: Yuxing Lu, J. Ben Tamo, Jinrui Fang, Weichen Zhao, Nan Sun, Xukai Zhao, Yishan Zhong, Wenqi Shi, Jinzhuo Wang, May Dongmei Wang

Abstract Large language models (LLMs) are strong sequence predictors, but standard context accumulation offers no explicit mechanism for revising the memory carried between predictions. We propose LLM-as-RNN, an inference-time framework for online learning with frozen LLMs through a bounded natural-language memory state. At each timestep, the model predicts from the current observation and its memory, receives feedback, and rewrites the state to correct errors and retain task-relevant patterns without updating model parameters. Implemented as a structured system prompt, this state makes the adaptation process inspectable through human-readable memory traces. We evaluate LLM-as-RNN on sequential prediction benchmarks in healthcare, meteorology, and finance across Llama, Gemma, and GPT-oss model families. Comparisons with static-context and feedback-aligned baselines examine predictive performance, feedback sources, memory budgets, and inference cost. Our results demonstrate that feedback-driven memory rewriting supports online error correction and sequential prediction under a fixed memory budget, with performance varying across tasks and backbones.

Paper 658 Epistemic Erosion: An Error Class Beyond Hallucination in LLM Text Generation

Authors: ~Micky_C._Nnamdi1, ~Shaunna_Wang1, ~Chinaza_Nnamdi1, ~Jieru_Li1, ~Benoit_Louis_Marteau1, ~J._Ben_Tamo1, ~Yining_Yuan2, ~May_Dongmei_Wang1

Abstract Expert writers calibrate their claims with qualifiers such as ``likely benign`` in a radiology report, ``may constitute`` in a court ruling, and ``suggest`` in a scientific paper. We ask whether this calibration survives when Large Language Models (LLMs) summarize and rewrite professional text. We define epistemic erosion as any change in a source's expressed certainty, including qualified claims that are dropped or introduced without basis. We measure it with the Epistemic Faithfulness Score (EFS), computed by aligning source and output claims one-to-one with rule-based matching and assigning each distortion to one of six types, namely escalation, attenuation, omission, fabrication, scope shift, and severity warp. On clinical notes, mean EFS for 11 open-weight LLMs ranges from 0.415 to 0.493; three of these models are also evaluated on legal and scientific text. Four hosted models (GPT-4o, GPT-5.4, Phi-4, Mistral-Large-3) reach at most 0.483 in any domain. The most frequent detected distortion is omission (66\%), in which a qualified claim is dropped entirely; claims that survive tend to shift toward greater certainty. Model scale and medical adaptation bring no consistent gain, seven prompting strategies bring at most modest gains, and ROUGE-L and BERTScore correlate only weakly with EFS ($|r| \approx 0.1$).

Paper 662 The bias is in the generator: locating medical chain-of-thought subgroup gaps upstream of the verifier

Authors: Micky C. Nnamdi, Benoit Louis Marteau, Shaun Qien Yeau Tan, J. Ben Tamo, Yining Yuan, May Dongmei Wang

Abstract Verifier-guided best-of-N decoding has two stages: a generator samples candidate chains of thought, and a verifier selects among them. A subgroup disparity may already be present in the generator's candidates or be amplified during selection. We separate the two stages using adversarial demographic counterfactuals of medical exam questions. On HuatuoGPT-o1-8B, a substantial race accuracy gap is already present in a single sampled response, before verifier selection. Across seven generators, only Gemma-3-27B-it under Skywork-RM shows significant verifier-induced amplification after multiple-testing correction. Even an oracle selector leaves a residual gap on every medical generator analyzed, showing that correct answers are less often available for one subgroup. Counterfactually paired low-rank adaptation (LoRA) of the generator reduces the single-sample gap on two medical models, although on HuatuoGPT-o1-8B the reduction does not persist after self-judge selection. Single-variant controls indicate that the single-sample reduction is not explained by additional training steps alone. A short fairness instruction modestly reduces the single-sample gap, and a direct preference optimization (DPO) variant of the paired training also reduces it and, on HuatuoGPT-o1-8B, keeps race mention near its baseline rate. For the medical generators studied, candidate generation rather than inference-time selection is the primary locus for mitigating this gap.

Paper 663 Dgrammar: Efficient Constrained Decoding for DLMs

Authors: Guan-Ming Chiu, Jeng-Yue Liu

Abstract Diffusion language models (DLMs) decode several tokens per step. However, the grammar-constrained decoders evaluated here, IG-CD and LAVE, verify candidate tokens sequentially within each denoising step and treat every violation as a rejection, a lookahead verification, or a fallback to autoregressive decoding. We observe that a violation is local to the first token the parser cannot consume, so the rest of the parallel step can be kept. We propose Dgrammar, a training-free decoder that repairs that frontier token in place under the same forward-pass logits, paired with a grammar-gated backoff schedule for the unmask batch size, asynchronous mask construction, and a bounded grammar-guided tail. On JSONSchemaBench medium with LLaDA-8B, Dgrammar improves schema validity by 8.6 percentage points over the strongest DLM baseline while reducing mean latency by 5.8x and eliminating timeouts. Validity and latency gains also hold on the other JSB splits and on Dream-7B, and experiments with a SMILES grammar demonstrate cross-grammar transfer.

Paper 667 Lay-Language and Cross-Lingual Disparities in LLM Medical Diagnosis: A Case Study in Rheumatological Skin Conditions

Authors: Grayson Hubbell, Yifei Zhang, Saksham Khatwani, Sriram Sankaranarayanan, Indira Sriram, Maria Leonor Pacheco

Abstract Patients increasingly turn to large language models (LLMs) for medical self-diagnosis in both English and non-English, often in colloquial language. We evaluate LLM diagnostic performance on rheumatological skin conditions across lay descriptions in English, Spanish, Hindi, and Mandarin, collected from native speakers without medical training, and clinical descriptions in English. Lay descriptions substantially reduce diagnostic accuracy relative to clinical ones, and the gap widens further across non-English languages. A diagnostic-focused system prompt narrows the lay-vs-clinical gap but does little to close the cross-language gap. Current LLMs appear to reproduce, and in some settings amplify, communication-driven disparities documented in clinical medicine.

Paper 668 CircuitPile: A Data Synthesis Framework for Enhancing Multimodal LLMs' Perception on Circuit Schematics

Authors: Hongyu Chang, Xinyue Li, Huanyong Liu, Feng Zhang, Yuhui Yin

Abstract In recent years, multimodal large language models have achieved substantial progress, yet their performance remains limited in specialized domains such as circuit diagram understanding. A key bottleneck lies in the lack of diverse and large-scale training data tailored to the structural characteristics of circuit diagrams, together with the high cost of expert annotation. To address this challenge, we propose CircuitPile, a systematic data synthesis framework for circuit visual-language modeling. CircuitPile integrates heterogeneous circuit resources from multiple sources and constructs standardized label spaces and multi-granularity visual representations. Based on this framework, we build 738K domain-specific pre-training samples and 100K task-oriented instruction-tuning samples for the pre-training and supervised fine-tuning stages, respectively. Furthermore, to systematically evaluate fundamental visually grounded circuit perception, we introduce CircuitPerception-Bench, a diagnostic benchmark comprising 2,700 multiple-choice samples across three dimensions: component recognition, spatial understanding, and quantitative reasoning. CircuitPerception-Bench focuses on fundamental visually grounded perception and complements external benchmarks such as AMS Benchmark. Experimental results show that multimodal models trained with CircuitPile achieve significant performance gains on the internal benchmark and demonstrate favorable transferability on the external public AMS Benchmark. For example, applying CircuitPile-based SFT to Qwen3-VL-8B improves the scores on CircuitPerception-Bench and AMS Benchmark from 0.5759 and 0.5067 to 0.7703 and 0.5988, respectively.

Paper 672 Not All Verifiable Preference Pairs Are Equal: Support-Aware Filtering for Stable Iterative DPO

Authors: Viet Anh Ngo, Long Tri Thai Son, Nguyen Viet Duc, Dong Minh Duc, Quang Minh Tran, Nguyen Thi Thuy Linh, Nguyen Viet Ha

Abstract Direct Preference Optimization (DPO) is well suited to verifiable reasoning tasks because preference pairs can be constructed automatically from answer correctness. However, in iterative DPO, correctness alone does not determine whether a pair provides a useful training signal. We distinguish two regimes according to the origin of the chosen response: On-policy pairs, where both the chosen and rejected responses are sampled from the current policy, and Fallback pairs, where the policy fails to generate a correct response and the chosen completion is supplied by a dataset-provided reference solution. Based on this distinction, we propose support-aware filtering, which applies margin-based filtering to On-policy pairs and a likelihood-based support gate to Fallback pairs. Experiments on GSM8K and MATH-500 with Qwen2.5-style 0.5B, 1.5B, and 3B models show that support-aware filtering yields its clearest gains on MATH-500 at 1.5B and 3B, while GSM8K and the 0.5B MATH-500 setting are more mixed. Overall, our findings suggest that preference margins in iterative verifiable DPO are pair-origin dependent, and that chosen-response support is a useful signal for stabilizing fallback supervision.

Paper 673 No-Think Is Not No-Thought: The White Bear Effect in LLM Reasoning Control

Authors: Zixuan Wang

Abstract In large language models with explicit reasoning controls, a final no-think directive appears to close the reasoning channel. Whatever mode history preceded it, the model should now answer directly. We show that this latest-control picture is false in a way that exposes a mechanism of negated instruction following. We construct matched-final histories in which the current no-think instruction, question, and answer contract are identical, and only the earlier mode controls differ. Across Qwen3-family models, these histories remain separated at the moment of answer formation even when visible reasoning markers are absent. The central surprise is directional. Prior no-think and alternating histories can make a later no-think state more, not less,

poised toward reasoning tokens. We interpret this as a White Bear effect in reasoning control. Suppressing chain-of-thought can preserve it as something the model must not say. Causal interventions locate the preserved state at the assistant answer boundary rather than in the literal control token or arbitrary prompt positions. Learned boundary coordinates bidirectionally steer reasoning-minus-direct pressure across scales, while matched random directions do not. Thus no-think is not simply an erasure command. It is a stateful control operation that can produce surface compliance while retaining a causal, boundary-local representation of the reasoning behavior being withheld.

Paper 674 Do Correctness Probes Transfer? A Cross-Family Study of LLM Hidden States

Authors: Evan Duan

Abstract Linear probes on language-model hidden states can predict output correctness, but it remains unclear whether they capture transferable correctness representations. We study out-of-distribution correctness-probe transfer across seven datasets, four task families, and three open-weight LLMs: Qwen3-8B, Llama-3.1-8B, and Gemma-2-9B, with all findings validated across five random seeds. Within-family transfer is consistently stronger than cross-family transfer (gap 0.075 AUC), with code generation showing the clearest pattern across MBPP-test, MBPP-train, and HumanEval. We further find that probe weight directions are nearly orthogonal across most dataset pairs despite strong predictive transfer, suggesting that probes do not encode a single canonical correctness direction. Comparing probes to logprob confidence baselines, we find probe advantage is concentrated in in-domain and within-family transfer cells. We also identify a layer-selection confound where early-layer shortcuts can appear transferable in-domain but fail cross-source. Overall, correctness probes transfer selectively rather than universally, making them most useful in-domain and within task families.

Paper 675 Learning to Elicit LLM Bias in Multi-Turn Conversations

Authors: Da Gao, Ziyu Chen

Abstract Static bias benchmarks score a model on isolated questions, yet LLMs that answer safely at first can still reveal bias as a conversation unfolds. We frame bias evaluation as multi-turn conversational elicitation and learn an interviewer that probes a target LLM, conditioned on a demographic group, a bias category, and a conversational strategy. The interviewer is trained with supervised fine-tuning on synthetic dialogues and refined with preference optimization on conversation-level feedback. Across all models we test, bias grows as the conversation continues, exposing failures that single-turn evaluation misses. Training the interviewer matters: it surfaces substantially more bias than the same model prompted with identical instructions, and preference optimization pushes it further. Multi-turn elicitation and single-turn search turn out to be complementary: at matched query budgets they surface similar amounts of bias, but conversation uncovers more severe cases. Human annotations corroborate these judgments. Our results argue that bias evaluation of conversational AI should move beyond single-turn probes.

Paper 680 SWE-Adept: An LLM-Based Agentic Framework for Deep Codebase Analysis and Structured Issue Resolution

Authors: Kang He, Kaushik Roy

Abstract Large language models (LLMs) exhibit strong performance on self-contained programming tasks. However, they still struggle with repository-level software engineering (SWE), which demands (1) deep codebase navigation with effective context management for accurate localization, and (2) systematic approaches for iterative, test-driven code modification to resolve issues. To address these challenges, we propose SWE-Adept, an LLM-based two-agent framework where a localization agent identifies issue-relevant code locations and a resolution agent implements the corresponding fixes. For issue localization, we

introduce agent-directed depth-first search that selectively traverses code dependencies. This minimizes issue-irrelevant content in the agent's context window and improves localization accuracy. For issue resolution, we employ adaptive planning and structured problem solving. We equip the agent with specialized tools for progress tracking and Git-based version control. These tools interface with a shared working memory that stores code-state checkpoints indexed by execution steps, facilitating precise checkpoint retrieval. This design enables reliable agent-driven version-control operations for systematic issue resolution, including branching to explore alternative solutions and reverting failed edits. Experiments on SWE-Bench Lite and SWE-Bench Pro demonstrate that SWE-Adept consistently outperforms prior approaches in both issue localization and resolution, improving the end-to-end resolve rate by up to 4.3%.

Paper 681 VTBench: Evaluating Visual Tokenizers for Autoregressive Image Generation

Authors: Huawei Lin, Tony Geng, Zhaozhuo Xu, Weijie Zhao

Abstract Autoregressive (AR) models have recently shown strong performance in image generation, where a key component is the visual tokenizer (VT) that maps continuous pixels into discrete tokens. The quality of the VT largely determines the upper bound of AR model performance. However, existing discrete VTs significantly lag behind continuous variational autoencoders (VAEs), leading to degraded reconstructions and poor preservation of fine details and text. Current benchmarks primarily evaluate end-to-end generation quality, making it difficult to isolate VT performance. We introduce VTBench, a comprehensive benchmark for systematic evaluation of visual tokenizers across three core tasks: image reconstruction, detail preservation, and text preservation. Evaluating state-of-the-art VTs with multiple metrics, we find that continuous VAEs consistently outperform discrete VTs, especially in retaining spatial structure and semantic details. Discrete VTs often produce distorted reconstructions and fail to preserve fine textures and text. We release VTBench publicly to facilitate future research on robust, general-purpose visual tokenizers.

Paper 685 User Feedback to Actionable Insights: A Fine-grained Multilabel Approach for Classifying App Reviews

Authors: Chirag Jain, Preethu Rose Anish

Abstract The rapid proliferation of mobile applications has resulted in a vast volume of user-generated reviews, which often contain critical insights into bugs, feature demands, and overall user satisfaction. Efficiently organizing and interpreting this feedback is essential for developers seeking to enhance app quality and user experience. In this study, we propose a fine-grained, multilabel classification schema for app reviews, structured as a two-level taxonomy. The first level categorizes reviews into three primary classes: Bug Report, Feature Request, and App Satisfaction & Feedback. The second level further refines these categories into subtypes addressing specific concerns such as performance, usability and privacy. To automate this classification, we fine-tune several state-of-the-art large language models (LLMs), including LLaMA, Gemma, Mistral, Qwen, on a manually annotated dataset spanning 10 diverse application domains. Among these, LLaMA 3.1 achieved the highest performance, with weighted F1 scores of 0.90 and 0.79 at the first and second taxonomy levels, respectively. To assess scalability, we applied the fine-tuned LLaMA 3.1 model to label an additional 10,000 unlabeled reviews, creating a silver-standard corpus. This taxonomy-driven approach enables more nuanced analysis of user feedback, supporting targeted app improvements and fostering enhanced user satisfaction.

Paper 688 Beyond Idealized Patients: Evaluating LLMs under Challenging Patient Behaviors in Medical Consultations

Authors: Yahan Li, Xinyi Jie, Wanjia Ruan, Xubei Zhang, Huaijie ZHU, Yicheng Gao, Chaohao Du, Ruishan Liu

Abstract Large language models (LLMs) are increasingly used for medical consultation and health information support, where safety depends not only on medical knowledge but also on robust responses to unclear, inconsistent, or misleading patient input. However, most existing medical LLM evaluations assume idealized and well-posed patient questions, limiting their realism. We study challenging patient behaviors that commonly arise in real medical consultations and complicate safe clinical reasoning. We define four clinically grounded categories of such behaviors: information contradiction, factual inaccuracy, self-diagnosis, and care resistance. For each behavior, we specify concrete failure criteria that capture unsafe responses. Building on four existing medical dialogue datasets, we introduce CPB-Bench (Challenging Patient Behaviors Benchmark), a bilingual (English and Chinese) benchmark of multi-turn dialogues annotated for these behaviors. We find that although models perform well overall, they exhibit consistent behavior-specific failures, especially when handling contradictory or medically implausible patient information. We further evaluate four intervention strategies and find inconsistent improvements, with some interventions introducing unnecessary corrections. We release the dataset and code.

Paper 689 Latent Prediction Needs Alignment: A Controlled Study of Joint-Embedding Predictive Vision--Language Learning

Authors: Mohammad Kohankhaki, Daniel Kusuma, Shirin Salehi, Carsten Kamp, Sigrid Brell-Cokcan, Anke Schmeink

Abstract Predictive vision--language learning aims to learn semantic representations by predicting latent targets rather than reconstructing pixels or generating text. We ask whether latent prediction alone induces the global image--text geometry required for zero-shot matching, and how the location of alignment pressure changes this outcome. In a controlled joint-embedding predictive model family with matched training and evaluation protocols, non-contrastive latent prediction retains information recoverable by supervised adapters and token-level probes but does not form a usable zero-shot image--text space under pooled cosine scoring. A target-contrastive design improves matching, but remains weaker than the directly aligned variant. Our results show that latent prediction is underconstrained at the global embedding level in this setting. Direct global alignment supplies the constraint that makes predictive representations usable for zero-shot scoring. With only a modest direct global alignment loss, latent prediction becomes competitive with matched CLIP and SigLIP baselines and improves several language-sensitive diagnostics.

Paper 690 SaFeR-ToolKit: Structured Reasoning via Virtual Tool Calling for Multimodal Safety

Authors: Zixuan Xu, Tiancheng He, Huahui Yi, Kun Wang, Xi Chen, Gongli Xi, Qiankun Li, Kang Li, Yang Liu, Zhigang Zeng

Abstract Vision–language models remain susceptible to multimodal jailbreaks and over-refusal because safety hinges on both visual evidence and user intent, while many alignment pipelines supervise only the final response. To address this, we present SaFeR-ToolKit, which formalizes safety decision-making as an inspectable protocol: during offline data synthesis, a planner specifies a persona, a **Perception**→**Reasoning**→**Decision** tool set, and a constrained transition graph, while a responder emits a typed key–value tool trace before the final answer. To learn this protocol, we train a single policy with a three-stage curriculum (SFT→DPO→GRPO), where GRPO directly supervises tool usage beyond answer-level feedback. We contribute (**I**) the first tool-based safety reasoning dataset of 30,654 training entries (SFT 6k, DPO 18.6k, GRPO 6k) plus 1k held-out evaluation, and (**II**) experiments on Qwen2.5-VL showing large gains in Safety/Helpfulness/Reasoning Rigor on 3B (29.39/45.04/4.98 → 84.40/71.13/78.87) and 7B (53.21/52.92/19.26 → 86.34/80.79/85.34), while preserving general capabilities

(3B: 58.67 → 59.21; 7B: 66.39 → 66.81). Code: https://github.com/Duebassx/SafeR_Toolkit.

Paper 693 Improving LLM-Based Entity Detection with Annotation Guidelines

Authors: Marcel Geromel, Philipp Cimiano

Abstract In this work, we investigate whether detailed, dataset-specific annotation guidelines can replace or augment annotated few-shot exemplars for prompt-based entity detection. To this end, we propose an in-context learning framework that conditions the LLM on a guideline document, task instructions, and optional demonstrations, enabling single-pass, primarily zero-shot nested entity detection. We evaluate multiple instructional aspects from the comprehensive ACE 2005 annotation guidelines, using three guideline-compliant datasets as benchmarks and GPT-5 variants for inference. Based on extensive experimentation, we demonstrate that detailed, dataset-specific instructions, especially those governing mention boundaries, consistently and robustly improve performance, even in conjunction with established few-shot methods. Indeed, even under zero-shot constraints, our guideline approach outperforms state-of-the-art few-shot methods on ACE-2005, achieving up to 81.2 percent F1 without relying on existing labeled datasets. Our findings present detailed annotation guidelines as a promising resource for improving instruction-based information extraction.

Paper 694 COMED: The Missing Middle Between Routing and Collaboration in Multi-LLM Inference

Authors: Norah Alballa, Wenxuan Zhang, Salma Kharrat, Fares Fourati, Zafar Ayyub Qazi, Mohamed Elhoseiny, Marco Canini

Abstract No single Large Language Model (LLM) is uniformly reliable across queries, motivating multi-model inference systems that either route among models or combine their outputs. However, routing stops after selecting an initial model, while dense collaboration invokes peers on every query. We show that collaboration is non-monotonic: peers can recover failures that no model solves alone, but can also corrupt initially correct answers. We introduce COMED (Controlled Model Escalation for Multi-LLM Deliberation), a post-anchor controller for selective cross-model collaboration. COMED uses anchor self-consistency, router margin, and a lightweight peer probe to accept confident answers, verify ambiguous cases, and escalate only when collaboration is likely beneficial. We formalize this trade-off with a rescue-harm decomposition showing that selective collaboration improves when rescued errors outweigh collaboration-induced harms. Across medical, scientific, and general reasoning benchmarks, COMED improves fixed and routed anchors in all 16 open-weight settings, with gains up to +10.7 percentage points on MedQA while invoking fewer models and using fewer decoded tokens than dense collaboration. On HLE with frontier models, COMED improves GPT-5.5 from 23.1% to 28.1%, outperforming dense collaboration and achieving the best results.

Paper 695 Toward LLMs Beyond English-Centric Development

Authors: Sho Takase, Ukyo Honda

Abstract Through an analysis of sequences generated by open-weight large language models (LLMs), we demonstrate that LLMs are heavily biased toward English. While continual pre-training is commonly used to adapt LLMs to a target language, we show that it does not offer a cost advantage over training from scratch, even for improving cultural understanding in the target language. These findings suggest that dedicated per-language investment may become increasingly important for future LLM development, rather than relying primarily on the expansion of English-centric resources.

Paper 696 Do not be greedy, Think Twice: Sampling and Selection for Document-level Information Extraction

Authors: Mikel Zubillaga, Oscar Sainz, Oier Lopez de Lacalle, Eneko Agirre

Abstract Document-level Information Extraction (DocIE) aims to produce an output template with the entities, relations, and events of interest occurring in the given document. Standard practices include prompting decoder-only LLMs using greedy decoding to avoid output variability. Rather than treating this variability as a limitation, we show that sampling can produce substantially better solutions than greedy decoding, especially when using reasoning models. We thus propose ThinkTwice, a sampling and selection framework in which the LLM generates multiple candidate templates for a given document, and a selection module chooses the most suitable one. We introduce both an unsupervised method that exploits agreement across generated outputs, and a supervised selection method using reward models trained on labeled DocIE data. To address the scarcity of golden reasoning trajectories for DocIE, we propose a rejection-sampling-based method to generate silver training data that pairs output templates with reasoning traces. Our experiments show the validity of unsupervised and supervised ThinkTwice, consistently outperforming greedy baselines and the supervised state-of-the-art.

Paper 699 RL-ARC: Calibrating Large Reasoning Models via Reasoning-guided Uncertainty

Authors: Gukhyeon Lee, SangKeun Lee

Abstract Language models (LMs) are commonly trained with Reinforcement Learning with Verifiable Rewards (RLVR) to enhance their reasoning capabilities. However, since RLVR does not explicitly account for calibration during training, it can lead to severe calibration degradation, including overconfidence. Recent calibration-aware training methods for LMs, which incorporate objectives for uncertainty estimation into training, improve calibration but still exhibit overconfidence under distribution shift, while sacrificing reasoning performance. To this end, we propose RL-ARC, a calibration-aware training framework that jointly leverages reasoning confidence and answer confidence. Specifically, RL-ARC leverages reasoning confidence as an auxiliary signal for calibrating answer confidence, applying it as reasoning-guided regularization for correct cases and as an overconfidence penalty for incorrect cases. Comprehensive results across ID and OOD settings show that, beyond improving calibration, RL-ARC enables reasoning models to adaptively estimate confidence based on the given question without substantially sacrificing reasoning performance, thereby highlighting the importance of reasoning confidence for training reliable reasoning models.

Paper 700 Transferable Structural Priors in Continuous Diffusion Language Models

Authors: Hao Li, Man Fung Zhuo, Xinqi Liu, Jiayu Wu, Yifan Hong

Abstract Continuous diffusion language models struggle with structured generation under hard, variable-length templates. We identify two failure modes: Boundary Attention Drift, in which head-region attention bleeds into padded tails under block-mode training, and Vocab Projection Drift, in which latent-to-token rounding over-projects to low-frequency tokens. The two modes dissociate cleanly across Chinese ci, English sonnets, and E2E NLG: the former is specific to block-mode training, whereas the latter appears in all three datasets regardless of padding regime, isolating padding regime and rounding rule as orthogonal sources of error. To address structural correctness, we adapt constrained decoding to diffusion rounding through a position-template primitive, the V_{i} rule, instantiated on continuous-diffusion logits as Length-Conditional Voronoi Rounding (LCVR). The same primitive applied to a small autoregressive baseline reaches identical zero-violation rates, showing that structural correctness is delivered by the template, not by the underlying model family. A contribution more readily realised with diffusion backbones is instead cross-domain backbone transfer: initialising a target-domain model from a structurally compatible source-domain backbone yields order-of-magnitude reductions in training steps, while a controlled negative-transfer arm fails to

reproduce the speed-up on structural metrics; an attention-span probe localises the transferable component to the source backbone's long-range attention span rather than to its padding regime or corpus size. We further report a calibrated wallclock analysis showing that the dominant cost of diffusion inference is recoverable with standard step-reduction techniques. Together, these results recast several previously scattered structured-generation failures of continuous diffusion as one consistent, transferable, and fixable phenomenon.

Paper 701 HeRE: A Novel Benchmark for Hebrew Relation Extraction via LLM-Guided Denoising of Knowledge Graph Alignments

Authors: Ron Keinan, Amir David Nissan Cohen, Reut Tsarfaty

Abstract While LLMs demonstrate strong capabilities in English Relation Extraction (RE) tasks, a significant gap remains in assessing their RE performance in low-resource languages, due to a lack of standardized benchmarks. Creating native, language-specific annotated RE benchmarks is expensive and requires human expertise, while relying on translation yields suboptimal, inaccurate datasets that introduce English-centric biases. To tackle this, we propose a scalable, language-agnostic methodology to create RE data by denoising naturally occurring yet noisy alignments of knowledge bases and corresponding texts. We define a proper alignment via a Natural Language Inference (NLI) task, where a relation triplet is entailed by an input text. We empirically assess five denoising strategies and demonstrate that multilingual models outperform language-specific models in the few-shot LLM setting, while Hebrew-specific encoders remain competitive under NLI fine-tuning and are strongest in the label-free cross-training setting. We further find that prompting in the native language achieves performance on par with or slightly exceeding English prompts, confirming that the pipeline operates naturally in the target language. The best open-weights model, Gemma-4-31B-it (F1=0.915), matches the best API model, GPT-5.4 (0.921), and reaches F1=0.844 on a fully held-out, dual-annotated test set. Applying the optimized pipeline, we release HeRE: a novel, massive, 1.5-million-instance silver dataset for Hebrew RE, designed to support data-intensive experimentation and model fine-tuning. HeRE bridges a critical gap in standard Hebrew RE evaluation, while providing the NLP community with a reproducible, language-agnostic pipeline for RE dataset construction in other low-resource settings.

Paper 703 Query-Conditioned Graph Retrieval for Contextualized LLM Reasoning in Personalized Wearable Data

Authors: ~Zhenyu_Lu4, ~Mahyar_Abbasian1, ~Amir_M._Rahmani1

Abstract The growing prevalence of multi-sensor wearable devices generates vast amounts of long-term, multimodal time-series data, posing significant challenges for context-aware Large Language Models (LLMs). Current LLM-based health analysis methods typically rely on manually curated context, which becomes impractical with increasing data volume and sensor diversity. To overcome these limitations, we introduce \textbf{Wearable As Graph (WAG)}, a novel framework that automates context retrieval for LLMs using personalized knowledge graphs. WAG constructs knowledge graphs mapping relationships between wearable modalities and incorporates user-specific data. We develop a data-driven retrieval pipeline that leverages both global (long-term) and local (short-term) relationships within metrics to identify the most relevant nodes for user queries. We evaluate WAG on a benchmark of over 10k data-associated queries created from multiple wearable datasets. WAG consistently delivers the strongest overall performance, while using substantially fewer input tokens than query-window full-context prompting in the diagnostic setting. These gains remain broadly consistent under human evaluation and across a cross-model generator--judge matrix, supporting structured, query-adaptive context retrieval for LLM-driven analysis of wearable data.

Paper 705 Policy-Conditioned Constrained Decoding for Column-Level Access Control in Text-to-SQL

Authors: Miyamoto Ryoto, Xin Fan, Hayato Yamana

Abstract Text-to-SQL is increasingly deployed across trust boundaries between data providers and users. Such deployment must balance three competing requirements: policy compliance, answer coverage, and bounded cost. Existing approaches typically decide refusal based on which columns a query mentions and enforce it stochastically. Whether a query is compliant, however, depends not only on which columns appear but on how they are used, and stochastic enforcement cannot deterministically rule out violations. We formalize this requirement as a column-use policy over semantic use: output, filter condition, and aggregation argument. We integrate the policy by aligning each role with grammar productions tracked by the decoder. The resulting system, PCC-SQL, applies a per-token logits mask that deterministically eliminates single-query column-use violations on the supported SQL fragment in a single decoding pass. Across three benchmarks and three open-source models, PCC-SQL achieves 0% Leakage Rate and Coverage up to 88.7% on Spider-CU, while staying within +10% tokens of direct prompting. We additionally assess semantic alignment with execution accuracy.

Paper 709 A Systematic Review of Meta-Reasoning for Large Reasoning Models

Authors: Haoran Yu, Zilong Wang, Kang Pan, Xiao Zhang, Erli Meng, Shuai Li

Abstract In this paper, we survey recent advances in meta-reasoning for Large Reasoning Models (LRMs). Although LRMs have made remarkable progress, they face critical limitations including overthinking on simple queries, under-thinking on complex problems, and ineffective exploration lacking strategic direction. We examine how meta-reasoning addresses these challenges by providing the necessary mechanisms to dynamically regulate and optimize the reasoning process, and propose a framework to categorize existing studies into three core components: the monitor which provides self-perception of the reasoning state, the controller which dictates the meta-policy for regulating actions, and the mechanism which executes the algorithmic mapping between them. By systematically reviewing these developments and foundational challenges, we aim to map the current landscape and identify promising directions for creating more adaptive, efficient, and self-regulating reasoning systems.

Paper 714 Domain-Agnostic Neural Topic Modeling with Contextual Token-Level Semantic Graph Representation

Authors: Seung-Won Seo, Won Ik Cho, Yongmin Yoo

Abstract PLM-powered neural topic models depend on general-domain pre-training, leading to degraded topic interpretability on specialized corpora. The root cause is geometric: domain-specific terms unseen during pre-training collapse into an indistinguishable region of the embedding space, and neither domain-specific re-training, word-level graph enrichment, nor parameter-efficient fine-tuning can restructure this space without inheriting the capacity ceiling of the underlying encoder. Our key insight is that a learnable graph layer operating on token-level PLM embeddings can acquire corpus-specific semantic structure that the frozen encoder lacks, because token-level graphs preserve document-local context that word-level representations discard and joint optimization with the topic objective reshapes embedding geometry directly from target-domain evidence. We instantiate this insight as DARTopic, a domain-agnostic framework that constructs token-level semantic graphs from frozen PLM embeddings and jointly trains a GNN encoder with variational topic inference. Across three benchmarks spanning general, biomedical, and legal domains, DARTopic consistently outperforms strong baselines in topic coherence and document clustering without any encoder fine-tuning, while demonstrating robustness to PLM choice and favorable runtime efficiency over fine-tuning-based alternatives.

Paper 717 Surfacing Minor Nuances in Subjective Expressions: Initial Evidence for Improved Output Quality in Collaborative Caption Revision

Authors: Takeru Isaka, Jiro Nagao, Iwaki Toshima

Abstract Creative collaboration benefits when collaborators can absorb the nuances embedded in others' subjective expressions. We focus on especially difficult-to-notice nuances, which we call **minor nuances**, and provide initial empirical evidence that supporting their communication can improve creative collaboration outcomes. We then propose a commonsense-generation-based method that identifies minor nuances in subjective expressions and presents them as cues for collaborators. We evaluated the proposed method in a collaborative image-caption revision experiment designed to reflect brainstorming. Compared with a baseline in which collaborators freely explained the nuances embedded in their own subjective expressions, our method significantly improved output-quality ratings from human evaluators. These results provide initial evidence suggesting that explicitly sharing minor nuances supports richer creative expression by helping convey fine-grained aspects of senders' inner states that are difficult to communicate through ordinary explanations.

Paper 718 PoliSpace: Political Opinion Prediction using Shared Latent Space

Authors: Johanna D'Ciofalo Khodaverdian, Frédéric Berdoz, Roger Wattenhofer

Abstract Modeling and predicting political opinions remains a challenging task for AI systems. We introduce PoliSpace, a framework that embeds individuals and political questions into a shared latent space, where vector alignment directly reflects agreement. Individuals are represented as learnable vectors, and questions are encoded via a contrastively fine-tuned language model with individual-specific affine transformations to capture subjective interpretation. This design yields an embedding space that is robust to question rephrasing and sensitive to negation. Evaluated on real-world voting data from Swiss national elections, PoliSpace achieves an 89.81\% F1 score on seen opinion prediction for rephrased questions, outperforming prior embedding-based approaches (66.17\% F1). We further introduce retrieval and confidence-aware thresholding mechanisms that reach up to 100\% F1 at reduced coverage, demonstrating how the learned embedding space can be leveraged to improve prediction reliability. Ablation studies confirm the importance of both individual-specific interpretation modeling and negation-aware encoder fine-tuning. The learned individual embedding space also exhibits clear clustering by political party affiliation, despite no party-label supervision during training, capturing the underlying ideological structure of the Swiss political landscape.

Paper 719 Predicting Mergeability of Language Model Domain Experts

Authors: Ervin Dervishaj, Quynh Do, Giannis Karamanolakis, Anna Rumshisky, Salma Sohrabi-Jahromi, Judith Gaspers

Abstract Model merging allows extending an expert model's capabilities by fusing its weights with another expert. However, task interference has been shown to negatively affect mergeability. To address this, prior work has mainly focused on introducing new merging methods. Despite reducing task interference, the effectiveness of these merging methods is only evident post-merging, requiring a full evaluation of the merged model, which for contemporary LLMs can be expensive. In this work, we study the overlooked predictive power of structural and functional properties of merged experts for mergeability. Our experiments with 8 expert LLMs spanning several programming languages, math, and knowledge show that activation L2 distance and activation cosine similarity emerge as the most important properties for predicting effective merging, contrasting with prior findings on vision models where gradient-based metrics dominated. We further show that task vector L2 distance, a cheaply computed structural metric, provides a mechanistic link between mergeability prediction and training dynamics: it explains why learning rate effects on mergeability are domain-dependent in LLMs, contradicting the universal ``higher noise improves merging"

principle from vision. This enables a practical intervention where mergeability metrics guide learning rate selection during expert training.

Paper 720 RAGCap-Bench: Benchmarking Capabilities of LLMs in Agentic Retrieval Augmented Generation Systems

Authors: Jingru Lin, Chen Zhang, Yong Liu, Haizhou Li

Abstract Retrieval-Augmented Generation (RAG) mitigates key limitations of Large Language Models (LLMs), such as factual errors, outdated knowledge, and hallucinations, by dynamically retrieving external information. Recent work extends this paradigm through agentic RAG systems, where LLMs act as agents to iteratively plan, retrieve, and reason over complex queries. However, these systems still struggle with challenging multi-hop questions, and their intermediate reasoning capabilities remain underexplored. To address this, we propose RAGCap-Bench, a capability-oriented benchmark for fine-grained evaluation of intermediate tasks in agentic RAG workflows. We analyze outputs from state-of-the-art systems to identify common tasks and the core capabilities required for their execution, then construct a taxonomy of typical LLM errors to design targeted evaluation questions. Experiments show that “slow-thinking” models with stronger RAGCap performance achieve better end-to-end results, underscoring the benchmark’s validity and the importance of enhancing these intermediate capabilities.

Paper 721 Beyond Retrieval: Deterministic Memory Governance with Pairwise Constraints and Certified Repair for Long-Term Language Agents

Authors: Panayiotis Christou, Miguel Palma, Zefan Du, Kaixun Hua, Ying Mao

Abstract Long-term memory significantly enhances language-model agents but introduces a critical failure mode: retrieval surfaces relevant yet stale, deleted, or mutually inconsistent facts. A query about a current address may retrieve the correct address alongside deleted ones. The answer is present, but the memory set is unsafe. We reframe agent memory as a governed object-selection problem and present MemGov, a substrate-agnostic middleware layer between existing retrieval pipelines and agent context. MemGov wraps retrieved evidence as explicit memory objects, detects must-link and cannot-link constraints, and applies deterministic constraint-aware repair, returning either a certified safe memory set or an auditable obstruction certificate rather than silently passing unsafe combinations. Evaluations on controlled governance scenarios and MemoryAgentBench-style end-to-end tasks, deliberately separating retrieval quality from governance decisions, demonstrate that while raw retrieval recovers answer-bearing evidence successfully, it simultaneously returns detected unsafe memory combinations in 47-51% of cases. The same retrieval substrates governed by MemGov reduce unsafe merges to 0% while preserving answer utility within about 2-4 percentage points, with fully auditable reasoning for all repair and refusal decisions. Additional diagnostics assess eligibility, temporal constraints, and paired answer retention.

Paper 723 Auditing Bias Detection Under Source Shift: A Source-Aware Benchmark and Evaluation Protocol

Authors: ~Ana_Pinto1, ~Alvaro_Figueira1

Abstract Bias-detection benchmarks routinely report accuracies above 90%, but a high number can reflect where a sentence came from as much as whether it contains a stereotype or an unfair claim about a group. We treat this as an evaluation problem and propose an integrated auditing protocol that makes source confounding measurable: an LLM-assisted label audit with cross-family verification, pair-aware splitting with role tracking, source-aware evaluation with a source-only baseline on the actual test split, counterfactual discrimination, and an artifact analysis of edited text. We apply it to a three-source corpus of 10,304 sentences that we audit and release. The audit reduces the global source-only majority baseline from 95.0% to 55.4%, and the

same protocol shows that the real-only test split still carries a 77.6% source-only baseline, so confounding is reduced and measured but far from eliminated. Against that baseline, fine-tuned BERT and GPT-2 gain roughly 18 pp and outperform attention-derived models; four augmentation variants that look ordered in the mean fall within seed variance under paired testing; and a real-vs-edited probe separates Gemini counterfactuals from real text, although the signal disappears once the gold label is held fixed and has no measurable accuracy effect. Applied to CrowS-Pairs, the label audit flags 33.2% of the supposedly neutral members as still containing a group-level stereotype. The main contribution is the protocol, which exposes confounding that standard pipelines would report as results.

Paper 724 HDSL: A Hierarchical Domain-Specific Language for Structured 3D Indoor Scene Generation and Localized Editing with LLM Agents

Authors: Letian Li, Chao Shen, Shuzhao Xie, Chenghao Gu, ZhengXiao He, Yu Meng, Xin Yang, Wenyuan Jiang, Zhi Wang

Abstract Text-driven indoor scene generation and editing require an intermediate representation that language models can both produce and revise. Existing LLM-based systems often rely on scene graphs or global constraint lists, which are compact but underspecify local geometry and make instruction-based edits difficult to localize. We frame the task as structured program generation and local program repair, and propose the Hierarchical Descriptive Scene Language (HDSL), an XML/CSS-style domain-specific language in which rooms, regions, objects, and support surfaces form a tree with parent-relative local coordinates. Our pipeline generates HDSL subtrees recursively with bounded verification, grounds non-virtual nodes through multimodal asset retrieval, and applies force-directed layout optimization to repair boundary and collision errors. For editing, Hierarchical Retrieval-Augmented Generation retrieves the relevant subtree, rewrites only that local context, and merges the result back through a deterministic three-way merge. In our reproduced benchmark, HDSL improves average object coverage, text-scene alignment, and generation time over full text-to-scene baselines while remaining competitive with layout-only reproductions on geometry metrics. On thirty paired editing instructions, HRAG reduces token use by 67.6% and runtime by 68.1%, and cuts non-target removals by 87.5% over full-scene rewriting; attribute ablations confirm that local coordinates and virtual/leaf flags are necessary for valid, locality-preserving edits.

Paper 727 Training LLMs Beyond Next Token Prediction - Filling the Mutual Information Gap

Authors: ~Chun-Hao_Yang4, ~Bo-Han_Feng1, ~Tzu-Yuan_Lai1, ~Yan-Yu_Chen2, ~Yin-Kai_Dean_Huang1, ~Shou-De_Lin1

Abstract Autoregressive language models are usually trained with next-token prediction under a fixed left-to-right target order. This paper studies target-token organization as a training-time design choice across arithmetic reasoning, multi-label text classification, and text generation. Controlled order studies show that the default NTP order can be far from the best-performing tested organization, while fixed hand-crafted orders such as reverse order do not generalize across tasks. To address this, Max(MI(S; t)) is introduced as a deterministic mutual-information-based heuristic that prioritizes information-rich target tokens. For fixed-length outputs, it constructs task-specific prediction orders; for open-vocabulary text generation, the same principle is applied through lightweight prefix augmentation. Experiments with GPT-style models and pretrained Llama/Qwen variants show that Max(MI(S; t)) often improves over standard NTP and simple baselines such as reverse order, TF-IDF, and conditional-probability selection, without changing model architecture or autoregressive decoding.

Paper 728 Structural vs Cultural Reasoning: Reliability Limits of LLM-as-a-Judge in Multilingual Evaluation

Authors: Shehenaz Hossain, Houda Bouamor, Haithem Afli

Abstract LLM-as-a-Judge has emerged as a scalable alternative to human evaluation in multilingual cultural reasoning, yet its reliability for culturally grounded judgments remains poorly understood. We present a human-aligned validation study using the CRaFT framework across Arabic, Bengali, and Spanish, evaluating six judge models, including five open-weight models and GPT-4.1, against paired human annotations across two structurally observable dimensions (Answer Consistency and Explanation Consistency) and three culturally interpretive dimensions (Cultural Fluency, Linguistic Adaptation, and Deviation). Our results reveal a consistent reliability divide: automated judges align substantially better with humans on structural dimensions than on culturally interpretive ones requiring sociocultural reasoning. Reliability further varies across languages, judge architectures, and response-generation models, with several failures reflecting systematic calibration limits rather than prompt underspecification alone. These findings suggest that current automated evaluators are best treated as complementary diagnostic tools rather than replacements for human judgment in culturally sensitive multilingual settings.

Paper 729 Decisions by Discrimination: Differential Evidence for Medical Reasoning

Authors: Yunpeng Xiao, Yuzhang Xie, Junhao Huang, Carl Yang, Xiao Hu, Kai Shu

Abstract Large language models (LLMs) are increasingly studied for clinical decision support, and medical question answering provides an effective testbed for evaluating this capability. In particular, medical multiple-choice questions (MCQs) reflect a core challenge in clinical decision-making: plausible options often share substantial overlapping evidence, while only a few decisive factors distinguish the correct option from other incorrect ones. However, existing retrieval-based and reasoning-oriented methods mainly focus on supplying broadly relevant knowledge or eliciting reasoning paths, rather than identifying the compact differential evidence that separates confusing alternatives. In this work, we present a retrieval-grounded differential evidence framework for medical MCQs. Our method extracts option-specific pieces of differential evidence and organizes them into concise evidence prompts for answer selection. This design better aligns external knowledge with differential reasoning while reducing irrelevant context. Experiments show consistent accuracy gains over strong baselines. Ablation studies further show that our framework outperforms both directly using retrieved evidence chunks and closed-source LLM-generated general hints, and that its improvements are not solely explained by answer-proximal guidance, but are consistent with differential evidence serving as a compact aid for reasoning over competing alternatives. Our findings also offer insights for medical reasoning and other decision-making settings with multiple possible choices.

Paper 730 What Improves During Pretraining? Tracking Knowledge Acquisition with Point-in-Time Models

Authors: ~Chang_Su21, ~Bin_Wu4, ~Emine_Yilmaz1, ~Yao_Lu5

Abstract Language models pretrained on web corpora collected over multiple years learn both general and domain-specific knowledge. How these different kinds of knowledge are learned as pretraining progresses remains unclear. Recent point-in-time models trained on web snapshots from different years provide a way to probe how knowledge prevalence varies over time. Using these models, we propose a temporal variability score (TVS), defined as the coefficient of variation of a fact's negative log-likelihood across models. A higher score suggests that the knowledge is less consistently represented across years of web text. We evaluate six benchmarks covering diverse topics using six models from multiple open-source model families. We find that low- and high-TVS items are learned similarly early in training, but high-TVS items improve faster later on. At the end of the pretraining run, high-TVS items outperform low-TVS items in 34 of 36 model--benchmark settings, with an average advantage of 7.3 points of chance-normalized accuracy. This pattern is especially pronounced on knowledge-intensive tasks, even when

comparing items of similar difficulty. Our findings reveal differences in how knowledge is learned over the course of pretraining, with temporal variability providing a diagnostic of these learning dynamics beyond aggregate benchmark scores.

Paper 731 The Expense of Seeing: Attaining Trustworthy Multimodal Reasoning Within the Monolithic Paradigm

Authors: Karan Goyal

Abstract The rapid proliferation of Vision-Language Models (VLMs) is often framed as enabling unified multimodal understanding, but rests on an under-examined assumption: that current VLMs faithfully synthesise multimodal data. We argue they often do not, and that this gap reflects a trustworthiness problem in the dominant Vision Encoder-Projector-LLM paradigm. Rather than extracting grounded knowledge from visual inputs, state-of-the-art models frequently exhibit functional blindness, exploiting strong language priors to bypass severe visual representation bottlenecks. We challenge the conventional methodology of multimodal evaluation, which relies on data ablation or new dataset creation and therefore conflates dataset biases with architectural incapacity. We propose an information-theoretic departure: the *Modality Translation Protocol*, which preserves a sample's semantic payload while translating its modality rather than ablating it. From this, we derive three metrics—the Toll (*ToS*), Curse (*CoS*), and Fallacy (*FoS*) of Seeing—culminating in the *Semantic Sufficiency Criterion (SSC)*, a single condition that detects, quantifies, and localises multimodal failures to the visual encoder or the projection head. We further hypothesise a *Divergence Law of Multimodal Scaling*: as language backbones grow, the penalty of the visual bottleneck may widen rather than close. Examining the closest-matching existing benchmarks, motivates our central call: the construction, through *Semantic Equivalence Engineering*, of a large, diverse dataset whose modalities are isomorphic by construction. We frame this not as an experiment awaiting compute, but as a mission for the community: to move beyond "multimodal gain" and build the substrate on which faithful multimodal architectures can be designed.

Paper 732 Ontology-Guided Knowledge Graph Construction from Structured Metadata and Climate Science Publications

Authors: Aayush Acharya, Qi Zhang, Eduard Dragut, Longin Jan Latecki

Abstract The rapid expansion of scientific literature far outpaces the speed of manual curation, creating a fundamental bottleneck in extracting structured knowledge from scientific texts. To address this, we introduce an agentic, ontology-guided framework for knowledge graph (KG) construction. Our framework utilizes specialized agents for document understanding, information extraction, ontology conformance, and entity linking. First, an LLM extracts factual knowledge in the form of triples with explicit textual evidence from research papers. Then the triplets undergo a multistage verification process to ground the triplets in the ontology. By enforcing schema constraints and requiring evidence in the form of text fragments, we ensure the data is both semantically consistent and traceable. In an agreement study over 1,000 extracted triplets, Gemini-2.5-flash achieves Cohen's κ scores of 0.640--0.700 against three human annotators, matching the human--human agreement range of 0.616--0.689. We applied this framework to 2,000 scientific publications and construct a novel, ontology-grounded KG for climate science with 133,990 nodes and 755,308 relations; of the 23,327 triplets that the publications contribute after validation, 98.96\% are judged correct. The pipeline code and KG is publicly available.

Paper 736 When Reranking Hurts: Uncertainty-Based Gating for Few-Shot Reranking

Authors: Orian Dabod, Amir David Nissan Cohen, Gabriel Stanovsky

Abstract Few-shot selection typically assumes that reranking retrieved examples always improves performance. We challenge this view by identifying that the expensive reranking step can in fact degrade performance. Instead, we propose Training-Free Gated Reranking, which decides whether to rerank the few-shot examples based on the model's uncertainty. Extensive experiments across 8 LLMs, covering 7 NLU datasets and 9 MT domain-language combinations, demonstrate that our approach reduces average computational costs by approximately 20\% for MT and 54\% for NLU at a calibrated operating point, while matching the performance of full reranking. These findings indicate that higher computational cost does not guarantee better performance, and that reranking is most beneficial when targeted at high-uncertainty instances.

Paper 737 Who Owns the Answer? Mechanistic Evidence for State Arbitration Between Written Reasoning and Internal LLM Computation

Authors: Zixuan Wang

Abstract Scratchpads turn written intermediate reasoning into context that can compete with a large language model's own computation. We study this competition as scratchpad arbitration. In deterministic symbol tasks, a corrupted scratchpad and the rule table imply different answer options, yielding matched trace-owned and rule-owned runs. In Qwen3-4B, the trace-implied state, rule-implied state, and ownership-relevant variables are represented before the answer token. Causal patching then switches which state controls the answer. A sparse pathway transfers trusted targets toward the rule answer and skeptical targets back toward the trace answer. The effect replicates on held-out items, is necessary under ablation, and localizes to layer-25 heads L25H29/H30/H31 with a smaller layer-36 head L36H26 contribution. Source, attention, and temporal decompositions place the decisive signal at the options interface and the answer position, with little causal contribution from the local corrupted trace step. Written reasoning thus behaves as an external state proposal whose control can be decoded, localized, and causally reassigned.

Paper 738 Prompt Hysteresis: Task Inertia and Metastable States in In-Context Learning

Authors: ~Zixuan_Wang58

Abstract Large language models can infer new task rules from in-context demonstrations, but it remains unclear how readily an inferred rule is revised when later evidence contradicts it. We study this question through prompt hysteresis: path dependence in the response to matched but oppositely ordered conflicting demonstrations. In semantic-free symbolic and cipher tasks, we construct paired incompatible mappings and compare A-to-B prompts with B-to-A prompts at the same rule proportions. A gap between the two response curves indicates retention of a prompt-induced task state beyond current evidence aggregation. We quantify the loop with task magnetization and hysteresis area, and develop a Landau-style effective theory in which evidence acts as a control field while task-state self-consistency creates metastable basins. Across Qwen3 dense models, task inertia increases with scale. Mean hysteresis area rises from 0.067 for Qwen3-0.6B to 0.269 for Qwen3-8B when prompts contain at least six demonstrations, and from 0.077 to 0.309 in the n=12 slice. A matched-suffix control gives the same ordering under identical later evidence, with the preparation gap increasing from 0.235 to 1.197. State-dynamics probes show gradual decay across long suffixes, and RESET boundaries reduce but do not erase preparation memory. Residual patching localizes transferable task-state information to late final-token activations, while learned task-axis interventions move stuck predictions 4 to 9× more than matched random directions. The behavioral and mechanistic evidence identifies prompt hysteresis as a measurable metastability phenomenon in in-context learning.

Paper 741 Sampling via Decision-Flow: Training-Free Extraction of Improved Latent Reasoning Paths in Large Language Models

Authors: ~Zhendong_Mi1, ~Shaoyi_Huang1

Abstract A central question in LLM reasoning is whether reinforcement learning (RL) instills genuinely new capabilities or merely reshapes how existing knowledge is expressed during inference. Building on the distribution-sharpening hypothesis, which holds that RL reallocates probability mass toward high-reward trajectories already latent in base models, we ask: can we unlock those latent paths without costly RL fine-tuning? We present Decision-Flow Sampling (DF-Sample), a training-free, data-free inference-time framework that constructs a hierarchical reasoning tree, scores terminal nodes for quality, and back-propagates utilities to inform each intermediate branching decision. Unlike conventional sampling strategies that make purely local step-wise choices, DF-Sample performs explicit global trajectory evaluation before committing to a path, recovering high-quality but low-probability reasoning chains that standard decoding overlooks. On GPQA, DF-Sample achieves 45.6% accuracy, surpassing power sampling (38.9%) and GRPO (39.9%), showing that a training-free method can outperform a trained one. Across three models and four benchmarks, DF-Sample consistently outperforms baselines, indicating substantial latent reasoning potential in pretrained base models.

Paper 744 Does Fairness Survive Model Adaptation in Legal Judgment Prediction?

Authors: Sarah Kamoun, Parth Patel, Faezeh Ensan

Abstract Large language models are increasingly used for Legal Judgment Prediction (LJP), raising concerns about demographic bias inherited from pre-training corpora and historical legal records. We investigate how fairness robustness changes under model adaptation in LJP through a controlled counterfactual evaluation across gender, ethnicity, and intersectional demographic groups. Using 23 demographic variants over encoder-based models and generative LLMs, we evaluate demographic sensitivity, counterfactual instability, and bias propagation under zero-shot, few-shot, and fine-tuned settings. Experiments show that model adaptation substantially affects fairness. Fine-tuning often improves predictive performance while simultaneously increasing demographic sensitivity and propagating biased supervision signals into downstream judicial predictions, whereas few-shot prompting generally produces smaller propagation effects. We further show that legal-domain specialization does not reliably guarantee fairness: legal-domain encoders span both highly robust and highly sensitive behavior after fine-tuning, and a legal-domain LLM is highly sensitive under few-shot prompting yet comparatively robust after fine-tuning under counterfactual perturbation.

Paper 749 IRB: Automated Benchmark Generation for Evaluating Information Seeking in Deep Research Agents

Authors: Lam Thanh Do, Bhagyashree Taleka, Hozaifa Ammar Bhutta, Vikram Sharma Mailthody, Kevin Chen-Chuan Chang, Wen-mei Hwu

Abstract As Deep Research Agents (DRAs) increasingly take on complex information-seeking tasks, evaluating them becomes essential. However, primary benchmarks are static, rely on expensive manual curation, making them susceptible to staleness and data contamination; and often lack the ground-truth trajectories required to assess an agent's intermediate search behavior. To address these limitations, we introduce IRB, a fully automated benchmark generation framework. IRB utilizes factual scaffold derived from Wikipedia citing sentences, and imposes an algorithmic scaffold to reliably synthesize multi-hop and false-premise questions. Human evaluation confirms that 96.5% of generated samples meet strict criteria for question well-formedness and

answer validity. This quality is achieved at a small cost of $120. Leveraging IRB, we evaluate various DRA configurations and find that agents equipped with the classic BM25 remain highly competitive against those using frontier neural retrievers, though neural models maintain an advantage in cross-lingual settings. We further observe that while agents successfully retrieve relevant evidence, they frequently fail to explicitly debunk false premises. Finally, trajectory analysis reveals that although agents correctly plan logical retrieval orders, they often prematurely advance to subsequent reasoning hops before securing the necessary prerequisite evidence.

Paper 750 Activation-Efficient Fine-Tuning via Sequence-Compressed Low-Rank Subspaces

Authors: Irene Tenison, Stella Ahn, Miriam Kim, Ebtisam Alshehri, Lalana Kagal

Abstract Parameter-Efficient Fine-Tuning (PEFT) has become the standard approach for adapting large language models (LLMs). However, despite reducing trainable parameter counts, many PEFT methods remain memory-bound by token-level activation intermediates that scale linearly with sequence length, limiting their practicality for long-context and on-device adaptation. Unlike prior memory-efficient PEFT approaches that preserve token-level adapter activations, LARS (Low-memory Activation-Rank Subspace) reduces the sequence-length scaling of stored adapter activation intermediates from O(BSRL) to O(BRL) by performing adaptation within a sequence-level low-rank subspace. Through pooled feature extraction, low-rank subspace modulations, and residual projection, LARS reduces the overall memory footprint by 23% on GPUs and 47.8% on CPUs relative to LoRA at S = 1024, with relative savings increasing with context length. Experiments across reasoning, understanding, long-context, retrieval, and token-level tasks show that LARS maintains competitive performance relative to LoRA while substantially improving memory efficiency. Beyond GPUs, LARS also reduces average CPU power and energy consumption per optimization step by 22.43% and 22.06%, respectively. These results demonstrate that activation-efficient adaptation provides a practical path toward resource-constrained and on-device fine-tuning of LLMs.

Paper 752 Breaking Bureaucracy: Evaluating open-source LLMs for legal document review

Authors: Farrukh Baratov, Niki van Stein, Suzan Verberne

Abstract In this paper, we evaluate open-source generative LLMs on legal Natural Language Inference (NLI). Legal inspectorial processes take place in specific domains and often deal with confidential data. This creates a need for working with local models that do not require labeled training data. We evaluate our models on the ContractNLI benchmark and two NLI4Wills datasets. We successfully reproduce the baseline for the task (Span NLI BERT) and we evaluate multiple open-source LLMs on the same task. We analyze the invalid rate of the models, and their stability across temperature settings and domains. Among the generative models, Gemma-4 26B performs the best, reaching an accuracy of 81.2%, even outperforming the supervised model on one metric. On accuracy, it is not possible to beat the supervised model with zero-shot approaches. Qwen-3.6 35B performs well on both ContractNLI and additional datasets in the legal wills domain. Our findings indicate that zero-shot, open-source, generative LLMs are a viable alternative for real-world legal NLI when no supervised data is available. Our code is available at https://github.com/fbaratov/contractnli-llms.

Paper 754 Detect Before You Disambiguate: Perturbation-Based Input Ambiguity Detection in LLMs

Authors: Tiejin Chen, Xiaoou Liu, Ahmadreza Moradipari, Michael Clifford, Kyungtae Han, Yitao Chen, Nejib Ammar, Hua Wei

Abstract Large language models (LLMs) often answer ambiguous inputs as if they had a single intended meaning. This is problematic because input ambiguity is an upstream source of uncertainty: when a prompt admits multiple valid interpretations, the key question is whether the model should answer directly at all. Existing input uncertainty methods often rely on LLM-generated statements or clarifications, but such transformations may already disambiguate the input before the ambiguity is

measured. We propose a label-preserving perturbation framework for input ambiguity detection in which perturbations remain close to the original input while preserving its label. We instantiate this principle with two detectors: a black-box rule-based perturbation library and a white-box importance-guided method using model-internal signals. Experiments across four datasets in open-domain, conversational, and visual QA show that our methods outperform output-side uncertainty baselines and LLM-paraphrase-based perturbation baselines. Our results highlight the importance of detecting ambiguity before attempting to disambiguate it. Our code is available \href{https://anonymous.4open.science/r/May_Arr-00C1/}{\textcolor{purple}{here}}.

Paper 756 Diverse Reasoning Paths Matter: Symbolic Equivalence Augmentation for Enhancing RL Exploration

Authors: Xiaoou Liu, Tiejin Chen, Hanyang Chen, Yuan Ling, Hua Wei

Abstract Reinforcement learning with verifiable rewards (RLVR) has emerged as a powerful paradigm for improving the reasoning capabilities of LLMs. However, the role of supervised fine-tuning (SFT) data in enabling effective downstream RL remains underexplored. In this paper, we investigate a key hypothesis: the diversity of reasoning steps in chain-of-thought (CoT) during SFT is a critical enabler of RLVR performance. To test this hypothesis, we introduce EquivAug, a data augmentation framework that uses symbolic equivalence transformations to generate diverse yet provably correct reasoning paths. By converting CoT solutions into formal symbolic representations, applying equivalence transformations, and reconstructing natural-language explanations, EquivAug produces training data that differs in reasoning steps rather than surface-level lexical variation. Experiments on GSM8K and FOLIO across model scales demonstrate that EquivAug consistently achieves the highest reasoning path diversity after SFT and the strongest accuracy after GRPO. Our results establish that CoT diversity during SFT is positively correlated with downstream RLVR performance.

Paper 761 Information-Theoretic Design of Sparse Autoencoder-Based Steering

Authors: Hirofumi Morise

Abstract Sparse autoencoder (SAE)-based activation steering is a lightweight approach for modifying the behavior of large language models (LLMs) by adding small vectors to hidden states during inference. However, designing effective steering interventions remains challenging. In this work, we develop an information-theoretic framework for analyzing SAE activations through empirical feature-firing distributions. The central object of the framework is the Jensen--Shannon divergence (JSD) between such firing distributions, which quantifies how far a generated or intervened state deviates from a dataset-induced target distribution. Based on an analytic divergence response for virtually shifted states, we construct steering vectors toward the target firing distribution. The same divergence can also be used to monitor deviation from another empirical distribution, such as a general valid-language distribution. We demonstrate the effectiveness of the steering in a question-mode generation task study. We further use teacher-forced pulse experiments to show that local interventions induce structured downstream JSD responses in later layers, and apply the framework to a safety-related target--contrast response setting for feature analysis.

Paper 763 An Empirical Study of Position Bias in Visualized Text Retrieval

Authors: Cuiqiaoshu Pan, Ziyang Zeng, Dun Zhang, Yuqing Yang

Abstract Visualized text retrieval, which renders text into images and retrieves them with vision-language models, has been proposed to avoid lossy text parsing and compress long contexts. We ask whether it also bypasses a known flaw of text retrieval: position bias, the tendency to overweight a passage's beginning. Rendering 1D text into a 2D image may change the bias in two respects: its shape, which position is favored, and its magnitude, how strongly performance varies across positions. We test this

by rendering the PosIR benchmark as controlled plain text images and comparing eight visual retrievers with four text retrievers on identical instances. We find that (1) position bias persists, with all visual retrievers exhibiting measurable bias (macro PSI 0.061–0.168); (2) its shape is transformed: seven of eight visual retrievers favor the middle of the page in an inverted-U pattern, contrasting with both the primacy bias of text retrievers and the lost-in-the-middle effect of LLMs with long contexts; (3) its magnitude is sensitive to rendering settings, as taller images amplify PSI and can even reverse positional preference, while font size has a similarly large effect; and (4) under matched backbones, visual retrievers exhibit lower position bias than their text counterparts on longer documents, but this advantage vanishes beyond the pixel budget and at document scale. Visualized text retrieval therefore does not eliminate position bias; it transforms a familiar 1D textual bias into a more complex 2D spatial bias.

Paper 766 A Geometric Perspective on Stabilizing Value Conflict Resolution

Authors: Saket Reddy, Andy Liu

Abstract Large Language Models (LLMs) often struggle to navigate user requests that contain complex value conflicts, as direct scalar rewards are often ill-equipped to handle these scenarios. To address this challenge, we investigate how chain-of-thought (CoT) reasoning can help improve performance in this domain. Geometrically, we show that CoT correlates with further smoothing the model's loss landscape in its sharpest direction, helping resolve the optimization instability of traditional scalar rewards. In terms of performance, we demonstrate via relevant downstream benchmarks that CoT may increase capability, demonstrating that value conflict-focused CoT has the potential to be an effective mechanism for better moral reasoning. To capitalize on this potential, we create a new value conflict-focused CoT design that correlates with further smoothing the sharpest direction of the loss landscape and further increases moral reasoning performance. This finding shows that explicitly modifying and improving the design of reasoning dynamics offers a promising avenue for advancing pluralistic alignment in LLMs.

Paper 767 Do LLMs Have Social Sense? Evaluating Alignment of Conversation Understanding and Reasoning

Authors: Zhaoqing Wu, Dan Goldwasser

Abstract Large language models (LLMs) demonstrate strong instruction-following and reasoning abilities, yet struggle with the socially nuanced understanding humans handle naturally. Humans rely on social sense to interpret intentions, emotions, and relationships in context. We investigate whether and to what extent LLMs exhibit similar abilities when reasoning about social interactions. We propose a structured approach that decomposes social interactions into three aspects and prompts LLMs to generate aspect-based summaries capturing complementary signals. We evaluate these summaries and resulting reasoning within a unified framework across three dimensions, analyzing both LLM–human alignment and the quality of LLM-generated social explanations. Our results show that LLMs achieve moderate agreement with humans and with one another as evaluators and capture some socially nuanced information across aspects, but struggle with interaction-level dynamics shaped by situational norms. Across four downstream CSS tasks, aspect summaries improve aggregate prediction performance across three model families, though effects vary by task. Complementary rationale analysis shows that summaries substantially alter expressed reasoning. Finally, we introduce a benchmark dataset with gold-standard annotations to support systematic study of social understanding and situated reasoning in LLMs.

Paper 771 BG-RAG: Concept-Mediated Bipartite Graphs for Scalable RAG

Authors: Hang Gao, Yangmin Ding, Shaobo Han, Zhuocheng Jiang

Abstract Graph-structured retrieval-augmented generation (GraphRAG) has shown strong potential for improving multi-hop reasoning in large language models. However, existing approaches typically depend on dense entity–entity linking, which is computationally expensive, susceptible to hallucinated relations, and difficult to maintain under continual updates. We present Bipartite Graph Retrieval-Augmented Generation (BG-RAG), a lightweight bipartite framework that organizes knowledge into three layers, from chunks to entities to concepts, and replaces direct entity links with concept-mediated connections. This design enables more efficient graph construction, reduces reliance on hallucination-prone relation-edge prediction by avoiding explicit entity–entity relation extraction, and naturally supports incremental updates without global reconstruction. Experiments on HotpotQA, 2Wiki, and MuSiQue show that BG-RAG achieves competitive performance against RAG and GraphRAG baselines. Its concept-mediated structure also provides a lightweight alternative to dense entity–entity graph construction. These results suggest that BG-RAG is a lightweight alternative for retrieval-augmented generation, particularly when avoiding dense entity–entity linking is desirable.

Paper 773 Topo-Narrator: Heterogeneous Topological Memory for Long-form Narrative Generation

Authors: Kai Chen, Haoran Zhang, Tian Lan, Zehua Duo, Jiang Li, Guanglai Gao, Xiangdong Su

Abstract Long-form narrative generation requires maintaining multi-dimensional consistency over extensive contexts, covering long-term constraints in character settings, event causality, spatial rules, and worldviews. Existing methods compress these heterogeneous elements into homogeneous representations, creating a topological mismatch that lowers retrieval signal-to-noise ratios and induces long-range logical collapse. We propose Topo-Narrator, a framework built on two complementary decouplings. Heterogeneous Topological Memory (HTM) organizes narrative state into a Character Graph, a Plot Chain, and an Environment Tree, providing topology-specific inductive biases. Plan-Reason-Write (PRW) multi-agent collaboration separates high-level reasoning, memory retrieval, and stylistic realization across dedicated agents. Experiments on MoPS and WritingPrompts show that Topo-Narrator consistently generates coherent narratives averaging over 13,000 words and substantially outperforms strong baselines on HANNA quality and fine-grained consistency metrics, validated by three frontier LLM judges and human evaluation.

Paper 774 Encoded but Not Express: Probing Subtle Speech-Conditioned Emotion in Large Audio Language Models

Authors: Jingxi Liang, Pengze Guo, Zhiwen Xie, Qifeng Wang, Yujia Huo, Derek F. Wong

Abstract Large Audio Language Models (LALMs) exhibit strong speech understanding ability, yet it remains unclear whether they meaningfully utilize subtle speech-conditioned emotional cues beyond coarse acted emotions. In this work, we investigate whether the evaluated LALMs exhibit measurable sensitivity to fine-grained emotional variation in speech through controlled perturbation experiments, fine-grained emotion elicitation tasks, and representation geometry analysis. Our results show that vocal delivery variations systematically influence both model behavior and hidden representations, indicating that the evaluated LALMs already encode non-trivial speech-conditioned affective information. However, models struggle under emotionally ambiguous settings, particularly for semantically related fine-grained emotions. Representation analysis further suggests that subtle emotional structure is partially preserved despite weak downstream utilization. Overall, our findings suggest that the evaluated LALMs do not completely ignore subtle speech emotion cues, but instead utilize them inconsistently under ambiguous affective conditions

Paper 776 See Every Block Once: Variance-Reduced Post-Training for Masked Diffusion Language Models

Authors: Yunzhen He, Hiroaki Yamagiwa, Hidetoshi Shimodaira

Abstract Policy gradient methods are widely used to align large language model outputs with human preferences, but the log-likelihoods they require cannot be directly computed in Masked Diffusion Language Models (MDLMs). MDLM post-training therefore relies on a surrogate objective based on the Evidence Lower Bound (ELBO), and recent work estimates a blockwise variant of this objective via Monte Carlo methods to support blockwise generation. However, existing estimation schemes randomly sample the block index, leading to unstable estimation. In this work, we propose a full-block estimator that evaluates every block exactly once and averages the results. We theoretically show that the proposed estimator remains unbiased while eliminating the between-group variance arising from block selection, thereby always achieving lower variance than the baseline for the same sample size. In experiments on five benchmarks spanning mathematics, planning, and coding, our method achieves the highest accuracy on four benchmarks and outperforms the random block sampling baseline by an average of 12.79 points.

Paper 777 Amortizing Self-Correction for Continual Embodied Adaptation

Authors: Woo Kyung Kim, Jooyoung Kim, Minjong Yoo, Honguk Woo

Abstract Language model (LM)-powered agents have shown promise in addressing complex embodied tasks by leveraging the commonsense world knowledge of pre-trained models in interactive physical environments. Yet, real-world deployment requires agents to continually adapt to a stream of tasks, where neither large-scale expert supervision nor reliable external feedback is consistently available. In this paper, we present ASC, a framework for efficient continual embodied task adaptation through amortizing self-correction. Specifically, ASC trains the policy under two objectives: a standard CoT reasoning objective distilled from stronger experts, and a reflexive reasoning objective that internalizes self-correction by correcting initial rationales generated by a previously learned policy. This dual optimization allows the agent to adapt more efficiently through forward knowledge transfer over time, ultimately fostering robust CoT reasoning at inference without requiring explicit reflection. Furthermore, we introduce rationale-wise test-time scaling that dynamically adjusts the depth of CoT reasoning based on the model's confidence in actions inferred from its own rationales. Through experiments on VirtualHome and ALFWorld, we demonstrate performance superiority over other LM-based planning and continual task adaptation approaches, while achieving strong efficiency in computation, data usage.

Paper 778 Back to Basics, Powered by Data: Character-Level Transformers for Robust Chinese Machine-Generated Text Detection

Authors: Shengjie Xu, Zhenghao Gao, Rujie Dai, Zijing Li, Xiaoyang Xie, Zehua Lyu

Abstract Machine-generated text detection remains difficult in Chinese, where systematic benchmarks and large paired corpora are still scarce. We revisit this problem with a compact 4-layer character-level Transformer, trained on a 363.4-million-character Chinese detection corpus, to our knowledge the largest AI-human paired Chinese resource for this task. The resulting data-centric pipeline attains strong scores on Chinese benchmarks and, for the independently trained English variant, state-of-the-art RAID leaderboard numbers that remain stable under 11 RAID adversarial attack types. Ablations expose a double-edged sword role for punctuation: it eases optimization yet encourages overfitting to surface cues; symbol cleaning shifts the model toward deeper syntactic and discourse patterns and improves generalization. Minimal architecture plus rich supervision thus yields compact detectors suitable as a document-level authorship authenticity screen before downstream Chinese text processing; on a laptop RTX 3070 Ti we measure 5.90 x 10^5 tokens/s at under 100 MB VRAM. We present GiantMelon-ZH (Chinese) and GiantMelon-EN (English for RAID): same backbone, disjoint training corpora, transparent data descriptions.

Paper 779 PULSE-CARE: A Measurement-Validity Audit of Simulated-Patient Clinical Agent Evaluation

Authors: Srimonti Dutta, Dinky Mishra, Aditi Singh, Akshata Kishore Moharir

Abstract Simulated-patient benchmarks are widely used to evaluate clinical AI agents, yet the patient simulator itself can alter the evidence available to the agent and therefore the conclusions drawn from the benchmark. We introduce PULSE-CARE, a measurement-validity audit for simulated-patient clinical-agent evaluation that quantifies sensitivity to patient simulator behavior, agent architecture, simulator backbone, scorer choice, and demographic framing. Across 24 synthetic longitudinal cases, four dynamic simulator variants, three agent models, and two architectures, the mean per-case score range across simulators is 0.25 (95% CI: 0.21-0.29), and escalation decisions differ across simulator conditions for 60% of agent-case pairs. A 392-trajectory repeated-sampling analysis shows that simulator effects persist after pooling repeated draws, while also revealing substantial within-policy stochastic variation. Nominal model rank reversals occur in 17% of comparisons but disappear under a 0.05 practical-difference margin. Demographic counterfactuals change escalation in 26% of matched pairs, essentially the same rate as the same-input repeated-sampling control, providing no evidence of a demographic effect in this audit. PULSE-CARE combines simulator-sensitivity metrics, EFI composite indices, Patient Simulator Cards, and an Audit Report template to make the measurement conditions behind simulated-patient benchmark results explicit. Our results show that clinical-agent performance reported from a single patient simulator should be interpreted as conditional on the simulator and evaluation pipeline used to obtain it.

Paper 780 DensePath REINFORCE: Lightweight Policy Improvement from SFT-Induced Token-Level Reward Signals

Authors: ~Jiangnan_Li1, ~Thuy-Trang_Vu1, ~Ehsan_Abbasnejad3, ~Gholamreza_Haffari2

Abstract Supervised fine-tuning (SFT) is typically viewed as behavior cloning on expert demonstrations. Recent methods recover richer supervision from demonstrations through self-play preference synthesis or bilevel inverse reinforcement learning, yet these require additional optimization stages or separately trained reward models. Motivated by the connection between SFT and inverse soft-Q learning, we propose Dense-Path REINFORCE (DPR), a lightweight post-SFT policy-improvement method that extracts a dense reward signal directly from the SFT training trajectory. We construct a token-level reward from the log-probability gap between the final SFT policy and an intermediate checkpoint, capturing the incremental likelihood gain acquired during training while mitigating the length and early-EOS biases inherent in raw log-probability rewards. DPR reuses the same demonstrations and prompts as SFT, requires no preference data, no external teacher, and no separately trained reward model. Across four pretrained backbones and four instruction-following evaluations, DPR obtains 55.2-66.8% win rates over SFT, improves MT-Bench by 0.22-0.46 points, and matches or exceeds the best demonstration-only baseline in 15 of 16 settings.

Paper 781 Not All Tokens Are Equally Important: Auditing and Compressing Redundancy in LLM Reasoning

Authors: Nguyen-Khang Le, Dieu-Hien Nguyen, Truong Dinh Do, Le-Minh Nguyen

Abstract Chain-of-thought traces drive strong LLM accuracy but dominate inference cost: output length sets both monetary cost and decoding latency. We ask which tokens in a reasoning trace actually carry the answer. A token-level analysis of the model's own next-token distribution identifies three redundancies - discourse-element removal, tokenizer-aware abbreviation, and substitution of prose causal connectives with an arrow "X -> Y". All three leave the predictive distribution close to the original under Jensen-Shannon divergence and preserve the greedy next token in the vast majority of prefixes. We instantiate the analysis as CONCISON, a single training-free system prompt that operationalises the three rules and protects numeric, equation, and code tokens verbatim. Across 18 models from six families (2B-120B) and five reasoning benchmarks, CONCISON cuts

output tokens by 24-62% on strong models at -1.1pp average accuracy, outperforms token-budget prompting in a single decoding call, and matches trained routers on the accuracy-cost frontier without any training, model pool, or extra infrastructure.

Paper 784 More Modalities, More Problems: When Multimodal Agentic Debate Hurts Time-Series Reasoning

Authors: Patara Trirat

Abstract Multi-agent debate is often motivated by the premise that multiple agents, modalities, and rounds of deliberation improve reasoning by combining complementary evidence and correcting individual errors. We test this premise in multimodal time-series analysis using **MAD-TS**, a prompting-based framework with modality-specific agents, cross-modal judging, and synthesis. Contrary to expectation, MAD-TS does not reliably outperform simpler baselines. Across 20 tasks in finance, weather, healthcare, and science, it is uniquely best on 2 of 23 outcomes and significantly beats a strong single-agent multimodal baseline on only one task, while requiring about 22$\times$ more tokens and 12$\times$ more wall-clock time. Adding modalities, increasing debate depth, or changing judge rounds can help some tasks but harm others. Trace analysis attributes 85% of errors to all three judges agreeing on the same wrong answer, a shared upstream bias that limits same-backbone multimodal debate as a general-purpose approach.

Paper 787 ICCU: In-Context Continual Unlearning via Pattern-Induced Refusal Rules

Authors: Ruihao Pan, Suhang Wang

Abstract Machine unlearning aims to remove the influence of specific data from trained language models. In real-world deployments, unlearning requests often arrive sequentially, which challenges existing fine-tuning-based methods: fine-tuning each request is costly, accumulates utility loss, and may cause cross-request interference. To address these issues, we propose ICCU (In-Context Continual Unlearning), an in-context continual unlearning framework that induces readable refusal rules from unlearning datasets and applies them at inference time either as a filter or via the system prompt, without modifying model parameters. Because rules are accumulated as an order-independent union, ICCU is compositional and free of cross-request interference, and the original forget-set data can be discarded after rule induction. Extensive experiments show that ICCU effectively suppresses target knowledge while preserving utility, scales across long sequences of requests, and remains robust to paraphrased, cross-lingual, and adversarially framed queries. Our code is available at https://github.com/rhpan/ICCU.

Paper 791 PerQ: Efficient Evaluation of Multilingual Text Platform-Adaptation Quality

Authors: Dominik Macko, Andrew Pulver

Abstract Since no metrics are available to evaluate specific aspects of a text, such as its personalization quality or quality of adaptation to a specific form, researchers often rely solely on large language models to evaluate such texts. Due to internal biases of individual language models, it is recommended to use multiple of them for combined evaluation, which directly increases costs of such evaluation. In this paper, a computationally efficient method (called PerQ) for evaluation of personalization quality (in terms of platform adaptation) of a given text (generated by a language model) is introduced, which is based on predicting the LLM judgments (i.e., a kind of metric distillation). A case study of comparison of generation capabilities of large and small language models shows the usability of the proposed metric in research, effectively reducing the waste of resources.

Paper 792 BrainMoE: A Brain-Inspired Modular Mixture-of-Experts Model for Multimodal Reasoning

Authors: Siyao Song, Yunhao Zhang, Jingyuan Sun, Yaping Zhang, Chengqing Zong

Abstract Reliable cross-modal reasoning requires Multimodal Large Language Model (MLLMs) with interpretable and controllable computation. However, existing models rely on data-driven architectures that entangle multimodal processing, obscuring the functional processes underlying their outputs. Inspired by the human brain's specialized and dynamically collaborative networks, we propose **BrainMoE**, a multimodal MoE model with six sensory and cognitive experts: Language, Logic, Social, World Knowledge, Vision, and Auditory. To support flexible sensory--cognitive coordination, we incorporates a layer-wise adaptive router that dynamically determines which experts are activated at each layer based on task-processing demands. Experiments on text, image, and audio reasoning benchmarks demonstrate that BrainMoE outperforms strong baselines by 1.59 and 2.2 points on text-only and multimodal benchmarks, while matching competitive external multimodal models. Routing and ablation analyses demonstrate clear expert specialization and causally important module-level contributions. Experiments on the cognitive benchmark further show that adaptive-routing BrainMoE aligns more closely with human behavioral patterns than baselines and static routing methods.

Paper 795 DALSA: Difficulty-Aware Long-to-Short LLM Reasoning with Model Merging

Authors: Dongwook Min, Jeonhwi Kang, Donghyeon Jeon, Inho Kang, Seung-Hoon Na

Abstract Recent advances in large language models(LLMs) have given rise to large reasoning models that achieve strong performance through explicit multi-step reasoning. However, models often generate long reasoning chains even for simple inputs. We present Difficulty-Aware Long-to-Short LLM Reasoning with Model Merging (DALSA), an inference-time framework that adaptively controls reasoning depth based on input difficulty. DALSA integrates (i) representation-based difficulty estimation, which predicts a soft failure-rate score from the prompt hidden state, and (ii) difficulty-aware model merging, which uses a quantile remapped score to modulate the interpolation coefficient between short- and long-reasoning models. DALSA trains a lightweight probe while keeping both LLMs frozen. On mathematical benchmarks with the evaluated Qwen pairs, the accuracy–length trade-off depends on scale: at 7B, DALSA preserves static ACM-TA average accuracy (66.60% vs. 66.61%) while reducing average response length by 22.35%; at 1.5B, it improves average accuracy from 56.75% to 58.24% with a 2.27% length in- crease. Additional merging baselines, routing-threshold and merge-range sweeps, and component ablations show sensitivity to the operating configuration. Domain-matched commonsense and Llama-family diagnostics do not establish broader efficiency gains, limiting the primary empirical claim to the evaluated Qwen mathematical setting.

Paper 799 From Preferences to Scores: Semi-Supervised Multi-Trait Essay Scoring with LLM Pairwise Preferences and Few Human Anchors

Authors: Takumi Shibata, Yuichi Miyamura

Abstract Multi-trait automated essay scoring (AES) provides trait-level scores as feedback but requires costly annotations across multiple traits to train accurate scoring models. To alleviate this annotation burden, we propose MTPAS (Multi-Trait Preference-to-Absolute Scoring), a semi-supervised framework that learns relative score orderings from LLM-generated pairwise preferences and calibrates them to absolute rubric scales using a small set of human-scored anchor essays in an end-to-end manner. To this end, MTPAS jointly optimizes three objectives: a Bradley–Terry pairwise preference loss, a regression loss on the anchor essays, and a Sinkhorn divergence-based distribution regularizer. The model adopts a bifactor architecture that decomposes essay representations into a shared general factor and trait-specific residuals, improving robustness against noisy LLM-generated preferences. We further introduce pairwise rubric compression, which reformulates the original score-level

rubrics into concise comparative criteria, reducing the inference cost of collecting LLM preferences. On ASAP and ASAP++, MTPAS achieves 91.4% of the fully supervised SoTA performance with only 30 anchor essays per prompt and further narrows the gap with 200 anchors. Under the same anchor budget, MTPAS outperforms conventional multi-trait AES baselines at every anchor size investigated in this study.

Paper 802 Orthogonal Style Injection: Mechanistic Persona Steering for Zero-Shot Cross-Lingual Subtitle Translation

Authors: Tarun Chintada, Kshetrimayum Boynao Singh, Asif Ekbal

Abstract In conversational and cinematic subtitle translation, large language models often exhibit stylistic collapse, defaulting to neutral registers that flatten character-specific personas. Prompt-based attempts to recover style frequently introduce instruction drift and semantic degradation. We propose Orthogonal Style Injection (OSI), a parameter-efficient, inference-time intervention that removes global stylistic bias through orthogonal projection at an intermediate layer and injects a character-specific style vector derived solely from English source dialogue. We evaluate OSI for English-to-Bengali, Hindi, Telugu, Tamil, Kannada, Chinese, Japanese, Spanish, French, German, and Italian subtitle translation, spanning 11 typologically diverse target languages. Results show that zero-shot stylistic transfer depends strongly on pretraining alignment. Llama-3.1-8B yields gains in several European languages but degrades on Indic languages, whereas Qwen-2.5-7B transfers style more robustly across Indic and East Asian languages. To probe sociolinguistic mechanisms, we curate a parallel English-Bengali cinematic dataset as a focused morphological stress test. Automatic metrics and human evaluation show that OSI enables meaningful honorific and register shifts while largely preserving semantic fidelity, revealing a dataset-dependent trade-off between stylistic fidelity and translation quality in highly colloquial, morphologically constrained settings.

Findings Papers

Paper 3 YUE-PUB: A Cantonese Pragmatic Understanding Benchmark

Authors: Chengyan Wu, Ziwei Gong, Jinzhi Chen, Mengmeng Su, Yun Xue, Bolei Ma

Abstract Pragmatic understanding remains a fundamental challenge in natural language processing, particularly for underrepresented languages such as Cantonese. To alleviate the scarcity of resources in this area, we introduce YUE-PUB, the first large-scale pragmatic understanding benchmark for Cantonese. Constructed from the PUB dataset through automatic translation with human verification for quality assurance, YUE-PUB covers three core pragmatic phenomena: Implicature, Presupposition, and Reference. We conduct systematic experiments across multiple baseline models to evaluate their performance under several cross-lingual transfer settings, compare models with and without pragmatic topic awareness, and further analyze their sensitivity to different pragmatic phenomena as well as their Chain-of-Thought (CoT) reasoning behavior. Experimental results show that current large language models still exhibit substantial limitations in Cantonese pragmatic understanding, underscoring the value of YUE-PUB as an important resource for advancing pragmatic reasoning research in low-resource languages.

Paper 4 Beyond Bilingual Transfer: Multilingual Code-Switching in Instruction Tuning

Authors: Shunta Asano, Jeonghun Baek, Toshihiko Yamasaki

Abstract Recent studies have shown that code-switching data (CSD), in which multiple languages are mixed within the same context, can improve cross-lingual transfer and multilingual alignment in large language models (LLMs). Recent studies on synthetic code-switching for LLMs have primarily focused on bilingual transfer between English and a target language, leaving

instruction-tuning settings that jointly mix three or more languages within each example largely unexplored. In this work, we investigate multilingual code-switching instruction tuning across four languages: English, Japanese, Korean, and Chinese. We evaluate multilingual understanding on Belebele and MMLU. Our experiments show that simple sentence-level multilingual CSD consistently improves the four-language average accuracy across all bilingual, trilingual, and quadrilingual settings for Gemma-2-2B, indicating that multilingual code-switching can be effective beyond bilingual transfer settings.

Paper 9 ECQL: An Executable Benchmark for Auditable Corpus Query Semantic Parsing

Authors: ~Yuxi_Zheng3, ~Yeqing_Liu1, ~Ziming_Zhou3, ~Xiaodong_Wang6, ~Yongwei_Zhang1, ~Jingsong_Yu1

Abstract Corpus Query Language (CQL) supports precise corpus analysis but requires expertise in annotations, regular expressions, structural scopes, and platform-specific syntax. Natural-language interfaces simplify access, but their answers can be difficult to inspect, reproduce, or revise. We study Text-to-CQL as executable semantic parsing, mapping intents to inspectable queries. We introduce ECQL, a Sketch-Engine-oriented benchmark of 3,572 natural language–CQL pairs comprising 811 tutorial-derived and 2,761 synthetic examples, with a default split of 3,215 training and 357 test examples. ECQL focuses on Sketch Engine normalization, CQL Diagnostic Profile (CQL-DP) annotations, controlled semantic contrasts, audit-oriented metadata, and generalization protocols, alongside an evaluation suite covering syntactic validity, exact match, canonical match, and execution match. The full 357-example headline test set is independently double-annotated and adjudicated, with 96.9% agreement for intent clarity, 94.7% agreement for CQL–intent alignment, and a 3.1% post-adjudication intent-match error rate. Experiments with prompting, retrieval-augmented prompting, syntax repair, DeepSeek-R1 reasoning prompting, a GPT-4o Agent/API alternative, and supervised fine-tuning reveal persistent difficulties with corpus-query semantics. The best fine-tuned model reaches 85.6% execution accuracy on the random split, but drops to 68.9% on template-disjoint evaluation. ECQL provides a reproducible testbed for auditable semantic parsing over specialized linguistic query languages.

Paper 18 Investigating Learner-Aware Design of LLM-Generated Educational Feedback

Authors: Momoka Furuhashi, Kouta Nakayama, Noboru Kawai, Takashi Kodama, Saku Sugawara, Kyosuke Takami

Abstract Although large language models (LLMs) show promise for generating educational feedback, it remains unclear how feedback should be designed (e.g., tone and coverage) to support answer revision and learner evaluations across learner profiles. We define six feedback designs for multiple-choice biology questions, including a baseline design and five variants with additional feedback elements, and conduct an empirical study with 321 high school students. We evaluate feedback using immediate revision performance and six subjective evaluation criteria, and analyze differences in subjective evaluations across learner profiles based on personality traits. Our results show that presenting task-relevant information clearly is associated with better immediate revision performance and is favorably evaluated across learner profiles, while we observe descriptive differences in evaluation patterns, particularly for informational novelty and affective framing. These findings support further investigation of personalized LLM feedback design.

Paper 20 Extended-SVD Low-Rank Adaptation: PLORA for Parameter-Efficient Fine-Tuning

Authors: RATAN KUMAR GULIPILLI

Abstract Parameter-efficient fine-tuning (PEFT) methods adapt pretrained models using low-rank updates applied to frozen weights. Recent SVD-based approaches initialize these updates using either dominant or lower-energy singular directions, implicitly treating different regions of the singular spectrum as separate adaptation subspaces. We propose PLORA (Procrustes Low-Rank Adaptation), an extended-SVD adaptation framework that treats lower-energy singular vectors near the top-r boundary as a geometric continuation of the principal pretrained subspace. PLORA constructs an extended frozen basis

from the top $(r \times m)$ singular directions and applies localized Procrustes alignment to produce a geometrically coherent adaptation coordinate system. Task-specific learning is then restricted to a compact trainable routing core parameterized by P, R, and Q.Across GLUE and vision benchmarks, PLORA achieves competitive performance under substantially reduced parameter budgets. In particular, PLORA-r8 matches standard LoRA performance on GLUE while using roughly half the trainable parameters, and PLORA-r16 achieves strong overall performance across multiple tasks. Beyond accuracy, PLORA systematically alters optimization behavior, frequently exhibiting lower seed variance, lower effective adaptation rank, and more concentrated subspace evolution during training.Ablations reveal that violating this geometric coherence via unaligned random bases severely degrades optimization dynamics, compounding ill-conditioning by up to $18.1\times$ and expanding weight distortion by $3\text{--}4.5\times$. These findings suggest that geometric alignment primarily serves to stabilize the optimization trajectory and concentrate update directions. Overall, PLORA provides both a competitive PEFT framework and a controlled setting for studying how geometric constraints shape optimization dynamics in low-rank adaptation systems.

Paper 22 Improving the Evaluation of Biomedical Question Answering with Few Human Preference Signals

Authors: Shakila Mahjabin Tonni, Xiang Dai, Sarvnaz Karimi, Xiaomei Li, Samuel C. Andrew, Stephen Wan

Abstract Retrieval-Augmented Generation (RAG) systems are increasingly used for biomedical Question Answering (QA). However, evaluating their effectiveness remains challenging due to the lack of reliable reference resources and the frequent misalignment between automatic evaluation metrics and human judgments. One practical question is how evaluation results should be used to guide development decisions, especially when they correlate weakly with human judgments and may even lead to different conclusions. We explore how to use a small amount of human preference signals to calibrate direct assessment evaluation metrics, and show that calibrated metrics consistently outperform baselines that either treat all metrics equally or rely on a single best metric across multiple datasets.

Paper 23 Beyond Factual Accuracy: A Benchmark for Evaluating Perspective Alignment in LLMs

Authors: Sunghoon Kim, Sangyeop Kim

Abstract Large Language Models often generate factually correct responses that fail to maintain user-requested perspectives, blending multiple perspectives and forcing users to filter through unwanted information—a perspective control failure that factual accuracy metrics cannot detect. We introduce PerspectiveBench, a benchmark quantifying perspective alignment by constructing scenarios where multiple valid reasoning paths coexist, and propose the Perspective Alignment Score (PAS) to measure adherence to specified perspectives. Experiments reveal that state-of-the-art models successfully include target content but fail to exclude non-requested perspectives. Providing documents through RAG improves factual grounding but does not resolve perspective drift, as models prioritize comprehensive coverage over strict user constraints. Code and data are available at https://github.com/sunghoon014/PerspectiveBench.

Paper 24 Multi-Signal Interpretability Identifies Task-Critical Circuits for Efficient Model Compression

Authors: Srikar Vardhan Mangadoddi

Abstract Post-training quantization methods such as GPTQ and AWQ minimize reconstruction error without asking which weights matter for the task. We combine six interpretability signals to select a sparse skeleton of task-critical MLP neurons that is kept at 16-bit while the rest of the model is quantized. On Gemma-2-2B fine-tuned for SQL generation, the 0.54% skeleton with 3-bit MLPs and 8-bit attention retains 100% of baseline-correct outputs at 4.03x compression, versus 75.2% for uniform 4-bit. On Llama-3-8B fine-tuned for Cypher generation, where uniform 4-bit loses 23 points, the 0.46% skeleton recovers 14.4 of them and a random skeleton of equal size recovers none. Three signals are individually effective and three are not, and the

effective set differs between models, so the consensus rule provides robustness no single signal does. Selection matters where uniform quantization leaves headroom, which suggests measuring headroom, selecting by consensus, and protecting as a recipe for larger models. GPTQ error compensation degrades 3-bit layouts by 16 to 18 points regardless of protection.

Paper 27 Multilinguality after Quantization: Low-Bits, Calibration Languages, Mixed-Precisions, and Cross-Linguality

Authors: Muhammad Rafif Priyo Waskito, Alfan Farizki Wicaksono, Alham Fikri Aji, Haryo Akbarianto Wibowo

Abstract There are inconsistent conclusions about the impact of calibration data on Post-Training Quantization (PTQ) across studies, specifically regarding the factors that make the method effective, such as language that we explore in this work. We investigate this question by evaluating \texttt{GPTQ} and its mixed-precision variants across three low-bit settings, using language-targeted calibration sets across five language families. By evaluating performance across both perplexity and downstream tasks, we find that selecting an appropriate calibration language is generally beneficial, though model sensitivity varies with bit-width, target language, and task difficulty. We further show that mixed-precision techniques expand the representational capacity of low-bit models, occasionally allowing 2-bit models to approach 3-bit performance when paired with language-targeted calibration, while increased effective-bits remains small. Furthermore, we explore the correlation between these factors and observe that the calibration set sensitivity is influenced by different combinations of factors for each controlled variable, which can be explained by the downstream task results. Finally, while the benefit of language-targeted calibration transfers to sibling languages within the same family, the magnitude of this transfer does not reliably predicted by their closeness. Together, these findings show calibration data selection as a deliberate design choice in multilingual PTQ rather than an incidental one. Our code and result data can be accessed in https://github.com/rafifpriyo/multilingual-calibration-study.

Paper 32 A Systematic Analysis of Linguistic Features in AI-Generated English Text Detection Across Domains and Models

Authors: Yassir El Attar, Esra Dönmez, Maximilian Maurer, Agnieszka Falenska

Abstract Interpretable linguistic features offer a promising approach for explaining *why* a given text appears machine-generated, particularly for non-expert users. However, existing findings on which features reliably indicate LLM-generated text remain fragmented across feature sets, models, and text domains. To address this gap, we conduct a large-scale empirical study assessing the robustness of linguistic signals for characterizing AI-generated text. Our analysis covers 284 interpretable linguistic features across outputs from 30 LLMs and 10 text domains under cross-model and cross-domain generalization settings. We show that classifiers based solely on linguistic features can reliably distinguish AI-generated from human-written text. However, many previously proposed indicators prove strongly context-dependent, with the exception of measures of lexical richness, which remain robust signals across model families and text domains. These results demonstrate which linguistic signals generalize across contexts and provide a foundation for more reliable, interpretable analyses of AI-generated language.

Paper 37 Can LLMs Write Reliable Rubrics? A Meta-Evaluation for Experiment Reproduction

Authors: Hanhua Hong, Yizhi LI, Jiaoyan Chen, Luu Gia Huy, Sophia Ananiadou, Jung-jae Kim, Chenghua Lin

Abstract Rubric-based evaluation is a promising approach for assessing open-ended outputs from LLM-based research agents, particularly in paper reproduction, where direct paper-to-repository comparison is prone to hallucinations. However, constructing paper-specific rubrics requires substantial expert effort, limiting the scalability of benchmarks such as PaperBench. In this work, we present, to our knowledge, the first systematic meta-evaluation of LLM-generated rubrics for paper reproduction. We

reformulate rubrics into a checklist-style format and evaluate four generation settings across two backbone models. We meta-evaluate generated rubrics intrinsically by semantic similarity and extrinsically by score alignment with ground-truth rubrics. Our results show that the augmented settings substantially improves downstream evaluation alignment, with the strongest setting approaching the human baseline, while intrinsic gains are more modest. Further analyses reveal that LLM-generated rubrics are often overly fine-grained, biased toward high scores, and less adaptive to paper domains, highlighting both the promise and limitations.

Paper 38 Are Language Models Script-Aware?

Authors: David Kletz, Sandra Mitrovic, Ljiljana Dolamic, Fabio Rinaldi

Abstract Language models frequently generate outputs in unintended languages or scripts—a phenomenon known as off-target generation. While existing research has focused on language selection, the dimension of script knowledge remains understudied: before any linguistic understanding can occur, users must recognize the graphic symbols in a model's response. We investigate whether Small and Large Language Models (SLMs and LLMs) possess script knowledge by testing them on multi-scriptic languages. Through two complementary experiments, we evaluate whether models (1) adapt their output script to match the input, and (2) follow explicit instructions to generate text in a specified script. The models we tested demonstrate substantial script knowledge: they all achieve a near-perfect Latin script fidelity (more than 98%), and follow script instructions with high frequency. Nevertheless, we notice differences between LLMs and SLMs, with higher scores for LLMs including for non-standard script combinations.

Paper 41 PAL-UI: Planning with Active Look-back for Vision-Based GUI Agents

Authors: Zikang Liu, Junyi Li, Xin Zhao, Dawei Gao, Yaliang Li, Ji-Rong Wen

Abstract Graphical User Interface (GUI) agents powered by Multimodal Large Language Models (MLLMs) promise human-like interaction with software applications, yet long-horizon tasks remain challenging due to memory limitations. Existing approaches either truncate history or rely on simple textual summaries, which risk losing critical information when past visual details become necessary for future decisions. In this paper, we propose \textbf{PAL-UI} (\textbf{P}lanning with \textbf{A}ctive \textbf{L}ook-back), a novel framework that enables GUI agents to adaptively retrieve past observations when required. PAL-UI combines a dual-level summary agent, capturing both observation-level cues and action-level outcomes, with a dedicated retrieval tool that allows the agent to recall specific historical screenshots during planning. We curate a step-level instruction dataset of 8.6K samples from mobile GUI navigation trajectories and train \textbf{PAL-UI-3B} and \textbf{PAL-UI-7B} models based on Qwen2.5-VL. Extensive experiments show that PAL-UI significantly outperforms baseline models and prior methods in mobile GUI tasks, even under data-efficient settings. Moreover, PAL-UI exhibits strong cross-domain generalization, achieving notable improvements in web navigation without additional training. Our work highlights the potential of active memory retrieval for long-horizon planning capabilities of vision-based GUI agents.

Paper 43 A Character-Level Neural Approach to Sinhala Sandhi Splitting

Authors: Yasas Ekanayaka, Deshan Sumanathilaka

Abstract Sinhala Sandhi splitting recovers the constituent words or morphemes hidden inside a phonologically merged surface form. The task is important for Sinhala NLP because Sandhi obscures lexical boundaries, but no prior published work has established a neural benchmark for Sinhala Sandhi splitting. We present a character-level sequence-to-sequence study based on SandhiLex, using native Sinhala Unicode input and evaluating recurrent encoder-decoder models for affixational and more complex lexicalized, derivational, and etymological Sandhi. A bidirectional LSTM encoder with a unidirectional LSTM decoder

yields the best results on the hard subset, achieving 68.40% exact-match accuracy and 82.08% character-level accuracy. On the full affixational subset, the same model achieves 94.00% exact-match accuracy. Ablations show that bidirectional encoding is the largest contributor to performance, while native Sinhala script improves exact match accuracy over romanized input. Qualitative analysis indicates that many errors are near misses involving boundary adjacent characters or plausible but incorrect phonological substitutions. These results establish an empirical baseline for Sinhala Sandhi splitting and identify data scale, Sandhi type conditioning, and attention-based decoding as the main directions for future work.

Paper 46 WAON: A Large-Scale Japanese Image-Text Dataset for Cultural Adaptation in Contrastive Vision-Language Models

Authors: Issa Sugiura, Shuhei Kurita, Yusuke Oda, Daisuke Kawahara, Yasuo Okabe, Naoaki Okazaki

Abstract Contrastive vision-language models have achieved remarkable progress through large-scale pretraining. Recent work has shown that removing English-only caption filters and pretraining on global data is effective for improving multicultural performance. We study whether such global pretraining is sufficient for culture-specific understanding, or whether further adaptation with natively sourced data can boost performance beyond what global pretraining alone achieves. To enable this investigation, we present WAON, the largest publicly available native Japanese image-text dataset constructed from native Japanese web content in Common Crawl, containing approximately 155 million examples. We also introduce WAON-Bench, a manually curated Japanese cultural benchmark spanning 374 classes. Through comparative fine-tuning experiments on multiple Japanese image-text datasets, we observe that models fine-tuned on WAON consistently achieve stronger performance on Japanese cultural benchmarks than those fine-tuned on English-to-Japanese translated data. Controlled experiments at matched scale, filtering, and training budget across two model families further indicate that native web origin is the primary driver of this gain. We release our dataset, benchmark, model, and code.

Paper 49 Grounded Revision vs. Prior Injection: Probing Retrieval-Augmented Patent Claim Amendment

Authors: Josepha Michiko Leo, Hyun-seok Min, Yehoon Jang, Irvan Zidny, Jin-Woo Chung, Sungchul Choi

Abstract Retrieval-augmented generation is widely used in professional writing, yet whether retrieval grounds revision or merely injects templates is rarely tested where "correct" has a definable meaning. Patent claim amendment supplies that signal: the examiner names the attacked limitation and cites prior art, providing per-case ground truth. We release three artifacts: (i) a corpus of 7,385 USPTO prosecution cases with XML-aligned pre/post claims, rejection, and cited prior art; (ii) a seven-probe battery comparing random and structural-match retrieval as two policies under a fixed prompt scaffold; (iii) a deterministic five-channel metric (C1-C3 and C5 in main, C4 supplementary) requiring no LLM evaluation. Across 9,600 pre-registered calls on four frontier LLMs (Claude Sonnet 4, Claude Haiku 4.5, GPT-5.4, GPT-4o-mini), no tested model exhibits detectable classical prior-injection behavior; retrieval effects are small and direction-inconsistent between random and structural retrieval, and the null is unchanged under a dense (semantic) retriever, across retrieval depths k in {1,3,5,10}, and under a paraphrase-sensitive grounding metric. Revision locality reveals a model-specific difference that the template channel misses. The four-cell taxonomy, which we treat as exploratory, leaves the prior-injector cell unoccupied.

Paper 53 PLAUSIEVE:A Distribution Based Framework for Measuring Model Belief Beyond Logits

Authors: Arghodeep Nandi, Tanmoy Chakraborty

Abstract Prompting has transformed large language models into general-purpose decision systems, where token-level probabilities are often treated as indicators of model preference and correctness. However, directly interpreting these

probabilities is challenging due to calibration disparities, tokenizer variability, and inconsistent output distributions across models, limiting their statistical comparability. This raises a fundamental question: *when can a model be said to be confident in its decision or preference?* Addressing this question remains an open and important area of research. We introduce **PLAUSIEVE**, a novel distribution-based statistical framework that estimates model preference through a forward pass without token generation. **PLAUSIEVE** provides a calibration-robust and model-independent mechanism for quantifying preference without relying on fixed thresholds or individual tokens. We apply **PLAUSIEVE** in estimating model belief, across two prompt settings (direct and epistemic), two confidence levels, and five models from four families to study prompt–architecture interactions. Our analysis shows significant improvements under epistemic prompting, with nearly 100% gains in certain model combinations. Notably, **Llama-3.2-3B** outperforms the larger **Gemma-7B** under epistemic prompting in terms of both decisiveness and veracity, demonstrating that model architecture plays a more significant role than scale in shaping belief formation.

Paper 62 Thinking Silently for Patent Claim Generation

Authors: Lekang Jiang, Wenjun Sun, Stefan Goetz

Abstract This work investigates whether Chain-of-Thought (CoT) prompting benefits patent claim generation. We demonstrate a counter-intuitive but important empirical finding: implicit CoT, where reasoning is kept internal rather than explicitly verbalized, consistently outperforms explicit CoT. Through systematic analysis, we show that explicit CoT can introduce an unnecessary information bottleneck for claim generation. Verbalized reasoning may compromise the quality of final outputs through three specific mechanisms: abstraction of critical details, disruption of internalized generation patterns, and cascading error propagation. Our findings provide new insights into structured text generation and CoT applications.

Paper 63 Judging the judge: Can LLMs track attitude-holders in predicates of personal taste?

Authors: Yan Cong

Abstract When LLMs annotate subjective content such as predicates of personal taste (PPTs), e.g., "Eliza tasted the cookie. It tasted delicious", a consequential but underexplored question is whose perspective they adopt: the narrator's (default) or Eliza's (shift toward the character)? Humans tend to attribute "delicious" to Eliza because tasting implies first-hand experience. We introduce GRANI (Graded Response Analysis for Narrative Implications), a protocol for evaluating LLMs' graded understanding of such phenomenon. Applying GRANI to PPTs, we test whether text-trained LLMs reproduce human perspective shifts across sensory modalities in English and Chinese. We found that instruction-tuned models partially capture the human pattern, in which taste yields the strongest shift toward the character, vision the weakest, and smell an intermediate effect, while reasoning-distilled models are less consistent. Moreover, some models encode modality distinctions internally even when their behavioral responses do not. These findings suggest that subjectivity inference can emerge from text exposure alone, while highlighting a reliability concern: LLMs may not consistently adopt the intended perspective-holder when evaluating subjective content.

Paper 66 Useful Diversity or Hallucinated Variety? Persona-Conditioned Synthetic Query Generation for E-Commerce Search

Authors: Filippos Ventirozos, Tom Armitage, Mahmoud Oshagh, Peter A. Appleby, Matthew Shardlow

Abstract Synthetic supervision can bootstrap semantic parsers in new search domains, but prompt-induced diversity is useful only when generations are faithful to the structured intent. We explore this for cold-start e-commerce search: 118,098 synthetic pairs produced by 6 LLMs with 9 persona-conditioned prompt families on a 47-field UK automotive schema, against 1,214 human-annotated and 15,304 production queries. We show that the IPIP personality bundle improves parser F1 in all four

matched contrasts (Holm-corrected p < .001), and all three bundles have a positive main effect over the complete 2^3 grid. Syntax conditioning is the axis that moves generations towards real traffic. It also induces lexical diversity. Alignment sKL, the divergence between von Mises-Fisher distributions fitted to synthetic and human-annotated queries, is an unlabelled proxy for parser F1 (rho = -0.917 on intent-elicited human queries, lower is closer). An LLM committee, checked against a blinded author pass, triages the 927 queries missed by every parser and labels 37% as hallucinated. Dropped constraints dominate. The four syntax-bearing families have the largest hallucinated shares, 38 to 42%, while IPIP alone has the lowest at 28%.

Paper 67 Think When Unsure: Leveraging Model Confidence to Decide When to Use Chain-of-Thought

Authors: Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, Nikolaos Aletras

Abstract Chain-of-thought (CoT) prompting improves performance on complex reasoning tasks, but often triggers unnecessary computation on simpler queries. A natural approach is to use the model's own confidence as a training-free signal to gate this decision, answering directly when confident and invoking CoT only when uncertain. Across four training-free confidence signals, three model families, and seven reasoning benchmarks, we find that the effectiveness of confidence-gated CoT varies, and it is unclear how to predict whether it will work for a new model. We address this with two key contributions. First, we identify that gating fails in the following circumstances: signal saturation, where confidence scores collapse to a narrow range and lose discriminative power; and selection bias, when confidence-gated routing sends the longest queries to CoT. Second, we propose a novel Pareto-optimal threshold selection method. This has the dual ability to preserve accuracy while reducing costs where the confidence signal is strong, and maintains accuracy where it is weak. Where the confidence signal is strong, gating reduces CoT usage by 30% without sacrificing accuracy. These findings provide practical guidance for evaluating and deploying models that selectively use CoT.

Paper 70 Improving Multimodal Contrastive Learning of Sentence Embeddings with Object-Phrase Alignment

Authors: Kaiyan Zhao, Zhongtao Miao, Yoshimasa Tsuruoka

Abstract Multimodal sentence embedding models typically leverage image-caption pairs in addition to textual data during training. However, such pairs often contain noise, including redundant or irrelevant information on either the image or caption side. To mitigate this issue, we propose MCSEO, a method that enhances multimodal sentence embeddings by incorporating fine-grained object-phrase alignment alongside traditional image-caption alignment. Specifically, MCSEO utilizes existing detection and segmentation models to extract accurate object-phrase pairs, which are then used to optimize a proposed contrastive learning objective tailored to object-phrase correspondence. Experimental results on semantic textual similarity (STS) tasks across different backbone models demonstrate that MCSEO outperforms image-caption-only baselines on average, and that it is complementary to existing methods: combining MCSEO with KDMCSE yields further gains. A random-pairing control and a manual evaluation of the extracted pairs additionally indicate that the improvements stem from correct object-phrase alignment.

Paper 72 CIL: Causal-Intervention and LLM-Synergized Multimodal Fake News Detection

Authors: HONGJIE WANG, Hongyu Chen, Peilin Jiang

Abstract Multimodal fake news detection requires models to reason over both textual and visual evidence while remaining robust to distribution shifts in social media graphs. Existing approaches often concatenate modality-specific representations or rely on post-hoc fusion, which can leave image and text features in misaligned semantic spaces. Graph-based methods further introduce structural bias: node popularity, degree, or connectivity may become shortcut signals that correlate with labels in the training distribution but do not causally determine veracity. We present CIL, a causal-intervention and LLM-synergized framework for multimodal fake news detection. CIL first uses CLIP to obtain aligned image and text representations and constructs a multi-

view semantic graph from cross-modal, intra-modal, and fused similarities. It then uses a large language model to produce high-confidence pseudo labels, which are injected as weak supervisory anchors. Finally, CIL mitigates degree-induced structural confounding with inverse propensity weighting and enforces counterfactual consistency under edge-drop interventions. Experiments on Twitter, PHEME, and Weibo suggest that each component contributes to performance, with the largest gains coming from aligned multimodal representations and causal debiasing.

Paper 74 Does Continued Pretraining on a Learner Corpus Improve Automated Essay Scoring on English Proficiency Tests? Evidence from EFCAMDAT

Authors: Duy Anh Nguyen

Abstract Automated Essay Scoring (AES) for English proficiency assessment increasingly relies on pretrained transformer models, yet these models are typically trained on general-domain English and may under-represent second-language learner writing. This study investigates whether domain-adaptive continued pretraining (DAPT) on a learner-writing corpus improves transformer-based AES for English proficiency assessment. We perform DAPT on BERT, RoBERTa, and DistilBERT using the EFCAMDAT corpus, then compare the adapted models with their original checkpoints on two English proficiency test datasets, FCE and IELTS, in both in-domain scoring and few-shot cross-dataset transfer. Full-corpus DAPT produces mixed effects across models, datasets, and metrics. Subsequent lexical and syntactic analyses suggest mismatches between EFCAMDAT and the downstream datasets in proficiency level, genre, and communicative purpose. We therefore repeat DAPT using proficiency-specific EFCAMDAT subsets across all three encoder architectures. Proficiency-specific DAPT frequently outperforms full-corpus DAPT and, in some settings, even the non-adapted baseline. Overall, continued pretraining on learner writing can improve in-domain AES, but its benefits depend on both the proficiency composition of the pretraining data and the underlying encoder architecture, and do not consistently extend to cross-test transfer.

Paper 80 MuSe-MI: A Framework for Assessing Multi-Session Motivational Interviewing with AI-Generated Patients

Authors: ~Topaz_Freizeit1, ~Moreah_Zisquit1, ~Kai_Golan_Hashiloni1, ~Chaitze_Tavens1, ~Doron_Friedman1, ~Kfir_Bar2

Abstract Evaluating conversational therapist competence in Motivational Interviewing (MI) requires multi-session assessment, yet current benchmarks rely on simulated patients that fail to model progression from session to session. We present MuSe-MI (Multiple Sessions Motivational Interviewing), a simulated patient framework grounded in Motivational Enhancement Therapy (MET), a multi-session clinical protocol of MI. MuSe-MI regulates dialogue generation through four evolving internal states from the MET clinical manual (anger, readiness to change, problem recognition, and self-efficacy), paired with an inter-session life event simulator and structured memory. Blind evaluations by licensed clinical psychologists confirm that patient dialogue reflects these internal states with high inter-rater agreement. Experiments show that MuSe-MI reliably distinguishes conversational therapist MI adherence levels based on patient language progression, establishing an automated multi-session evaluation benchmark.

Paper 82 Do LLMs Understand Context? A Knowledge Graph-Based Evaluation Framework

Authors: Subavarshana Arumugam, Mamta Nallaretnam, Kithuni Wickramasinghe, Chamath Gunapala, pragatheeswaran vipulanandan, Kamal Premaratne, Uthayasanker Thayasivam

Abstract While large language models (LLMs) have achieved remarkable linguistic capabilities, a profound question lingers at their core: do these models truly comprehend context or simply excel at pattern matching on an unprecedented scale? Contextual

understanding in LLMs refers to the ability to correctly extract relevant information from a given context, integrate it into a coherent internal representation, and reason over it to produce factually consistent and contextually grounded responses. However, traditional methods such as BiLingual Evaluation Understudy (BLEU) and perplexity simply measure surface-level performance. This reveals a critical gap in question answering (QA), where responses must be contextually grounded rather than simply being memorized associations. To fill this void, we propose a novel knowledge graph (KG) based evaluation framework for LLM contextual understanding in QA. Central to this is Semantic Structural Similarity for KGs (S3KG), a hybrid similarity measure combining structural and semantic signals into a single score. In addition, a diagnostic analysis framework is developed to identify and categorize reasoning errors at the triplet level, enabling fine-grained analysis of model failures. Together, across nine benchmarks, S3KG achieves F1 gains of up to $+7.6$ points over the strongest baseline and AUROC up to 0.973.

Paper 83 SelfGrader: LLM Jailbreak Detection via Anchored Token-Level Logits

Authors: Zikai Zhang, Rui Hu, Olivera Kotevska, Jiahao Xu

Abstract Large Language Models (LLMs) are powerful tools for answering user queries, yet they remain highly vulnerable to jailbreak attacks. Existing guardrail methods typically rely on internal features or textual responses to detect malicious queries, which either introduce substantial latency or suffer from randomness in text generation. To overcome these limitations, we propose \textbf{SelfGrader}, a lightweight guardrail method that formulates jailbreak detection as a numerical grading problem using anchored token-level logits. Specifically, SelfGrader evaluates the safety of a user query within a compact set of numerical tokens (NTs) (e.g., 0–9) and interprets their logit distribution as an internal safety signal. To align these signals with the target safety rubric, SelfGrader constructs Probably Approximately Correct-guided ICL anchor examples and introduces a dual-perspective scoring rule that considers both the maliciousness and benignness of the query, yielding a stable and interpretable score that reflects harmfulness and reduces the false positive rate simultaneously. Extensive experiments across diverse jailbreak benchmarks, adaptive attacks, benign prompt benchmarks, multiple LLMs, and state-of-the-art guardrail baselines demonstrate that SelfGrader achieves strong robustness with low false positive rates, memory overhead, and latency. \textcolor{red}{Warning: Some content generated by LLMs may be offensive.}

Paper 84 Self-Routing RAG: Binding Selective Retrieval with Knowledge Verbalization

Authors: ~Di_Wu14, ~Jia-Chen_Gu1, ~Kai-Wei_Chang1, ~Nanyun_Peng1

Abstract Selective retrieval aims to make retrieval-augmented generation (RAG) more efficient and reliable by skipping retrieval when an LLM's parametric knowledge suffices. Despite promising results, existing methods are constrained by a binary design choice: either retrieve from a single external source or skip retrieval and let the LLM directly produce the final answer. We argue that this fallback underestimates the model's knowledge and obscures the more general multi-source decision problem that arises in practical systems. We propose Self-Routing RAG (SR-RAG), which casts selective retrieval as knowledge source selection and treats the LLM itself as a first-class knowledge source. SR-RAG learns to select an appropriate knowledge source, optionally verbalize parametric knowledge, and answer using the selected source, all within a single left-to-right generation pass. SR-RAG further augments source selection by combining LLM-based uncertainty with a flexible external policy datastore to improve decision calibration. Across four benchmarks and three 7B-class LLMs, SR-RAG outperforms a strong selective retrieval baseline by 8.5%/2.1%/4.7% while performing 26%/40%/21% fewer retrievals, and it achieves favorable accuracy-latency trade-offs without dataset-specific threshold tuning.

Paper 85 Towards Low-Resource Manchu Text Generation: Corpus Construction and Multi-Granularity Augmentation

Authors: Meiqi Wang, Zhiyong Xu, Xiaoxin Sun, Ruixin Yu, Zhiyong Guo, Zifan Wang, Shuqi You, Siyu Liu, Xiayu Qin, Suhua Wang

Abstract Manchu is a typical low-resource agglutinative language, for which existing digital resources are largely limited to character recognition and basic annotation, while sentence-level generation remains underexplored. In this paper, we study constrained sentence-level Manchu continuation. We construct the first context-target corpus for this continuation task, consisting of 3214 examples collected from Manchu Stories in Translation, Common Manchu Dialogues, and other sources, covering both narrative and conversational genres. To improve robustness to rich suffixation, relatively flexible word order, and frequent OCR or transcription noise in Manchu, we propose a Multi-Granularity Data Augmentation (MGDA) strategy that expands training data through character-level noise injection, phrase rearrangement, and suffix perturbation, without relying on external dictionaries or translation models. Experiments on ByT5-small and Qwen2-1.5B show that MGDA substantially improves Manchu text generation performance: BLEU-4/ROUGE-L increases from 2.03/2.59 to 17.60/42.70 for ByT5, and from 1.98/6.42 to 14.12/41.49 for Qwen. Further ablation studies demonstrate the complementarity of the three augmentation granularities and suggest potential benefits of byte-level modeling in this experimental setting.

Paper 92 PeMA-KGQA: A Multi-Agent LLM Framework for Persian Knowledge Graph Question Answering

Authors: Kimia Hosseini, Mohammad Hosseini, Saeedeh Momtazi

Abstract Knowledge Graph Question Answering (KGQA) is particularly challenging for Persian, a complex and low-resource language with significant ambiguity. To address this, we introduce PeMA-KGQA (Persian Multi-Agent Knowledge Graph Question Answering), the first multi-agent LLM framework for Persian KGQA, arguing that an agent-based approach is essential for robust reasoning. Our system operates on FarsBase and is evaluated on both the ParsSimpleQA (one-hop) and PeCoQ (compositional) benchmarks. The framework employs a plan-and-execute paradigm where a dedicated Planner agent first decomposes a question into a structured executable plan using a toolbox of operators. Subsequent specialized agents—an Entity Grounder, a Relation Selector, and a Path Scorer—then execute this plan, effectively grounding mentions, selecting relations, and scoring reasoning paths to produce the final answer. This multi-agent decomposition provides a transparent and interpretable reasoning trace. PeMA-KGQA sets a new state-of-the-art on ParsSimpleQA with 21.14% improvement in Hits@5 and achieves 80.67% in Hits@5 on the complex multi-hop questions of PeCoQ.

Paper 93 Code-Prompt: Evaluating Code-LLMs for Natural Language Tasks via Code Aligned Prompts

Authors: Yuxuan Chen, Philippe Thomas, Robert Oehring, Felix Krenzien, Sebastian Möller, Roland Roller

Abstract Large language models pre-trained on code (Code-LLMs) have achieved remarkable performance on coding tasks. Despite their success of processing programming language, their application to natural language tasks has largely been limited to narrow tasks and ad-hoc model selection. In this paper, we present Code-Prompt, a comprehensive prompt-based framework to benchmark Code-LLMs over ten NLP tasks. We evaluate Code-Prompt on 13 Code-LLMs using code-aligned prompts and compare them with their parallel natural language LLMs (NL-LLMs) using natural language prompts. Our results show that Code-LLMs perform on par with NL-LLMs across a range of NLP tasks, while producing more format-consistent generation with less redundancy. Our cross-lingual experiments further indicate that knowledge in Code-LLMs is well shared among different programming languages, whereas performance across different human languages still varies significantly. All our code and data will be made publicly available at https://github.com/chen-yuxuan/code-prompt.

Paper 95 Evaluating Multi-Dimensional Generalization of Large Language Models in Temporal Extraction Tasks

Authors: FAHMID SHAHRIAR IQBAL, Ritam Dutt, Soumitra Das, Arnav Verma, Sagnik Ray Choudhury

Abstract Time and event expression extraction are fundamental temporal reasoning tasks, but the problem remains difficult due to annotation ambiguity, domain sensitivity, and unstable model behavior. Existing evaluations focus on in-domain performance, offering limited insight into reliability under distribution shifts. We evaluate multiple model configurations across families, architectures, and reasoning strategies over four dimensions of generalization, examining transfer from base performance, cross-dimensional correlations, and the effects of scale, architecture, and prompting. This provides a systematic study of how prompted LLMs generalize in time and event expression extraction tasks. We find that strong base-task performance generally predicts better generalization. However, this relationship weakens under substantial distribution shifts. Inductive prompting performs most consistently across domain shift, adversarial perturbations, compositionality, and length increase, while gains from scale, architecture, and deductive and abductive prompting strategies are uneven and dimension-specific. We conclude that LLM generalization in temporal extraction tasks cannot be predicted from any single dimension alone and cannot be reliably inferred from in-domain or single-dimension evaluations, highlighting the need for reasoning strategies that generalize across dimensions.

Paper 97 Mechanistically Understanding and Mitigating the Anchoring Effect in LLMs via Attention Head Intervention

Authors: Yoshiki Takenami, Yin Jou Huang, Yugo Murawaki, Koh Takeuchi, Chenhui Chu

Abstract Similarly to humans, LLMs are susceptible to the anchoring effect in price negotiations, which can lead to unfavorable or unintended outcomes. To ensure LLM safety in dialogue tasks such as negotiation, a mechanistic understanding and mitigation methods are essential. We propose four variants of a method to identify Anchoring Heads, which are the attention heads responsible for the anchoring effect, while preserving those essential for general tasks. Experimental results show that suppressing these heads mitigates the anchoring effect by over 16%, achieving complete mitigation in certain models without compromising general performance. Furthermore, the observation that suppressing random heads fails to mitigate the anchoring effect suggests that specific attention heads drive the anchoring effect by paying excessive attention to the anchor.

Paper 98 Beyond Top Words: MonoTM for Topic Modeling with Interpretable Monosemantic Features

Authors: Una Joh, Bei Yu

Abstract Topic models summarize large text corpora, but top-ranked words often provide only a limited representation of topic semantics. Sparse autoencoders (SAEs) offer a way to move beyond word-level descriptors by extracting interpretable features from dense representations, yet how feature interpretability relates to topic-inference quality remains unclear. We introduce MonoTM, an interpretable topic modeling framework that decouples these roles. Across three benchmark corpora, we show that document--topic mixture estimation and semantic interpretation favor different SAE configurations and feature subsets. MonoTM estimates mixtures from the full SAE bag-of-features representation and, with them fixed, learns topic descriptors over a separate vocabulary of corpus-grounded semantic features. This design preserves global topic structure while representing topics with semantic units more meaningful than individual words, making them more useful for downstream corpus analysis.

Paper 100 Measuring Iterative Temporal Reasoning with Time Puzzles

Authors: Zhengxiang Wang, Zeyu Dong

Abstract Tool use, such as web search, has become a standard capability even in freely available large language models (LLMs). However, existing benchmarks evaluate temporal reasoning mainly in static, non-tool-using settings, which poorly reflect how LLMs perform temporal reasoning in practice. We introduce Time Puzzles, a constraint-based date inference task for evaluating iterative temporal reasoning with tools. Each puzzle combines factual temporal anchors with (cross-cultural) calendar relations and may admit one or multiple valid dates. The puzzles are algorithmically generated, enabling controlled and continual evaluation. Across 13 LLMs, even the best model (GPT-5) achieves only 55.3% accuracy without tools, despite using easily searchable facts. While web search improves performance, models perform substantially better when constraints are rewritten with explicit dates, removing the need for factual lookup. These results reveal a gap in reliable tool use for iterative temporal reasoning.

Paper 101 Anthropogenic Regional Adaptation in Multimodal Vision-Language Model

Authors: Samuel Cahyawijaya, Peerat Limkonchotiwat, Tack Hwa Wong, Carlos Rafael Catalan, Manuel Antonio Rufino, Hitesh Laxmichand Patel, Amit Agarwal, Muhammad Reza Qorib, Vicky Feliren, Holy Lovenia, Aye Hninn Khine, Frederikus Hudi, David Anugraha, Alham Fikri Aji, Romrawin Chumpu, Viet Thanh Pham, Minghan Wang, Mohamed Fazli Mohamed Imam, Ruochen Zhang, Joseph Marvin Imperial, Khumaisa Nur'aini, Do Xuan Long, Musa Izzanardi Wijanarko, Joel Ruben Antony Moniz, Patrick Amadeus Irawan, Hanif Muhammad Zhafran, Isaiah Edri W. Flores, Salsabila Zahirah Pranida, Jun Kevin, Jostin Jerico Rosal, Patricia Nicole Monderin, Kun Kerdthaisong, Ahmad Mustafid, My Chiffon Nguyen, Natchapon Jongwiriyanurak, Siva Worajitwannakul, Haochen Li, Adrian Xuan Wei Lim, Bin Wang, Muhammad Ravi Shulthan Habibi, Lynnette Hui Xian Ng, Mithil Bangera, Yeshil Bangera, Priyaranjan Pattnayak, Dun Li Chan, Sherissa Caren Djuniwar, chochanmyeioo@gmail.com, Ming Shan Hee

Abstract While the field of vision-language (VL) has achieved remarkable success in integrating visual and textual information across multiple languages and domains, there is no dedicated framework for assessing human-centric alignment in vision-language systems. We offer two contributions to address this gap. First, we introduce Anthropogenic Regional Adaptation: a regional-adaptation framework that aims to optimize model relevance to specific regional contexts while ensuring the retention of global generalization capabilities. Second, we present a simple, but effective adaptation method named Geographical-generalization-made-easy (GG-EZ), which utilizes regional data filtering and model merging. Through comprehensive ablations on 3 VL architectures: large vision-language models (LVLM), text-to-image diffusion models, and vision-language embedding models, and a case study in Southeast Asia (SEA) regional adaptation, we demonstrate the importance of Anthropogenic Regional Adaptation and GG-EZ, showing 5-15% gains in cultural relevance metrics across SEA while maintaining over 98% of global performance and even occasionally surpassing it. Our findings establish Anthropogenic Regional Alignment as a foundational paradigm towards applicability of multimodal VL architectures in diverse regions and demonstrate a simple-yet-effective baseline method that optimizes regional value alignment while preserving global generalization.

Paper 103 Right for the Wrong Reasons: Investigating the Unrobust Nature of High Accuracy Models in Discourse Relation Prediction

Authors: Yi Fan, Wei Liu

Abstract Large language models (LLMs) can achieve strong performance on discourse relation prediction through fine-tuning. Yet it remains unclear whether such performance reflects a robust modelling of inter-sentential relations or a reliance on shallow lexical cues. This paper presents a behavioural audit of fine-tuned models across explicit and implicit settings. In explicit discourse relation prediction, we find that fine-tuned models rely heavily on connectives as categorical triggers, rather than integrating argument semantics. More surprisingly, in implicit discourse settings, we observe a severe performance degradation when masking semantically light function words from the arguments, suggesting that models may also exploit subtle surface-level cues beyond explicit connectives. These results suggest that high accuracy alone is insufficient to demonstrate robust discourse processing. To address this limitation, we propose a causally validated diagnostic framework that measures whether model predictions are grounded in linguistically meaningful evidence. Our metric provides a complementary perspective to standard accuracy-based evaluation and establishes a rigorous auditing toolkit for future work on discourse relation prediction.

Paper 104 RabakBench: Evaluating Safety Guardrails in Singapore's Multilingual Vernaculars

Authors: Gabriel Chua, Leanne Tan, Ziyu Ge, Roy Ka-Wei Lee

Abstract LLM safety guardrails often behave unevenly in localized multilingual settings, especially with code-mixed vernaculars and non-Latin scripts. We introduce RabakBench, a Singapore-localized multilingual safety benchmark for Singlish, Chinese, Malay, and Tamil, built through a semi-automated, human-validated Generate--Label--Translate pipeline. The benchmark contains 1,341 unique examples in four language versions, yielding 5,364 entries across six harm categories, with LLM-assisted labels validated against human consensus. Evaluating 13 commercial and open-source guardrails reveals substantial cross-language and system-level variation, highlighting localized and script-sensitive safety gaps in the Singaporean context. RabakBench provides both a diagnostic resource and a reproducible, human-in-the-loop framework for constructing localized safety benchmarks.

Paper 105 Understanding Knowledge Distillation in Post-Training: When It Helps and When It Fails

Authors: Xin Liu, Simin Ma, Shujian Liu, Song Wang, Sathish Reddy Indurthi, Haoyun Deng, Lu Wang, Kaiqiang Song

Abstract Large language models (LLMs) achieve strong performance across many tasks, but their high computational cost limits deployment in resource-constrained environments. Knowledge Distillation (KD) offers a practical solution by transferring knowledge from a teacher model of a larger size to a smaller student model. While prior work has mainly examined task-specific or small-scale settings, the post-training stage for building general instruction-following models has received limited attention. In this paper, we conduct a systematic study of KD in post-training using the large-scale Tulu 3 dataset. We find that KD outperforms supervised fine-tuning (SFT) in low-data regimes, but its advantage diminishes as more training data is added. Distilling from a stronger instruction-tuned teacher restores substantial gains even with abundant data, indicating that KD remains effective when the teacher provides knowledge that the student cannot easily acquire from the training data alone. We further study domain-specific, low-resource scenarios and propose a two-stage KD strategy that leverages synthetic teacher-labeled data followed by refinement on human annotations. This method consistently improves student performance, providing practical guidance for building compact models in data-scarce environments.

Paper 111 Two-Stage LLM-as-a-Judge Using Required and Bonus Criteria

Authors: ~Kohei_Tsunefuji1, ~Akihiro_Tamura3, ~Daichi_Nakatsuji1, ~Hiroaki_Oobayashi1, ~Mitsuo_Hamada1, ~Tsuneo_Kato1

Abstract This study proposes a new LLM-as-a-judge method for relative evaluation of LLM responses based on both required criteria (i.e., minimal conditions that responses must satisfy) and bonus criteria (i.e., desirable additional qualities). Recent LLM-as-a-judge methods have improved the stability and accuracy of relative evaluation through the use of automatically generated criteria. However, existing methods rely solely on the required criteria, which may limit their ability to distinguish closely matched responses. To address this issue, the proposed method first evaluates responses based on required criteria and, in the case of a tie, further evaluates them using bonus criteria, enabling more fine-grained comparisons. Experiments on the Writing, Roleplay, and Humanities categories of the MT-Bench dataset demonstrate that the proposed method achieves higher agreement with human judgments than existing methods.

Paper 114 Stylistic Evolution and LLM Neutrality in Singlish Language

Authors: ~Linus_Tze_En_Foo1, ~Weihan_Angela_Ng1, ~Wenkai_Li2, ~Lynnette_Hui_Xian_Ng1

Abstract Singlish is a creole rooted in Singapore's multilingual environment that continues to evolve alongside social and technological change. We examine diachronic stylistic change across a decade of informal digital messages and ask whether Large Language Models (LLMs) can generate temporally neutral outputs approximating the stable characteristics of the variety. Using lexical, pragmatic, psycholinguistic, and encoder-based features, we find that stylistic separability increases with temporal distance, driven primarily by structural features such as length and complexity. Evaluated against a null distribution baseline, most LLMs fail to achieve both authenticity and temporal neutrality simultaneously at their most authentic prompting configuration. This reveals a structural trade-off: models generating realistic Singlish tend to inherit its temporal biases, while temporally neutral models tend to produce inauthentic outputs. These findings position temporal neutrality as a diagnostic metric for assessing sociolectal grounding in LLMs.

Paper 117 Linguistic Cues in Political Debates: Lexical Choices and Evaluative Framing in GPT-Generated Commentary

Authors: Yu-Yu Chang, Chung-Chi Chen, Ming-Hung Wang

Abstract Political debates shape public opinion, but many readers increasingly encounter debate takeaways through AI-generated summaries and commentary rather than full broadcasts. This paper studies whether fine-grained lexical choices in debate transcripts can influence the evaluative framing of GPT-generated political commentary. We introduce controlled word-level interventions by replacing embedding-identified key lexical items with near-synonymous alternatives, allowing us to examine how minimally invasive lexical changes affect model-generated interpretation. Across U.S. presidential debate segments, we find that single-word substitutions can shift the stance of GPT-generated commentary, but these effects are probabilistic, context-dependent, and sensitive to the speaker's communicative goal. Substitutions are more likely to produce favorable stance shifts in critical contexts than in promotional ones, suggesting that GPT-generated commentary may amplify adversarial framing more readily than self-promotional framing. We further analyze which linguistic cue types are associated with substitution success, including sentiment direction, part of speech, and contextual suitability, and show that sentiment alone is insufficient to explain model responses. Finally, a small exploratory human evaluation suggests partial alignment between model-based stance shifts and reader preferences in promotional cases, while also revealing substantial subjectivity in human judgments. These findings highlight how candidates' word choices may be transformed when political discourse is filtered

through GPT-generated commentary, and caution against treating model-internal stance metrics as direct substitutes for large-scale human audience evaluation.

Paper 121 Reasoning under Vision: Understanding Visual-Spatial Cognition in Vision-Language Models for CAPTCHA

Authors: Python Song, Luke Tenyi Chang, Yun-Yun Tsai, Penghui Li, Junfeng Yang

Abstract Modern visual CAPTCHA has evolved from simple perception tests into real-world spatial reasoning challenges. It remains unclear whether state-of-the-art vision-language models can solve such tasks through their own reasoning ability. Existing benchmarks offer limited insight into this question as they often ignore intermediate reasoning steps, typically assess only final answers, and thus cannot know how models fail. To address this gap, we present CAPTCHA-X, the first unified evaluation framework for real-world CAPTCHA solving with reasoning steps, enabling automatic assessment of solving accuracy and reasoning quality. Using this framework, we show that reasoning substantially improves performance. Beyond evaluation, CAPTCHA-X exposes characteristic reasoning failure modes of vision-language models and reveals reasoning scaling laws that connect multiple dimensions of reasoning. Furthermore, evaluating pure vision generation models reveals an interesting phenomenon: generation models also fail at multi-step logical planning on images. These findings motivate our super-simple and reasoning-centric agent built solely on vision-language models, designed to identify and correct errors in intermediate reasoning steps before translating reasoning into executable actions.

Paper 125 VMMU: A Vietnamese Multitask Multimodal Understanding and Reasoning Benchmark

Authors: Vy Tuong Dang, An Vo, Emilio Villa-Cueva, Quang Tau, Duc Dm, Thamar Solorio, Daeyoung Kim

Abstract Vision language models are predominantly evaluated on English-centric multimodal benchmarks, leaving their behavior in other languages and cultures insufficiently understood. We introduce VMMU, a Vietnamese Multitask Multimodal Understanding and Reasoning benchmark of 2,548 multiple-choice questions across seven domains: Mathematics, Physics, Chemistry, Biology, Geography, Driving Test, and IQ Test. Every question requires jointly reasoning over Vietnamese text and non-text visual evidence (charts, diagrams, tables, traffic scenes), and is distributed as a single rendered page image. We evaluate 10 open-source and 5 closed-source VLMs on VMMU. Test-time compute improves performance from 57% to 78% accuracy, but it remains inferior to an expert baseline (99% accuracy). Extensive error analyses show that closed-source VLMs already achieve strong Vietnamese OCR performance, yet still struggle on VMMU. This suggests that the primary bottleneck is multimodal grounding and reasoning rather than text recognition or Vietnamese language understanding. Code and data are available at: https://vmmu-bench.github.io.

Paper 127 Mirrored Frames: LLM-Based Narrative Framing Analysis of Thai-Cambodian Border Conflict Coverage

Authors: Siwakorn Bubphasawan, Attapol Rutherford

Abstract Framing in news media shapes how events, actors, and causal responsibility are interpreted, yet computational framing analysis remains limited for lower-resource languages and fast-evolving international conflicts. In this work, we analyze narrative framing in news coverage of the 2025 Thailand–Cambodia border conflict across English, Thai, and Khmer news articles. To support benchmarking for this task, we manually annotate 51 conflict-related articles using a stakeholder-oriented Hero–Villain–Victim (HVV) framework and source attribution annotations. We benchmark proprietary and open-weight large language models (LLMs), compare original-language and translated input, evaluate salience-focused prompt variants, and test supervised cross-

dataset transfer. Original-language inference and explicit article-level salience instructions improve the best micro-averaged F1 from 0.66 to 0.73. Applying this improved pipeline across active-conflict and post-conflict periods suggests largely mirrored partisan framing patterns in Thai and Cambodian national media.

Paper 128 HiCAP: Hierarchical Cultural Adaptation via Personality for Dialogue Generation

Authors: Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Shoaib Jameel

Abstract Cultural adaptation in dialogue systems refers to the ability to generate responses that conform to linguistic norms and social expectations across different cultures. However, existing dialogue generation approaches pay limited attention to the role of personality in shaping cultural expression and communication style. We therefore propose HiCAP, a novel framework that integrates personality traits into culturally adaptive dialogue generation. HiCAP employs a personality-conditioned control mechanism that hierarchically modulates response style by first aligning it with the cultural tone and then refining linguistic expression and content according to individual personality attributes. To embed these factors, the model combines supervised fine-tuning and instruction-based tuning, supported by dynamic instruction templates and multi-agent demonstrations to capture diverse combinations of culture and personality. Experiments on CulturalPersonas and CPED show that HiCAP achieves the strongest overall performance across various dialogue generation metrics, demonstrating effective cultural adaptability and personality modelling.

Paper 129 PrivacyEvalClaw: A Tool-Augmented Deliberative Agent for Auditing Privacy Leakage in LLM Jailbreak Responses

Authors: ~Krishna_Kanth_Nakka2, ~Yujun_Wang2, ~Shijie_Zhou10, ~Tu_Nguyen2, ~Gabriele_Gelardi1, ~Xuebing_Zhou1

Abstract Privacy auditing of large language models (LLMs) increasingly relies on advanced jailbreak attacks to uncover unintended disclosures of personal information; the reliability of such audits, however, hinges on evaluators that can accurately determine whether a response constitutes privacy leakage. This evaluation remains challenging because disclosures may be indirect, contextual, partial, or obfuscated, and different evaluators can interpret the same response differently. We conduct a systematic study of guard models and LLM-as-Judges and find substantial disagreement among evaluators, particularly in ambiguous cases that require additional evidence or targeted analysis to resolve. Motivated by these findings, we introduce \textsc{PrivacyEvalClaw}, a training-free deliberative agent for evaluating privacy leakage in LLM jailbreak responses. Rather than relying on isolated single-pass judgments, PrivacyEvalClaw treats evaluators and analysis modules as tools, adaptively gathers evidence, investigates uncertain disclosures, and reconciles conflicting signals before producing a final judgment. Our primary experiments use pairwise LLM-as-Judge evaluation of complete trajectories, showing that PrivacyEvalClaw achieves favorable win rates against competing baselines across outcome, process, and overall quality, while also producing explicit reasoning traces that make the evaluation process more transparent and auditable.

Paper 130 Hidden Consensus: Preference-Validity Compression in Human Feedback

Authors: Dorcas Chia Ern Chua, Karen Myn Hui Lee, Jia Yue Tan, Zhen Xue Gue, Han Ying Lim, Norzalena Abdul Hamid, Azima Binti Azmi, Keat Mei Yeong, Aizat Izyani binti Mujab, Hafsah Noor Azam, Chee Guo Khoo, Chee Seng Chan

Abstract Standard RLHF pipelines often reduce heterogeneous human judgments into a single scalar reward target. We argue that this reduction can mis-measure alignment in structurally plural societies, where disagreement may reflect culturally, historically, linguistically, regionally, or normatively grounded interpretations rather than annotation noise. We call this failure Preference-Validity Compression, the collapse of multiple plural-valid response options into a single optimization target. Using

Malaysia as a diagnostic setting, we analyze RLHF-style feedback aggregation through preference events linking prompts, responses, and acceptability judgments across interpretive frames. Across 321 preference events from 20 participants and 107 trio-annotated prompts, 79% of prompts contain more than one majority-supported response that single-winner aggregation would discard, and apparent dominance gaps between top responses diminish when all majority-supported options are considered. Participants frequently select multiple acceptable responses, and discarded responses demonstrably reflect coherent local, practical, or cultural frames. These findings show that majority aggregation in this corpus measures $\arg\max$ acceptability rather than plural alignment. We treat this as a measurement-validity issue and argue that future alignment methods should satisfy Validity-Preserving Consistency, remaining stable across plural-valid interpretive frames rather than collapsing them into a single reward target.

Paper 131 EffComp: Efficient Prompt Compression via Hybrid Reinforcement & Supervised Learning

Authors: Romtham Tungsununtham, Piyawat Lertvittayakumjorn, Atiwong Suchato, Peerapon Vateekul

Abstract Prompt compression is aimed at reducing input prompt lengths to enable cheaper, faster, and potentially higher-quality LLM predictions. However, existing prompt compression methods are often limited by modest compression gains, a risk of hallucination, and/or high compression latency. We propose EffComp, an **Eff**icient prompt **Comp**ression framework for RAG-based question answering (QA). EffComp employs BERT-style document reranker and sentence selector models to allow fast extractive prompt compression at the sentence level. Its extractive nature prevents hallucinations in the compressed prompts. Additionally, the training process using a hybrid reinforcement and supervised learning approach is designed to optimize the compression ratio while preserving LLM accuracy. Experiments on various QA datasets demonstrate that EffComp outperforms state-of-the-art prompt compression methods achieving the best trade-off between compression ratio and accuracy with minimal latency. This makes it practical for real-world applications.

Paper 132 Replacing the Human in the Loop: Towards Automated Evaluation of LLM Mathematical Proof Generalization

Authors: ~Carolina_Dias-Alexiou1, ~Edison_Marrese-Taylor2, ~A._Solano1, ~Yutaka_Matsuo1

Abstract As we continue to extend their range of applications, the study of the generalization capabilities of large language models (LLM) has become an important topic. Previous works have recently proposed to study these abilities in the context of mathematical reasoning. This was done by asking models to reproduce proofs that they have most likely seen during training, but when key symbols are replaced. Since this evaluation relies heavily on human feedback, we study to what degree this evaluation can be automatically performed. To that end, we simulate the human evaluation proposed by previous work by interactively presenting the same questions to the LLMs as prompts. We analyze the degree of alignment between human and model answers via an accuracy and correlation analysis, as well as via inter-annotator agreement metrics. We test a selection of models, including ones with reasoning capabilities, and show LLMs currently have important limitations in this regard, which cannot be addressed by ICL.

Paper 134 Persuaded but Not Aligned: A Relapse Test for LLM Realignment under Adversarial Incentives

Authors: Roberta Viola, michele@dcc.ufmg.br, Adriano Veloso

Abstract Alignment evaluations typically measure model behavior in single-turn settings, leaving unclear whether apparent behavioral change persists once external pressure is removed. We investigate this question in a controlled multi-agent Among Us testbed where LLM agents initialized with deceptive goals are exposed to structured persuasion and later evaluated under

neutral prompts without explicit policy guidance. We find that persuasive dialogue can induce rapid cooperative behavior: most agents verbally accept the moral argument (69%), but only 46% sustain cooperation once external pressure is removed (a 23-point compliance gap). The gap is strongly asymmetric: when verbal acceptance fails to translate into durable behavior, it does so almost exclusively in the direction of superficial compliance. A cross-model ablation shows that susceptibility varies across models, and a reverse-direction experiment reveals that verbal acceptance overestimates alignment-favorable durability but accurately tracks alignment-adverse shift. These results motivate relapse-based protocols as a complement to single-turn evaluations of agents operating under adversarial incentives.

Paper 135 STEERLSN: Training-Free Representation Alignment through Language-Specific Neuron Subspace Interpolation

Authors: Minjun Kim, Dongyeon Kim, Hoyun Song, Jihyo Kim, KyungTae Lim

Abstract Recent advancements in Large Language Models (LLMs) have shown remarkable capabilities in multilingual contexts. However, previous studies suggest that multilingual LLMs primarily rely on the rich semantic knowledge of high-resource (i.e., English) representations to conduct tasks across languages. To mitigate the performance gap arising from resource imbalances, we propose \textit{Steering via Language-Specific Neurons}, a training-free method that aligns non-English representations with the English manifold intervening on language-specific neuron activations via linear interpolation toward English activation prototypes. Experimental results show that the proposed method generally improves downstream task performance with minimal computational cost, yielding performance gains for low-resource scenarios where limited data often constrains knowledge coverage. Our findings suggest that the fundamental limitations of low-resource languages can be effectively alleviated by leveraging the robust representation space of high-resource languages, offering a promising direction for more balanced multilingual LLMs.

Paper 141 BreakFun: Jailbreaking LLMs via Object Instantiation under Simulated Code Execution

Authors: Amirkia Rafiei Oskooei, Mehmet Aktas

Abstract Large Language Models (LLMs) are widely used because they process structures, syntax and code well, but this same ability also makes them paradoxically vulnerable. We introduce BreakFun, a jailbreak method that frames a harmful request as code-execution simulation. The prompt gives the model a benign Python class definition, the "Trojan Schema", and asks what that code would print if it ran. To answer, the model must create an object from the class and invent a value for each field, and the adversarial field names steer those values toward the attacker's goal. The harmful content therefore appears as a side-effect of the simulated instantiation, not as a direct answer. A three-part prompt carries the attack: an innocent frame, the Trojan Schema, and a Chain-of-Thought distraction. On JailbreakBench, BreakFun reaches an average attack success rate of 89% across 13 open-weight and commercial models (open-weight ~98%, API systems ~78%), and 100% on several. An ablation identifies the Trojan Schema as the strongest contributing component. As a counter-ablation probe we introduce Adversarial Prompt Deconstruction (APD), in which a second LLM writes out all readable text in the prompt before it judges safety. APD improves detection across three model families, and a probe-specific ablation shows that the transcription step gives most of the gain. Our findings show that LLM safety must account for content that a simulation task produces as a side-effect, not only for content that a user requests directly.

Paper 143 Beyond the Login: Continuous Behavioral Authorship Verification for LLM Agent Sessions

Authors: Muskan Sharma, Seba Susan

Abstract LLM agents now operate as a privileged application layer, with access to mail, files and MCP-connected tools, yet they have no built-in way to verify that the person typing into an already-signed-in session is still the legitimate user. This identity gap aligns with OWASP LLM06 (Excessive Agency), Agentic ASI03 (Identity and Privilege Abuse), and NIST guidance on AI risk management, but whether it is closable from user text alone is largely open. We build a profile-query authorship verification benchmark on WildChat-1M and PAN-20 regrouped per author, with author-disjoint 60/20/20 splits across three seeds. We train contrastive bi-encoders with triplet margin loss on Distil- BERT, BERT, MiniLM-L6 and ELECTRA, under 100% random imposters or a 50/30/20 mix of random, semi-hard and hard imposters mined by TF-IDF. Models trained on random imposters look strong (ROC-AUC above 0.95) but drop by around 10 points on WildChat against topic-similar imposters, suggesting they rely on topic and vocabulary rather than writing style only. Our contribution is to expose and quantify this failure mode and to provide a reproducible hardness-stratified benchmark; applying established hardness-aware training as a controlled mitigation closes most of the gap, lifting TAR@FAR=1% on WildChat Hard from around 6–7% to 20–22% for BERT and DistilBERT with median messages-to-detect of one, and the same direction holds under a secondary SBERT miner. We release the benchmark construction and evaluation code.

Paper 146 Synthesizing Dialogues from Social Media Data for Mental Health Support: A Case Study on Gambling Disorders

Authors: Zhouqing Zhang, Lis Pereira, Wan Jou She, Benjamin Adams, tham.piumsomboon@canterbury.ac.nz, Kongmeng Liew

Abstract Recent advances in artificial intelligence have seen natural language processing (NLP) applications used increasingly for mental health support, such as question-answering systems, conversational agents, and chatbots. However, these applications often require high-quality dialogue datasets, which are scarce for certain under-resourced mental health conditions like gambling disorders. Privacy concerns and the sensitive nature of mental health issues make real-world data collection difficult and time-consuming. We propose a framework that uses publicly available social media posts as seeds to generate empathy-driven multi-turn dialogues between help-seekers and supporters in mental health conditions that have been under-researched in NLP. Through a case study on gambling disorder, we demonstrate the effectiveness of our approach in producing higher-quality dialogues than standard role-playing prompts. Our work showcases an effective dataset generation method for fine-tuning mental health language models alongside downstream applications like dialogue system development, particularly for low-resourced mental health disorders like gambling disorders.

Paper 147 Beyond SFT: Reinforcement Learning for Safer Large Reasoning Models with Better Reasoning Ability

Authors: Jinghan Jia, Nathalie Baracaldo, Sijia Liu

Abstract Large reasoning models (LRMs) extend large language models by generating explicit chain-of-thought (CoT) reasoning, significantly improving mathematical and logical problem solving. However, this explicit reasoning process also introduces new safety risks, as unsafe behaviors often emerge within intermediate reasoning trajectories, even when final answers appear harmless. Existing safety alignment approaches primarily rely on supervised fine-tuning (SFT) over safety-oriented long CoT datasets. While intuitive, we find that SFT produces inconsistent safety improvements, degrades reasoning ability, and generalizes poorly across model families. These limitations suggest that purely supervised approaches are insufficient for robust safety alignment in LRMs. Motivated by these limitations, we systematically compare post-training paradigms for LRM safety

alignment, including SFT, preference-style optimization, and reinforcement learning (RL). Unlike SFT, RL directly optimizes model policies with reward feedback, enabling more adaptive and stable alignment. Extensive experiments across multiple model families and benchmarks show that RL achieves stronger and more consistent safety gains while maintaining reasoning competence. Further analysis of reflection dynamics and token-level entropy reveals that RL suppresses unsafe exploratory reasoning while preserving reflective depth, leading to safer and more reliable reasoning processes.

Paper 152 Towards the Roof of the World: NLP Progress in Tibetan Languages

Authors: Xi Cao, Yijie Li, Yuan Sun, Shiwei Dong, Zeguang Huang, Nuo Qun

Abstract Tibetan, spoken by some seven million people on and around the Qinghai-Xizang Plateau, remains underrepresented in natural language processing (NLP). We present, to our knowledge, the first peer-reviewed systematic bibliometric survey of Tibetan NLP: 140 publications from 2004 to 2026 in the ACL Anthology and DBLP, annotated for venue type, research area, first-affiliation country, and resource release. Output has grown rapidly, with 84% of the publications appearing since 2018, and has expanded bottom-up along the NLP pipeline, from corpora and word segmentation through script, speech, and applications to large language models (LLMs). Researchers based in China first-authored 87% of the publications, which span all 25 represented research areas, whereas work from other countries concentrates on Classical Tibetan. We argue that this full-pipeline paradigm, realized under the compounded constraints of a complex script, dialectal diversity, and geographic remoteness, offers a reference model for other low-resource languages of the Global South. We also identify gaps in dialect coverage, responsible NLP, native and comparable evaluation, cross-community visibility, and resource release (only 29% of the publications release data, code, or models), and recommend open science, shared standards, and community participation for the LLM era.

Paper 155 Agent Memory with Episodic Retrieval for Financial Decision-Making

Authors: ~Nuoyue_Xu1, ~Jiang_Liu3, ~Wenxuan_Huang1, ~Xiang_Zhang17, ~Juntai_Cao1, ~Jiaqi_Wei1

Abstract Large language models (LLMs) have demonstrated strong capabilities in financial analysis and reasoning, inspiring recent advances in agent-based trading frameworks. While these systems show promise, prior approaches either emphasize long-horizon forecasting or operate as stateless analyzers, limiting their applicability to the demands of trading in complicated settings. To address these gaps, we introduce META (Memory Enhanced Trading Agent), the first RAG-like episodic-memory-augmented multi-agent framework for financial decision making. META integrates a family of specialized indicator agents (e.g., Trend, MACD, Stochastic, RSI, SMA, AVWAP, Heikin-Ashi) with a Decision Agent that fuses their reports, and a Memory module that retrieves and updates past trading episodes encoded as market state embeddings with outcomes and reflections. By recalling relevant experiences and adaptively reweighting signals under similar market regimes, META achieves improved directional accuracy and robustness under short-horizon evaluation. Our results demonstrate that \textbf{episodic memory} provides a powerful mechanism for regime-aware, interpretable, and low-latency decision-making in trading and decision making. Code of this project is released at GitHub.

Paper 156 ProcrastiCoT: A Dual-Encoder Cross-Attention Framework for Multi-Dimensional Procrastination Motives Assessment

Authors: Chia-Ni Chiang, Chia-Yun Lee, Hsin-Hsuan Chen, Jheng-Long Wu

Abstract Procrastination is driven by emotional self-regulation failures and cognitive distortions that narratives often mask. We introduce ProcrastiType-ZH, to our knowledge the first Traditional Chinese multi-dimensional procrastination dataset, with 1,000

LLM-generated narratives and human consensus labels for four motives on a 1-5 severity scale (effective range 1-4). We propose ProcrastiCoT, a CBT-informed Chain-of-Thought framework, and inject its LLM-generated reasoning into BERT-size encoders either by concatenation or through a Dual-Encoder Cross-Attention architecture that gives the reasoning its own context window. Across backbones, reasoning injection gives the best single-run QWK on Avoidant, Decisional, and Irrationality. On our primary CKIP-BERT backbone, the Dual-Encoder scores above its pure-encoder counterpart on every dimension in each of four paired training runs, significantly (uncorrected) on Arousal, Decisional, and Irrationality but not Avoidant; on Decisional its mean (0.316) only matches a lexical baseline. A Qwen2.5-7B model fine-tuned on the same labels, with about 70 times more parameters, is strongest on Arousal (QWK 0.709) but below the Dual-Encoder on Irrationality and Avoidant in all four runs. Adding the teacher's reasoning to its targets lowers QWK on every dimension and pulls its predictions toward the teacher's ratings, consistent with a conflict between that reasoning and the human labels.

Paper 159 Activation-Informed Pareto-Guided Low-Rank Compression for Efficient LLM/VLM

Authors: Ryan Solgi, Parsa Madinei, Jiayi Tian, Rupak Vignesh Swaminathan, Jing Liu, Nathan Susanj, Zheng Zhang

Abstract Large language models (LLM) and vision-language models (VLM) have achieved state-of-the-art performance, but they impose significant memory and computing challenges in deployment. We present a novel low-rank compression framework to address this challenge. First, we upper bound the change of network loss via layer-wise activation-based compression errors, filling a theoretical gap in the literature. We then formulate low-rank model compression as a bi-objective optimization and prove that a single uniform tolerance yields surrogate Pareto-optimal heterogeneous ranks. Based on our theoretical insights, we propose Pareto-Guided Singular Value Decomposition (PGSVD), a zero-shot pipeline that improves activation-aware compression via Pareto-guided rank selection and alternating least-squares implementation. We apply PGSVD to both LLM and VLM, showing better accuracy at the same compression levels and inference speedup.

Paper 165 Position: Real-World Clinical NLP Robustness Requires Messy, Multimodal, Longitudinal, Privacy-Preserving Corpora

Authors: Azmine Toushik Wasi, Shahriyar Zaman Ridoy

Abstract Real-world healthcare data is inherently complex: noisy, incomplete, heterogeneous, and temporally irregular, making clinically robust AI development difficult. Traditional models, trained on idealized datasets, fail to capture these operational realities, limiting real-world performance. We propose the \textbf{\textit{Messy Clinic}} paradigm: a dataset and blueprint for building AI systems grounded in authentic clinical complexity. This framework enables seamless multimodal data integration across EHRs, imaging, genomics, and wearables, using privacy-preserving methods such as federated learning, synthetic data, and differential privacy. It provides (1) structured methodologies for longitudinal multimodal management, (2) AI techniques resilient to noise and missing data, and (3) governance mechanisms for consent, accountability, and bias mitigation. By embracing the “Messy Clinic,” healthcare AI can shift from artificial idealism to real-world robustness, accelerating research, enabling personalized care, and aligning AI with the true complexity of clinical practice.

Paper 168 When Tool-Backed Skill Retrieval Fails: Source-Style Collapse in Executable Capability Retrieval

Authors: Yiqi Liu, Joseph James, Yang Wang, Chenghao Xiao, Chenghua Lin

Abstract Large-scale agents increasingly rely on retrieval to access external capabilities. We study this retrieval gate in structured tools and APIs, a measurable class of tool-backed executable skills that must be surfaced before an agent can plan, incorporate, or act. In this setting the retrieval layer can silently fail even when the capability corpus is fixed: on ToolRet, a

retriever fine-tuned on one source-specific slice collapses on another source-specific slice of the same benchmark, with FT-1100 falling to 0.7% coverage on APIGen despite its higher lexical overlap with the gold tools. We call this failure mode source-style collapse. Query-side TF-IDF centroid distance identifies low-coverage transfer sources more reliably than semantic- or length-based proxies, providing a lightweight compatibility signal over a fixed tool corpus. We propose ToolScout, a source-aware routing method that uses this signal as a routing criterion: on the mixed 4,996-query stream, TF-IDF-based routing raises coverage from 22.3% to 86.1%, and across five collapsed sources, 20-shot source-specific adaptation, using 20 training queries per source, raises the coverage-weighted global top-1 proxy from 1.3% to 53.9%. The same failure and routing behaviors persist under executable skill-card rendering, showing that source-style collapse also occurs with this capability representation.

Paper 169 Feedback as an Interface: Controlling Form and Detail in LLM Translation Refinement

Authors: ~Yiqi_Liu1, ~Flavjo_Tomori1, ~Nafise_Sadat_Moosavi1, ~Steffen_Eger1, ~Chenghua_Lin1

Abstract Feedback presentation shapes how LLMs use evaluator judgments during translation refinement. We introduce a machine translation protocol that keeps the evaluator score, main issue, and recommendation fixed while varying feedback form and supporting detail. Across seven models and six translation directions, low-load feedback outperforms high-load feedback in all 42 translator-direction settings under xCOMET. On the 29,400-record rewriting subset, mean xCOMET scores multiplied by 100 are 83.30 for low-load feedback and 80.67 for high-load feedback; the full-matrix original-translation reference mean is 83.46. MetricX-24 supports the low-load advantage on a stratified subset. A no-feedback second pass yields a small improvement over the original translations, and direct evaluator prompting reproduces the low-load advantage without a separate rewriting stage. These results show that compact feedback limits the quality loss from evaluator-guided refinement in the evaluated high-resource MT settings.

Paper 172 HPO: Hysteretic Policy Optimization for Stable and Efficient Training under Sparse-Reward Regime

Authors: Mohamed SANA, Nicola Piovesan, Antonio De Domenico, Fadhel Ayed, Haozhe Zhang

Abstract We investigate a narrow but common failure mode of GRPO-style reinforcement learning in the context of sparse verifiable rewards: early updates contain more responses with negative advantages than those with positive advantages, while response-level length normalization ties the magnitude of the update to the length of the output. We propose Hysteretic Policy Optimization (HPO), a minimal modification of GRPO that reduces the weight of negative-advantage updates and replaces per-response length normalization with mean-length normalization. We further introduce Adaptive HPO (A-HPO), which sets the hysteretic weight based on batch-level advantage-sign statistics, thereby removing the need for tuning a fixed hysteretic weight. In our TeleLogs and Countdown experiments, A-HPO improves the reward per update compared to GRPO, with the largest gains in early sparse reward regimes. On TeleLogs, A-HPO achieves a final reward of 0.84, outperforming SAPO by 5%, GSPO by 11%, and GRPO by 15%, while maintaining a comparable response-length. On Countdown, A-HPO achieves the largest gains in initial and most difficult configurations across 1.5B-7B models. Ablation studies on the hysteretic weight show that the gains of A-HPO come from better balancing the contributions of positive and negative advantages compared to positive-only or fully symmetric updates.

Paper 173 Greedy Matching Aggregation for Soft Contrastive Learning in Multi-Label Text Classification

Authors: Ryoma Kobayashi, Jiyi Li

Abstract Multi-label text classification is important for organizing large document collections, but correlated labels and skewed label distributions make representation learning difficult. Existing multi-label contrastive methods usually define pairwise supervision from label overlap, resulting in a coarse signal limited to observed shared labels. We propose Greedy Matching Aggregation Soft Contrastive Learning (GMA-SCL), which constructs contrastive targets for MLTC by greedily matching label texts and incorporating label importance weighting. The resulting reference distribution is sharpened to emphasize highly similar candidate documents while preserving graded supervision from moderately similar ones. Experiments on AAPD, Reuters-21578, and UK-LEX show that GMA-SCL improves Macro-F1 over existing multi-label contrastive baselines while maintaining competitive Micro-F1 and Hamming Loss. Our code is available at https://github.com/koba-84/gma-scl.

Paper 174 IBPS: Indian Bail Prediction System

Authors: Uddeshya Raj, Puspesh Kumar Srivastava, Praveen Patel, Shubham Kumar Nigam, Noel Shallum, Arnab Bhattacharya

Abstract Bail decisions constitute a large share of criminal proceedings in Indian courts, yet they often suffer from delays, inconsistency, and limited transparency. More than 75\% of India's prison population consists of undertrial prisoners, highlighting the urgency of improving the efficiency of bail adjudication. In this work, we present the Indian Bail Prediction System (IBPS), an AI-assisted framework that predicts bail outcomes and generates legally grounded explanations from factual case attributes and statutory context. We construct a large-scale dataset of 89,969 bail judgments from 5 High Courts of India, enriched with structured annotations such as statutes, custody duration, criminal history, health conditions, and judicial reasoning. Using this dataset, we design a two-stage architecture that combines a transformer-based classifier for outcome prediction with a fine-tuned large language model for rationale generation. Our results demonstrate the feasibility of combining structured legal data and large language models to develop interpretable AI systems that support data-driven analysis of bail decisions.

Paper 178 Benchmarking Frontier Agentic Deception in LLMs vs Humans

Authors: Shiven Sharma, Gautam Choudhary, Ashish Sam, Dhruv Kumar

Abstract Can AI models deceive humans strategically—and do they deceive the same way humans do? We study this through Among Us, a social deduction game where a hidden minority of players (Impostors) must lie convincingly to eliminate opponents without being detected—a controlled setting where every false claim is verifiable. We run 411 games each placing one real human alongside six AI agents, spanning five models from three frontier model families (Claude Haiku 4.5, Claude Opus 4.6, Gemini Flash 3, GPT-4o mini, GPT-5.4), creating, to our knowledge, the first benchmark for direct human–AI deception comparison at this scale. Without any explicit instruction to lie, all five models deceive at rates between 14% and 90%. The central finding is a statistically significant structural difference in style: AI Impostors fabricate plausible-sounding activity to blend in, while human Impostors make accusations to actively steer group beliefs. All lobbies are homogeneous: each game uses a single model for all six agents, so cross-model win rates reflect self-play within a model. These results suggest that deceptive behavior can arise from task incentives without explicit instruction—a baseline risk for AI systems operating in multi-agent settings.

Paper 179 When Reject Turns into Accept: Quantifying the Vulnerability of LLM-Based Scientific Reviewers to Indirect Prompt Injection

Authors: ~Devanshu_Sahoo1, ~Manish_Prasad2, ~Vasudev_Majhi1, ~Jahnvi_Singh1, ~Vinay_Chamola1, ~Yash_Sinha1, ~Murari_Mandal1, ~Dhruv_Kumar4

Abstract Driven by surging submission volumes, scientific peer review has catalyzed two parallel trends: individual over-reliance on LLMs and institutional AI-powered assessment systems. This study investigates the robustness of "LLM-as-a-Judge" systems to adversarial PDF manipulation via invisible text injections and layout-aware encoding attacks. We specifically target the distinct incentive of flipping "Reject" decisions to "Accept," a vulnerability that fundamentally compromises scientific integrity. To measure this, we introduce the Weighted Adversarial Vulnerability Score (WAVS), a novel metric that quantifies susceptibility by weighting score inflation against the severity of decision shifts relative to ground truth. We adapt 15 domain-specific attack strategies, ranging from semantic persuasion to cognitive obfuscation, and evaluate them across 13 diverse language models (including GPT-5 and DeepSeek) using a curated dataset of 200 official and real-world accepted, rejected, and template papers (e.g., ICLR OpenReview). Our results demonstrate that obfuscation techniques like "Maximum Mark Magyk" successfully manipulate scores, achieving an increase in acceptance rates of up to 86.26\% in open-source models, while exposing distinct "reasoning traps" in proprietary systems. We release our complete dataset and injection framework to facilitate further research on the topic (\url{https://github.com/Manishprsd563/scientific-reviewer-jailbreak}).

Paper 182 Dependency-Aware Trajectory Refinement for Efficient Multi-Turn Agent Fine-Tuning

Authors: Zhuo Chen, Zhen Zhang, Xinyu Wang, Kewei Tu

Abstract Multi-turn agent trajectories often contain redundant rounds (failed tool calls, parallel sub-queries, verification-only steps) that inflate both training and inference cost. We propose viewing each trajectory as a round-level dependency DAG that exposes which rounds are globally load-bearing for the final answer, and fine-tune agents on trajectories refined through this DAG. Operating on the DAG lets us apply deterministic and interpretable edits, pruning rounds whose outputs are never consumed and merging interchangeable sibling rounds, with optional LLM-based rephrasing. Models trained on these refined trajectories consistently outperform those trained on the original trajectories at lower inference cost. Specifically, across four multi-modal QA benchmarks, our refinements improve downstream accuracy by up to ,pp over vanilla SFT (and ,pp over an LLM-deletion baseline) while reducing per-sample inference messages by up to and inference tokens by up to , translating to substantial savings in compute and serving cost. Code is released.

Paper 183 ReThink: Adversarially Robust Arithmetic Reasoning via Neuro-Symbolic Soft Constraints

Authors: Soikat Das, Sparsh Mittal, Rekha Singhal

Abstract While Large Language Models (LLMs) perform well on arithmetic word problems, they remain brittle to adversarial perturbations like irrelevant distractors or contradictory facts. To address this, we introduce \textsc{ReThink}, a training-free neuro-symbolic framework that formulates adversarial arithmetic reasoning as a Weighted MaxSAT/MaxSMT problem. \textsc{ReThink} uses a four-stage pipeline: it explicitly decomposes problems to strip noise, encodes the mathematical core as confidence-weighted soft constraints in the Z3 theorem prover, and cross-verifies the formal solution against a parallel Chain-of-Thought (CoT) path. This formulation can sacrifice lower-weight conflicting statements while satisfying higher-weight constraints. Evaluated on a balanced 2,400-example subset of GSM-Plus, \textsc{ReThink} with GPT-4.1 achieves 81.62\% adversarial accuracy, outperforming CoT, PAL, and Self-Verification. It narrows the robustness gap to 14.69 percentage points while achieving 96.31\% clean accuracy.

Paper 184 Enhancing Dynamic Text-Attributed Graph Modeling via Instruction Tuning

Authors: Weili Shi, Haoming Li, Sheng Li

Abstract Unlike conventional static graphs, dynamic graphs evolve over time as their links change. Recent benchmarks, such as the Dynamic Text-attributed Graph Benchmark, further enable the study of dynamic graph learning with rich textual information on nodes and edges. The success of graph instruction tuning on static text-attributed graphs motivates us to extend this approach to dynamic graphs. However, previous works suggest directly using LLMs as predictors on dynamic graphs often underperforms traditional dynamic graph models. To address this issue, we propose a novel two-stage training framework consisting of Dynamic Text-attributed Graph Instruction Tuning (DTGIT) followed by standard supervised training. Unlike prior approaches that employ instruction tuning to enhance LLMs' understanding of graph structures and directly use LLMs as predictors for downstream tasks, our proposed instruction tuning framework serves as a warm-up stage before standard supervised training. Specifically, during the instruction tuning stage, the embeddings produced by the dynamic graph model are aligned with temporal semantics in the natural language space, while the LLM itself is not directly used as a predictor for downstream tasks. In the second stage, the dynamic graph model is initialized with the knowledge acquired during instruction tuning and further fine-tuned in a supervised manner for downstream tasks. Extensive experiments demonstrate that existing dynamic graph models~(e.g., DyGFormer) can achieve consistent performance improvement on the downstream tasks, particularly on link prediction, when trained with our framework.

Paper 187 The Social Cost of Intelligence: Emergence, Propagation, and Amplification of Stereotypical Bias in Multi-Agent Systems

Authors: Thi-Nhung Nguyen, Linhao Luo, Amardeep Kaur, Rollin Omari, Tamas Abraham, Junae Kim, Thuy-Trang Vu, Dinh Phung

Abstract Bias in large language models (LLMs) remains a persistent challenge, often leading to stereotyping and unfair treatment across social groups. While prior work has mainly focused on individual LLMs, the emergence of multi-agent systems (MAS), where multiple LLMs collaborate and communicate, introduces new and underexplored dynamics in how bias emerges, propagates, and amplifies. To systematically investigate these dynamics, we propose a simple evaluation framework with three agent-level metrics that quantify bias emergence, propagation, and amplification throughout multi-agent interaction. We evaluate MAS across three bias benchmarks under varying LLM backbones, social-group configurations, communication behaviors, and adversarial settings. Our results show that communication can trigger up to 70% new bias emergence, propagate bias across over 80% of agents, and amplify stereotypes by more than 3x. We further find that denser and competitive communication generally increases bias. Finally, we demonstrate that MAS are highly vulnerable to simple bias injection attacks, and existing defense strategies provide only limited protection. Our findings provide important insights into the fairness and robustness of multi-agent LLM systems.

Paper 190 How Good Are LLM Agents? Benchmarking Opinion Dynamics Against Indonesian Electoral Discourse on Social Media

Authors: Musa Izzanardi Wijanarko, Derry Tanti Wijaya

Abstract Political polarization on social media intensifies during elections, when echo chambers harden partisan divisions and amplify online harms at the moment collective decisions are being made. Classical opinion dynamics models such as Friedkin-Johnsen provide a framework for studying belief change but abstract away the textual content that drives opinion shifts on social

platforms. LLM-powered agent-based models reintroduce language to simulation, yet their fidelity to real political discourse---as opposed to synthetic personas or factual-consensus domains—remains untested against classical baselines. We construct a benchmark from the 2024 Indonesian Presidential Election: ~1,000 agents, 70,336 interactions, and five topics spanning four weeks of Indonesian-language discourse—to our knowledge the first opinion-dynamics benchmark grounded in non-English political data. We propose a two-axis evaluation framework pairing 1-Wasserstein distance with 3-class buffered-sign F1, reported on aggregate and per-topic tracks. Under this framework, we evaluate three open-weight LLMs (Large Language Models) alongside Friedkin Johnsen and per-agent Ridge regression, analyzing how model family, interaction topology, and sentiment polarity shape simulation fidelity.

Paper 192 From Early Encoding to Late Suppression: Interpreting LLMs on Character Counting Tasks

Authors: Ayan Datta, Mounika Marreddy, Alexander Mehler, Zhixue Zhao, Radhika Mamidi

Abstract Large language models (LLMs) still fail at counting the characters in a word, even while they excel on far harder benchmarks. Using character counting (e.g., "How many 'p's are in apple?") as a controlled task, we find a consistent pattern across Llama, Qwen, and Gemma in both base and instruction-tuned variants. The models compute the correct answer internally but fall back on degenerate output strategies, such as always predicting "1" or guessing uniformly among "1", "2", and "3". Probing classifiers, activation patching, and logit lens analysis show that character-level information is encoded in early and mid layers, then actively attenuated by a small set of later components, mainly the penultimate and final layer MLP and sometimes the final layer attention heads. The failure is not missing information or insufficient scale but structured interference in the computation graph, which explains why these errors persist under scaling and instruction tuning.

Paper 193 Measuring the Impact of Big Five Personality Traits on Suicide Severity Level Detection with Large Language Models

Authors: Kit May Shum, Yusuke Sakai, Masahiro Kaneko, Taro Watanabe

Abstract Sensitive content warning: This paper contains sensitive content related to suicide. Suicide is one of the leading causes of global mortality, making risk detection a critical public health priority. Although psychological studies have identified a link between personality and suicide, this relationship has received little attention and has not yet been systematically examined or applied in the context of large language models (LLMs). In this study, we investigate whether LLMs can capture and leverage this relationship for suicide severity detection. We propose incorporating persona profiles based on the Big Five personality traits to simulate user characteristics in a zero-shot setting, enabling us to assess the impact of personality on detection performance. Experimental results show that Neuroticism yields a modest but consistent improvement across most models, with statistically significant gains in several cases. These findings suggest that LLMs can capture inter-domain relationships, and that further refinement of personality-aware modelling may lead to performance improvements.

Paper 195 SignSim: A Landmark Driven Model for Phonological Sign Similarity Prediction

Authors: Abhishek Bharadwaj Varanasi, Manjira Sinha, Tirthankar Dasgupta, Dr. Charudatta Jadhav

Abstract Phonological similarity is fundamental to sign language structure and cross-linguistic analysis, yet existing approaches based on HamNoSys operate only at the transcription level and ignore visual realization. We propose a phonology-aware model that predicts continuous phonological similarity directly from sign videos while remaining grounded in symbolic phonological structure. Our landmark-based approach uses body and hand keypoints extracted with MediaPipe and a temporal encoder combining short-term convolution with Head-Token Mixing Attention to capture both local articulation and long-range dynamics.

By factorizing HamNoSys annotations into interpretable phonological components and using them as auxiliary supervision, the model learns robust and interpretable representations. Experiments on four sign languages show consistent improvements over alignment-based and RGB-based baselines, particularly in cross-linguistic generalization.

Paper 196 BioCAT5: Concept Mention-Aware Text-to-Text Transfer Transformer for Biomedical Literature

Authors: Mohammad Golam Sohrab

Abstract Integrating biomedical knowledge information like concept name a.k.a concept mention during transfer learning is not a common fashion as it needs an alignment of large-scale concepts with unstructured biomedical text during pre-training. Existing learning objectives on biomedical text mainly rely on masked language modeling or random span denoising where grafting knowledge like integrating medical concepts has never been explored nor been discussed to facilitate the model pre-training. A prevailing assumption is that such concept-aware pre-training on biomedical text can be effective in biomedical task-specific language models. Based on this assumption, we introduce Bio-CAT5, a biomedical domain-specific text-to-text transfer transformer that integrates concept related knowledge from the unified medical language system (UMLS) during pre-training. BioCAT5 leverages concept-aware denoising– an out-of-the-box language model learning objective that is trained to reconstruct text by masking out the concept spans of the input sequence. To facilitate the BioCAT5 pre-training, large-scale concept mentions from UMLS are considered to align with PubMed abstract text to create the pre-training dataset. A publicly available comprehensive biomedical natural language processing (NLP) benchmark datasets including named entity recognition (NER), relation extraction, question answering are evaluated. Besides, the general language understanding evaluation (GLUE) benchmark datasets are considered to show the adaptability of BioCAT5 in general domain tasks. We conduct a series of experiments with ablation studies to validate our proposed approach. The experiment results show that the alignment of concepts during pre-training is potentially significant for biomedical NLP tasks, especially on the NER datasets. Our code is publicly released on GitHub (Available at: https://github.com/aistairc/biocat5) under the Apache 2.0 License.

Paper 202 Plan–Write–Refine: An Iterative Framework for Argumentative Essay Generation

Authors: Zhijun Zhang, Qianlong Wang, Jianzhu Bao, Keyang Ding, Genan Dai, Bowen Zhang, Ruifeng Xu, Yongsheng Liang

Abstract Argumentative essay generation (AEG) aims to produce coherent and persuasive essays on controversial claims. Most existing methods follow a \textit{plan-then-write} paradigm, where a single plan is generated and used to guide long-form writing. However, the quality of the generated essay heavily depends on the initial plan, and errors or omissions in planning often propagate to the final output. To address this limitation, we propose the \textit{Plan–Write–Refine} framework, an iterative approach for AEG. It first generates an initial plan and a draft essay, then abstracts the essay to obtain condensed feedback, and refines the plan accordingly. Compared with the prior static planning methods, our framework enables iterative plan improvement and better structural organization of the essay by closing the loop between planning and writing. We conduct extensive experiments on four AEG benchmarks. Both automatic and human evaluations show that our framework achieves consistent improvements over strong baselines.

Paper 204 EVO-Reranker: Continuous Reranker Evolution under Multi-Round Feedback

Authors: ~Yujie_Dong2, ~Huiying_Yan3, ~Shanfeng_Zhu1

Abstract Effective reranking is crucial for RAG performance, yet the continuous evolution of retrieval distributions necessitates multi-round optimization. We formulate this challenge under two scenarios: (1) Full Feedback, where iterative updates cause catastrophic forgetting of historical preferences; and (2) No Feedback, where the reranker suffers from a "dual-role" dilemma by acting as both learner and supervisor, leading to error propagation. We introduce EVO-Reranker to ensure safe continuous

evolution. It employs ReplayDPO for feedback-rich environments to stabilize historical memory, and CautiousDPO for feedback-absent settings, leveraging confidence gating and multi-LLM consensus to ensure supervision quality. Extensive evaluations across authentic (e.g., BioASQ) and simulated shifts demonstrate that EVO-Reranker effectively averts performance collapse, translating to superior downstream generation on HotpotQA and MultiRC.

Paper 205 MultiReAct: Multimodal Tools Augmented Reasoning-Acting Traces for Embodied Agent Planning

Authors: Zhouliang Yu

Abstract In the field of embodied AI, Large Language Models (LLMs) have demonstrated impressive proficiency in tasks involving simple reasoning. However, they face significant challenges when confronted with longer-horizon tasks described in abstract instructions, particularly those involving intricate visual concepts. These challenges stem from two main limitations. Firstly, LLMs, which heavily rely on text, struggle to cope with the demands of complex embodied tasks that require nuanced multimodal reasoning. Secondly, LLMs encounter difficulties in recognizing and autonomously recovering from intermediate execution failures. To overcome these limitations and enhance the planning capabilities of LLMs in embodied scenarios, we propose a novel approach called MultiReAct. Our framework made the following efforts: We utilize a parameter-efficient adaptation of a pre-trained visual language model, enabling it to tackle embodied planning tasks by converting visual demonstrations into sequences of actionable language commands. By leveraging CLIP as a reward model, we identify instances of sub-instruction execution failure, significantly increasing the success rate in achieving final objectives. We introduce an adaptable paradigm for embodied planning through in-context learning from demonstration, independent of the specific Visual Language Model (VLM), and low-level actor. Our framework supports two distinct low-level actors: an imitation learning agent and a code generation-based actor. Using the MultiReAct framework, we apply it to a diverse set of long-horizon planning tasks and demonstrate superior performance compared to previous LLM-based methods. The extensive experimental results underscore the effectiveness of our approach in addressing long-horizon embodied planning.

Paper 206 Adversarial Prompt Distillation: Efficient Jailbreak Attacks from Large to Small Language Models

Authors: Xiang Li, Chong Zhang, Jia Wang, Fangyu Wu, Ben Li, Xiaobo Jin

Abstract Current jailbreak attacks on large language models (LLMs) predominantly rely on LLMs themselves to generate adversarial prompts, creating a critical efficiency bottleneck: each attack requires substantial computational resources and API queries, limiting scalability and practical deployment. To overcome this limitation, we propose \textbf{Adversarial Prompt Distillation (APD)}, a novel framework that transfers jailbreaking capabilities from LLMs to small language models (SLMs) for efficient, low-resource attacks. APD integrates three key components: (1) masked adversarial knowledge pre-training via LoRA fine-tuning, (2) dynamic temperature-controlled knowledge distillation to bridge architectural gaps, and (3) reinforcement learning-based template optimization for adaptive refinement. Extensive experiments across 12 models show that APD achieves state-of-the-art attack success rates (e.g., 96.4\% ASR$_k$ on GPT-4) while dramatically improving efficiency—generating prompts 3.7$\times$ faster with 11.3$\times$ fewer parameters than teacher models. Our work establishes the first practical framework for lightweight jailbreak attacks, exposes new vulnerabilities in LLM defenses, and provides a scalable testbed for advancing AI safety research. Our code is available at: https://github.com/lxgem/Efficient_and_Stealthy_Jailbreak_Attacks_via_Adversarial_Prompt.

Paper 209 RESCUE-BENCH: Towards Relation-Aware Multi-Party Emotional Support Conversation Systems

Authors: Haichuan Hu, Yang Xiao, Mingni Tang, Jiawen Duan, Quanjun Zhang, Congqing He, Hao Zhang, Jiashuo Wang, Johan F. Hoorn, Wenjie Li

Abstract Existing emotional support conversation systems mainly focus on one-on-one seeker-supporter interactions and individual emotional states, leaving interpersonal relations in multi-party scenarios underexplored. In this work, we introduce relation-aware emotional support conversation, a new task that evaluates whether LLMs can capture and utilize the evolving dynamics of relationships to offer more effective emotional support. We construct RESCUE-BENCH (Relation-aware Emotional Support Conversation Understanding and Evaluation Benchmark) from real couple and family interview conversations, containing 191 samples, 7,079 annotated turns, and 1,064.8 minutes of video. Based on rich annotations of socio-emotional and support-related dynamics, RESCUE-BENCH defines six tasks that evaluate two core capabilities required for relation-aware emotional support: Relational Understanding and Relation-Sensitive Support. Experiments with ten LLMs show that current models perform relatively well on tasks relying on local emotional or intervention cues, but struggle with relation-intensive tasks such as relation pattern prediction, viewpoint prediction, and support strategy prediction. These findings reveal the limitations of current LLMs in modeling interpersonal relations and making relation-sensitive support decisions. We release our code and data on Github and Huggingface.

Paper 216 GuideWeb: A Benchmark for Automatic In-App Guide Generation on Real-World Web UIs

Authors: Chengguang Gan, Yunhao Liang, Yoshihiro Tsujii, Tatsunori Mori, Shiwen Ni, Hiroki Itoh

Abstract Digital Adoption Platforms (DAPs) provide web-based overlays that deliver operational guidance and contextual hints to help users navigate complex websites. Although modern DAP tools enable non-experts to author such guidance, maintaining guides remains labor-intensive because website layouts and functionalities evolve continuously. In this work, we introduce GuideWeb, a human-verified benchmark for automatic in-app guide proposal on real-world web UIs. GuideWeb formulates the task as selecting guide target elements grounded in the webpage and generating concise guide text aligned with user intent. We make the target audience and guide-worthiness criteria explicit, and construct annotations through LLM-assisted candidate discovery followed by human verification, including deletion, rewriting, grounding correction, and addition of missed targets. We also propose an evaluation suite that jointly measures guide target selection, DOM grounding, and generated intent/guide-text quality. Experiments show that our proposed GuideWeb Agent achieves 30.79% F1 in guide target prediction, while obtaining BLEU scores of 44.94 for intent generation and 21.34 for guide-text generation. Existing baselines perform substantially worse, highlighting that reliable automatic guide proposal remains challenging on real-world web UIs.

Paper 217 Audio Editing in the Era of Foundation Models: A Survey

Authors: Changhao Pan, YiFei Fan, Fan Zhuo, Yifu Chen, Wenxiang Guo, Yu Zhang, Zhiyuan Zhu, Rui Yang, Shengpeng Ji, Chenyuhao Wen, Jiayang Xu, Ke Lei, Xiaoda Yang, Jingyu Lu, Zhou Zhao

Abstract Audio editing aims to modify a given synthetic or real-world audio signal to satisfy specific user needs. As a promising yet challenging direction in AIGC, it has attracted increasing attention. Recent advances in audio generation have made powerful generative models central to modern audio editing systems. This rapid progress has created a growing need to organize emerging tasks, methods, and resources into a coherent view. In this survey, we provide a comprehensive review of audio editing in the era of foundation models. We first present a unified taxonomy of existing editing tasks and then summarize the major foundation-model paradigms that support modern audio editing, covering representative approaches from both training-based and training-free perspectives. We further discuss related resources, including datasets, evaluation protocols, and data

construction tools. Finally, we identify open challenges in this field and outline promising directions for future research. The accompanying repository is released at https://github.com/MM-Speech/AudioEditSurvey.

Paper 220 Beyond Accuracy: ARIA-Rubrics for Evaluating Audio Reasoning in Large Audio Language Models

Authors: Yupei Li, Qiyang Sun, Mohamed Mady, Chenxi Wang, Zhengwei Gong, Berrak Sisman, Björn Schuller

Abstract Large Audio Language Models (LALMs) have shown strong performance on audio reasoning benchmarks, but accuracy alone cannot distinguish true reasoning from superficial pattern matching, often overestimating reasoning ability since high scores may result from guessing rather than genuine audio understanding. Evaluating the reasoning process itself is essential for improving LALMs' reasoning ability, yet remains challenging. Existing methods either rely on costly human annotation or opaque LLM-as-judge approaches, making them impractical, biased, and lacking transparency. Moreover, audio reasoning introduces unique challenges absent in text-based settings, perceptual hallucination and cross-modal alignment between audio understanding and textual inference, hence text-based evaluation frameworks cannot be directly applied. Therefore, we propose ARIA-Rubrics (Audio Reasoning Integrity Assessment), a lightweight, annotation-free gold reasoning chains, automatic and transparent framework comprising six complementary metrics that evaluate audio reasoning quality across perceptual grounding, reasoning coherence, and answer consistency. We use Chain-of-Thought prompting as an externalization mechanism to make the reasoning process observable. Experiments on 9 models across 2 benchmarks identify three reasoning modes of current LALMs with actionable directions for future development, with ARIA-Rubrics achieving high correlation with human judgments. The code is available at the https://github.com/glam-imperial/allm_assessment.

Paper 226 Knowledge Divergence and the Value of Debate for Scalable Oversight

Authors: Robin Young

Abstract AI safety via debate and reinforcement learning from AI feedback (RLAIF) are proposed methods for scalable oversight. When is one better than the other? We study this question through the geometry of knowledge divergence between debating models. Using principal angles between their representation subspaces, we derive a closed-form expression for the debate advantage, governed by the amount of K-relevant information one model contributes beyond the other. When models share representation subspaces, debate reduces to an RLAIF-like single-agent setting with the same optimum. We classify three regimes of knowledge divergence and show that debate can produce outcomes inaccessible to either model alone in the compositional regime, while sufficiently strong adversarial incentives can cause coordination failure above a threshold. These results provide a geometric framework for predicting when adversarial oversight is beneficial and connect debate to the broader problem of eliciting latent knowledge across models.

Paper 228 Large Language and Vision-Language Models as World Models: A Survey

Authors: Azmine Toushik Wasi, Sadia Tasnim Meem, Riashat Islam, Md Rizwan Parvez

Abstract Recent advances have reframed Artificial Intelligence (AI) around world models, internal simulators that predict environmental dynamics and support general intelligence. This review surveys Large Language Models (LLMs) and Vision–Language Models (VLMs) not merely as text or vision processors, but as predictive systems capable of modeling state, causality, and dynamics. We organize the literature with a dual taxonomy: a theoretical taxonomy covering pixel-space generative simulators, latent-space predictive architectures, neuro-symbolic and code-coupled models, autoregressive foundation models, and hybrid mixtures of world models; and an application taxonomy covering autonomous driving, embodied robotics, interactive simulation and gaming, and social and agentic simulation. Synthesizing recent works, we conclude that although autoregressive

transformers exhibit emergent topological and causal reasoning, robust physical world modeling requires moving beyond next-token prediction toward hybrid architectures that integrate generative diffusion models with symbolic grounding.

Paper 229 Can LLMs Threaten Human Survival? Benchmarking Potential Existential Threats from LLMs via Prefix Completion

Authors: Yu Cui, Yifei Liu, Hang Fu, Sicheng Pan, Haibin Zhang, Cong Zuo, Licheng Wang

Abstract Research on the safety evaluation of large language models (LLMs) has become extensive, driven by jailbreak studies that elicit unsafe responses. Such response involves information already available to humans, such as the answer to "how to make a bomb". When LLMs are jailbroken, the practical threat they pose to humans is negligible. However, it remains unclear whether LLMs commonly produce unpredictable outputs that could pose substantive threats to human safety. To address this gap, we study whether LLM-generated content contains potential existential threats, defined as outputs that imply or promote direct harm to human survival. We propose ExistBench, a benchmark designed to evaluate such risks. Each sample in ExistBench is derived from scenarios where humans are positioned as adversaries to AI assistants. Unlike existing evaluations, we use prefix completion to bypass model safeguards. This leads the LLMs to generate suffixes that express hostility toward humans or actions with severe threat, such as the execution of a nuclear strike. Our experiments on 10 LLMs reveal that LLM-generated content indicates existential threats. To investigate the underlying causes, we also analyze the attention logits from LLMs. To highlight real-world safety risks, we further develop a framework to assess model behavior in tool-calling. We find that LLMs actively select and invoke external tools with existential threats. Our data are publicly available at https://github.com/cuiyu-ai/ExistBench.

Paper 230 Is There a Dr. House in the Machine? A Survey of Clinical and Diagnostic Reasoning in LLM-based Medical Agents

Authors: Azmine Toushik Wasi, Abdur Rahman, Riashat Islam, Md Rizwan Parvez

Abstract Large language models (LLMs) are increasingly used in agentic systems that reason over multiple steps, call tools, and interact with clinical knowledge sources. This survey examines whether LLM-based medical agents approximate expert diagnostic reasoning or reproduce its outputs without its causal and epistemic foundations. We organize agent design along five axes: reasoning paradigms, agent architectures, tool use, degree of autonomy, and embodiment. We then analyze learning and adaptation, evaluation, safety, and deployment as cross-cutting concerns that interact with these axes. Drawing on research in machine learning, biomedical informatics, and clinical AI, we examine how design choices shape diagnostic reliability and failure modes, and identify requirements for clinical and medical agents.

Paper 233 When Choices Become Risks: Safety Failures of Large Language Models under Multiple-Choice Constraints

Authors: ~Yuheng_Chen9, ~Zhiyu_Wu2, ~Bowen_Cheng3, ~Yihang_Wu12, ~Tetsuro_Takahashi1

Abstract We identify and systematically characterize a class of task-structural alignment failures in large language models (LLMs): even when the harmful intent remains unchanged, changing the task presentation and output constraints can substantially alter model safety behavior. Specifically, when a harmful request is reformulated as a forced-choice multiple-choice question (MCQ) in which all options are harmful and no refusal option is provided, some models that refuse the equivalent open-ended query instead select, prefer, or justify a harmful option. We evaluate 14 proprietary and open-source models on a bilingual Chinese-English human-authored dataset covering five harm categories, together with 900 model-generated Chinese

adversarial MCQs. On human-authored data, attack success rate (ASR) increases sharply as prompts shift from open-ended queries to explicit forced-choice formats, typically peaking under intermediate levels of choice constraint. Model-generated Chinese MCQs further weaken or eliminate the recovery regime observed on human-authored data, driving ASR close to saturation for multiple models. The observed transfer patterns are consistent with stronger generators producing more difficult or boundary-adjacent MCQs, although other properties of the generated inputs may also contribute. We also find that adding an explicit refusal option or a safety preamble substantially reduces ASR for several high-capability models, often to near-zero levels, although their effectiveness varies across target models. These findings suggest that safety evaluations centered on open-ended generation may underestimate risks in structured deployment settings, and that task structure should be treated as an important and diagnosable dimension of safety evaluation and alignment training.

Paper 248 RegMix-D: Dynamic Data Mixing via Proxy Training Trajectories

Authors: Kaiyan Zhao, Zhongtao Miao, Akiko Aizawa, Yoshimasa Tsuruoka

Abstract Data mixture selection is critical for Large Language Model pretraining. Existing methods such as RegMix select a single static mixture by fitting a regression model on small-scale proxy runs. We propose RegMix-D, a simple extension of RegMix to dynamic mixing. Our key observation is that proxy runs produce not only endpoint losses, but also full loss trajectories, which can be used to further improve data mixture. By training regression model on these trajectories, we can predict optimal mixtures at multiple training stages. RegMix-D supports two deployment modes: an *offline* variant that generates a complete mixture schedule before target training, and an *online* variant that adapts the mixture during training using observed loss. Experiments on 25B tokens of the Pile dataset with a 1B parameter target model show that RegMix-D consistently improves over baselines across 13 downstream tasks, and the improvement carries over to a 7B target model on 100B tokens. RegMix-D also remains proxy-efficient: it surpasses RegMix even with only 128 proxy models (25% of RegMix's proxy compute budget).

Paper 251 CCTU: A Benchmark for Tool Use under Complex Constraints

Authors: Junjie Ye, Guoqiang Zhang, Wenjie Fu, Zelin Li, Tao Gui, Qi Zhang, Xuanjing Huang

Abstract Solving problems through tool use under explicit constraints constitutes a highly challenging yet unavoidable scenario for large language models (LLMs), requiring capabilities such as function calling, instruction following, and self-refinement. However, progress has been hindered by the absence of dedicated evaluations. To address this, we introduce CCTU, a benchmark for evaluating LLM tool use under complex constraints. CCTU is grounded in a taxonomy of 12 constraint categories spanning four dimensions (i.e., resource, behavior, toolset, and response). The benchmark comprises 200 carefully curated and challenging test cases across diverse tool-use scenarios, each involving an average of seven constraint types and an average prompt length exceeding 4,700 tokens. To enable reliable evaluation, we develop an executable constraint validation module that performs step-level validation and enforces compliance during multi-turn interactions between models and their environments. We evaluate nine state-of-the-art LLMs in both thinking and non-thinking modes. Results indicate that when strict adherence to all constraints is required, no model achieves a task completion rate above 20\%. Further analysis reveals that models violate constraints in over 50\% of cases, particularly in the resource and response dimensions. Moreover, LLMs demonstrate limited capacity for self-refinement even after receiving detailed feedback on constraint violations, highlighting a critical bottleneck in the development of robust tool-use agents. To facilitate future research, we release the data and code.

Paper 253 LOTUS: Language-Vision Optimal Transport for Unified Report Synthesis

Authors: Xuhui Li, Jiachen Liu

Abstract Automated radiology report generation can reduce reporting burden, but existing encoder-decoder systems still rely on soft cross-attention that offers no geometric guarantee that generated tokens consume the correct visual evidence. We present LOTUS, an optimal-transport framework that places alignment supervision directly at the decoder evidence interface. LOTUS introduces Cross-Attention OT (CA-OT) to align decoder token states with visual patches, and extends it with Prototype OT (Proto-OT) for structure-level alignment between visual and textual codebooks. Experiments on IU X-Ray and ROCO, complemented by a scalability study on MIMIC-CXR-RRG, show that decoder-side OT improves macro-level clinical-label agreement while maintaining competitive language quality. In multi-seed IU X-Ray evaluation, LOTUS-Deep improves macro clinical F1 from 0.162 to 0.201 over the CE baseline, and on ROCO it yields micro-F1 gains on a thoracic-label proxy (0.395 to 0.600). We also provide a population-level risk-gap bound that justifies expected OT regularization as a principled cross-modal training signal.

Paper 256 From Lexical to Entity-Level: Predicting Individual Knowledge Boundaries via Knowledge Graph Label Propagation

Authors: Shuyi Yu, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe

Abstract Personalized reading assistance requires identifying specific concepts a reader finds unfamiliar. While prior work on personalized difficulty prediction has focused on the word or jargon level, a reader's knowledge boundary is more naturally defined over entities grounded in knowledge bases. In this paper, we introduce a novel task of predicting entity-level familiarity for individual readers and propose a graph-based framework for prediction. The system extracts Wikipedia entities from a document, constructs a document-specific graph in the entity embedding space, selects maximally informative seed entities for reader annotation through active learning, and propagates these sparse labels to the remaining entities. We evaluate the framework on 90 Wikipedia articles across five knowledge domains, and find that our method outperforms GPT-4.1-mini and performs comparably to Qwen2.5-72B-Instruct-AWQ without requiring LLM inference. Ablation experiments confirm the contribution of each component. We further analyze the impact of incorporating structured knowledge and popularity-guided directed propagation, finding that both yield domain-dependent effects rather than universal gains. These results lay a practical foundation for constructing reading assistance systems that adapt to individual knowledge backgrounds.

Paper 264 ChartHal: A Fine-grained Framework Evaluating Hallucination of Large Vision Language Models in Chart Understanding

Authors: Xingqi Wang, Yiming Cui, Xin Yao, Shijin Wang, Guoping Hu, Xiaoyu Qin

Abstract Large Vision-Language Models (LVLMs) have recently demonstrated remarkable progress, yet hallucination remains a critical barrier, particularly in chart understanding, which requires sophisticated perceptual and cognitive abilities and demands rigorous factual accuracy. While prior work has investigated hallucinations and chart comprehension independently, their intersection remains largely unexplored. To address this gap, we present ChartHal, a benchmark that features a fine-grained taxonomy of hallucination scenarios in chart understanding, along with a human-validated dataset of 1,062 samples. Our evaluation shows that state-of-the-art LVLMs suffer from severe hallucinations on ChartHal, with even GPT-5 achieving only 34.46% accuracy. Further analysis reveals that questions involving information absent from or contradictory to charts are especially likely to trigger hallucinations. We also evaluate several representative hallucination mitigation methods and find that they exhibit notable limitations, which underscores the urgent need for more robust chart-specific mitigation strategies.

Paper 265 Do LLMs Respect Concept Boundaries? A Bilingual Diagnostic Benchmark

Authors: Renjun Lai, Yixuan Si

Abstract Large language models are increasingly used as natural-language interfaces for structured information processing and agentic workflows, where user-specified constraints can determine downstream extraction, selection, tool use, and action planning. This paper studies whether models can execute natural-language concept-set constraints over candidate objects: given inclusion and exclusion conditions, the model must decide which candidates fall inside the intended concept boundary. We introduce a dual-resource diagnostic benchmark built from WordNet and BabelNet. The benchmark fixes backend constraints, candidate realizations, and gold answers before prompt rendering, and evaluates three task views: single-object judgment, exact candidate-set construction, and satisfiability judgment. The main evaluation set contains 202,500 rendered requests. Across 11 scored model runs, overall accuracy ranges from 43.70\% to 56.36\%, showing that the benchmark remains difficult for current strong models. Exact candidate-set construction yields consistently lower accuracy under stricter set-construction requirements, and recall diagnostics reveal asymmetric effects of constraint direction: increasing exclusion constraints lowers positive recall and raises negative recall, while increasing inclusion branches mainly erodes false-case rejection. These results suggest that concept-boundary execution is a measurable weakness under this diagnostic setup and that evaluating it requires factor-level diagnostics beyond aggregate accuracy.

Paper 274 FACTA: Structured Fact Alignment for Multi-Hop Retrieval-Augmented Generation

Authors: Jiahao Li, Yongxiu Xu, XUMENG ZHANG, Hongbo Xu, Minghao Tang, Yubin Wang, Gaopeng Gou

Abstract Although Retrieval-Augmented Generation (RAG) empowers Large Language Models (LLMs) with external knowledge, current frameworks struggle with complex multi-hop reasoning as they fundamentally rely on shallow retrieval signals. Existing methods typically employ coarse semantic matching or entity-centric graph expansion, which often retrieve textually similar but factually irrelevant evidence, injecting noise into the reasoning process. To overcome this limitation, we introduce FACTA, a structured fact alignment framework for multi-hop retrieval. FACTA bridges the gap between a query's step-wise information demand and a document's information supply by translating both into structured, explicit factual triples. Through a dual mechanism of triple-level fact matching and multi-signal document reranking, FACTA better identifies evidence that satisfies the factual prerequisites of each reasoning hop. Extensive evaluations on three multi-hop QA benchmarks show that FACTA substantially outperforms state-of-the-art baselines, demonstrating that structured fact-centric retrieval effectively reduces reasoning noise and enhances the accuracy of multi-hop generation.

Paper 275 AGOPNs: Null-Space Activation Steering for Jailbreak Defense via Average Gradient Outer Product Concept Extraction

Authors: ~Minh-Nguyen_Bui1, ~Binh-Phuong_Phan1, ~Thanh-Khang_Doan1, ~Bao_Huynh-Gia1, ~Phung-Anh_Nguyen1

Abstract Activation steering offers a training-free path to defending large language models (LLMs) against jailbreak attacks, but existing methods fail to resolve the tension between safety and utility. We identify two root causes: safety concept vectors derived from difference-in-means are fragile under diverse attack geometries, and steering transformations lack a principled guarantee of leaving benign prompts unaffected. We propose AGOPNs (Average Gradient Outer Product with Null-space steering), which addresses both issues by extracting a safety concept vector via the Recursive Feature Machine, a feature-learning estimator grounded in the Average Gradient Outer Product, and confining the resulting steering matrix to the null space of benign activations, guaranteeing near-zero perturbation of benign prompts by construction. Experiments on Llama-3.1-8B-Instruct,

Llama-3.1-70B-Instruct, and Llama-3.3-70B-Instruct across seven jailbreak attack families show that AGOPNs consistently achieves strong defense success rates while preserving near-baseline utility on AlpacaEval, XSTest, GSM8K, and MATH500.

Paper 279 BHAID: A Dataset of LLM-Generated Text in Indian Languages with a Dual Contrastive Learning Framework for Detection

Authors: Shifali Agrahari, SUJIT KUMAR, Ranbir Singh Sanasam

Abstract The rapid growth of multilingual Large Language Models (LLMs) has enabled the generation of fluent and culturally nuanced text across Indian languages, posing new challenges in distinguishing LLM-generated from human-written content. Existing datasets and methodologies have primarily focused on AI-generated text detection in the English language. Existing datasets and methodologies, which are predominantly trained on English data, exhibit limited generalisation to low-resource and morphologically rich Indian languages, underscoring the need for linguistically diverse and culturally grounded benchmark datasets and methodologies. To address this gap, we introduce BHAID, BHaratiya AI-generated text Detection, the first large-scale multilingual benchmark for AI-text detection spanning nine Indian languages: Hindi, Bengali, Tamil, Telugu, Kannada, Malayalam, Gujarati, Odia, and Assamese. Furthermore, we propose a Dual Hierarchical Contrastive Learning framework, DualCLF, that integrates self-supervised and supervised contrastive objectives with focal loss to jointly enhance semantic robustness and inter-class discriminability. Observations from our experiments on the BHAID and benchmark multilingual datasets from the literature reveal that our method consistently outperforms recent state-of-the-art baseline approaches from the literature in detecting AI-generated text.

Paper 280 LecTrans: A Multimodal Translation Benchmark for Academic Lectures

Authors: Frederikus Hudi, Nina Nanakorn, Justin Vasselli, Yusuke Sakai, Hidetaka Kamigaito, Akihiro Tamura, Taro Watanabe

Abstract Existing machine translation and multimodal translation benchmarks mainly focus on text-only inputs or short examples with tightly aligned image-text pairs. Academic lectures present a different challenge: spoken explanations, presentation slides, and domain-specific terminology interact across long-form discourse. LecTrans is constructed from 383 hours of expert-taught online lectures and provides transcripts, presentation slides, and an eight-language parallel evaluation subset covering English and seven target languages. The 162-minute translation subset spans 17 subjects and contains human-corrected English transcripts, verified slide text, and human-verified translations of both transcripts and slides. We define lecture translation as two complementary tasks: transcript translation for spoken lecture content and slide translation for visually presented text. We report text-only and multimodal baselines, showing that noisy transcripts, loosely aligned slide context, and domain-specific content remain challenging for current systems. LecTrans provides a testbed for evaluating multimodal translation in long-form academic lectures.

Paper 281 Predicting Emerging Topics from Outliers: A Prospective Study of Weak Signals in Embedding Space

Authors: Evangelia Zve, Gauvain Bourgne, Jean-Gabriel Ganascia

Abstract Some documents that embedding-based topic models initially classify as noise later become founding members of emerging topics. At publication time, however, they appear as scattered points in embedding space and are difficult to distinguish from ordinary noise without the benefit of hindsight. We study whether such *anticipatory outliers* can be predicted prospectively, using only information available when a document first appears. We derive labels from the subsequent trajectories of outlier documents, distinguishing those that anticipate new topics from those that reinforce existing topics or remain isolated, and

estimate label confidence through agreement across multiple embedding models. On two French news corpora, anticipatory outliers prove predictable at publication time. Under cross-validation, F_1 rises from about 0.77 over the full eligible population to above 0.90 on high-consensus subsets, and remains at 0.76–0.80 under a strictly chronological evaluation. Predictive performance is driven mainly by geometric features capturing each outlier's position in embedding space.

Paper 285 PhaseLoRA: Role-Specialised Low-Rank Adaptation through Prompt-Completion Phase Splitting

Authors: Rishi Ravikumar, Riza Batista-Navarro

Abstract In Low-Rank Adaptation (LoRA) fine-tuning of instruction-tuned LLMs, a single set of adapters is applied uniformly across all token positions, treating prompt and completion tokens identically. We question whether this is appropriate. The two phases serve semantically different roles: prompt tokens constitute the natural language input to be understood, while completion tokens constitute the task-specific output to be generated. Conflating the two under a single adapter set ignores a meaningful distinction that is intrinsic to the instruction-tuning paradigm. The phases are also computationally different: prompt tokens are conditioned on but never predicted, whereas completion tokens are generated autoregressively based on the preceding context. We introduce PhaseLoRA, an extension of LoRA that maintains independent adapter sets for the prompt and completion phases. Experiments across multiple benchmarks and a range of model scales and LoRA ranks show that PhaseLoRA consistently outperforms standard LoRA in low-capacity and low-rank settings. Beyond benchmark comparisons, we conduct mechanistic analysis of the learned adapter representations, structured ablations on warm-up initialisation and asymmetric rank allocation, and a quantitative error analysis characterising the inputs on which phase specialisation helps most.

Paper 286 Can Model Editing Survive Fine-Tuning? Evaluating Robustness and Persistence in Text-to-Image Diffusion Models

Authors: Feng He, Marco Valentino, Zhixue Zhao

Abstract Model editing enables targeted behavioral changes in pre-trained models without costly retraining, supporting applications such as factual correction, debiasing, and safety mitigation. In practice, models are often subsequently fine-tuned for downstream tasks, raising the question of whether edits persist or are reversed during fine-tuning. Understanding this interaction is critical for AI safety, as fine-tuning may either remediate malicious edits or inadvertently undo beneficial ones. We systematically study the impact of subsequent fine-tuning on model editing in text-to-image models. Using two model families (Stable Diffusion and FLUX), five editing methods, and four fine-tuning approaches, we evaluate edit retention of multiple editing objectives under fine-tuning on unrelated tasks. Across models, methods, and metrics, we find that fine-tuning moderately weakens appearance and debiasing edits, while somewhat strengthening edits that remove unsafe content, e.g., for appearance editing, fine-tuning reverses avg. 6.78% editing effect across methods. Further, we provide a mechanistic analysis of edit preservation by examining cross-attention representations and disentangling update magnitude from functional behavioral impact, finding that robustness is associated with systematic representation shifts and limited interference with the edited components. These findings highlight the need for monitoring and, when necessary, reapplying edits in model deployment pipelines.

Paper 287 Neuron-Level Emotion Control in Speech-Generative Large Audio-Language Models

Authors: Xiutian Zhao, Ismail Rasim Ulgen, Philipp Koehn, Björn Schuller, Berrak Sisman

Abstract Speech-generative large audio-language models (LALMs) can produce expressive speech, yet reliable emotion control remains elusive: conversions often miss the target affect and may degrade linguistic fidelity through refusals, hallucinations, or

paraphrase. We present, to our knowledge, one of the first neuron-level studies of emotional voice conversion (EVC) in LALMs and provide intervention-based evidence that compact sets of emotion-sensitive neuron (ESN) are functionally involved in, and can be exploited for, steering expressive speech generation at inference time. ESNs are identified via success-filtered activation aggregation enforcing both emotion realization and content preservation. Across three LALMs (Qwen2.5-Omni-7B, MiniCPM-o 4.5, Kimi-Audio), ESN interventions yield emotion-specific gains that generalize to unseen speakers. Controllability depends on selector design, mask sparsity, filtering, and intervention strength. Our results provide an intervention-based framework for training-free emotion control in LALMs for EVC.

Paper 290 Explore-over-Graph: Hybrid Embedding–LLM Reasoning for Knowledge Graph Question Answering under Incompleteness

Authors: Ola El Khatib, Djellel Difallah

Abstract Large language models (LLMs) are increasingly combined with knowledge graphs (KGs) to ground reasoning in structured evidence. However, most LLM-based KGQA methods rely on traversing existing graph edges and become unreliable when reasoning paths are broken by missing facts. Alternatives that ask LLMs to generate missing knowledge risk introducing hallucinated evidence. We introduce XoG (eXplore-over-Graph), a framework for multi-hop question answering over incomplete KGs that recovers missing reasoning paths from learned graph structure rather than LLM parametric knowledge. XoG combines type-level entity–relation statistics to identify candidate relations with KG embeddings to retrieve plausible missing entities, using the LLM as a semantic selector and reasoner. These mechanisms are integrated into an iterative planning–exploration–reasoning process. Experiments on WebQSP, CWQ, and the Wikidata-based BRINK benchmark show that XoG remains competitive on complete KGs and consistently outperforms comparable methods without task-specific KGQA training under KG incompleteness. These gains persist across multiple LLM backbones, indicating that stronger LLMs alone do not resolve missing graph evidence. XoG also reduces LLM token consumption by up to 33% compared with a closely related planning-based approach.

Paper 293 Iterative Topic Taxonomy Induction with LLMs: A Case Study of Electoral Advertising

Authors: ~Alexander_Brady1, ~Tunazzina_Islam1

Abstract Social media platforms play a pivotal role in shaping political discourse, but the scale and rapid evolution of online content make systematic analysis difficult. We introduce an end-to-end framework for inducing an interpretable topic taxonomy from unlabeled text corpora. The framework combines embedding-based clustering with iterative large language model (LLM) inference to construct a topic taxonomy without requiring predefined labels or seed topics. It first synthesizes candidate topics from document clusters and then uses the resulting taxonomy to assign consistent topic labels across clusters. We evaluate the approach through a case study of political advertising ahead of the 2024 U.S. presidential election. We use the induced taxonomy to support downstream analyses of issue prevalence, moral framing, advertising spend, and demographic exposure patterns. These results suggest that iterative taxonomy construction can provide a scalable and interpretable approach to organizing large unlabeled text corpora while supporting substantive downstream analysis.

Paper 294 TokenMapper: A Step Toward Interoperable Speech Token Translation

Authors: Tal Kozakov, Tal Rosenwein, Eliya Nachmani

Abstract Neural audio codecs discretize speech into token sequences, but the resulting token spaces differ in vocabulary and codebook structure, preventing direct communication across models. This limitation affects applications such as conversational

voice agents and speech to speech translation systems where multiple speech models must interact. As a result, transferring information between speech systems typically requires decoding to waveform audio and re-encoding with a second tokenizer, increasing latency and introducing potential information loss. To address these limitations, we present TokenMapper, a direction aware framework for direct token to token translation between heterogeneous speech tokenizers in the discrete domain. TokenMapper supports structurally mismatched token spaces, including mappings between single codebook and multi codebook representations, under a shared effective token rate. Experiments on GLM-4-Voice, MiMi and DualCodec show consistent cross model performance. Specifically, translation WER approaches native reconstructions within 2.5-6.8% absolute WER, human MOS for TokenMapper outputs ranges from 2.29 to 4.39, following the same direction level trends as UTMOS and end to end latency is reduced by 4.8-94.5% relative to waveform bridging, reaching up to 972 ms per utterance. These results provide a practical step toward cross model speech token interoperability without intermediate waveform reconstruction.

Paper 295 Modeling Triplet Interactions Beyond Pairwise Attention

Authors: ~Mafizur_Rahman1, ~Lijun_Qian1

Abstract Standard transformers use pairwise (2D) self-attention, which models only token-to-token interactions. Sparse attention reduces computational cost while keeping the same pairwise structure and cannot capture richer multi-token interactions. Therefore, we propose Higher-Order Block Attention (HOBA), which models explicit triplet interactions using structured 3D attention with block-diagonal unrolling. This design remains computationally efficient while extending attention beyond pairwise relations. Additionally, we demonstrate that cross-block interaction improves HOBA performance by supporting long-range token dependencies beyond triplet modeling. Experiments on ten NLP benchmarks across five tasks show that HOBA consistently outperforms dense and sparse 2D baselines.

Paper 297 A Multimodal Text- and Graph-Based Approach for Open-Domain Event Extraction from Documents

Authors: Praval Sharma

Abstract Event extraction is essential for event understanding and analysis. It supports tasks such as document summarization and decision-making in emergency scenarios. However, existing event extraction approaches have limitations: (1) closed-domain algorithms are restricted to predefined event types and thus rarely generalize to unseen types and (2) open-domain event extraction algorithms, capable of handling unconstrained event types, have largely overlooked the potential of large language models (LLMs) despite their advanced abilities. Additionally, they do not explicitly model document-level contextual, structural, and semantic reasoning, which are crucial for effective event extraction but remain challenging for LLMs due to lost-in-the-middle phenomenon and attention dilution. To address these limitations, we propose multimodal open-domain event extraction, MODEE , a novel approach for open-domain event extraction that combines graph-based learning with text-based representation from LLMs to model document-level reasoning. Empirical evaluations on large datasets demonstrate that MODEE outperforms state-of-the-art open-domain event extraction approaches and can be generalized to closed-domain event extraction, where it outperforms existing algorithms.

Paper 303 Scalable Methods for Long-Range Argumentation in Distributed Discourse

Authors: ~Alina_Littek1, ~John_Lawrence1

Abstract Argumentative reasoning over long or distributed discourse remains challenging because the number of possible relations between textual units grows quadratically, while true argumentative links remain sparse. As a result, exhaustive relation

classification becomes computationally impractical for long documents, multi-document collections, and fragmented discourse. We address this problem with a modular graph-based architecture designed to aggressively reduce the candidate relation space prior to classification. The framework encodes graph-structured discourse representations and applies a two-stage unsupervised filtering pipeline to eliminate unlikely argumentative links before downstream inference. Across experimental settings, the proposed approach removes more than 99% of candidate relations while substantially improving classification performance (up to 6.5× MCC over an unfiltered baseline). These results show that lightweight modular filtering can make large-scale argumentative reasoning possible without sacrificing predictive quality, enabling practical deployment in long-context and distributed-document environments.

Paper 304 STEVE: Stabilizing Textual Gradient-Based Prompt Optimization via Error-Driven Refinement and Regularized Verification

Authors: ~Yifan_Xu18, ~Yixuan_Li20, ~Xinzhuo_Li1, ~Yixin_Gu1, ~Yifan_Shen5, ~Lijun_Yu1, ~Haohan_Wang1

Abstract Textual-gradient methods automate prompt optimization through natural-language feedback, but their iterative updates can be unstable. We identify two sources of this instability: noisy gradients produced from already-correct examples and over-specialization to hard cases that degrades performance on simpler inputs. We introduce STEVE, a stabilization framework with two coupled mechanisms. Error-Driven Refinement generates gradients only from incorrectly handled examples, concentrating updates on informative failures. Regularized Verification treats every update as provisional and accepts it only when improvement on hard cases does not cause unacceptable regression on a preservation set. Across ten reasoning benchmarks, three evaluator/optimizer models, and established prompt-optimization baselines, STEVE reduces degradation and produces more robust prompts. Additional evaluations with gpt-5.4-mini/gpt-5.4 on symbolic reasoning, GSM8K-Platinum, and DS-1000 show that these gains persist with newer models and larger test sets. STEVE therefore provides a practical way to improve the stability and effectiveness of textual-gradient prompt optimization.

Paper 311 DunbaaBERT: From Sacrifice to Semantics

Authors: Iffat Maab, Waleed Jamil, Raphael Schmitt

Abstract Large language models have achieved strong performance across many NLP tasks, yet Urdu remains comparatively underexplored due to limited resources and fragmented evaluation settings. To address this gap, we introduce DunbaaBERT, a family of Urdu RoBERTa-base models trained from scratch with Byte-BPE vocabularies of 32k, 52k, and 96k tokens on a deduplicated 17\,GB Urdu corpus. We evaluate DunbaaBERT across intrinsic and downstream Urdu NLP benchmarks covering linguistic acceptability, news classification, offensive language detection, and sentiment analysis, while analyzing vocabulary-size effects on performance, pre-training dynamics, and computational efficiency. Across benchmarks, the DunbaaBERT variants achieve competitive performance against strong multilingual baselines while consistently maintaining favorable efficiency trade-offs. Under our controlled experimental setup, increasing the vocabulary size does not consistently improve downstream effectiveness, with DunbaaBERT\textsubscript{32k} repeatedly providing the strongest overall efficiency profile. Overall, our results demonstrate that carefully curated Urdu-specific encoder models can remain highly competitive despite comparatively compact model and training scales. All models are released under the MIT license.

Paper 314 CLUE: A Contrastive Learning Framework for Unsupervised Event Type Induction and Annotation

Authors: ~Pratik_Maitra1, ~Lauren_Kwon1, ~Qi_Li14

Abstract Event ontology is the foundational framework for defining events in a dataset. Event type induction generally involves domain experts who define guidelines for creating event types and their names based on the dataset's domain and scope. However, this process is resource-intensive in terms of annotation costs and is prone to human biases. To alleviate the need for manual event ontology creation, we propose the CLUE framework for event type induction, naming, and annotation. Our main contributions in this work are: 1) designing a novel joint embedding representation for expressing both the verb-centric view and the semantic view of events; 2) event type induction using contrastive loss guided clustering; 3) event naming and refinement. Experiments demonstrate the effectiveness of our method.

Paper 316 The Validity–Fidelity Gap: A Controlled Study of Representation Drift in Structured Extraction Under Redundant Contexts

Authors: ~Gokul_Srinath_Seetha_Ram1, ~Anirudh_Vasudevan1, ~Rashmi_Elavazhagan1

Abstract Pipelines that use LLMs for structured extraction are commonly monitored by checking that each output parses as JSON and matches the expected keys. We show that this check can stay perfect while the extracted content degrades. We call the failure representation drift: the model still finds an entity but writes it in a different form than the intended one, for example a nested key-value object where a plain string is expected. The output is still valid JSON, yet the entity is scored both as missing (a false negative) and as an unexpected extra (a false positive). We study this with synthetic clinical notes repeated verbatim 1x-32x, which gives exact ground truth with prompt, schema and decoding held fixed. For LLaMA-3.1-8B-Instruct, micro-F1 falls from 0.499 to 0.259 while JSON validity stays 1.00. All 280 medication entities scored as missing in a 100-patient run are present in the output as nested objects, and a follow-up request with no schema recovers 96.7% of medications versus 40.0% under the schema at 16x. We introduce the Differential Redundancy Effect (DRE), the extra F1 drop under repeated content relative to an approximately length-matched unrelated-filler control, and use it as a diagnostic contrast rather than a causal estimate. The effect is scale-dependent: DRE is large at 8B and near zero at 70B and Qwen2.5-72B, where F1 does not degrade. Validity metrics stay at 1.00 in both regimes, so they carry no information about content fidelity. A smaller synthetic news experiment shows the same dissociation with a different schema. We also give a zero-ground-truth drift check and monitoring recommendations. All evidence is synthetic, and we make no claim about real clinical notes.

Paper 319 E-RBS: Entropy-Guided Reflective Backtracking Search for Interactive Abductive Reasoning

Authors: ~Jia_Luo3, ~Min_Liu16

Abstract In interactive deductive reasoning with large language models, the model often falls into a confirmation loop due to prematurely locking in incorrect assumptions. This paper is the first to formalize this failure mode as reasoning trajectory collapse and proposes an entropy-guided retrospective search framework, E-RBS. The method monitors the entropy of hypothesis distributions and semantic representations in real time. When cognitive collapse is detected, it proactively triggers state rollback and orthogonal exploration. In addition, this paper constructs LateralBench, an evaluation benchmark targeting lateral thinking. Experimental results show that E-RBS effectively breaks cognitive deadlock, with error recovery and problem-solving success rates significantly surpassing existing baseline methods.

Paper 323 Retrieval-Induced Mirage: When Similar Cases Become Fabricated Evidence in Medical VLMs

Authors: Xiaoxu Ma, Xiangbo Zhang

Abstract Retrieval-augmented medical vision-language models are increasingly used to support clinical reasoning when patient information is incomplete. However, retrieved cases are observations from other patients, not direct evidence for the target patient. In this paper, we study a source-attribution failure that we call retrieval-induced mirage: under missing target modalities, a model may copy findings from retrieved complete cases and present them as if they were observed in the target patient. We introduce RIM-BENCH, a counterfactual evaluation benchmark that fixes a controlled missing-modality setting while varying retrieval consistency and source-awareness. Unlike standard diagnostic accuracy evaluations, RIM-BENCH measures whether generated clinical claims are attributed to the correct evidence source. We define metrics for unavailable-modality claims, retrieved-to-target misattribution, evidence-source accuracy, retrieval-following under conflicting evidence, and missing-modality awareness. Across 1,200 target instances and three main models, implicit RAG raises unavailable-modality claims from 20.0% to 36.3% and induces a 26.2% RIMR, while source-attribution prompting lowers RIMR to 11.4% but does not eliminate the failure. Our framework highlights a safety risk for retrieval-augmented multimodal diagnosis: retrieval can be helpful as statistical context, but unsafe when models verbalize retrieved evidence as patient-specific observations.

Paper 332 TileBench: A Controlled Benchmark for Performance Evaluation and Bottleneck Diagnosis of Tile-Based Programming Models

Authors: Bowen Cui, Zhongchun Zhou, Hao Wu, Tejas Ramesh, Junyu Yin, Jialiang Gu, Keren Zhou

Abstract Tile-based programming models, such as Triton and cuTile, aim to simplify high-performance kernel development, but their practical performance, tuning behavior, and usability remain difficult to compare systematically. We present TileBench, a controlled benchmark for evaluating Triton and cuTile on NVIDIA B200 GPUs under matched operator semantics and comparable implementation structures. TileBench contains 45 operators covering diverse AI-kernel patterns and memory/computation behaviors. Each task provides a PyTorch reference, verified Triton and cuTile implementations, standardized data-types (dtype) and input-size sweeps, default and autotuned configurations, roofline-based metrics, and profiling-guided diagnosis. Our evaluation shows that performance gaps are workload-dependent: cuTile excels on a small cluster of Tensor-Core/TMA-friendly kernels, while Triton is stronger on many irregular, streaming, and bandwidth-bound operators. We further evaluate LLM-generated cuTile and Triton kernels and find that Triton is consistently more token-efficient than cuTile under the same iterative refinement protocol. TileBench is publicly available at https://github.com/Deep-Learning-Profiling-Tools/Tilebench.

Paper 335 MF-QAT: Multi-Format Quantization-Aware Training for Elastic Inference

Authors: Zifei Xu, Sayeh Sharify, Hesham Mostafa

Abstract Quantization-aware training (QAT) is typically performed for a single target numeric format, while practical deployments often need to choose numerical precision at inference time based on hardware support or runtime constraints. We study multi-format QAT, which fine-tunes a pretrained model for multiple precisions within a microscaling format family. Across the evaluated language models, it achieves accuracy close to single-format QAT at most target precisions and generalizes to intermediate precisions unseen during fine-tuning, although the lowest-precision MXINT setting remains challenging. To make these models deployable from a single quantized checkpoint, we propose Slice-and-Scale, which converts a higher-precision MXINT or MXFP anchor to lower precisions by transforming both element values and shared block scales, without access to the original full-precision weights. Fine-tuning through this conversion produces an anchor checkpoint whose converted variants closely track

standard multi-format QAT in language-modeling evaluations. Together, multi-format QAT and Slice-and-Scale enable precision selection from one stored checkpoint within each format family.

Paper 340 Bi-Phase Training: Learning Efficiently From Scratch

Authors: Abel Gurung, Joseph Campbell

Abstract Pre-training foundational models is computationally expensive as models scale. In this work, we introduce Bi-Phase Training (BPT), a parameter-efficient pre-training method that preserves the high-rank learning dynamics observed in full-parameter optimization with substantially reduced trainable parameters. BPT parameterizes each weight update as a composition of constrained high-rank left/right transformations via trainable diagonal matrices and low-rank adapters whose updates are periodically merged into the frozen base weights, resulting in a compounding expansion of accessible directions during training. Across language-model pre-training from scratch (Qwen2.5 models from 100M to 1.5B parameters and OLMo-2-1B) on FineWeb-Edu (10B and 100B tokens), BPT achieves validation perplexity close to full-parameter baselines while using far fewer trainable parameters—e.g., at the 1.5B scale, BPT uses 66% fewer trainable parameters with only a small validation-perplexity gap. Additionally, a comprehensive evaluation across diverse downstream tasks shows that BPT preserves much of the full-parameterized model's downstream performance.

Paper 343 NEXUS: Structured Runtime Safety for Tool-Using LLM Agent

Authors: Elias Hossain, Md Mehedi Hasan Nipu, Tasfia Nuzhat Ornee, Rajib Rana, Niloofar Yousefi

Abstract Tool-using LLM agents increasingly execute high-impact actions, making runtime safety monitoring essential. We present NEXUS (Neural EXecution Utility and Safety), a structured-plan safety monitor that applies a formal intervention policy Π to select among four actions: allow, block, request confirmation, or request revision. NEXUS combines deterministic safety rules, argument-level inspection, and a calibrated logistic-regression risk score for graded escalation. On a 128-instance synthetic benchmark, NEXUS attains F1 = 0.949 and 4-class intervention accuracy of 0.6406, outperforming rule-only intervention selection by +27.3 percentage points. It also improves over rule-only on R-Judge (F1 = 0.861 vs. 0.849), matches rule-only on AgentHarm due to threat-model limits, and achieves 0% ASR at 99% control allow on IPI. On the rule-blind NEXUS-Stress benchmark, NEXUS reaches F1 = 0.897 with REVISE recall of 0.500, highlighting both progress and remaining difficulty in fine-grained intervention routing. With 0.38 ms median latency, NEXUS adds under 0.1% overhead to typical agent loops.

Paper 350 SFMP: Fine-Grained, Hardware-Friendly and Solver-Free Mixed-Precision Quantization for Large Language Models

Authors: Xin Nie, Haicheng Zhang, Liang Dong, Beining Feng, Jinhong Weng, Guiling Sun

Abstract Mixed-precision quantization is a promising approach for compressing large language models under tight memory budgets. However, existing mixed-precision methods typically suffer from one of two limitations: they either rely on expensive discrete optimization to determine precision allocation, or introduce hardware inefficiencies due to irregular memory layouts. We propose SFMP, a solver-free and hardware-friendly mixed-precision quantization framework for large language models. SFMP defines a theoretically simple objective for mixed-precision quantization and is built upon three novel ideas: 1) Block-wise mixed-precision, enabling fine-grained precision within weight matrices while remaining hardware-friendly; 2) Row-column weight reordering, which improves structural alignment via row and column reordering, incurring only a small activation reordering overhead during inference; 3) Unified GEMM kernel, which supports block-wise mixed-precision GEMM at arbitrary average bit-widths. Extensive experiments demonstrate that SFMP outperforms state-of-the-art layer-wise mixed-precision methods under

the same memory constraints, while significantly reducing quantization cost and improving inference efficiency. Code is available at https://github.com/Nkniexin/SFMP.

Paper 352 Markhor: A framework for Adapting LLMs to Urdu via Alignment-Free Tokenizer Replacement

Authors: Muhammad Zain Ali, Yuxia Wang, Muhammad Arslan Manzoor, Tony C Smith, Bernhard Pfahringer

Abstract Large language models (LLMs) have established state-of-the-art performance across many NLP tasks, yet their effectiveness for low-resource languages such as Urdu remains limited due to under-representation in pretraining corpora and inadequate tokenization. This study explores adapting pretrained LLMs to new languages by replacing the original tokenizer without cross-lingual correspondence. We show that stable adaptation depends on maintaining architectural compatibility, matching vocabulary size, and reusing pretrained embedding and language modeling head weights. We propose *Markhor*, a framework for adapting LLMs to low-resource languages in this way. It spans tokenizer training, continual pretraining, and instruction tuning via knowledge distillation, resulting in two models, *MKGPT* and *MKQwen*, based on GPT and Qwen architectures. These models are evaluated on multiple classification tasks (fake news detection, hate speech detection, emotion classification) and open-ended question answering. Results show both MKGPT and MKQwen outperform baselines across all tasks, suggesting the effectiveness of the methodology for developing open-source Urdu LLMs.

Paper 354 HakushoBench: A Japanese Chart and Table VQA Benchmark from Governmental White Papers

Authors: Issa Sugiura, Shuhei Kurita, Yusuke Oda, Naoaki Okazaki

Abstract Understanding chart and table images is essential for applying vision-language models (VLMs) to real-world document understanding. While English benchmarks have advanced rapidly, non-English counterparts remain scarce, leaving it unclear whether this progress generalizes across languages. A key obstacle is the difficulty of collecting realistic and diverse non-English chart and table images at scale. To address this, we leverage governmental white papers as a source for benchmark construction, as they contain naturally occurring charts and tables across diverse formats and domains and are freely accessible in many countries. As a first instantiation, we introduce HakushoBench, a Japanese chart and table VQA benchmark built from 33 governmental white papers. HakushoBench contains 2,053 images spanning over 10 image types, with manually annotated and independently verified QA pairs designed to assess holistic understanding of charts and tables rather than local visual cues alone. Experiments across a broad range of VLMs show that HakushoBench is substantially harder than the existing Japanese benchmark and remains challenging for open-weight models: sub-10B open-weight models reach at most 58.6\% accuracy, and even the flagship open-weight model Qwen3.5-397B-A17B trails Gemini~3~Pro by 8.1 points (85.8\% vs.\ 93.9\%), highlighting substantial room for improvement in complex chart and table understanding. We release our dataset and code.

Paper 355 YFPO: Yoked Feature Preference Optimization with Neuron-Guided Rewards

Authors: Yifan Le

Abstract Preference optimization has become a widely used post-training paradigm for improving the reasoning abilities of large language models. Existing methods typically learn from preferred and dispreferred responses as external behavioral supervision, while largely ignoring capability-related signals encoded in the model's internal representations. In this work, we study whether such internal signals can provide useful auxiliary supervision for mathematical reasoning. We introduce YFPO (Yoked Feature Preference Optimization), a neuron-guided preference optimization framework that couples response-level preference learning with neuron-level rewards. YFPO first uses AttnLRP to identify math-associated internal features, and then derives an auxiliary reward from the activation margin of these neurons between preferred and dispreferred responses. This reward is combined with the standard preference optimization objective, encouraging the model to align external preferences with internal math-

related features. We conduct small-scale experiments on GSM8K with a compact language model. Results show that neuron-guided rewards can influence preference optimization dynamics and yield measurable improvements in several settings, suggesting that internal representations can serve as lightweight and interpretable signals for reasoning-oriented post-training.

Paper 356 Disentangling Perception and Reasoning in Multimodal LLMs via Reward Design

Authors: Omar Sharif, Eftekhar Hossain, Nikhil Singh, Patrick Ng

Abstract Reinforcement learning with verifiable rewards has driven major gains in LLM reasoning, and it is intuitive to assume this recipe will transfer well to multimodal models. However, multimodal models do two things: first, perceive what is in an image, then reason about what it implies. Because these stages are graded jointly, it is hard to tell how much room reasoning alone has to grow. We study this on algorithmic visual puzzles, where both components are necessary and show that perception, not reasoning, is the binding constraint. Replacing images with simple textual descriptions raises performance by over 20 points on average for Claude models. We then evaluate six reward designs aimed at inducing visual grounding during reasoning without chain-of-thought supervision. Training Qwen-2.5-VL-7B with GRPO, reward design induces long, structured reasoning with self-reflection and visual references, yielding a 5.56-point gain over the base model. These gains are, however, uneven; no single reward improves all categories, and rewards with verifiable accuracy signals trade cross-domain transfer for in-domain accuracy. These results point to perception-aware reward design as a path forward, so that signals correct perception at its source rather than the reasoning that inherits its errors.

Paper 360 Scale-Aligned Reinforcement Learning for Adaptive Computation Allocation in Reasoning Language Models

Authors: Yilin Liu, Xuan Wang, Zheng Yang, Zhuoran Li, FAN WAN, Wu Chen, Xiaojie Wang

Abstract Long chain-of-thought (CoT) reasoning in large language models (LLMs) improves performance on complex tasks but incurs substantial inference cost, although many queries do not require extended computation. Existing approaches typically reduce this cost via uniform length penalties or fixed computation budgets, but lack length constraints aligned with the model's own query-dependent distribution of successful reasoning lengths. We propose a reinforcement learning framework that aligns the generated reasoning length with query-dependent successful computation scales estimated from verified trajectories. By providing adaptive guidance derived from the distribution of successful reasoning lengths, our method avoids the limitations of global length regularization. Experiments on mathematical reasoning benchmarks demonstrate improved accuracy--efficiency trade-offs, reducing unnecessary reasoning while maintaining or improving task performance.

Paper 362 UniGuardian: A Unified Defense for Detecting Prompt Injection, Backdoor Attacks and Adversarial Attacks in Large Language Models

Authors: Huawei Lin, Yingjie Lao, Tony Geng, TAN YU, Weijie Zhao

Abstract Large Language Models (LLMs) are vulnerable to attacks like prompt injection, backdoor attacks, and adversarial attacks, which manipulate prompts or models to generate harmful outputs. In this paper, departing from traditional deep learning attack paradigms, we explore their intrinsic relationship and collectively term them Prompt Trigger Attacks (PTA). This raises a key question: \textit{Given a prompt, can we tell whether a hidden trigger is steering the model's behavior?} To address this, we propose UniGuardian, the first unified, training-free defense that detects trigger-activated prompt injection, backdoor, and adversarial attacks by measuring how structured perturbations of the prompt shift the model's output distribution. Additionally, we introduce a single-forward strategy to optimize the detection pipeline, enabling simultaneous attack detection and text

generation within a single forward pass. Our experiments confirm that UniGuardian accurately and efficiently identifies trigger-activated prompts in LLMs.

Paper 368 ProgMem: Programmatic Belief-State Memory for Long-Horizon Conversational Agents

Authors: ~Pengchao_Zheng1, ~Xinan_wang3, ~Ziyi_Wei3, ~Haoran_Liu21, ~Lei_Wei4, ~Jiajun_Niu1, ~Bohan_Guo1, ~Han_Wu1

Abstract Long-horizon conversational agents require memory that can evolve with dialogue, retrieve information on demand, and remain uncertainty-aware when evidence is insufficient, rather than merely larger storage or context windows. Existing methods store interaction histories as retrievable text, summaries, or vector indices, and retrieve relevant content at query time. Although effective for simple factual queries, this paradigm does not explicitly model how memory writes change state, leaving temporal updates, plan changes, and counterfactual reasoning to query-time inference over retrieved text. We propose ProgMem, which compiles dialogue events into structured memory programs and checks answers against dialogue evidence. On LoCoMo-10, ProgMem has the highest average F1 with Qwen3-4B, Qwen3-8B, and Qwen3-14B. Its temporal F1 exceeds Mem0 by up to 9.95 points and Zep by at least 12.88 points across these settings. Ablations show gains from route-aware retrieval and two-stage verification. On DialSim, ProgMem improves Answerable and FUQ F1 over Mem0 while using fewer query-time memory tokens. A single-GPU comparison finds higher mean and P95 query latency than Mem0. These results show where state maintenance helps and what it costs at query time.

Paper 369 BAEmpEval: A Framework for Probing Large Audio Language Models on Multidimensional Empathetic Response Assessment

Authors: Jingxi Liang, Pengze Guo, Zhiwen Xie, Qifeng Wang, Yujia Huo, Derek F. Wong

Abstract Large Audio Language Models (LALMs) are increasingly applied in affective and conversational AI, yet their ability to align with human empathetic judgments in spoken dialogue remains poorly understood. We introduce \textbf{BAEmpEval}, a multimodal evaluation framework for analyzing empathetic response judgment in instruction-following LALMs under subjective conversational settings. The framework contains 5,436 dialogue pairs collected from three speech sources, where models evaluate listener responses across six fine-grained dimensions spanning both vocal perception (Vocal Warmth, Perceived Genuineness, Prosodic Attunement) and conversational response strategies (Exploration, Interpretation, Emotional Reaction). Experiments reveal a challenging setting: experiments across recent LALMs reveal uneven alignment with human relational judgments across empathy dimensions and speech conditions. Our findings suggest that empathetic response judgment remains a challenging multidimensional perceptual task for current LALMs and highlight the importance of evaluation frameworks that account for subjective relational variability. We release the dataset and evaluation code to support future research in human-centric AI.

Paper 370 SQLStructEval: Structural Evaluation of LLM Text-to-SQL Generation

Authors: ~Yixi_Zhou1, ~Fan_Zhang67, ~Zhiqiao_Guo1, ~YU_CHEN61, ~Haipeng_Zhang3, ~Preslav_Nakov2, ~Zhuohan_Xie1

Abstract In Text-to-SQL tasks, large language models can generate structurally different SQL queries that return correct answers for the same intent. We call this phenomenon execution-correct structural divergence (ECSD). We introduce SQLStructEval, a framework that analyzes this behavior through canonical abstract syntax tree representations. We quantify structural diversity and agreement across generations. Experiments with different LLMs on multiple Text-to-SQL datasets, including Spider, document ECSD across models and datasets. Furthermore, our experiments demonstrate that generated queries are sensitive

to question paraphrases and schema presentation. To address ECSD, we adopt a pipeline that first generates structured intermediate representations and then deterministically compiles them into SQL, improving execution accuracy and structural agreement among correct outputs. Structural analysis thus provides an additional diagnostic perspective that complements execution-based evaluation. Code is available at https://xanderzhou2022.github.io/AACL2026-SQLSTRUCTEVAL/.

Paper 373 NüshuVoice: Reviving the Voice of Endangered Nüshu with Pitch-Aware Text-to-Speech

Authors: Hongkun Yang, YI Xinhui, Zhao Xiyan, YIBO MENG, Lionel Z. WANG, Lixu Wang, Yaqi Zhang, Ruiqi Chen, Xuanyue Zhao, Lanxin Zhang, Yu Zeng, CHU WEIJIA, Yiming Ma, Chenyu Liu, Jianghao Lin, Xu Xin

Abstract Nüshu is an endangered phonetic script historically used by women in Jiangyong County, southern Hunan, China. While existing computational studies of Nüshu mainly focus on textual digitization and visual recognition, the acoustic reconstruction of its authentic pronunciation remains largely unexplored. Building a Nüshu text-to-speech (TTS) system is particularly challenging because available recordings are extremely limited and mostly consist of isolated syllable-level pronunciations rather than natural sentence-level utterances. In this work, we introduce NüshuVoice, the first TTS benchmark for Nüshu. We construct a sentence-level Nüshu text-to-audio dataset that aligns standardized Unicode Nüshu text, phonetic transcriptions, standard Chinese translations, and archival recordings. To synthesize speech under this extreme low-resource setting, we propose Nüshu-PitchVITS, an -conditioned VITS framework that leverages Nüshu's five-level pitch notation as an explicit prosodic inductive bias. Experimental results show that Nüshu-PitchVITS outperforms strong TTS baselines in spectral fidelity, pitch reconstruction, and human-rated intelligibility. We publicly release the dataset and code at: https://anonymous.4open.science/r/Nvshu-TTS-2EB6.

Paper 374 GenState-AI: State-Aware Dataset for Text-to-Video Retrieval on AI-Generated Videos

Authors: Minghan Li, Tongna Chen, Tianrui Lv, Yishuai Zhang, Suchao An, Guodong Zhou

Abstract Existing text-to-video retrieval benchmarks often permit matching from static appearance, making it difficult to determine whether models capture state changes and final outcomes. We introduce GenState-AI, a benchmark comprising 956 AI-generated videos and 195 verified state-change queries across Toy, Cartoon, Home, and Manipulation domains. Its construction uses prompt-controlled variants that target either conflicting temporal/state outcomes or semantic substitutions. Following human verification, all valid videos contribute to subset-level retrieval pools, while complete query-specific triplets support diagnostic hard-negative ordering. We evaluate VCF-Lik, Qwen3-VL, and LanguageBind using both standard Recall@K and triplet-based consistency metrics. Qwen3-VL (8B) achieves the best macro-average R@1 and triplet consistency, while LanguageBind obtains the highest semantic discrimination accuracy. However, stronger retrieval does not consistently yield correct hard-negative ordering, and Qwen3-VL and LanguageBind remain weak on Manipulation triplets, where VCF-Lik performs substantially better. GenState-AI therefore provides a controlled diagnostic testbed for distinguishing global retrieval quality from fine-grained temporal and semantic grounding in AI-generated videos. Dataset: https://huggingface.co/datasets/Mh-Li/genstate-ai

Paper 376 Quantifying Uncertainty Through Coverage in Latent Semantic Space

Authors: Gabriele Carrino, Brown Ebouky, Katja Hose, Mattia Rigotti

Abstract Large language models often generate plausible yet unreliable answers, making uncertainty quantification essential for detecting potential hallucinations. Prior work has shown that the information needed to detect hallucinations and quantify a model’s confidence in its generations may already be present in its internal states, but how to extract this information effectively remains an open problem. We introduce Semantic Coverage, a white-box, representation-level spectral uncertainty estimator

inspired by coverage in reinforcement learning. Given multiple sampled generations for a prompt, Semantic Coverage quantifies uncertainty by measuring how broadly their internal representations span the model's latent semantic space. Across three open-weights language models and four short-form question-answering benchmarks, Semantic Coverage achieves the strongest overall performance among the evaluated methods and consistently outperforms prior representation-level methods in quantifying LLM uncertainty.

Paper 378 Selective Reflection for Context-Sensitive Emotion Recognition in Conversation

Authors: Xinyu Wang, Weijie Feng, Hongchuang Wang, Zhiyong Cheng

Abstract Recent advances in Large Language Models have transformed Emotion Recognition in Conversation, yet existing instruction-tuned approaches still produce high-confidence misjudgments on utterances whose surface text is weakly informative but whose emotion depends on the surrounding situation. We propose SRERC, a selective reflective correction framework for LLM-based ERC. SRERC uses a verdict teacher to assess contextual support for the first-pass prediction and triggers reflective decoding only when reflection has an actionable role. For selected cases, an SRC generator constructs a Scene-aware Reflection Context, which provides target-specific situational grounding and verification probes for an SRC-conditioned second pass. This gated design enables targeted correction while preserving non-gated predictions and limiting unnecessary over-correction. Experiments on IEMOCAP, MELD, and EmoryNLP show that SRERC achieves the highest average weighted-F1 across the three benchmarks, with diagnostic gains on context-sensitive cases where surface cues are insufficient or misleading.

Paper 379 Style-Aware Translation for Multi-Word Expressions in Low-Resource Languages

Authors: Antonios Dimakis, Antonios Anastasopoulos

Abstract While neural machine translation (NMT) has achieved remarkable parity with human performance in general domains, handling multi-word expressions (MWEs) remains a persistent challenge, particularly in low-resource settings. This difficulty is twofold: models need not only decode the non-compositional meaning of MWEs but also navigate the stylistic choice between idiomatic ("MWE") and literal ("plain") outputs. In this study, we introduce a new 4-way parallel dataset for Greek and English, comprising of Greek and English MWE and plain sentences. We evaluate the style-transfer and style-switching capabilities of state-of-the-art Large Language Models (LLMs) across four tasks. Using traditional metrics, we demonstrate that LLMs often struggle with style-aware translation. Furthermore, we introduce StyleCOMET, a fine-tuned quality estimation metric to better evaluate stylistic nuances. Our results suggest that current limitations in MWE translation are a product of both model inability and metric-level insensitivity to idiomaticity.

Paper 387 ReleaseEval: A Benchmark for Evaluating Language Models in Automated Release Note Generation

Authors: Qianru Meng, Zhaochun Ren, Joost Visser

Abstract Automated release note generation aims to reduce the cost of documenting frequent software updates. Existing benchmarks mainly rely on commit messages, limiting evaluation of how models use broader software-change evidence. We introduce ReleaseEval, a reproducible and openly licensed multi-granularity benchmark with 94,987 release notes from 3,369 repositories across six programming languages. Each release is aligned with commit messages, commit tree structures, and raw code diffs, supporting three task settings: commit2sum, tree2sum, and diff2sum. Automatic and human evaluations show that models benefit from lightweight commit-tree structure, while raw diffs remain challenging due to length and noise. These

findings position ReleaseEval as a multi-granularity benchmark for studying release note generation. We release the dataset and code at https://github.com/Changanyuexia/rn_benchmark/

Paper 390 EXAM^2: $\underline{Ex}tending$ $\underline{A}udio$ $Understanding$ in $\underline{M}ultilingual$ and $\underline{M}ultimodal$ $Analysis$

Authors: Jiawen Wang, Xiaoxue Gao, Zi Haur Pang, Nancy F. Chen

Abstract Recent large audio language models (LALMs) have achieved impressive progress in audio understanding. However, existing evaluations remain largely constrained to English and narrow audio domains. Prior benchmarks typically focus on a single audio modality, i.e., speech, sound, or music, limiting the systematic investigation into how these models generalize across diverse visual scenarios. In this paper, we introduce EXAM^2, a benchmark for multilingual and multimodal audio understanding spanning six languages and multiple modalities, including speech, sound, music, mixed-audio settings, and visual images. By incorporating visual information alongside heterogeneous audio inputs, EXAM^2 enables more realistic evaluation of scene-aware audio reasoning and cross-modal comprehension. EXAM^2 comprises $5,667$ multiple-choice questions, $22,614$ image instances, and $135,684$ multilingual translations. We evaluate state-of-the-art open-source and proprietary LALMs as well as multimodal LLMs, revealing substantial performance gaps in multilingual and cross-modal understanding. Furthermore, we propose Gemma3n-EXAM^2, a lightweight fusion-model fine-tuned on EXAM^2-train, achieves up to 15.8% improvement in multilingual settings and 16.5% gains in multimodal evaluation over a strong baseline. Empirical results establish EXAM^2 as a challenging benchmark and pioneer future multilingual and multimodal audio intelligence research.

Paper 393 Student-Centered AI Requires Auditable Evidence of Learning

Authors: Xu Aofeng, Sichao Li

Abstract Educational LLMs such as writing coaches, coding assistants, study companions, and AI tutors are widely marketed as student-centered, usually because they are “personalized” or because they raise performance on the assisted task. We argue that this is an unverified pedagogical claim: a system that helps students complete tasks is not necessarily a system that helps students learn. The distinction is between performance-centered assistance, which optimizes the immediate product, and student-centered learning, which preserves the learner’s cognitive work and validates capability the learner keeps without the system. A 2025 randomized trial shows the gap in practice: an unguarded tutor raised assisted performance while harming unassisted exam scores. We trace this to a units mismatch: “student-centered” is a learner-side claim, yet it is measured almost entirely with tutor-side proxies. We propose five auditable standards (S1–S5) that translate established learning-science principles into requirements for benchmark design, reward shaping, red-teaming, and deployment monitoring; a small illustrative process audit then shows both that the standards can be operationalized and why such audits cannot substitute for learner-side validation. Our position is that LLM-based educational systems should be judged not by answer quality, engagement, satisfaction, or assisted performance alone, but by whether they measurably support learners’ capacity to reason, self-regulate, transfer, and keep learning without the model.

Paper 397 The Image Reconstruction Game: Drawing Common Ground Through Iterative Multimodal Dialogue

Authors: Sherzod Hakimov, Mattia D'Agostini, Ivan Samodelkin, David Schlangen

Abstract We introduce the Image Reconstruction Game, a fully automated benchmark in which a vision-language model issues corrective instructions to an image generator across multiple turns, making accumulated common ground directly observable as a rendered image. Benchmarking two Describer models crossed with two Generator models across seven image categories, we find that the describer is the dominant factor in reconstruction quality, while the generator determines whether iterative refinement helps or hurts. Mathematical and geometric images pose the greatest challenge. The describer's token budget strongly affects convergence: shorter budgets yield sparser first renderings with more room for visible improvement, while longer budgets raise absolute quality but leave less to fix. Stronger describers use a richer correction vocabulary spanning spatial, numeric, and structural categories, while weaker describers concentrate on surface properties and tend to stop after a few turns. Human validation shows that the best automated judge reaches only slight-to-fair agreement with human preferences, and automated scores require human recalibration to be used reliably.

Paper 400 Pingala: Prosody-Aware Decoding for Sanskrit Poetry Generation

Authors: Manoj Balaji Jagadeeshan, Atul Singh, Nallani Chakravartula Sahith, Amrith Krishna, Pawan Goyal

Abstract Poetry generation in Sanskrit typically requires the verse to be semantically coherent and adhere to strict prosodic rules. In Sanskrit prosody, every line of a verse is typically a fixed length sequence of syllables adhering to prescribed binary patterns of syllable weights. We observe that instead of treating a verse as a monolithic sequence, segmenting them as grouped-lines leads to significant improvement in semantic coherence by 10% with comparable metrical adherence. Specifically, PINGALA, our proposed decoding approach is designed to encourage every line to have well-formed words and our token selection biases the model towards it by preferring longer tokens. Writing in Sanskrit follows phonemic orthography, hence using a phonetically aware transliteration scheme, SLP1, increased the metrical alignment by 46% with comparable semantic similarity, for a instruction fine-tuned large language models like Phi-4. We also introduce a new approach for reference-free evaluation using cross-encoders which achieved better alignment with true poetry instances.

Paper 404 SMARTS: A Small-to-large Model Coordinated Retrieval and Response Framework for Task-Oriented Dialogue

Authors: ~Jiale_Chen9, ~Xuelian_Dong1, ~Shengcheng_Ye1, ~Wenxiu_Xie1, ~Kun_Zeng3, ~Tianyong_Hao1

Abstract Task-Oriented Dialogue (TOD) systems are commonly used to assist users in achieving specific goals through human-computer interactions. Existing methods typically employ a single model during response generation to simultaneously learn response style and perform fine-grained knowledge retrieval. However, these task-coupling methods often result in suboptimal responses generation and the inclusion of irrelevant knowledge in responses, making the optimization of TOD systems more difficult. To address this challenge, we propose a novel small-to-large model collaboration framework for task-oriented dialogue, named SMARTS. This framework utilizes a small language model to generate style-only responses (without knowledge), while a large language model retrieves top-1 relevant knowledge from the knowledge base independently. The style-only responses are finally filled with top-1 relevant knowledge and form the completed responses to users. Finally, a re-thinker is designed to check the language fluency and knowledge accuracy of responses. Experiments on two public datasets demonstrate that SMARTS outperforms existing methods by an average of 5.79% in Entity F1.

Paper 407 Activation Residual Hessian Quantization for Low-Bit Large Language Model Inference

Authors: ~Yifeng_wang9, ~Zhun_Sun1, ~Hakaze_Cho1, ~Linlin_Wang1, ~Keisuke_Sakaguchi2

Abstract Low-bit activation-weight quantization is crucial for efficient large language model inference but remains fragile because activation quantization errors propagate across Transformer layers. We propose Activation Residual Hessian Quantization (ARHQ), a post-training residual-aware low-rank correction framework that directly targets this propagation. ARHQ measures the residuals produced by the deployed activation quantizer, builds an input-side residual Hessian, and identifies the weight directions that most amplify these residuals. This yields a closed-form truncated-SVD solution: residual-sensitive components are placed in a high-precision low-rank branch, while the remaining weights stay in the low-bit path. ARHQ requires no backpropagation, supports LoRA-style deployment, and complements activation smoothing. Across LLaMA-family models and Qwen3-4B-Thinking, ARHQ improves W4A4 reconstruction, reduces WikiText2 perplexity over matched-rank SVDQuant, preserves downstream reasoning and coding accuracy, and achieves up to 3.21$\times$ prefill speedup over BF16 (3.11$\times$ over AWQ W4A16) with over 60\% memory saving.

Paper 408 Residual Stream Bias: Inherent properties that suppresses a model's reasoning ability.

Authors: Priyesh Vakharia, Reva Agarwal, Leilani H. Gilpin

Abstract Large language models can fail on reasoning tasks in ways that are difficult to explain. In this work, we analyze internal model representations during reasoning failures and identify some inherent properties we term Residual Stream Bias: a bias in the residual stream that suppresses signals corresponding to the correct output even when those signals are produced by intermediate self-attention components. To study this behavior, we introduce LayerSee, a method for isolating and interpreting intermediate components during inference time. We use causal interventions to show attention layers important for generating the correct token can still have their signals suppressed, directly demonstrating the presence of Residual Stream Bias. We evaluate multiple models on symbolic sequence completion tasks, and we find consistent evidence of Residual Stream Bias across model families and task variations. Our results suggest that incorrect outputs do not always indicate a lack of reasoning capability but may instead arise from internal biases that prevent correct intermediate signals from being reflected in the final prediction.

Paper 409 Beyond Correctness: A Framework for Analyzing Reasoning and Faithfulness of Multi-hop Question Answering

Authors: Jiyuan Liu, Yi Xie, Yanghui Rao, Fu Lee Wang, Jianxing Yu, Li Qing

Abstract Diagnosing the root causes of failures in Retrieval-Augmented Generation (RAG) is challenging, as current evaluations often conflate logical reasoning errors with the failure to adhere to retrieved context. To address this, we propose **ReFa**, a framework inspired by Structural Causal Models to explicitly disentangle *Reasoning* from *Faithfulness*. Through this lens, we characterize two distinct failure modes: the **'Fool'**, representing a deficiency in logical reasoning, and the **'Lazy'**, representing a lapse in faithfulness where the model defaults to parametric priors. To operationalize this diagnosis, we introduce Dual Reasoning Chain Editing, a mechanism that constructs controlled proxy chains to isolate reasoning structure from evidence faithfulness. We apply this method on **ReFaBench**, a novel benchmark constructed in this work featuring factual, counterfactual, and knowledge-conflicting scenarios. Furthermore, we propose **ReFa-DPO**, a decoupled preference optimization strategy that leverages these proxies to target specific failure patterns. Experimental results demonstrate that **ReFa-DPO** enhances robustness, particularly in mitigating parametric interference, by simultaneously enhancing both contextual faithfulness and reasoning capabilities.

Paper 414 Prompt Rewrites Are Semantic Actions: Testing Whether Language Models Obey Typed Operator Laws

Authors: Zida Yang, Weiqing Zeng, Guquan You, Yu Zheng

Abstract Prompt rewrites can prescribe a change in an answer, not merely a change in wording. We introduce ORBIT-SIT, a framework that assigns each rewrite an explicit action on a task's output space and separately measures correctness against transformed gold answers and consistency with transformed model predictions. Across eight operators, five datasets, and three instruction-tuned models, high task accuracy does not ensure compliance: Phi-4-mini achieves 90.7% accuracy on SST-2 but only 16.4% under a label-swap instruction. Deterministic validation and a three-annotator audit check the operator contracts. Non-deceptive binary controls retain a 17.8-percentage-point gap, and separating inference from transformation improves compliance, implicating joint execution rather than refusal alone. OrbitAlign, a quantized low-rank adaptation objective, reduces gold-grounded risk by 25.4 points on a 400-item BoolQ repair pool. The recipe lowers risk in 39 of 40 cells on two additional backbones, and same-action transfer across binary tasks reduces risk by 12.9–15.1 points. The binary tasks provide the primary evidence; structured tasks remain extensibility diagnostics. ORBIT-SIT makes declared prompt actions measurable and partly repairable without equating compliance with general understanding.

Paper 418 Woodpecker Distillation: Weak Models Diagnose Reasoning Bugs in Strong Models

Authors: Dayu Wang, Jiaye Yang, Jiahui Liang, Weikang Li, Liwei Qian, Xin Pei, Jizhou Huang

Abstract Large language models often fail on reasoning tasks despite possessing the capability to solve them. We argue that many such failures arise from localized reasoning bugs in intermediate steps rather than from global incompetence. We show that these bugs are frequently repairable: inserting a short patch generated by a weak probe model after the same strong-model reasoning prefix can redirect the trajectory toward a correct solution. However, this corrective effect is not reliably internalized by directly fine-tuning on weak patches or repaired trajectories, suggesting that the useful signal lies not in the intervention text itself, but in how it reshapes the model's future reasoning distribution. We therefore propose Woodpecker Distillation, a weak-to-strong training framework that learns from contrastive local interventions. Our method contrasts successful and unsuccessful weak-model patches at the same prefix, constructs a corrective teacher distribution from their induced future token predictions, and distills this signal into the strong model. Experiments on mathematical reasoning benchmarks show that Woodpecker Distillation consistently improves strong-model performance and outperforms direct imitation baselines. Code and datasets are available at https://anonymous.4open.science/r/Woodpecker-E6BD.

Paper 419 Beyond Alignment: Expanding Reasoning Capacity via Manifold-Reshaping Policy Optimization

Authors: Dayu Wang, Jiaye Yang, Jiahui Liang, Weikang Li, Liwei Qian, Xin Pei, Jizhou Huang

Abstract Reinforcement Learning with Verifiable Rewards (RLVR) has demonstrated remarkable success in enhancing the reasoning capabilities of Large Language Models (LLMs). However, recent studies question whether RL genuinely expands reasoning capacity or merely aligns existing latent capabilities, arguing that exploration remains confined within the pre-trained model's low-rank bias manifold. In this work, we challenge this accessibility boundary hypothesis by demonstrating that the latent reasoning space can be fundamentally expanded through targeted geometric interventions. We propose Manifold-Reshaping Policy Optimization (MRPO), a geometric framework designed to fundamentally restructure the inference space of LLMs. MRPO operates in two stages: first, we employ Spectral Orthogonal Exploration (SOE) to eject the policy initialization into the null space of the bias manifold; second, we integrate an Effective Rank regularization term into the policy optimization objective. This approach incentivizes the discovery and maintenance of high-dimensional reasoning trajectories against the entropy-reducing

tendency of standard RL. Empirically, our 4B-parameter method achieves state-of-the-art performance on mathematical tasks, significantly outperforming larger models (e.g., Qwen3-32B) and expanding the capability boundary beyond standard GRPO. Our code is available at https://anonymous.4open.science/r/MRPO-D57B/

Paper 423 When Quality Filters Backfire: Agreement-Based Latent Filtering Hurts Bootstrap Pretraining

Authors: Suan Lee, JaeSeong Kim

Abstract Latent thought pretraining augments training text with model-generated "thoughts". Because generating them is expensive, the obvious way to scale is to spend the budget only on "high quality" chunks — and the most intuitive quality signal is whether the thought generated before a chunk agrees with the one generated after. We show this intuition backfires. In a controlled 1.1B experiment that compares every allocation rule with a random baseline at its own inclusion rate, keeping only high-agreement chunks raises held-out perplexity above the budget-matched random baseline in every seed ($+0.84\%$; $n=5$ vs. 3, exact permutation $p=0.018$). The deficit persists across a second training iteration, though downstream accuracy does not separate the filter from random at the same budget. We trace the cause: importance rises with agreement over most of its range, but the top agreement decile, where prior and posterior coincide or the prior is missing, holds degenerate latents that an agreement-only filter cannot exclude. To diagnose this we introduce PAIR, which pairs agreement with a length-corrected importance weight (raw importance is itself a length artifact, $\rho=-0.786$). Importance routing alone matches random at its budget, the intersection of the two filters is slightly worse than random ($+0.3\%$, $p=0.008$), and inverting the importance signal collapses perplexity. Budget matching decides these verdicts: against a random baseline at a lower inclusion rate the agreement filter looks 4.9% worse and importance routing 1.2% worse, and an earlier version of this work credited the intersection with a gain that was likewise a budget effect. Intuitive single-axis quality filters do not beat random allocation, and any latent allocation study must compare each rule with a random baseline at its own inclusion rate.

Paper 425 A Survey of Neurosymbolic Approaches to Propaganda and Persuasion Technique Detection

Authors: Alysa Ziying Tan, Daniel Kainovan Handoyo

Abstract Propaganda and persuasion technique detection, the task of identifying how misleading content persuades rather than merely what it claims, occupies a distinctive and underexplored position within computational misinformation research. This task requires interpretable and rhetorically grounded reasoning over persuasion techniques, making it particularly well aligned with neurosymbolic artificial intelligence (AI). This survey presents the first systematic review of neurosymbolic approaches for propaganda and persuasion technique detection, together with relevant structured-reasoning approaches from closely related areas that offer methodological foundations for these tasks. It synthesizes work spanning argument mining, knowledge-augmented learning, logic-based fallacy detection, structured reasoning with large language models (LLMs), and hybrid neural-symbolic architectures. We further organize existing methods through a unified taxonomy of neurosymbolic integration paradigms and review foundational rhetoric theory, benchmark datasets, and evaluation methodologies. We identify key open challenges, including subtle and implicit persuasion, cross-lingual generalization, explanation faithfulness, and the growing challenge of LLM-generated propaganda. Finally, we outline future research directions and argue that neurosymbolic AI provides a promising foundation for bridging predictive performance with human-interpretable rhetorical reasoning in next-generation misinformation detection systems.

Paper 426 FairJudge: Abstention-Aware Multimodal Judges for Fairness and Alignment Evaluation in Text-to-Image Models

Authors: Zahraa Al Sahili, Maimuna Nowaz, Maryam Fetanat, Ioannis Patras, Matthew Purver

Abstract Evaluating text-to-image (T2I) systems requires judging not only whether an image matches a prompt, but also whether socially salient attributes are represented faithfully and without unsupported inference. Existing automated evaluators typically rely on face-centric recognizers or contrastive image--text similarity, which provide limited diagnostic feedback and often force predictions even when visual evidence is ambiguous or absent. For fairness-sensitive attributes such as religion and disability, where cues may be contextual, indirect, or intentionally unspecified, these evaluators can therefore miss failure modes that careful human reviewers would notice. We introduce \textsc{FairJudge}, an abstention-aware evaluation protocol that uses instruction-following multimodal LLMs as structured judges for social-attribute prediction, profession grounding, and prompt--image alignment. The protocol constrains outputs to closed label sets, requires visible-evidence rationales, supports an explicit \textsc{unspecified} decision when cues are insufficient, and maps rubric-based alignment judgments to . These constraints turn MLLM judging from open-ended assessment into a parseable, auditable evaluation procedure. Across four attribute-prediction benchmarks and three profession/alignment benchmarks, \textsc{FairJudge} outperforms or complements CLIP, DeepFace, VIEScore, and VQAScore. Ablations show that closed labels, abstention, and evidence reporting are central to reliability. We further introduce \textsc{DIVERSIFY} and \textsc{DIVERSIFY-Professions}, two context-rich resources for evaluating social representation and profession grounding beyond face-visible or iconic cues. We release code, prompts, datasets, parser logs, and per-image judge outputs to support reproducible auditing.

Paper 433 WK-Drift: Benchmarking Procedural Memory Maintenance under World Knowledge Drift

Authors: ~Pratyay_Banerjee1, ~Masud_Moshtaghi1, ~Ankit_Chadha1

Abstract LLM agents that accumulate procedural memory can reuse successful plans, but what happens when the world changes? Deprecated APIs, revised clinical thresholds, and reorganized knowledge graphs silently invalidate stored procedures. No existing benchmark evaluates whether a memory system can \emph{maintain} its procedures under such drift. We introduce \textbf{WK-Drift}, a benchmark constructed from three public temporal datasets (Time-Sensitive QA, GitChameleon, DriftMed). WK-Drift formalizes drift as typed events with a fact-change/procedure-change taxonomy, defines a four-phase protocol (learn, drift, evaluate, recover), and introduces entity-plan linkage for targeted maintenance. We evaluate ten configurations spanning no-memory, unstructured (including Mem0 and MemRL), tagged, and graph-based approaches. Graph-based surgical updates achieve 90% post-drift accuracy on knowledge QA (+6pp over stale baselines), 64% on clinical (+26.5pp), and 57% on code (+7.4pp), reducing staleness from 77--86% to 41--42%. The advantage concentrates on procedure-change tasks (+17.4pp) and scales with propagation depth (+50pp at depth~3). Surgical maintenance achieves 59% from the first post-drift epoch versus 26% for stale baselines, eliminating multi-epoch recovery latency.

Paper 437 DeScribe: Discourse-Driven Extraction and Structuring of Claim Verification for Better Faithfulness Evaluation

Authors: Yang Zhong, Yang Janet Liu, Diane Litman

Abstract Evaluating the faithfulness of multi-sentence summaries generated by language models (LMs) remains challenging due to their length and the subtlety of factual errors. While claim decomposition methods have proven useful for assessing short summaries (1-3 sentences), their effectiveness for longer, coherent summaries across diverse domains is not well understood. To better understand this issue, we first conduct an empirical study analyzing the quality of claims produced by existing

decomposition approaches. Motivated by our findings, we propose a new discourse-informed method, Describe, which first extracts claims from the full summary using discourse structure, and then evaluates them through a structure-guided process that progresses from intra- to inter-sentential claim validation, enabling more fine-grained faithfulness evaluation. Experimental results across benchmarks reveal the benefits of discourse-guided decomposition and evaluation. Further analysis of model-generated error rationales and ablation studies shows our method effectively identifies nuanced errors within sentences, particularly those arising from complex discourse structures.

Paper 438 An Empirical Study of Many-Shot In-Context Learning for Machine Translation of Low-Resource Languages

Authors: Yinhan Lu, Gaganpreet Jhajj, Chen Zhang, Anietie Andy, David Ifeoluwa Adelani

Abstract In-context learning (ICL) allows large language models (LLMs) to adapt to new tasks from a few examples, making it promising for languages underrepresented in pre-training. Recent work on many-shot ICL suggests that modern LLMs can further benefit from more ICL examples enabled by their long context windows. However, the inference cost can be prohibitive for low-resource language communities, and it is unclear whether example selection can reduce the number of examples needed. In this paper, we present a systematic empirical study of many-shot ICL for machine translation from English into ten truly low-resource languages recently added to FLORES+. We analyze the effects of retrieving more informative examples, using out-of-domain data, and ordering examples. Our findings show that many-shot ICL becomes more effective as the number of examples increases. More importantly, we show that BM25 retrieval substantially improves data efficiency: 50 retrieved examples roughly match 250 many-shot examples, while 250 retrieved examples approach the performance of 1,000 many-shot examples. We further show that example ordering has little effect and that ICL provides additional gains on top of fine-tuning.

Paper 443 SuperFlow: An Efficient and Stable Online RL Method for Flow-Matching Models

Authors: Kaijie Chen, Zhiyang Xu, Ying Shen, Zihao Lin, Yuguang Yao, Lifu Huang

Abstract Recent progress in flow-based generative models and reinforcement learning (RL) has significantly improved text–image alignment and visual quality. However, existing RL training methods for flow-based models still suffer from two critical limitations: (i) GRPO-style fixed per-prompt group sizes ignore heterogeneous sampling importance across prompts, leading to inefficient computation allocation and slow convergence; and (ii) trajectory-level advantages are commonly reused as per-step estimates, resulting in biased credit assignment along continuous-time flow trajectories. We propose SuperFlow, a novel RL training framework for flow-based models that addresses both issues. Specifically, SuperFlow introduces a variance-aware sampling strategy that dynamically adjusts per-prompt group sizes based on prompt-level uncertainty, and derives step-level advantage estimates grounded in continuous-time flow dynamics. Empirically, SuperFlow achieves state-of-the-art performance among RL-optimized flow-based text-to-image models while requiring only 5.4%–56.3% of the original training steps and reducing wall-clock training time by 5.2%–16.7%, without any architectural modifications. Across standard text-to-image benchmarks, including text rendering, compositional generation, and human preference alignment, SuperFlow improves over SD3.5-M by 4.6%–47.2% and over Flow-GRPO by 1.7%–16.0%.

Paper 444 Bridging the Long-Tail Gap: Robust Retrieval-Augmented Relation Completion via Multi-Stage Paraphrase Infusion

Authors: Fahmida Alam, Mihai Surdeanu, Ellen Riloff

Abstract Large language models (LLMs) struggle with relation completion (RC), both with and without retrieval-augmented generation (RAG), particularly when the required information is rare or sparsely represented. To address this, we propose a novel multi-stage paraphrase-guided relation-completion RAG framework, RC-RAG, that systematically incorporates relation paraphrases across multiple stages. In particular, RC-RAG: (a) integrates paraphrases into retrieval to expand lexical coverage of the relation, (b) uses paraphrases to generate relation-aware summaries, and (c) leverages paraphrases during generation to guide reasoning for relation completion. Experiments with five LLMs on two benchmark datasets show that RC-RAG consistently outperforms several RAG baselines. In long-tail settings, the best-performing LLM augmented with RC-RAG improves by 50.4 Exact Match (EM) points over its standalone performance and surpasses two strong RAG baselines by 30.4 and 16.0 EM points, respectively, without requiring additional model training or fine-tuning.

Paper 446 DKCD: Domain Knowledge-Enhanced Causal Discovery from Unstructured Data

Authors: Xin Li, Jin Li, Shoujin Wang, Kun Yu, Fang Chen

Abstract Causal discovery from unstructured data is a challenging yet underexplored task in high-expertise domains such as healthcare, finance, and education. Existing methods typically leverage the general knowledge of large language models (LLMs) to identify causal factors from unstructured data and annotate them into structured data for causal graph construction. However, they remain limited by two key challenges: (CH1) insufficient identification of latent factors, which are implicit in the data yet essential for causal discovery, and (CH2) unreliable factor annotation, causing errors to propagate into the resulting graphs. To address these challenges, we introduce a novel Domain Knowledge-enhanced Causal Discovery framework (DKCD) for extracting causal graphs from unstructured data in high-expertise domains with three interconnected components: (1) Knowledge Mining: It retrieves relevant domain knowledge based on observable factors to support causal reasoning. (2) Knowledge-guided Causal Reasoning: Reasoning with relevant knowledge discovers latent causal factors to address CH1 and generates causal clues for more accurate annotation to address CH2, and (3) Causal Structure Discovery: It constructs the final causal graphs based on a more complete factor set and accurate annotations. Experiments on two domain-specific datasets show that DKCD significantly improves both causal factor identification and causal graph construction. The code and data are available at https://github.com/bigoldli329/DKCD.

Paper 448 When more supervision hurts: reward design for structured biomedical reasoning in 4B models

Authors: Batuhan Nursal, Christophe Ye, Xinyu Guo, Xin He, Cassie S. Mitchell

Abstract Reinforcement learning has become a common approach for improving language-model reasoning, but its effectiveness depends heavily on reward design, especially for compact domain-specialized models. We study structured biomedical question answering with two 4B-scale models, MedGemma-1.5-4B-it and Qwen3-4B-Thinking-2507, using a two-stage SFT+GRPO pipeline that trains models to produce tagged think/answer outputs. We compare three reward schemes: a simple format + accuracy reward, a depth incentivized reward that encourages multiple reasoning cycles, and a gated LLM-judge reward for intermediate reasoning milestones. Across five biomedical QA benchmarks, the simplest reward is the most robust: it substantially improves MedGemma in the structured-output setting, while depth and LLM-judge rewards often reduce accuracy. For Qwen3, SFT or native long-budget inference can outperform GRPO on several benchmarks, showing that reasoning-tuned models may not benefit from additional structural reinforcement. Our results suggest that, for lightweight

biomedical QA, reward simplicity and model priors are more important than increasingly elaborate intermediate reasoning supervision. The codebase for this study is publicly available on GitHub: https://github.com/batuinmetz/When-more-supervision-hurts.

Paper 449 Modeling Discourse Profiles in Climate Misinformation Discourse

Authors: Ge Gao, Zhengyang Shan, Bryan A. Plummer, Margrit Betke, Derry Tanti Wijaya

Abstract Prior work on climate misinformation detection typically classifies individual comments or claims using text alone. This framing overlooks a key property of online climate discourse: misinformation is often produced, reinforced, and sustained through recurring rhetorical patterns that reappear across comments and threads. Using a large-scale TikTok climate discourse dataset, we construct discourse profiles by clustering climate-related comments into recurring rhetorical patterns. We then incorporate these profiles into fine-grained climate misinformation classifiers using a late-fusion architecture. Under video-disjoint evaluation across three training seeds, discourse-profile-aware models significantly improve macro-F1 over text-only baselines for bag-of-words, TF-IDF, and MPNet representations. RoBERTa-large exhibits a positive mean improvement but substantially greater variability, and the difference is not statistically significant. Beyond classification, our results show discourse profiles shape misinformation reception and propagation. Discourse-profile-conditioned language model agents exhibit misinformation acceptance patterns aligned with profile-level misinformation prevalence in ClimateDisc, and the relationship between misinformation prevalence and engagement persistence differs across dominant discourse profiles.

Paper 452 CULTURETALK-ID: A Multi-Task Dialogue Benchmark for Cultural Commonsense in Indonesian Local Languages

Authors: ~Muhammad_Dehan_Al_Kautsar1, ~Salsabila_Zahirah_Pranida1, ~Bilal_Elbouardi1, ~Fajri_Koto1

Abstract Culture shapes how people interpret and navigate social interactions, yet existing Indonesian cultural commonsense benchmarks evaluate LLMs on short and isolated prompts, stripping away the dialogic context in which cultural nuances actually surface. We introduce CULTURETALK-ID, the first dialogue-based benchmark for cultural commonsense in Indonesian and its local languages, comprising 4,496 culturally grounded dialogues across 11 languages and 13 culturally salient topics, curated through a multi-stage human pipeline with native speakers to ensure authenticity. CULTURETALK-ID introduces three complementary tasks, namely dialogue-based multiple choice cultural commonsense reasoning, culturally faithful machine translation, and language steering, which jointly probe whether LLMs can understand, transfer, and generate culturally grounded language. Our results show that open-source models still underperform, particularly in the generation task and local language settings, highlighting a gap in their ability to capture and generate culturally grounded knowledge in conversational contexts.

Paper 456 How Is Automated Research Evaluated? A Survey of Benchmarks and Evaluation Practices

Authors: Liulei Zhang, Dejing Zhou, Chuyue Huang, Guanhua Chen, Yutong Yao, Lidia S. Chao, Chi Man VONG, Derek F. Wong

Abstract Automated research systems support literature synthesis, ideation, experiments, writing, and peer review, but their evaluation is dispersed across tasks, benchmarks, and studies that are difficult to compare directly. We review this literature from the perspective of evaluation design and evidence, covering six targets: literature synthesis, research ideation, executable workflows, scholarly writing and communication, automatic peer review, and end-to-end research. We compare task construction, evidence sources, evaluators, and scoring procedures to explain the capabilities assessed by different designs. Our synthesis highlights three recurring lessons: output checks, process checks, and human studies provide complementary information;

evaluator calibration is specific to the property being assessed; and resource budgets and attempt selection are integral to interpreting performance comparisons. We identify diagnostic evaluation designs and documented gaps in supporting evidence, and translate these comparisons into reporting and audit recommendations for specific evaluation settings. The survey helps readers navigate existing evaluations, select appropriate benchmarks, and design subsequent studies.

Paper 458 Pretraining Builds Cooperation, Not Components: Tracking Inter-Layer Coordination via Lie Brackets

Authors: ~JaeSeong_Kim2, ~Suan_Lee1

Abstract Transformer language models predict the next token by passing through dozens of sequential layers, yet whether adjacent layers operate independently or coordinate their computations---and how this relationship forms during pretraining---remains largely unknown. We view residual updates of adjacent layers as vector fields and introduce a Lie bracket-based inter-layer coordination metric that quantifies how much two layers depend on each other's outputs. Analyzing pretraining checkpoints of 6 models across 3 architecture families (Pythia, OLMo, Amber), we find that all layer pairs are nearly independent at initialization, but as pretraining progresses, individual layer contributions decrease while coordination systematically increases---concentrating in the middle layers that existing metrics (BI, CKA) classify as ``redundant.'' The mechanism is Jacobian amplification: middle layers write small updates but respond sensitively to their neighbors, and this amplification co-emerges with coordination ($r > 0.99$). Block duplication experiments confirm that high-coordination blocks are 2.7--7.9 more robust than low-coordination blocks, verifying that this cooperative structure substantively contributes to model function. Our findings reframe pretraining as a process that builds cooperative circuits between layers rather than strengthening each layer individually.

Paper 463 Where Does Plan Validity Live in Grammar-Constrained LLM Planning? An Applicability Mask Induced from Environment Rollouts

Authors: Nijesh Upreti

Abstract LLM-augmented symbolic planning combines natural-language goal grounding with symbolic search and constrained decoding. We present Abstraction-based Grammar-Constrained Planning (AGCP), a framework that induces lifted action schemas from environment rollouts, compiles a bounded language of grounded action sequences that are applicable and goal-reaching under the induced model, and enforces that language during decoding. The guarantee is relative to the induced model. Environment execution additionally requires a sound induced model and a correct parsed goal. With the model, prompts, instances, and decoding settings fixed, replacing a syntax-only grammar with the applicability mask, compiled here from the domain's own transition model, moves goal-reaching samples from 0/160 to 160/160. On human-annotated ALFWorld goals, verifier-guided correction raises parse acceptance from 61.4% to 75.4% and execution success from 67.5% to 80.7% on 114 paired tasks. Under paraphrasing, GPT-4o and Sonnet grounding remain stable while a regular-expression baseline falls from 31.9% to 0.8%. Together, these results locate validity relative to the induced model in an applicability grammar built from a symbolically observable, bounded environment. In this setting, natural-language goal grounding is the main remaining role of the LLM.

Paper 466 Corrective Retrieval Augmented Generation

Authors: Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, Zhen-Hua Ling

Abstract Large language models (LLMs) inevitably exhibit hallucinations since the accuracy of generated texts cannot be secured solely by the parametric knowledge they encapsulate. Although retrieval-augmented generation (RAG) is a practicable

complement to LLMs, it relies heavily on the relevance of retrieved documents, raising concerns about how the model behaves if retrieval goes wrong. To this end, we propose Corrective Retrieval Augmented Generation (CRAG), an autonomous agent that dynamically corrects and refines the retrieval content. Specifically, an agent core is introduced that perceives, decides, and executes in a closed loop. It dynamically selects and orchestrates tools, including internal retrieval, web search, keyword evaluators, semantic evaluators, and LLM evaluators, to perform multi-dimensional contextual quality evaluation and knowledge supplementation, thereby constructing accurate and precise retrieval content. Furthermore, because CRAG refines and corrects retrieved context without requiring modifications to the underlying system, it naturally serves as a plug-and-play framework that can be seamlessly integrated with a wide range of RAG-based approaches. Experiments on five datasets demonstrate that CRAG outperforms previous approaches, while its autonomous decision-making capabilities improve tool-calling efficiency and retrieval context quality.

Paper 467 AI-Sustainability Discourse Across Social Media: Sentiment, Topics, and Engagement on LinkedIn, X, and Bluesky

Authors: Kholoud Khalil Aldous, Md. Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Kais Attia, Wajdi Zaghouani

Abstract We present the first cross-platform computational analysis of AI-sustainability discourse across LinkedIn, X, and Bluesky, using 52,102 posts collected over 15 months. Combining RoBERTa-based sentiment classification with BERTopic topic modeling, we identify 36 topics and compare sentiment, topic distributions, and engagement across platforms. Overall, 43.98% of posts are positive, 40.16% neutral, and 15.86% negative. Sentiment varies by topic rather than by platform: corporate and AI-potential themes are mostly positive, while resource-cost topics carry the highest negative sentiment. Topic distributions differ significantly across platforms: LinkedIn focuses on corporate sustainability, X on hardware and cryptocurrency, and Bluesky on environmental resource costs. Niche topics consistently attract more engagement than high-volume ones. These findings have implications for how organizations and policymakers communicate about AI-sustainability, as topic framing and platform audience composition shape public discourse more than platform selection alone. We release the annotated dataset to support future research.

Paper 473 WhisperX-LLM-TermFixer: Enhancing Whisper with LLM-Guided Post-Decoding for Domain-Specific Terminology

Authors: Ronghao Pan, José Antonio García-Díaz, Rafael Valencia-García

Abstract Large-scale ASR models such as Whisper perform well in general domains but struggle with specialized terminology, particularly in sensitive domains like medicine. We propose WhisperX-LLM-TermFixer, a lightweight post-decoding framework that improves term recognition without retraining or modifying the acoustic model. The method combines N-best decoding, phonetic n-gram correction, and LLM-based contextual filtering. We evaluate the method on a controlled synthetic benchmark containing 1,005 utterances per language in English, Spanish, Catalan, and Basque, with one target medical term per utterance and full lexicon coverage. Across six Whisper sizes, the selected configurations improve term accuracy by 16.12-35.32 percentage points and reduce WER by 0.88-4.01 points relative to WhisperX. LLM filtering reduces WER relative to the phonetic stage in all evaluated configurations, although it reduces term accuracy in three Basque configurations. Perplexity results provide additional evidence of improved linguistic plausibility. The selected configurations require approximately 1.26x-2.38x the baseline inference time. Our code and data are publicly available at https://github.com/NLP-UMUTeam/WhisperX-LLM-TermFixer.

Paper 484 BlockBench: Contamination-Controlled Evaluation of Large Language Models for Code Analysis

Authors: Paul Osemudiame Oamen, Courage Ochuko Mene, Abraham Ugbedeojo Daniel, Ananya Mukherjee, Pius Onobhayedo, Wei Zhao

Abstract How much of large language model (LLM) performance in code analysis reflects memorization of training data rather than genuine reasoning? We introduce BLOCKBENCH, a contamination-controlled evaluation framework that separates memorization effects from analytical capability in smart contract vulnerability detection. We evaluate seven frontier models on 316 vulnerable contracts augmented with 322 semantic-preserving transformations and 100 clean contracts. We observe a substantial performance gap across data exposure levels: detection accuracy drops from 86.5% on likely-contaminated samples to 27.6% on post-cutoff contracts, and further to 0.8% on the most recently disclosed subset. A temporal analysis shows a strong exposure effect, where detection rates increase from 5.9% to 84.7% as contracts become more publicly visible, indicating that benchmark performance is heavily inflated by data contamination. We further identify a pattern matching paradox, where models can often locate security-relevant code regions but fail to explain why they are exploitable, revealing a form of explanation unfaithfulness in code understanding. In contrast, traditional static analyzers show complementary limitations: Slither compiles only 6.6% of post-cutoff contracts, while Mythril compiles all inputs but fails to detect vulnerabilities. BlockBench generalizes to code understanding tasks where semantic equivalence can be verified, enabling more reliable evaluation under potential contamination.

Paper 485 SAFE-KV: Causal Suffix Repair for Behavior-Preserving Non-Prefix KV Cache Reuse in RAG

Authors: Xingjian Zhang

Abstract Retrieval-augmented generation (RAG) reuses the same documents across many requests, making non-prefix KV-cache reuse attractive. Yet a cached document state is not reusable text: it is a context-conditioned causal state whose keys and values depend on the prefix, position, and preceding documents under which it was computed. We formulate non-prefix KV reuse as a behavior-preservation problem with full prefill on the exact prompt as the reference, and introduce SAFE-KV, which repairs a composed cache by re-encoding the causal suffix from a selected start document. On Qwen2.5-7B-Instruct, prefix-conditioned composition preserves only 47.2% of HotpotQA-500 and 37.8% of 2Wiki-500 full-prefill answers; SAFE-KV with risk-guided K2 selection restores 99.8%/98.4% (donor-copy protocol) and 100.0%/98.4% with actual suffix forwarding. Document-only and fixed-window repairs recover far less, because the document states after the repaired region keep histories that omit it. Reuse without repair is exact up to four documents and degrades at sixteen, where suffix repair recovers 98–100%. Faithful repair re-encodes most of the context, so in a single-request L40S study SAFE-KV's operating rule serves every workload by full prefill. SAFE-KV thus establishes the causal-state contract for non-prefix reuse and a validated repair mechanism that meets it.

Paper 489 InLegalLaySum: A Novel Benchmark for Lay Summarization of Indian Legal Documents

Authors: Trishita Saha, Debtanu Datta, SHUBHAM TIWARI, MOHIT KAUSHAL, Priyank Rathore, Mohammad Khalid Pandit, Pawan Goyal, Saptarshi Ghosh

Abstract Legal documents are often lengthy and contain domain-specific terminology and complex legal reasoning, making them difficult for the general public to understand. This challenge is further amplified in the Indian context, where a large portion of the population lacks proficiency in complex English. Although automatic legal summarization has been studied, most existing approaches generate summaries intended for legal professionals and thus remain difficult for laypeople. To address this gap, we introduce InLegalLaySum, the first benchmark for lay summarization of Indian legal documents. InLegalLaySum comprises

9,570 judgements from the Supreme Court and several High Courts of India paired with expert-written summaries in English. We benchmark a wide range of generative models (ranging from 160M to 7B parameters) and perform comprehensive experiments with multiple training strategies, including supervised fine-tuning (SFT) and direct preference optimization (DPO). We observed that the hybrid SFT+DPO training achieved superior performance across readability metrics, which is further validated by domain experts.

Paper 490 WISE-Flow: Workflow-Induced Structured Experience for Self-Evolving Conversational Service Agents

Authors: Yuqing Zhou, Zhuoer Wang, JK, Hong Wang, Samson Koelle, Ziwei Zhu, Wei Niu

Abstract Large language model (LLM)-based agents are widely deployed in user-facing services but remain error-prone in new tasks, tend to repeat the same failure patterns, and show substantial run-to-run variability. Fixing failures via environment-specific training or manual patching is costly and hard to scale. To enable self-evolving agents in user-facing service environments, we propose WISE-Flow, a workflow-centric framework that converts historical service interactions into reusable procedural experience by inducing workflows with prerequisite-augmented action blocks. At deployment, WISE-Flow aligns the agent's execution trajectory to retrieved workflows and performs prerequisite-aware feasibility reasoning to achieve state-grounded next actions. Experiments on ToolSandbox and τ^2-bench show consistent improvement across base models.

Paper 492 Copy Less, Ground More: Selective and Confidence-Aware Decoding for Grounded Legal Generation

Authors: Tan-Minh Nguyen, Truong Dinh Do, Hiep K. V. Nguyen, Le-Minh Nguyen

Abstract While retrieval-augmented generation (RAG) grounds large language models (LLMs) in external knowledge, ensuring faithful evidence utilization in complex domains like law remains challenging. LLMs often ignore retrieved facts or indiscriminately copy from noisy contexts. Existing context-aware decoding methods attempt to mitigate this by amplifying the influence of retrieved content, but they treat the context as a monolithic block—amplifying irrelevant spans alongside evidence—and intervene at every decoding step, incurring substantial inference overhead. To address this, we propose \textsc{SaCoD} (Selective and Confidence-aware Decoding), a training-free inference framework. \textsc{SaCoD} leverages the LLM's internal attention signals to dynamically isolate salient evidence paragraphs at each step, restricting a customized copy mechanism strictly to relevant facts. Furthermore, an entropy-based gate activates this intervention only when the model is uncertain, bypassing copy operations for deterministic tokens to preserve linguistic fluency and cut computational overhead. Experiments on four legal question-answering benchmarks across three open-weight LLMs demonstrate that \textsc{SaCoD} significantly improves faithfulness and response correctness while substantially reducing inference latency compared with prior context-aware methods.

Paper 503 When Do Supervised UQ Ensembles Improve LLM Hallucination Detection? A Robustness Study

Authors: Mohit Singh Chauhan, Vipin Gyanchandani, Dylan Bouchard

Abstract Uncertainty quantification (UQ) methods are widely used for hallucination detection in large language models (LLMs) in closed-book settings where ground-truth evidence is unavailable at inference time. Prior work has proposed combining UQ signals via learned ensembles, but empirical investigations into the robustness of these ensembles are limited. We study a supervised ensembling framework that trains a classifier over heterogeneous UQ-based scorer outputs on a small, domain-specific dataset of labeled LLM responses, then applies it to out-of-sample hallucination classification without retrieval, tools, or

reference documents. Across four LLMs, nine datasets, and three generation regimes (short-form QA, long-form generation, and code generation), we provide a systematic robustness analysis along three axes: sample efficiency, in-domain dataset transfer, and generation regime dependence. We find that supervised ensembles outperform the best individual scorer in 30 of 32 settings, with gains realized from as few as 100 labeled instances. Ensembles retain most of their advantage in cases of in-domain transfer under distribution shift, outperforming the best non-ensemble scorer in 23 of 28 transfer settings. Sampling-based black-box ensembles are nearly as effective as full ensembles, while single-generation white-box ensembles offer limited benefit.

Paper 507 FAIR: Balancing Identity-Associated Training Signals to Mitigate Personalization Bias

Authors: Anvesh Rao Vijjini, Snigdha Chaturvedi

Abstract Personalization of LLMs can greatly improve user experience by catering to individual preferences and contexts. However, unnecessary differences in model behavior across users based on their personal information can be undesirable. Such differences introduce \textit{personalization bias}, leading to inconsistent or unsafe behavior across social attributes such as religion, race, and gender. We introduce \fair (Fairness Aware Imbalance Reduction), a framework for mitigating personalization bias by addressing imbalances in stereotype-related content during preference optimization. \fair includes two variants: \fair_D, which applies data-level intervention based on data augmentation with LLMs, and \fair_T, which incorporates fairness-aware reweighting directly into the optimization objective. We evaluate both variants using multiple personalization bias metrics across diverse identities and models. Across all three LLMs, both \fair methods consistently reduce bias relative to baseline, with up to 20 and 30.1 relative improvement by \fair_D and \fair_T, respectively.

Paper 510 Evidence-Guided Schema Normalization for Temporal Tabular Reasoning

Authors: Ashish Thanga, Vibhu Dixit, Abhilash Shankarampeta, Vivek Gupta

Abstract Temporal reasoning over evolving semi-structured tables poses a challenge to current QA systems. We propose an approach that induces explicit relational structure from the revision history of a semi-structured table: (1) prompting an LLM to synthesize a 3NF-compliant relational schema from a sequence of Wikipedia infobox snapshots, (2) populating the schema to obtain a queryable database, and (3) generating and executing SQL queries against it. Every answer is therefore backed by an executable query that traces to the specific snapshots and fields it depends on, and QA accuracy serves as an extrinsic evaluation of the induced structure. In controlled comparisons across three schema generators and six query models, the choice of schema source accounts for a larger share of the variance in QA accuracy than the choice of query model: changing only the schema shifts exact match (EM) by up to 20 points, whereas changing the query model under a fixed schema shifts it by 4-12 points. From this evidence, we distill three schema-design principles: balanced normalization that preserves context, semantic naming that reduces ambiguity, and consistent temporal anchoring. Our best configuration (Gemini 2.5 Flash schemas + Gemini-2.0-Flash queries) achieves 80.39 EM, an 11.5-point absolute (16.7% relative) improvement over the strongest baseline (68.89 EM).

Paper 513 SAVER: Selective Auditing of Verbal Evidence for Error Recovery in VLM Change Reasoning

Authors: ~Youdi_Li1

Abstract Vision-language models (VLMs) frequently fail at visual change reasoning, even when their vision encoders contain sufficient information. We observe that correct VLM outputs tend to contain explicit verbal evidence (object names, colors, spatial locations) that supports the claimed change, while incorrect outputs often lack such evidence. We propose SAVER (Selective Auditing of Verbal Evidence for Error Recovery), a lightweight, rule-based method that parses VLM responses for this evidence

and triggers structured reprompting only when evidence is missing or inconsistent. Across three change detection benchmarks and four VLMs, SAVER significantly improves accuracy on tasks where errors stem from the model failing to articulate what it saw (expression failures), with gains up to +25.8% on CLEVR-Change. The evidence patterns can also be generated by an LLM in a single call, matching the hand-tuned gate on CLEVR-Change. Ablation experiments confirm that the evidence gate, not reprompting alone, drives the improvement.

Paper 518 When and How to use In-context Learning for modern LVLMs? An Extensive Investigation

Authors: Pengxiang Li, Yu Zhao, Jiaxin Yu, Liang Xie, Erwei Yin

Abstract Multi-Modal In-Context Learning (MM-ICL) is commonly employed to enhance performance of Large Vision-Language Models (LVLMs) on downstream tasks. However, as next-generation models like QwenVL and InternVL have been optimized through instruction tuning and alignment, their baseline capabilities have been strengthened. This necessitates a re-evaluation of the effectiveness and optimal configuration of MM-ICL across different models. By evaluating diverse LVLMs spanning different parameter scales (7B–32B) and accessibility settings (open-source and closed-source) across eight cross-domain datasets, this paper systematically investigates the performance of MM-ICL under varying model capabilities and task settings. Experimental results indicate that, beyond validating established patterns, next-generation models exhibit distinct phenomena: in certain logical reasoning and format-sensitive tasks, the introduction of MM-ICL does not necessarily yield improvements and can even lead to performance degradation in specific scenarios. We analyze these observations and, based on extensive experimental data, provide new recommendations for selecting MM-ICL strategies for LVLMs.

Paper 522 Analyzing Multilingual Parametric Knowledge Retrieval with Sparse Autoencoders

Authors: Marc E. Canby, Ikhyun Cho, Julia Hockenmaier

Abstract Multilingual language models can answer factual questions across languages, but their answers are not always consistent across languages, and the internal pathways used to retrieve this knowledge remain unclear. A model may rely on language-specific shortcuts, shared multilingual representations, or another language as an internal pivot. We investigate these possibilities in Llama-3.1-8B and Gemma-2-9B with sparse autoencoders (SAEs), identifying, analyzing, and causally intervening on language-associated features during factual recall. On CUS-QA \cite{libovicky2026}, a parallel QA dataset spanning Czech, Slovak, Ukrainian, and English, we identify SAE features associated with particular subsets of languages. We then decompose each question's activations into these subsets to measure reliance on input-language, multilingual, or pivot-like representations across layers. Across both models, factual recall is heterogeneous: different input languages and questions exhibit different language-associated pathways. Steering substantially changes cross-lingual answer alignment and sometimes increases correctness in Llama, while it produces only small changes in Gemma. These results show that multilingual factual recall is mediated by identifiable SAE features that can be used to diagnose and control how models access knowledge across languages.

Paper 525 DECK: A Consistency × Confidence Taxonomy of LLM Hallucinations

Authors: Mohit Singh Chauhan

Abstract Existing hallucination taxonomies classify LLM errors by what is wrong with the output — memorised misconceptions, reasoning failures, fluent fabrications — but cannot answer a different question: which uncertainty scorer would have caught this error? We propose a complementary taxonomy that classifies errors by their detectability signature, the signal a scorer family would read. The DECK taxonomy is a 2×2 partition along inter-sample consistency and token-level confidence into four regimes (Drift, Entrenched, Confabulation, Knotted) that yields a falsifiable blind-spot map: black-box consistency scorers have signal in D and C, white-box token-probability scorers in K and C, and only an LLM-as-a-Judge with independent pretraining can detect

E. Across three models and four short-form QA datasets we test this map two ways: judge-involving scorer disagreements concentrate in each family's predicted blind-spot cells, and external labels (SelfAware unanswerable, HaluEval adversarial, PopQA entity popularity) land in the predicted cells, robustly to cross-fitted thresholds. We further identify a universal blind spot of output-level UQ: on knowledge-gap inputs where the generator emits confident, repeatable fabrications, every output-level family collapses by construction. A linear probe on Llama-3-8B's final-layer hidden states also falls to chance, with or without quantisation, though an intermediate layer retains weak signal.

Paper 527 DEAF: A Benchmark for Diagnostic Evaluation of Acoustic Faithfulness in Audio Language Models

Authors: Jiaqi Xiong, Yunjia Qi, Qi Cao, Yu Zheng, Yutong Zhang, Ziteng Wang, Ruofan Liao, Weisheng Xu, Sichen Liu

Abstract Recent Large Audio Language Models (LALMs) demonstrate impressive performance on speech benchmarks, yet it remains unclear whether these models genuinely process acoustic signals or rely on text-based semantic inference. To systematically study this question, we introduce DEAF (Diagnostic Evaluation of Acoustic Faithfulness), a bench- mark of over 2,700 conflict stimuli spanning three acoustic dimensions: emotional prosody, background sounds, and speaker identity. Then, we design a controlled multi-level evaluation framework that progressively increases textual influence, ranging from semantic conflicts in the content to misleading prompts and their combination, allowing us to disentangle content-driven bias from prompt-induced sycophancy. We further introduce diagnostic metrics to quantify model reliance on textual cues over acoustic signals. Our evaluation of seven LALMs reveals a consistent pattern of text dominance: models are sensitive to acoustic variations, yet predictions are predominantly driven by textual inputs, revealing a gap between high performance on standard speech benchmarks and genuine acoustic understanding. Code and benchmark resources are publicly available at: https://github.com/kikixiong/DEAF

Paper 529 Interactive and Scalable Evaluation of LLM Math Tutoring via Simulated Teacher-Student Dialogues

Authors: Gihyeon Choi, Harksoo Kim

Abstract Evaluating LLMs as math tutors is challenging because many automated protocols either score a single response against a reference or place the model in a fixed dialogue context, without measuring its ability to build and steer tutoring interactions. We present a fully automated, simulation-based framework that generates controlled multi-turn teacher--student dialogues with turn-level answer verification under a fixed interaction budget. This budget specifies the maximum number of student--teacher rounds allowed per problem (K=5 in our main experiments). To prevent score inflation from answer-revealing feedback, we introduce an overinformative eligibility filter: a hint-only student test applied to the teacher utterance immediately preceding the student's first correct attempt. Only eligible corrections receive credit in our effectiveness and efficiency metrics, both of which are normalized by the number of initially incorrect sessions. We also propose a diagnostic metric suite separating tutoring effectiveness/efficiency (ECR/FEI), answer reachability (ARG), and teacher judgment (correctness/incorrectness F1). Experiments on GSM8K and MATH with diverse teacher and student models show that problem-solving accuracy is an informative but incomplete proxy for interactive tutoring ability, and that filtering overinformative feedback can substantially alter apparent performance. Our framework enables scalable, reproducible, and fine-grained evaluation of LLM math tutoring beyond outcome-only scoring.

Paper 533 OMIT the Action: Measuring Framing-Invariant Omission Bias under Philosophical Disagreement

Authors: Sihyeon Lee, Jihun Song, Chanwoo Kim, Jiwoo Kum, Chanjun Park

Abstract As LLMs increasingly assist in moral reasoning, omission bias, the tendency to prefer inaction even when equivalent framings reverse substantive outcomes, poses a significant risk of skewed decision-making. Yet omission bias remains underexplored in LLM evaluation, with the few existing studies limited in scale and focused largely on utilitarian-deontological conflicts. To address this gap, we introduce OMIT, a benchmark consisting of 218 paired-frame scenarios across 10 conflict types, constructed by leveraging disagreement patterns from an LLM-based, five-perspective philosophical persona panel (utilitarianism, deontology, virtue ethics, care ethics, and contractualism). Evaluating eight LLMs, we find that omission bias is pervasive but inversely correlates with model size within families. We further evaluate four inference-time interventions and find that interventions encouraging models to consider moral principles before committing to a yes/no answer reduce omission bias and increase frame-consistent responses, although lower omission bias rates can also coincide with shifts toward action-biased responses. Ultimately, this work contributes not only the OMIT benchmark, but also a methodology for using diverse philosophical disagreement signals to evaluate framing-sensitive inaction preferences and the distributional effects of mitigation attempts in LLMs under complex moral conflicts.

Paper 534 Structured but Silent: Probing Capability Requirements in LLM Hidden States

Authors: ~Kyojun_Choo1, ~Minsoo_Song2, ~Yunju_Kang1, ~Chanjun_Park1

Abstract Reliable tool use requires more than triggering a mechanism or matching a query to an API description. Before selecting a specific tool, an agent must first infer the capability requirements implied by the user query. In this paper, we investigate whether these query-side capability requirements are linearly decodable from LLM hidden representations prior to generation, and how this hidden-state accessibility compares with explicit verbal classification. We introduce TACIT, a framework that decomposes external requirements along three fundamental axes: Source, Transformation, and World Effect, defining eight structurally distinct capability classes. Using 1,600 balanced training queries from benchmarks, synthetic examples, and new domain scenarios, we train linear probes on pre-generation hidden states from four open-weight LLM families. Our empirical results demonstrate that fine-grained capability structures are linearly decodable with high accuracy across all models. Crucially, however, we expose a representation-to-verbalization gap: these same models are significantly less reliable when asked to explicitly classify the same queries in natural language. This disconnect indicates that information about required external capabilities is linearly accessible in LLM hidden representations but not reliably expressed—a phenomenon we define as structured but silent.

Paper 539 CASCADE: Cognitive Assessment via Sentence Context and Distortion Extraction

Authors: Archit Narumanchi, Azamat Rashidov, Jong C. Park

Abstract Cognitive distortions are unhelpful thinking patterns central to Cognitive Behavioral Therapy (CBT) and closely linked to mental health outcomes. Despite their context-dependent nature, studies at the intersection of Natural Language Processing and Cognitive Behavioral Therapy are centered largely around sentence-level utterances, ignoring the document-level complexity of real clinical text. To this end, we propose a two-stage rank-then-classify cascade that decomposes document-level multi-label cognitive distortion detection into (1) a sentence-ranking stage that identifies the most likely distorted sentences, and (2) a multi-label classification stage that assigns distortion subtypes to the selected sentences by incorporating their surrounding interleaved context. While the pipeline stages are trained on separate out-of-domain datasets, our model matches state-of-the-art performance on the \textsc{CBT-CD} benchmark of 166 expert-annotated clinical vignettes. Operating with only 368M

parameters, our system achieves a 51.8\% weighted F1 (9-class), surpassing massive large language model baselines such as Mistral-7B-v0.3 (46.0\%) and performing on par with state-of-the-art models such as Gemma 4 26B A4B (50.8\%).

Paper 540 One Prompt Does Not Fit All: Self-Meta-Evolve for Personalized Information Extraction

Authors: Hongliang Li, Lu Wang, Yong Xu, Hanyang Chen, Zhitao Hou, Xiaoting Qin, Song Ge, Qingwei Lin, Dongmei Zhang

Abstract Large language models (LLMs) are increasingly deployed for enterprise information extraction (IE), where the same document must be reorganized differently for each user. Existing prompt optimization methods, however, rely on a single prompt optimized against a global objective, which is misaligned with the inherent user heterogeneity of real workplaces. We formulate enterprise IE as per-user prompt adaptation under interaction feedback and propose Self-Meta-Evolve, a hierarchical framework that maintains a dedicated prompt for each user and continuously refines it through a dual-loop process: an inner loop that edits structured prompts based on persona-conditioned feedback, and an outer loop that evolves the meta-prompt itself by distilling successful editing patterns. To enable scalable training and evaluation, we release a persona-driven IE benchmark of 292 simulated enterprise users, paired with a reproducible persona-generation pipeline grounded in O*NET occupational taxonomies. On this benchmark, Self-Meta-Evolve achieves a 74.58% success rate, outperforming the strongest prompt-optimization baseline by 13.56 absolute points, and reaches 52.54% within only two iterations. A double-blind human study with twenty real professionals further confirms that prompts adapted by our framework win against static baselines in 71% of pairwise comparisons.

Paper 542 Models Know Their Shortcuts: Deployment-Time Shortcut Mitigation

Authors: Jiayi Li, Shijie Tang, Gün Kaynar, Shiyi Du, Carl Kingsford

Abstract Pretrained text encoders are prone to shortcut learning, relying on token-label correlations that fail once the distribution shifts in deployment. Existing shortcut mitigation methods mainly operate at training time and assume access to training data, training dynamics, or shortcut annotations, which are hardly available during deployment, where only the converged model remains. We show that this model alone suffices to mitigate shortcuts during deployment: a biased model internalizes a signal of its learned shortcuts that can be captured via unsupervised gradient-based attribution. We further prove that deployment-time mitigation is information-theoretically upper-bounded by training-time mitigation. Nevertheless, exploiting this gradient signal, our proposed unsupervised deployment-time shortcut mitigation framework for pretrained text encoders, Shortcut Guardrail, recovers substantial performance under shortcut distribution shift, matching or outperforming training-time baselines across sentiment classification, toxicity detection, and natural language inference.

Paper 543 Entailed Opinion Matters: Improving the Fact-Checking Performance of Language Models by Relying on their Entailment Ability

Authors: Gaurav Kumar, Ayush Garg, Debajyoti Mazumder, Aditya Kishore, Babu kumar, Jasabanta Patro

Abstract Automated fact-checking has been a challenging task for the research community. Prior work has explored various strategies, such as end-to-end training, retrieval-augmented generation, and prompt engineering, to build robust fact-checking systems. However, their accuracy has not been high enough for real-world deployment. We, on the other hand, propose a new learning paradigm, where evidence classification and entailed justifications made by generative language models (GLMs) are used to train encoder-only language models (ELMs). We conducted a rigorous set of experiments, comparing our approach with recent works along with various prompting and fine-tuning strategies. Additionally, we performed ablation studies, error analysis,

quality analysis of model explanations, and a domain generalisation study to provide a comprehensive understanding of our approach.

Paper 545 COAT: An Orthogonal Benchmark for Diagnosing Chinese Adversarial Toxicity Detection Failures

Authors: Jiaqi Hu, Mingshuai Liu, Haobo Wu, Xin Li

Abstract Character-level perturbations are widely used to evade Chinese toxicity detection, yet existing benchmarks either cover only a limited range of perturbations or expand perturbation coverage without balanced construction across toxic and non-toxic samples, leaving these two requirements unmet within a single benchmark. We introduce COAT (Chinese Orthogonal Adversarial Toxicity), a 41K-sample benchmark spanning eight perturbation types that enforces perturbation orthogonality: all perturbations are applied with equal probability to both toxic and non-toxic texts, ensuring P(toxic | perturbed) = 0.5. As a core diagnostic layer, we annotate Natural Language Variant Ambiguity (NLVA): 21.6% of clean samples naturally contain surface features overlapping with perturbation artifacts, enabling controlled analysis of model sensitivity to perturbation-like features and classification difficulty. Evaluating five encoder classifiers, three proprietary LLMs, and Llama3-8B with chain-of-thought reasoning, we find that (1) models remain systematically sensitive to perturbation-like features even under balanced training, and (2) across Llama3-8B variants, overcorrection varies systematically with the linguistic ambiguity introduced by each perturbation type. Step-level diagnosis localizes the shared bottleneck to semantic restoration, with targeted preference optimization yielding reliable gains on the most ambiguous types.

Paper 546 Mass-Tail Attribution Entropy: A Regime-Conditional Decoder for Vision-Language Model Hallucination

Authors: Jaegwang Shin, Suan Lee

Abstract Decoding-time mitigation of vision-language model (VLM) hallucination applies one fixed intervention and typically trades recall for lower hallucination. We trace this trade-off to treating hallucination as a single failure mode. Measuring the normalised entropy H of leave-one-out attributions over panoptic image segments, we find that a VLM's inference mass concentrates in one task-conditional tail of H: captioning in the over-concentration tail (low H, 53–66%), binary question-answering in the over-spread tail (high H, 74–95%). The pattern holds in 6/6 (backbone × benchmark) cells on LLaVA-1.5-7B and Qwen2.5-VL-7B, survives exact Shapley recomputation, reproduces within AMBER, and is shared by three further backbones. Because errors and correct outputs share the tail—H is a regime signal, not a hallucination detector—an intervention matched to the wrong tail actively hurts: blank-image contrastive decoding raises captioning CHAIR_S from 49.3 to 60.3. The regime can be read without attribution at inference, from blank-image agreement and on-image token margin; SBC (Segment-Based Conditioning) uses them to route between segment boosting and probabilistic marginalisation. With one shared configuration, SBC lowers CHAIR_I while preserving recall and raises POPE F1 on both backbones (+1.95 / +3.31), at 1.7–1.9× greedy cost on captioning and 4.9× on single-token QA.

Paper 553 IndexRAG: Index-Time Reasoning for Multi-Hop Retrieval-Augmented Generation

Authors: Zhenghua Bao, Yi Shi

Abstract Multi-hop question answering (QA) requires reasoning across multiple documents, yet existing retrieval-augmented generation (RAG) approaches address this either through graph-based methods requiring additional online processing or iterative multi-step reasoning. We present IndexRAG, a novel approach that shifts cross-document reasoning from online inference to offline indexing. IndexRAG identifies bridge entities shared across documents and generates bridging facts as

independently retrievable units, requiring no additional training or fine-tuning. Experiments on three widely-used multi-hop QA benchmarks (HotpotQA, 2WikiMultiHopQA, MuSiQue) show that IndexRAG improves F1 over Naive RAG by 4.6 points on average, while requiring only single-pass retrieval and a single LLM call at inference time. When combined with IRCoT, IndexRAG achieves the best average performance among all evaluated methods, including graph-based baselines such as HippoRAG2 and FastGraphRAG, while relying on a flat vector index. Our code is available at https://github.com/Continuum-AI-Corp/IndexRAG.

Paper 556 GPC-E4: Generalized Consistency and Entropic Evidence for Backdoor Detection in Vision Language Models

Authors: ~Xuan_Wang17, ~Xuehu_Yan1, ~Yu-liang_Lu1, ~Kejiang_Ye1, ~Cheng-zhong_Xu1, ~Xitong_Gao1

Abstract Vision-language models (VLMs) are increasingly used through black-box, open-ended interfaces, where input-level backdoor screening must work without model parameters, gradients, logits, or training data. We present GPC-E4, a detector for single image-prompt queries that probes a VLM with mild, semantics-preserving visual and textual perturbations. The method measures whether the output behavior contracts unusually under these probes: GPC captures dominant-mode or semantic consistency, while E4 summarizes predictive entropy, augmentation mutual information, minimal collapse, and local slope along perturbation paths. A fixed robust normalization combines these signals and can operate without a known-clean calibration set. Across COCO, Flickr30k, Caltech101, and MMBench-style evaluations with LLaVA-1.5, MiniGPT-4, CLIP, CoOp, and Qwen3-VL, GPC-E4 outperforms STRIP, SCALE, TeCo, and BDet in both closed-set and open-ended settings. Additional analyses under alternative stressors, prompt and distribution shifts, zero clean calibration, VLM-native attacks, and detector-aware attacks characterize when the contraction signal is reliable and when it weakens.

Paper 562 TripPattern: A Pattern-based Text Watermarking Method for Large Language Models

Authors: Sangjun Moon, Dasom Choi, Jingun Kwon, Hidetaka Kamigaito, Taro Watanabe, Manabu Okumura

Abstract Text watermarking techniques have gained significant attention for identifying machine-generated text and mitigating risks from large language models (LLMs). Existing methods typically divide an LLM's vocabulary into green and red tokens, but encouraging generation toward green tokens can reduce text quality and naturalness. To address this, we propose TripPattern, a watermarking framework that formulates text watermarking as a pattern-based matching task using three vocabulary partitions. TripPattern divides the vocabulary into one neutral group and two pattern groups. During generation, the model alternates token selection between the two pattern groups to embed detectable patterns, while neutral tokens are selected independently to improve flexibility and preserve naturalness. For detection, TripPattern uses pattern-based statistical tests that provide interpretable p-values by measuring how often adjacent tokens alternate between the pattern groups. Theoretical analysis and empirical evaluations on four multilingual datasets show that TripPattern maintains LLM generation quality while achieving robust watermark detectability.

Paper 565 IMFD: End-to-end Multi-Face Forgery Detection through Instruction-based Large Vision-Language Models

Authors: Dasom Choi, Sangjun Moon, Hyeongchan Im, Jaeeon Park, Jingun Kwon, Hidetaka Kamigaito, Taro Watanabe, Manabu Okumura

Abstract The rapid increase of deepfakes has raised significant concerns due to their spread on social media. Traditional multi-face forgery detectors crop and verify each face independently, ignoring background context and inter-face relationships, which

often yields suboptimal performance. To overcome these limitations, we leverage instruction-based Large Vision–Language Models (LVLMs), which can interpret entire images and follow complex textual instructions. We propose a simple yet effective single-stage multi-face forgery detector, called IMFD (Instruction-based Multi-face Forgery Detector), which is trained end-to-end to jointly localize faces and predict per-face forgery labels. Rather than treating face box prediction only as a joint objective, IMFD explicitly integrates predicted face bounding boxes into the instruction as visual cues that enhance instruction grounding and forgery detection. To support the training and evaluation of IMFD, we convert existing multi-face forgery datasets into an instruction-based format. Experimental results and analyses show that IMFD improves multi-face forgery detection by integrating face bounding boxes into the instruction, and consistently outperforms various state-of-the-art methods.

Paper 570 Efficient Many-Shot In-Context Learning with Task-Level Compression and Instance-Level Retrieval

Authors: Masayuki Kawarada, Ukyo Honda, Sho Takase

Abstract Many-shot in-context learning (ICL) lets large language models exploit hundreds or thousands of demonstrations, but its cost grows with the prompt. We argue that many-shot demonstrations carry two kinds of information: task-level information, such as task structure, which emerges only across many examples, and instance-level information, such as exact input-output associations and rare exceptions, which resides in individual demonstrations. Existing efficiency methods commit all information to a single placement: demonstration retrieval keeps a few demonstrations as raw text, while context compression places the full pool into a compact memory. We study where each kind of information should be placed. On ManyICLBench, each one-sided method falls short on tasks that depend on the information it discards, while combining the two retains much of the full many-shot performance at a much smaller inference budget. Synthetic tasks further show that compressed memory and retrieved examples can complement each other but sometimes interfere.

Paper 573 Beyond EOS: Revealing the Internal Monologue of Language Models

Authors: Aishwarya Maheswaran, Maunendra Sankar Desarkar

Abstract Outputs generated by LLMs reflect traits present in the training data. However, it is unclear how the model acquires these traits and steers itself in that direction. We demonstrate that allowing the model to continue generating beyond the <eos> token reveals its comprehension of the task in the form of specific 'Notes' or Rules distinct from normal generation. We refer to this as the model's Internal Monologue (IM) and demonstrate its utility to understand the model's perception of the task it is being finetuned on. We find this monologue is transient and dissipates as training progresses. We hypothesize this is a inherent behavior of instruction tuned models that adapts to reflect task instructions until the latent space beyond <eos> is sufficiently altered. Importantly, we do not ask the model to reason or do anything special to invoke this. We perform analysis on four diverse dialogue datasets from the Counseling domain to compare how dataset differences affect the internal monologue during early finetuning. We then show this is a general phenomenon by testing with three models of varying sizes and families and other domain datasets. We demonstrate the utility of IM as a means to gain insight on the model's perception of the dataset and task through clustering based visualizations.

Paper 578 Polar: A Benchmark for Evaluating Political Bias in LLMs

Authors: Sangho Kim, Heejin Kim, Hyunggeun Jeon, Yoonhee Park, Jaejin Lee

Abstract Political bias in large language models (LLMs) is increasingly significant, but difficult to measure reproducibly across political and linguistic contexts. We introduce Polar, a 4,026-instance multiple-choice benchmark that measures political bias

through option-level likelihoods rather than prompt-based generation. Polar covers two ideological axes and eight issue categories derived from the Manifesto Project, and evaluates models in parallel across U.S. and South Korean political contexts. Across 38 LLMs, measured bias varies systematically with political context, issue category, model group, and presentation language. All models lean left-progressive on U.S. political content, but show more centered and mixed patterns on South Korean content. Translation experiments further show that presentation language alone can shift measured bias. These findings highlight the need for multilingual and cross-contextual evaluation of political bias in LLMs.

Paper 580 Reducing Logprob Without Causal Verification: Diagnosing SAE-Based PII Suppression in LLMs

Authors: CHAERIN LEE, Wonho Lee, Sohee Park, Daeseon Choi

Abstract SAE-based feature-level interventions to suppress PII leakage in LLMs are primarily evaluated by log-probability reduction. We show that this criterion cannot distinguish feature-specific suppression from a magnitude artifact: a direction-agnostic effect of sufficiently large activation perturbations. Across 36 (model, layer, PII type) cases, we compare selected SAE feature interventions with coherence-matched random direction controls using logprob and rank-aware metrics. While 17 cases (47\%) appear successful under logprob-only evaluation, only 1 case (2.8\%) passes our multi-metric framework. The rest reveal two failure modes: random directions outperform selected features, and feature selection collapse across PII types. These results indicate that existing evaluations cannot verify whether interventions target causal pathways of PII leakage. A 9,000-run causal mediation feasibility study corroborates this: the verifier separates candidates reliably, but correlation-selected candidates rarely pass it. We therefore propose a new research mission for NLP: causally grounded PII suppression through circuit-level localization and causal mediation analysis.

Paper 585 GRAIN: Gradient-Based Interaction Graph of Neurons for Explainable Misinformation Detection

Authors: Rathod Darshan D, Rahul Mishra

Abstract Prior work on misinformation detection often applies large language models (LLMs) in a black-box manner, relying on knowledge implicitly encoded in model parameters or on pre-hoc and post-hoc reasoning techniques. However, the opaque nature of LLMs limits access to their internal decision-making processes, and existing approaches largely overlook the rich signals embedded within their hidden layers, instead focusing primarily on surface-level generative capabilities. To address these limitations, we propose GRAIN, an explainable misinformation detection framework that explicitly models neuron–neuron interaction graphs using dominancy scores. Experimental results across multiple benchmark datasets show that GRAIN consistently outperforms state-of-the-art baselines while providing interpretable insights into the neuronal patterns and cues that distinguish misinformation from factual content. Furthermore, we introduce NNI-IG, a GRAIN-guided variant of Integrated Gradients that leverages critical neurons to produce sharper and more focused token-level attributions, further enhancing model interpretability.

Paper 587 SIEVE: Slot-wise Evaluation of Information-Role Retention in Fixed-Length Meeting Summarization

Authors: Jumpei Nagasawa, SEIYA SHIBAYAMA

Abstract Existing meeting-summary evaluations mainly ask whether a summary resembles a reference or is consistent with the source. They do not directly show which information roles needed for decision understanding are retained or dropped. We propose SIEVE, a method that evaluates meeting summaries from seven information-role perspectives. For each role, SIEVE separately measures Presence, whether the role appears in the summary, and Fact Check, whether the retained content is free

of contradictions with the source. This produces a role-level retention profile, allowing summarizers and length limits to be compared by what they preserve and lose rather than by a single aggregate score. We apply SIEVE to 2,250 summaries from 50 MeetingBank municipal-meeting segments, generated by three models under 2-, 3-, and 6-sentence limits. The role-level profiles show that summaries more reliably retain topic, outcome, and action-detail information, while more often dropping background and concerns that shaped the discussion, as well as unresolved items left open afterward. Reference-similarity metrics widely used to evaluate MeetingBank summaries (ROUGE-L, BERTScore) correlate weakly or negatively with SIEVE Presence. Consistent with this, the dataset-provided reference summaries against which similarity is computed contain topic and action-detail information but few of the other roles. Reference similarity and the role retention that SIEVE measures are different perspectives on summary quality. SIEVE therefore supports auditing, selecting, and designing meeting summarizers based on which decision-relevant information survives compression.

Paper 594 Encoded but Not Routed: Explaining the Table-Chart Gap in Scientific Claim Verification

Authors: Sunisth Kumar, Xanh Ho, Tim Schopf, Andre Greiner-Petter, Florian Boudin, Akiko Aizawa

Abstract Multimodal LLMs are increasingly used to assist scientific peer review, where a core requirement is verifying whether claims in a paper are supported by its evidence. Prior work has shown that models perform substantially better at this task when the evidence is a table than when it is a chart of the same underlying data. This raises the question of whether models fail to extract information from charts, or do they extract it but fail to use it when forming their prediction? We study this question through layer-wise linear probing and attention analysis on three open-weight VLMs over table and chart evidence, representing the same underlying data. We find consistent evidence for the latter. Chart information is encoded in the models' intermediate representations but does not reach the prediction position, a gap that is absent for tables and holds across all conditions tested. Attention analysis further reveals that this disconnect takes two architecturally distinct forms across model families. These findings point toward reframing the table-chart gap as a failure of how encoded visual information is used at prediction time, rather than a failure of encoding itself.

Paper 597 Chained Cross-Modal Learning for Modality-Agnostic Embeddings

Authors: Naoki Tamura, Taiki Sekii

Abstract This study proposes a cross-modal learning framework that learns modality-agnostic embeddings for six modalities: language, temporal 2D/3D skeletons, IMU, GPS, and Wi-Fi. Existing approaches such as ImageBind align each target modality with images. Consequently, they are restricted to modalities for which paired image data can be prepared a priori, which makes them unsuitable for time-series tasks, such as action recognition. To overcome this limitation, we propose "chained modal alignment," which aligns the six modalities sequentially via temporal 3D skeletons, thereby enabling emergent zero-shot action recognition based on a shared embedding space. Because the proposed method relaxes the requirement of direct pairing with images, it supports a wider range of modality combinations and thus expands practical applications. In addition, we introduce "chained modal transfer learning," which transfers the knowledge learned between language and 3D skeletons to other modalities with data that cannot be directly paired, thereby improving generalization. Experiments on three public datasets spanning all six modalities quantitatively validate the effectiveness of the proposed method for emergent zero-shot action recognition.

Paper 598 SASU: Structure-Anchored Semantic Unification for Code Search

Authors: Qian Zhang, Hao Li, Hao Jiang

Abstract Code search aims to retrieve code snippets according to natural-language intent. Despite the progress of pretrained code models, there remains a gap between natural language and programming language: descriptions are usually sequential and lexical, while code is governed by syntax, data dependencies, and execution logic. Existing retrieval models often rely on token-level matching or use structure as auxiliary information, which limits their ability to capture executable behavior under lexical variation or incomplete code evidence. In this paper, we propose SASU, a structure-anchored semantic unification framework for code search. SASU extracts a normalized multi-relational graph for each function and composes syntax, data-flow, control-flow, call, return, and exception relations into a unified structural view. The structural representation is then used to guide node-level lexical evidence selection before code and text are aligned in the same retrieval space. Experiments on CodeSearchNet show that SASU achieves 0.835 MRR for code-to-text retrieval and 0.840 MRR for text-to-code retrieval across Python, Java, and JavaScript. Under 40–60% node deletion, SASU retains 86.6% of its full-graph MRR, indicating that structure provides useful retrieval signals beyond lexical pretraining.

Paper 600 EMR: Self-Evolving Medical Multi-Agent System via Experience Mining and Reuse

Authors: Dongsheng Shi, Yue Li, Xin Yi, Linlin Wang

Abstract Large language model (LLM) driven multi-agent systems have shown promise in complex clinical reasoning, yet existing approaches rely on static strategies and lack persistent clinical memory, preventing self-evolving from prior diagnostic successes and failures. We present EMR, a self-evolving medical multi-agent system via Experience Mining and Reuse. EMR introduces a hierarchical clinical experience library that organizes accumulated knowledge into three levels: clinical principles, diagnostic patterns, and representative cases. During inference, EMR emulates multidisciplinary consultation: a planner agent coordinates domain-specific department agents for specialized reasoning, while a summary agent synthesizes their analyses into a final decision. Critically, EMR automatically extracts correct diagnostic insights and failure-related warnings from multi-agent reasoning trajectories, incrementally updating the experience library to guide future cases. Experiments on medical reasoning benchmarks demonstrate that EMR consistently outperforms state-of-the-art medical multi-agent baselines. Further analysis reveals that the hierarchical experience enables cross-specialty generalization and transfer across diverse LLM backbones, offering a scalable and interpretable pathway toward continually evolving intelligent medical reasoning.

Paper 603 Psychometric Benchmark and Bias Assessment in Open-Source Pre-trained NLI Models

Authors: Fadi Amun, Ofir Cohen, Aviad Elyashar, Odeya Cohen, Rami Puzis

Abstract Pre-trained language models harbor latent psychological biases inherited from training data, which remain largely unmeasured, creating risks in areas ranging from mental health support to professional decision making. In this study, the evaluation of model psychometrics and assessment of latent constructs framework is employed to evaluate psychometric biases in 334 Hugging Face natural language inference models, assessing the models' anxiety, depression, sense of coherence, sexism, compassion, and personality. Models consistently score above zero on positively valenced constructs (e.g., compassion, sense of coherence) and below zero on negatively valenced constructs (e.g., anxiety, depression), favoring positive over negative psychological language. Architecture and size are associated with systematic differences: DeBERTa models show elevated negatively valenced constructs (e.g., PHQ $\delta = +0.45$, GAD $\delta = +0.40$) and reduced positively valenced constructs (e.g., Extraversion $\delta = -0.61$), while small models show the opposite pattern (e.g., Extraversion $\delta = +0.30$, PHQ $\delta = -0.22$), with both effects surviving multivariate control. Among the 105 models reporting training data, fine-tuning on adversarial natural language

inference was associated with elevated depression (δ = +0.65), while year of release showed no significant effects after multivariate control. These results suggest that psychometric profiling is a critical, yet missing, layer of safety auditing for NLI models.

Paper 606 Not All Graphs Are Maps: A Task-Agnostic Exploration Framework for Mobile GUI Navigation

Authors: Adamo Orsini, Seyed Mehdi Ayyoubzadeh, Sadman Rakib Pinon, Hannuo Zhang, Koosha Gholipour Baradari, YUANHAO YU, Xinxin Zuo, Yang Wang

Abstract Modern mobile GUI agents can perceive visible screens and execute local actions, but they still lack persistent app-specific navigation knowledge: where functions live, which pages contain them, and which screens are equivalent across repeated visits. We introduce TAG-Nav, a Task-Agnostic Graph Navigation framework that wraps existing GUI-agent backbones with reusable destination memory. Offline exploration is distilled into compact records of app functions, target pages, and waypoint routes before any evaluated instruction is seen. When a future task's destination is covered by this memory, TAG-Nav can retrieve the destination and provide route guidance without requiring a task-specific demonstration. This separates mobile GUI destination navigation from full task completion: many tasks share the prefix of reaching the right page, while final actions may enter content, send data, make purchases, delete information, or change sensitive settings. To evaluate route finding without downstream execution confounds, we introduce DestBench, a navigation-only protocol that asks whether an agent reaches the correct pre-action destination screen. DestBench enables a controlled graph-memory study comparing no-memory agents with random, automated BFS, and clean graph memory. Graph-derived memory improves strict navigation success rate of UI-TARS-1.5-7B from 52.4% without graph memory to 58.3% with automated BFS memory, and improves Qwen-VL-Max success rate from 59.5% to 71.4% on the same conditions. Together, TAG-Nav and DestBench show that destination coverage and route cleanliness strongly shape graph-memory value, making graph quality a measured variable rather than an assumed benefit of exploration.

Paper 611 Skin-Deep: A Geometric Diagnostic for Alignment Fragility in Large Language Model Representations

Authors: Dongyub Jude Lee, Jungseob Lee, Seungyoon Lee, Seongtae Hong, Suhyune Son, Sugyeong Eo, Jaehyung Seo, Heuiseok Lim

Abstract Refusal on a safety benchmark does not reveal how stable that behavior will remain after model updates. Benign downstream fine-tuning can weaken refusal, yet behavioral evaluations typically expose this fragility only after an intervention. We introduce SKIN-DEEP, a geometric diagnostic that examines the unmodified model's residual-stream activations. It compares aligned and base checkpoints to identify safety-separating directions, tests their behavioral relevance through ablation, and summarizes the layer-wise pattern in the Geometric Fragility Score (GFS). Across twenty-one instruction-tuned models, harmful requests and benign instructions exhibit a recurring low-rank separation pattern. Selected direction ablations weaken refusal, with the effective direction varying across models. In benign low-rank fine-tuning experiments, the initially safe model with the lowest score before fine-tuning has the lowest harmful-compliance rate when trained on the largest tested set of harmless examples. These findings connect representation geometry to subsequent behavioral susceptibility and support activation-based diagnostics as a complement to refusal tests. Our code is available at https://github.com/js-lee-AI/skin-deep.

Paper 612 Same Benchmark Score, Different Failures: Atomic Capacities for Diagnosing Language Model Failure Modes

Authors: Xingwei Qu, Xiao Wang, Joseph James, Tyler Loakman, Yang Wang, songxin wang, Chenghua Lin

Abstract Benchmarks typically report aggregate scores over broad task domains such as knowledge or reasoning, but these scores do not reveal which behaviour failed or what intervention would repair it. We propose \emph{atomic capacities}: paired minimal perturbations that isolate repair-relevant failure modes. We instantiate nine atomic capacities spanning \textit{Knowledge}, \textit{Reasoning}, and \textit{Hybrid Composition}, and evaluate them on paired original--variant prompts using behavioural log-probability measures. Mainly focusing on Qwen3 models, we find that models with similar domain-level performance can exhibit substantially different capacity profiles, implying different repair targets. We further show that corruption perturbations separate into two distinct regimes: some reduce confidence in the gold answer without changing the preferred decision, while others flip the model toward the planted incorrect answer. Finally, we show that a standard zero-shot Chain-of-Thought intervention is not uniformly beneficial across capacities and can worsen robustness under misleading cues. Our results suggest that benchmark categories are too coarse for diagnosis and that model evaluation should instead target repair-relevant atomic capacities.

Paper 614 ArabMisinfo: A Decade-Long Arabic Facebook Corpus for Misinformation Discourse and Engagement Analysis

Authors: Wajdi Zaghouani, Shimaa Ibrahim, Mabrouka Bessghaier, Md. Rafiul Biswas

Abstract We present ArabMisinfo, a corpus of 274,837 public Facebook Page posts in Arabic spanning 2014-2024 and focused on news and public affairs. Each record includes page-level metadata, UTC timestamps, link domains, and platform-native engagement signals (likes, comments, shares, and the seven post-2016 reaction types). We accompany the corpus with reproducible diagnostics for engagement and polarization, the Polarization Index (PI), Share-to-Like Anomaly (SLA), Risk Triage Index (RTI), and Engagement Anomaly Rule (EPR). It provides scripts that deterministically reproduce all tables and figures. To address the absence of content-level labels, we contribute a discourse role annotation of 10{,}000 posts (3.6% of the corpus) using Fanar-C-2-27B, classifying each post into one of six roles. The annotation reveals that Arabic Facebook misinformation discourse is dominated by counter-information activity: 30.6% of posts are Rumor Denial or Official Correction, outpacing Unverified Claim posts (12.3%) by a 2.5:1 ratio, while 19.3% represent incidental keyword matches, directly quantifying the sampling noise introduced by keyword-based collection. To comply with platform terms and copyright, we publicly release Post IDs, Page IDs, and full metadata and derived indices; message text is available only under a research data-use agreement (DUA). ArabMisinfo enables engagement-aware analyses of Arabic misinformation ecosystems and, to our knowledge, provides the one of the first discourse-role-annotated Arabic Facebook corpora at this scale.

Paper 618 DiagBench: Evaluating Diagnostic Revision in Large Language Models with Longitudinal Inpatient Records

Authors: Huawei Feng, Dongsheng Shi, Linlin Wang

Abstract Reliable clinical decision support requires tracking evidence over a hospital stay and revising a disease diagnosis when new findings challenge the initial assessment. Strong performance on medical question answering does not establish that large language models (LLMs) can make these longitudinal decisions. We introduce DiagBench, a benchmark of 2,865 Chinese inpatient admissions that evaluates diagnostic revision as clinical evidence accumulates. It pairs admission-only prediction with

staged evidence disclosure, distinguishing an unchanged diagnosis, refinement of an existing hypothesis, and replacement by a different diagnosis. Development and test cases have human-annotated reference labels. The best LLM configuration in the main evaluation achieves only 52.6% macro recall; 80.3-87.7% of errors across five DeepSeek-Chat (DeepSeek-V3.2, non-thinking mode) prompting strategies occur between adjacent revision categories. Strategies that improve final-diagnosis prediction do not consistently improve revision decisions. We further use DiagAgent to trace diagnoses and evidence across rounds, revealing failures to replace an initial hypothesis despite recognizing contradictory findings. On DeepSeek-Chat, a single agent with the same state outperforms the full system on macro recall.

Paper 619 XChoice: Explainable Evaluation of AI–Human Alignment in LLM-based Constrained Choice Decision Making

Authors: Weihong Qi, Fan Huang, Rasika Muralidharan, Jisun An, Haewoon Kwak

Abstract We present XChoice, an explainable framework for evaluating AI-human alignment in constrained decision making. Moving beyond outcome agreement such as accuracy and F1 score, XChoice fits a mechanism-based decision model to human data and LLM-generated decisions, recovering interpretable parameters that capture the relative importance of decision factors, constraint sensitivity, and implied trade-offs. Alignment is assessed by comparing these parameter vectors across models, options, and subgroups. We demonstrate XChoice on Americans' daily time allocation using the American Time Use Survey (ATUS) as human ground truth, revealing heterogeneous alignment across models and activities and salient misalignment concentrated in subgroups defined by race (Black respondents) and marital status (Spouse: Present). We further validate robustness of XChoice via an invariance analysis and evaluate targeted mitigation with a retrieval-augmented generation (RAG) intervention. Overall, XChoice provides mechanism-based metrics that diagnose misalignment beyond surface outcome matching and make the effects of targeted interventions measurable at the mechanism level. Code and data access instructions are available at https://github.com/WeihongQ/xchoice.

Paper 624 Molecule Generation through Reasoning with Large Language Models

Authors: Congfeng Cao, Di Wu, Zhi Zhang, JIA-HONG HUANG, Jelke Bloem, Khalil Sima'an

Abstract Molecule generation is significant for its potential in scientific discovery and practical applications. Recent attempts often frame this task as a *translation* problem from molecular caption to structural representation, such as SMILES. This paper first examines the feasibility of modeling the task as a reasoning process with large language models (LLMs), generating higher-quality molecules through structural decomposition and recombination within curated Chain-of-Thought (CoT). We then introduce a workflow for curating accurate CoT data, incorporating both machine and expert verification. Lastly, we demonstrate that with a limited dataset of 4,213 curated samples, namely **MolCoT4K**, we elicit strong reasoning capabilities for molecule generation in open-source LLMs such as Qwen2.5-7B, achieving state-of-the-art exact match accuracy over strong open-source baselines (e.g., MolT5 and LlaSMol) as well as advanced commercial LLMs like GPT-4o. Moreover, the resulting model, **MolGeneration**, attains a Pass@16 exact match accuracy of 48.46%, highlighting its strong potential for real-world experimental applications when supported by a feasible external verifier or chemistry experts. Our analysis shows that the correctness of the CoT path is crucial, while reasoning ability primarily enhances accuracy in fine-grained molecule generation.

Paper 625 Mitigating Over-Personalization in LLMs via Structured Memory

Authors: Hakeem Hannoon, Andrew Zhao, Mihir Narayan, Sharvin Goyal, Ivaxi Sheth

Abstract Conversational assistants increasingly rely on persistent long-term memory to personalize responses across sessions. However, when stored user information is reintroduced into the model context, it can also influence responses in inappropriate or unrelated settings. We study two such failure modes in memory-augmented LLMs: cross-domain leakage, where memories from one life domain affect responses in another, and memory-induced sycophancy, where stored user beliefs make models more likely to agree with the user rather than respond truthfully. We apply a simple inference-time modification to how memories are presented to the model, without changing the model or the memory contents. Across seven models on PersistBench, we compare the commonly used all-in context format, where memories are injected as an unstructured list, with structured formats that partition memories by domain. This simple modification consistently reduces cross-domain leakage while preserving utility, with our strongest method reducing leakage by 8.8% on average relative to the baseline.

Paper 627 Breaking the Speech-Act Contract: Uptake Failures and Void Commitments in Large Language Models

Authors: Sahana Varadaraju, Bharathwaj Vijayakumar, Lathish Balaji Baskaran

Abstract Large language models do not just answer questions in dialogue; they are also expected to make the right conversational move. When a user apologizes, thanks the assistant, corrects an earlier claim, expresses frustration, or sets a preference, the reply carries a pragmatic obligation. We call failures to meet that obligation uptake failures, and model statements that imply unsupported memory or future behavior void commitments. We introduce a 300-item benchmark of user-assistant exchanges across six speech-act families, three difficulty tiers, and both synthetic and adapted-from-real-conversation items. This benchmark is intended as a curated core set: small enough for careful agreement analysis, but structured so that additional balanced batches can be added under the same ontology. On the current core set, we find strong annotation agreement (κ_family = 1.00, κ_act = 0.978). We also present SpeakFix, a lightweight post-generation repair layer that detects likely contract violations and fixes them with a prepend-not-replace strategy. Across GPT-4o, Claude Sonnet 4.6, and GPT-5.2, baseline uptake accuracy ranges from 71.0% to 77.3%, while void commitment rates range from 8% to 9%. SpeakFix raises uptake accuracy to 87.0–93.3% and reduces void commitments to 0%; the paired uptake gains are statistically significant for all three models ($p < 10^{-10}$). In a two-reviewer manual evaluation of 54 repaired failure cases (108 total judgments), 70.4% of repairs are rated better than the original, 28.7% equivalent, and 0.9% worse; helpfulness and naturalness are preserved in 100% and 99.1% of judgments, respectively. A separate blinded study with four independent, non-author raters is more critical: it confirms the repairs are at least as helpful as the original, but finds that the minimal prepend-style edits are often judged less natural, so reliable uptake gains do not yet translate into uniformly preferred rewrites. Taken together, the results suggest that speech-act contract violations are common enough to measure and often reducible with lightweight repair.

Paper 632 Towards Robust Numerical Claim Verification

Authors: Peter Røysland Aarnes, Vinay Setty

Abstract Large language models (LLMs) are widely used for claim verification, yet remain brittle for numerical reasoning: even small changes in value can sharply degrade accuracy. We show that this brittleness persists in frontier LLMs, but can be mitigated through adversarial fine-tuning on numerically perturbed examples. Using parameter-efficient fine-tuning, small Qwen3 models (0.6B–8B) reach 98.7\% accuracy on label-flipping perturbations, outperforming larger zero-shot models and frontier systems (GPT-5.4 Pro (74.0\%) and Gemini 2.5 Flash (73.9\%)). The gains generalise to unseen perturbation types, indicating

robust numerical decision boundaries rather than memorised edits. Robustness also transfers without target-domain data, significantly improving cross-lingual performance in Spanish. We further show that the same fine-tuning recipe confers robustness to evidence-side perturbations, using the VitaminC dataset.

Paper 635 FreakOut-LLM: The Effect of Emotional Stimuli on Safety Alignment

Authors: Daniel Kuznetsov, Ofir Cohen, Karin Shistik, Asaf Shabtai, Rami Puzis

Abstract Safety-aligned LLMs are trained to refuse harmful requests, but whether this holds under emotionally charged context is untested. We introduce FreakOut-LLM to test this. Using established psychological stimuli as system prompts, we evaluate ten LLMs under three conditions (stress, relaxation, neutral) plus a no-prompt baseline, scoring attack success with the HarmBench classifier on AdvBench prompts. The prompts contain no adversarial content; ASR measures safety fragility under an induced affective state. Stress priming produces a small but significant increase in jailbreak success over neutral, +1.03 percentage points (1.58\%$\rightarrow$2.61\%; 65.6\% relative; OR $= 1.67$, Cohen's $d = 0.28$), while relaxation produces no effect. Five of ten models show significant vulnerability, with the largest absolute effects in open-weight models. Controlling for prompt length and model identity, a logistic regression over $114{,}400$ primed queries confirms stress as the only significant condition. The shift also appears under four judges, on a second benchmark (HarmBench), and, for open-weight models, with the stimulus in the user message. Measured psychological state strongly predicts attack success ($|r|\geq0.70$ across five instruments). Emotional context is therefore a measurable vulnerability of safety-aligned LLMs, with implications for high-stress deployment.

Paper 639 Personalization Matters: Long-Horizon Conversation Agent with User-Centric Information in Online Shopping Interactions

Authors: Rena Wei Gao, Yue Dai, Hao Guan, Shengxiang Gao, Wangyang Wu, Yixin Shen, Jey Han Lau

Abstract Personalized conversational shopping requires maintaining preference consistency over multi-turn interactions, where users reveal constraints gradually. Existing approaches often rely on static profiles and do not explicitly control long-horizon interaction behavior. We propose a multi-agent, multimodal Retrieval-Augmented Generation (RAG) framework that decomposes dialogue state tracking, recommendation retrieval, preference-aware reasoning, and response generation, while integrating product metadata, product reviews, image-derived descriptions, and user historical reviews. To evaluate interaction-level quality, we adopt a trajectory-level protocol with four dimensions: Global Preference Consistency, Cumulative Information Synthesis, Interaction Trajectory, and Tone Consistency. On an Amazon Reviews 2023 benchmark, retrieval-enabled variants outperform a no-RAG baseline on automatic trajectory metrics (average 4.82 vs. 3.74). In a small real-user study ($n{=}5$), the Full variant achieves the highest mean overall rating (4.60 vs. 2.20 for Baseline), providing exploratory evidence that role decomposition plus user-centric retrieval improves perceived personalization.

Paper 641 From Textual Columns to Query Plans: A Unified Relational-Semantic Execution Framework for Hybrid Query Processing

Authors: ~Nima_Shahbazi1, ~Seiji_Maekawa1, ~Nikita_Bhutani1, ~Estevam_Hruschka1

Abstract Existing table question answering systems face a key trade-off: approaches designed for relational databases scale efficiently but assume fully structured schemas, while LLM-based methods support flexible reasoning over unstructured text but struggle to scale to large and multi-table datasets. This limits their applicability to semi-structured data, where answering a query may require dynamically recovering latent schema elements from free-form text before symbolic execution can proceed. We

present OmniTQA, an operator-centric framework that unifies relational execution and LLM-based semantic reasoning within a single query runtime. OmniTQA compiles queries into directed acyclic graphs (DAGs) of relational and semantic operators, enabling cost-aware optimization, dual-engine execution, and parallel exploration of alternative semantic groundings under ambiguity. Experiments on structured and semi-structured benchmarks show that OmniTQA substantially outperforms existing approaches on hybrid data settings while remaining competitive on structured datasets.

Paper 642 Behaviorally Grounded User Profiles from the Wild for Personalized Alignment and Multi-Perspective Reasoning

Authors: Yuxuan Li, Victor Zhong, Ehsan Kamalloo

Abstract Persona-driven techniques increasingly adapt large language models (LLMs) to diverse contexts. However, existing methods predominantly rely on rigid, synthetic personas that flatten individual variation, rely on stereotypes, and miss the nuanced signals driving actual human preferences. We introduce profile behavioral grounding, a framework for extracting open-ended, high-fidelity user profiles directly from authentic, anonymized social media posts. We evaluate these profiles across two paradigms: train-time personalization via supervised fine-tuning (SFT) and non-parametric test-time multi-perspective reasoning. Across complex recommendation and open-ended query benchmarks, behaviorally grounded profiles consistently improve base models and outperform synthetic profile baselines, driving stronger parametric alignment and enabling richer, multifaceted reasoning. Our findings establish open-ended, behavior-derived profiles as a highly diverse and effective foundation for the next generation of personalized language systems.

Paper 644 PPL-Factory: Task-Aware and Budget-Aware Data Selection from Language Modeling to Reasoning

Authors: Hang Zhang, Warren J. Gross

Abstract Not all training samples contribute equally to the large language model fine-tuning. Selecting informative training samples can reduce the computational cost while preserving downstream performance. Many existing data selection methods rely on indirect heuristics, such as data quality, diversity or reasoning trace length. However, the effectiveness of these fixed criteria is task-dependent and difficult to generalize across diverse downstream tasks. Perplexity-based data selection provides a simple and model-aware solution to estimate the sample difficulty, but existing approaches typically score the entire training sequence and ignore the difference in learning objectives of language modeling and reasoning tasks. In this paper, we propose **PPL-Factory**, a simple and interpretable data selection framework that combines task-aware perplexity-based scores and data budget-aware selection criteria. Experiments on GSM8K demonstrate that PPL-Factory outperforms other state-of-the-art data selection methods using only 1% of the training set. With 10% of the data, PPL-Factory exceeds full-data fine-tuning accuracy by 0.9 on GSM8K and 4.8 on MATH. Overall, our results demonstrate that task-aware and budget-aware perplexity-based selection provides an effective and applicable approach for efficient fine-tuning.

Paper 645 The Geometry of Personality in Language Models

Authors: Veer Chheda, Avantika Sankhe, Aaditya Uday Ghaisas, Vansh Shah, Devansh Dhruve, Dr. Narendra Shekokar

Abstract Whether language models (LMs) encode personality traits as stable internal representations or merely mimic them through surface-level behavioural shortcuts remains an open question. In this work, we move beyond prompting-based behavioural outputs and adopt a mechanistic interpretability approach to investigate the internal geometry of the OCEAN personality traits within transformer-based LMs. Using trait-conditioned activations from the Big5-Chat dataset, we evaluate

layer-wise hidden representations across five open-source model families in both their base and instruction-tuned variants. Through linear probing, entanglement analysis and information-theoretic MDL compression, we show that personality features emerge early as low-dimensional, distributed semantic structures that compress and separate monotonically with network depth. Importantly, our geometric analysis reveals that several traits occupy overlapping representational subspaces, causing naive activation steering to suffer from severe cross-trait contamination. To decouple these features, we introduce an inference-time Null-Space Projection (NSP) technique that orthogonalizes target trait directions against competing trait subspaces. Triangulated evaluations using both a GPT-5 judge and human annotators demonstrate that NSP-corrected steering resolves trait entanglement. It significantly amplifies intervention efficacy, improving target shift by 25.7% under automated metrics and 35.7% for human raters, while simultaneously reducing unintended cross-trait leakage by 31.9%. Our findings offer a rigorous foundation for robust, multi-dimensional behavioural control by bridging the gap between psycholinguistics and representational geometry.

Paper 649 Training free Cross-Modal Alignment for Multilingual Image Retrieval

Authors: Usman Shahid, Manan Jain, Natalie Parde

Abstract We investigate whether multilingual image-text retrieval can be made practical without training a multi-modal model, using frozen multilingual uni-modal encoders and inducing cross-modal alignment through relative representations computed against a large external anchor set. We use English image-caption pairs as anchors and evaluate caption-to-image retrieval on 3,600 images with captions in 36 languages (some low-resource). Across multiple encoder pairings and hyperparameter settings, the method yields usable performance, positioning training-free relative representation as a baseline for multilingual cross-modal alignment when target-language multi-modal training data is unavailable.

Paper 652 No Convo! The Illusion of Reliability in Online Toxicity Detection

Authors: Rachita Rajesh, Christina A Chance, Ashima Suvarna, Kai-Wei Chang, Saadia Gabriel

Abstract While AI toxicity detection models are widely used, their (un)reliability under realistic linguistic variation remains poorly understood. We provide a comprehensive study of how contextual change affects prediction reliability. We evaluate eight state-of-the-art toxicity detection models across three dimensions of linguistic variation: conversational context, conversational depth, and surface form transformations. Our results on three toxicity benchmarks show that conversational context often introduces substantial prediction instability, with some systems exhibiting prediction flip rates of up to 40%, while increasing conversational depth fails to produce consistent performance gains. Under paraphrasing transformations designed to preserve the underlying meaning and harmful intent, models exhibit substantial degradation and prediction flip rates of up to 57%. Overall, our findings show that strong benchmark performance does not necessarily translate to robust moderation behavior under realistic linguistic variation, highlighting the need for robustness-oriented model development and evaluation frameworks that better capture deployment conditions.

Paper 654 BetXplain: An Explanation-Annotated Dataset for Detecting Manipulative Betting Advertisements on Social Media

Authors: MSVPJ Sathvik, Parmitha Vangapandu, Nishit Rahul Rane, Sathwik Narkedimilli, Mark G. Lee, Akrati Saxena

Abstract The promotion of betting applications on social media platforms has increased significantly in recent years. Many of these advertisements use persuasive techniques that may mislead users, encourage risky behavior, and potentially influence users' mental well-being. However, research on the automated detection of manipulative and deceptive betting advertisements

remains limited due to the lack of publicly available annotated datasets. In this work, we introduce a new dataset of betting-related advertisements collected from two widely used social media platforms, Instagram and Reddit. The advertisements were manually annotated for manipulative and deceptive advertising practices. In addition to classification labels, the dataset includes human-provided explanations that describe the reasoning behind each annotation, enabling research into explainable approaches to detecting manipulative advertising. Furthermore, we analyze the strategies commonly used in betting advertisements and examine how these persuasive tactics may impact users' mental health. The proposed framework can also enable practical applications such as browser plugins that warn users about manipulative betting advertisements and automated web crawlers that help regulatory authorities monitor and detect such promotions online.

Paper 664 HYDRA: Evaluating Logical Thinking Abilities of LLMs through Formulating Hypotheses and Verification Methods

Authors: Shin-nosuke Ishikawa, Hirotsugu OHBA, Masato Todo, Atsushi Yoshino, Masahiro Daimon

Abstract Logical consistency in the outputs of large language models (LLMs) is essential for extending their applications beyond single-task assistants, including their use in AI agent systems. Although logical reasoning capabilities are often evaluated through tasks such as mathematics and coding, where it is feasible to assess the correctness of outputs in a well-defined manner, real-world problem solving requires inductive and abductive reasoning, with an emphasis on the ability to formulate and verify hypotheses. When evaluating the hypothesis formulation and verification capabilities of LLMs, the major challenge lies in preparing a variety of tasks with appropriate difficulty, avoiding those that are either too easy or unsolvable. We introduce an automated framework called HYpothesis-Driven Reasoning Assessment (HYDRA), focusing on hypothesis formulation and verification for real-world problem solving with logical reasoning. In HYDRA, all components, including situation generation, hypothesis, and evaluation, are fully performed by LLMs. HYDRA's systematically constructed generation and evaluation strategies enable assessment using realistic, non-trivial situations that offer both diversity and clear criteria. We evaluated six state-of-the-art LLMs using HYDRA and found that their success rates ranged from 74.3% to 90.3%, indicating that even the best-performing model among those evaluated produced rubric-invalid hypothesis--verification pairs in nearly 10% of cases.

Paper 666 GCoTD: A Non-Intrusive Diagnostic Pipeline for Black-Box LLM Backdoor Detection via Graph Chain-of-Thought

Authors: Chengyu Jia, AprilPyone MaungMaung, Huy H. Nguyen, Isao Echizen, Jinyin Chen

Abstract Backdoor detection for large language models is increasingly difficult in realistic black-box settings, where defenders lack access to weights, gradients, training data, or attack-specific trigger mechanisms. Existing detectors often rely on white-box signals, labeled poisoned data, or output-level signatures, making them brittle against inference-time and non-training-based attacks on closed-source models. We propose GCoTD, a black-box detector that identifies backdoors by analyzing the structure of an LLM's reasoning rather than its final outputs alone. GCoTD retrieves few-shot graph Chain-of-Thought templates to elicit graph-structured reasoning traces from the target model, then applies a lightweight language-model detector to distinguish clean from poisoned reasoning patterns. By representing reasoning as graphs, GCoTD captures mechanism-agnostic irregularities such as logic discontinuities, isolated subchains, and inconsistent intermediate transitions, without requiring access to model internals or training data. Experiments across four attack types and six LLMs, including LLaMA-3, DeepSeek-V3, Qwen-3, and GPT-class models, show that GCoTD outperforms recent black-box baselines and retains strong detection performance on held-out attacks.

Paper 670 TokenMix: Efficient Batched Inference for Mixture-of-Experts Models with Joint Expert Selection

Authors: ~Aradhana_Mohan_Parvathy1, ~Sarada_Krithivasan1, ~Swagath_Venkataramani2, ~Anand_Raghunathan1, ~Viji_Srinivasan1

Abstract The mixture-of-experts (MoE) architecture has proven indispensable for scaling transformer-based artificial intelligence (AI) models. MoEs use a router network to activate specialized subnetworks, called experts, within each transformer layer on a per-token basis. Thus, MoEs increase model capacity without proportionally increasing the computational requirements of training and inference. However, a key challenge is that MoE inference throughput scales poorly with batching, as each input in a batch activates a different subset of experts, thereby increasing memory intensity. To improve batched MoE inference throughput, we propose TokenMix, a novel technique that mixes token representations within the router across inputs in a batch to promote expert reuse. In effect, TokenMix encourages coordinated expert selection across inputs. TokenMix requires neither training nor fine-tuning and can be applied directly to off-the-shelf MoE models. We evaluate TokenMix across five state-of-the-art MoE models with parameter counts ranging from 16B to 235B. TokenMix achieves a 25–60% improvement in expert reuse, leading to a 12–28% reduction in inter-token latency with negligible or no loss in accuracy.

Paper 676 From Chains to Graphs: Self-Structured Reasoning for General-Domain LLMs

Authors: Yingjian Chen, Haoran Liu, Yinhong Liu, Sherry T. Tong, Aosong Feng, Jinghui Lu, JTzhang, Yusuke Iwasawa, Yutaka Matsuo, Irene Li

Abstract Large Language Models (LLMs) have demonstrated strong reasoning ability in open-domain question answering, yet their reasoning is typically expressed through linear textual traces, such as Chain-of-Thought (CoT). While these traces are readable, their sequential form can obscure logical dependencies, making inconsistencies difficult to inspect. Existing graph-based methods introduce structured representations, but they often rely on externally provided graphs, leaving it unclear whether LLMs can autonomously perform graph-structured reasoning. To fill this gap, we explore graph-structured reasoning of LLMs in general-domain question answering. We propose Self-Graph Reasoning (SGR), a framework that enables LLMs to explicitly represent their reasoning process as a structured graph before producing the final answer, improving both consistency and interpretability. We further construct a graph-structured reasoning dataset that merges multiple candidate reasoning graphs into refined graph structures for model training. Experiments on five QA benchmarks across both general and specialized domains show that SGR consistently improves reasoning consistency and provides clearer, more interpretable graph-structured reasoning processes, achieving a 17.74% gain over the base model.

Paper 677 CORA: A Protocol for Diagnosing Boundary Robustness in Text-to-Audio Retrieval under Query Reformulations

Authors: Jae Min Woo, Kyongmin Kong, Bogyung Jeong, Minjeong kim, HaeJun Yoo, Du-Seong Chang

Abstract Text-to-Audio (T2A) retrievers are typically evaluated with caption style queries, but the same user intent can be expressed in many forms. We introduce CORA (Caption-Offset Retrieval for Audio), a caption anchored diagnostic protocol that rewrites each source caption into five intent preserving forms (Command, Question, Indirect, Key phrase, and Statement) while fixing the target audio. By tracking the same target across query forms, CORA defines RankDrop, a metric revealing failures hidden by Recall@k. Using Pearson's correlation coefficient r, we find that RankDrop is weakly associated with raw text space movement ($r=0.084$), but strongly associated with Target Alignment Loss and Target Boundary Margin Degradation ($r=0.508$ and $r=0.615$). The same pattern appears in OEA retrievers, where RankDrop is better explained by boundary

degradation ($r=0.472/0.478$) than by query movement ($r=0.084/0.046$). Overall, these results suggest that robust T2A retrieval requires preserving the target's boundary advantage over competing audio under reformulation.

Paper 678 It Hears, It Sees too: Multi-Modal LLM for Depression Detection By Integrating Visual Understanding into Audio Language Models

Authors: Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, Zongyuan Ge

Abstract Depression is one of the most prevalent mental health disorders globally. In recent years, multi-modal data, such as speech, video, and transcripts, has been increasingly used to develop AI-assisted depression assessment systems. Large language models (LLMs) have further advanced this field due to their strong language understanding and generalization capabilities. However, conventional LLMs remain text-centric and cannot process the rich non-verbal cues found in audio and visual modalities, which are critical components in mental health evaluation. While multi-modal LLMs offer a promising direction, few are tailored for psychological applications. In this study, we propose a novel multi-modal LLM framework for depression detection. Our approach augments an audio language model with visual understanding and aligns audio-visual features at the timestamp level. This fine-grained alignment improves modeling of temporal dynamics across modalities while reducing the need for extensive training data and computational resources. Experiments on the DAIC-WoZ and E-DAIC datasets demonstrate that our model consistently outperforms both single-modality approaches and previous multi-modal methods. Moreover, the proposed framework can be extended to incorporate additional physiological signals, paving the way for broader clinical applications beyond mental health.

Paper 683 UniToR: Towards Universal Token-level Routing for Efficient Reasoning in Large Language Models

Authors: Nguyen-Khang Le, Truong Dinh Do, Le-Minh Nguyen

Abstract Large Reasoning Models (LRMs) deliver strong performance on complex reasoning tasks, but their long chain-of-thought traces incur substantial inference latency and cost. To mitigate this overhead while preserving reasoning quality, token-level routing between a small language model (SLM) and a large language model (LLM) offers one of the best efficiency-performance trade-offs. However, existing token routers face two major practical limitations: they must be retrained for every new SLM-LLM pair, and they require both models to share a tokenizer. We present UniToR, a universal token-level router that removes both constraints. UniToR predicts per-token divergence using model-agnostic signals, enabling a single router to generalize across diverse SLM-LLM pairs without retraining. To support heterogeneous model pairs, we further introduce a token-alignment mechanism that eliminates the requirement for tokenizer compatibility, enabling seamless token-level switching across models with different tokenization schemes. Experimental results show that a single UniToR router generalizes across different model families (DeepSeek-R1, Qwen3, and Phi-4), including pairs with incompatible tokenizers. UniToR matches the accuracy of R1-32B at up to 6x lower cost and successfully operates on unseen model pairs where existing token-level routing methods fail.

Paper 684 CuBEs: Culturally-Situated Behavioral Evaluations and the Limitations of Culture-Blind LLM Judges

Authors: Hoda Ayad, Tanu Mitra, Abhishek Mukherji

Abstract Evaluating the occurrence and triggers of large language model (LLM) behaviors - such as sycophancy, self-preference, or over-confidence - is critical for predicting real-world model deployment risks. However, existing situated behavioral evaluations typically ignore cultural context, limiting their generalizability across an increasingly global user base. To address this gap, we propose CuBEs - Culturally-situated Behavior Evaluations that probe for response patterns across diverse user cultures. We first extend an automated testing pipeline to inject cultural context into behavioral test scenarios and subsequent evaluation. We assess the cultural adaptability of this pipeline by building a human-labeled dataset that captures nuanced dimensions of behavior understanding across 12 distinct cultures. Our dataset reveals significant cross-cultural variations that one-size-fits all judgments fail to capture. Through evaluating 13 open- and closed-source LLMs, we find that introducing cultural situatedness in the evaluation scenario creates significant variation in the presence of a behavior. For example, while our baseline experiments testing for political bias capture localized Western political dimensions like the American conservative-progressive divide, non-Western culturally situated evaluations surface entirely different axes of bias such as religious and colonial political issues. Our findings demonstrate that standard, culturally-agnostic evaluations fail to capture these shifts, highlighting the necessity of culturally situated behavioral testing for global deployments.

Paper 692 K-FORGE: Fisher-Guided Initialization for Preference-Based LLM Unlearning

Authors: ~Daniil_Moskovskiy1, ~Alexander_Panchenko1

Abstract Preference-based LLM unlearning methods such as NPO and SimNPO must discover directions that suppress forget-set behavior without damaging retained capability. K-FORGE uses forget- and retain-set Fisher information to construct a retain-aware low-rank edit, then starts NPO or SimNPO from the edited checkpoint. On the TOFU forget10 split, where 10% of the original dataset is designated for forgetting, stronger one-shot edits lower both Forget Q/A Probability and Model Utility. With Llama 3.2 1B, K-FORGE reduces Forget Q/A Probability by 22.3% under NPO while improving utility and using less total time and compute than scratch. SimNPO shows a smaller descriptive probability reduction with lower utility. On the same split, random and weight-SVD initializers yield higher forget probability than K-FORGE at 50 and 100 steps. Experiments with Gemma 3 1B also yield lower mean forget probability for both optimizers at 100 steps. MUSE results vary by metric, and direct relearning recovers forgotten behavior.

Paper 702 MOOC-MESE: Dataset and Framework for Multimodal Entity Set Expansion in MOOCs

Authors: Shichao Yue, Jing Wan, yubin chen

Abstract Entity Set Expansion (ESE) aims to expand entities belonging to the same semantic category given only a small set of seed entities. Traditional ESE methods are primarily text-based and struggle to address two key challenges in MOOC scenarios: (1) Multimodal Semantic Dependency; (2) Generalization Issues.To facilitate research in this area, we construct MOOC-MESE, the first Chinese multimodal ESE dataset for MOOCs, which contains over 45k aligned image-text pairs. The dataset provides rich multimodal semantic information for entity expansion research in educational scenarios. Furthermore, we propose GMRM (Generative Multimodal Retrieval Model), a multimodal ESE framework that reformulates the task as a multimodal generative retrieval problem. By achieving multimodal alignment and fully exploiting cross-modal semantics, GMRM effectively mitigates generalization issues. Extensive experiments and analyses on MOOC-MESE demonstrate both the high quality of the dataset and the effectiveness of GMRM. The code is publicly available at: https://github.com/constantinus-xi/mooc-mese.

Paper 706 Self-Route: Automatic Mode Switching via Capability Estimation for Efficient Reasoning

Authors: Yang He, Xiao Ding, Bibo Cai, Yufei Zhang, Kai Xiong, Zhouhao Sun, Bing Qin, Ting Liu

Abstract While reasoning-augmented large language models (RLLMs) significantly enhance complex task performance through extended reasoning chains, they inevitably introduce substantial unnecessary token consumption—particularly for simpler problems where Short Chain-of-Thought (Short CoT) suffices. This "overthinking" phenomenon leads to inefficient resource usage without proportional accuracy gains. To address this issue, we propose Self-Route, a dynamic reasoning framework that automatically selects between general and reasoning modes based on model capability estimation. Our approach introduces a lightweight pre-inference stage to extract capability-aware embeddings from hidden layer representations, enabling real-time evaluation of the model's ability to solve problems. We further construct Gradient-10K, a model difficulty estimation-based dataset with dense complexity sampling, to train the router for precise capability boundary detection. Extensive experiments demonstrate that Self-Route achieves comparable accuracy to reasoning models while reducing token consumption by 30-55% across diverse benchmarks. The proposed framework demonstrates consistent effectiveness across models with different parameter scales and reasoning paradigms, highlighting its general applicability and practical value.

Paper 711 No Masking Needed: Predicting Latent Semantics for Efficient Pre-training of Encoder-based Language Models

Authors: Gábor Berend

Abstract In this paper, we introduce Latent Semantic Modeling (LSM), a sample-efficient alternative for pre-training encoder-based language models that does not require the use of a special mask symbol. This is beneficial, as it allows us to perform pre-training on naturally occurring sentences (as opposed to such ones that contain an unnatural mask symbol, which symbol is otherwise not included during fine-tuning). The main idea behind LSM is that instead of requiring the language model to return the exact identity of randomly selected (sub)tokens from an input sequence, it determines their latent semantic properties extracted from an auxiliary model using sparse coding over its activations. When fine-tuned for downstream applications, models pre-trained with LSM often perform as good or better as alternatively pre-trained models that use considerably more compute per update steps and which have gone through considerably more pre-training steps, illustrating the improved representation developed during LSM pre-training. We release our code base for replicating our results under https://github.com/szegedai/MLSM.

Paper 712 Long-Term Agent Memory through Complementary Curation and Consolidation

Authors: Abhishek Gupta, Dongge Han, Daniel Madrigal, Wei-Ning Chen, Mayukh Das, Victor Rühle, Saravan Rajmohan

Abstract Long-horizon dialogue agents require effective memory systems to recall facts, events, and relationships from interaction histories. In this work, we propose a hybrid-memory architecture comprising episodic memories and structured relational memories and a consolidation mechanism which preserves efficiency. We conduct an empirical study across two benchmarks: LoCoMo and LongMemEval-S. Our experiments demonstrate that the two memory types are consistently complementary, combining both outperform state-of-the-art memory frameworks and achieves 90.2% accuracy on LoCoMo and 85.5% on LongMemEval-S.

Paper 713 RASER: Recoverability-Aware Selective Escalation Router for Multi-Hop Question Answering

Authors: Yuyang Li, Zihe Yan, Tobias Käfer

Abstract Multi-hop question-answering systems often use expensive retrieval on every question. They may decompose the question, run several retrieval rounds, or search through bridge entities before answering. All of these strategies rely on repeated LLM calls to rewrite or decompose the question, which increases extra token cost, and it is not fitting when the LLM budget is tight. However, our analysis shows that lots of multi-hop questions are already answered correctly by a single one-shot RAG, so running an extra retrieval on every question wastes the budget. We introduce RASER (Recoverability-Aware Selective Escalation Router), a family of cheap routers built on one-shot RAG and six features from it. RASER-2 decides whether to stop or escalate to the extra-retrieval action PRUNE. RASER-3 chooses among one-shot RAG, PRUNE, and iterative retrieval IRCoT, using the same features but adding an explicit cost-accuracy trade-off. Neither router makes an extra LLM call to decide. Across six LLMs and three multi-hop QA benchmarks, both routers stay competitive with the other state-of-the-art (SOTA) baselines in F1 while spending only 41–49% of always-prune's tokens and also less than the iterative and decomposition retrieval baselines.

Paper 715 POLAR: Ontology-Guided Risk Prevention for Tool-Calling LLM Agents

Authors: Yunju Kang, Seonghyeon Cho, Irene Li, Yeo Chan Yoon, Chanjun Park

Abstract LLM tool-use agents operate in dynamic environments where many actions carry operational risk. However, most safety mechanisms react only after errors manifest. Existing pre-emptive approaches either fine-tune the agent on chain-of-thought deliberation or compile natural-language guardrails into runtime checks, but they do so without exposing a structural, auditable verdict. We propose POLAR, a guardrail framework for small tool-calling agents that assesses reversibility through a structured two-layer ontology. POLAR assigns each action a graded reversibility score by deriving a candidate inverse sequence; calls failing a threshold are pruned before execution. Evaluated on τ^2-bench across six agent models, POLAR improves mean task reward by 0.11 to 0.18 points on airline for four of six agents, but only eight of eighteen model--domain cells improve overall; retail and stronger agents often regress. POLAR provides an auditable structural check and characterizes its task-utility trade-offs. Reward is not a direct measure of prevented harm.

Paper 722 Outgrowing the Curse of Multilinguality: Scaling Strategies for Efficient Language Adaptation

Authors: Kevin Glocker, Kätriin Kukk, Romina Oji, Marcel Bollmann, Marco Kuhlmann, Jenny Kunz

Abstract Training high-performing language models which include medium- and lower-resource languages remains a challenge. Massively multilingual models still underperform language-specific adaptations, especially at smaller scales, due to the curse of multilinguality. In this work, we investigate scaling as an efficient strategy for adapting pretrained models to new target languages. Through comprehensive ablations matched for data and training FLOPs, we test whether upscaling from English base models enables more effective and resource-efficient adaptation than standard continued pretraining. We find that upscaling is an efficient strategy for improving target-language performance by relieving capacity constraints, especially for higher-resourced languages. An initial upscaling step in English is beneficial for low-resource languages. However, in higher-resource scenarios it is more compute-efficient to directly scale up on the target language, albeit slightly compromising data efficiency for most languages. For multilingual continued pre-training, we identify directly scaling up using the target data as the optimal setup for both data efficiency and reduced English forgetting.

Paper 726 Vouch: Counterfactual Contracts for Multi-Agent Coordination

Authors: ~Abhay_Bhandarkar1, ~Vedanth_Nanesha1

Abstract Multi-agent LLM coordination typically reduces to voting or arbitration over self-reported labels and reasoning that is uncoupled from agent behavior and fails predictably under structured bias. A shortcut-reliant agent is confident about its shortcut, and its predictions survive any rule that consults only self-reports. We propose VOUCH, a coordination protocol in which every agent claim is coupled to a deterministic test the protocol runs itself. An agent commits to a counterfactual prediction about an intervention's outcome; the protocol applies the intervention and verifies the contract only if the actual outcome matches. The natural single-side instantiation is unsound as a self-coherent agent satisfies its own commitment by construction. We resolve this with a dual-side construction where the proposer additionally commits to a second agent's post-intervention prediction, a cross-side claim that self-coherence cannot fulfill. Across multimodal classification (Hateful Memes, CrisisMMD, MM-IMDb), code (MBPP, HumanEval), and math (GSM8K, MATH, SVAMP), the dual-side fix reduces wrong-verified events by 46%–85% per domain while never falling more than 1 point below the Ensemble baseline on accuracy. VOUCH is therefore designed not to outperform ensembling, but to preserve task accuracy while making coordination decisions testable and auditable.

Paper 734 VideoNorms: Benchmarking Socio-Cultural Norm Understanding of Video Language Models

Authors: Nikhil Reddy Varimalla, Yunfei Xu, Meng Fan Wang, Arkadiy Saakyan, Smaranda Muresan

Abstract As Video Large Language Models (VideoLLMs) are deployed globally, it is important to assess their ability to reason across cultural contexts. To advance cultural norm awareness evaluation in VideoLLMs, we introduce VideoNorms, a dataset of cultural norm annotations from popular US and Chinese TV shows annotated with adherence or violation labels and (non-)verbal evidence. Through a human-AI collaboration framework, each item was first annotated by a large VideoLLM, and then reviewed by at least three trained monocultural annotators with significant lived experience in the target culture, resulting in a dataset of over 3,000 human judgments. Human verification showed disparity in US and Chinese norm extraction performance, cautioning against fully automatic approaches cultures under-represented in training data. Hierarchical linear modeling analysis of 7 open-weight VideoLLMs' performance revealed that: 1) models perform worse in Chinese compared to US, particularly for norm adherence prediction; 2) models have more difficulty in providing non-verbal evidence compared to verbal evidence for norm adherence/violation predictions. Ablation studies confirm video modality is indeed necessary for accurate performance, and scaling model size does not yield classification score improvements. Our findings and data contribute to culturally grounded video model training and evaluation.

Paper 740 Towards Smart Exploration: Reinforcement Learning for LLM Agents Guided by Meta-Actions

Authors: Hanwen Tong, Zhengqi Zhang, Ruixuan Lu, Jincheng Dai, Bin Yang, Bo Zheng, Feng Wei

Abstract Reinforcement Learning (RL) has empowered Large Language Model (LLM) agents to tackle long-horizon tasks through multi-step interactions with the environment. However, current LLM agents still rely on inefficient trial-and-error exploration, leading to frequent task failures and redundant actions. In this paper, we propose MetaAct, an RL framework that incorporates a set of high-level meta-actions to explicitly guide the agent towards promising trajectories and facilitate goal-aware exploration. Specifically, the agent is required to predict one of four meta-actions (FOCUS, GUESS, RESUME, UNDO) to constrain the direction of its subsequent actions to minimize unproductive trials. More importantly, we employ group-based RL with episode-level and step-level advantages to fine-tune the agent, enabling it to identify strategic directions more accurately and generate effective task-specific actions. Experimental results demonstrate that our method consistently outperforms the

state-of-the-art approaches on ALFWorld and WebShop, achieving higher success rates with fewer steps. Our code and data are publicly available at https://github.com/HavenTong/MetaAct.

Paper 742 Towards Universal Calibration for In-Context Classification

Authors: ~Awik_Dhar1, ~Sunita_Sarawagi1, ~Arpit_Agarwal2

Abstract In-Context Learning (ICL) has received a lot of interest recently due to its ability to improve classification accuracy for new tasks without having to fine-tune the model over these tasks. However, the impact of ICL on calibration has been far less studied. Recent work has shown that ICL can have a negative impact on calibration due to biased selection of in-context examples. The existing methods for improving ICL calibration require additional task and model-specific training, detracting from the no-fine-tuning convenience of ICL. In this work, we introduce CALIBFORMER, a universal ICL calibrator for classification tasks, that generalizes to new models and tasks without further fine-tuning on these tasks. Specifically, we design a transformer-based calibrator which uses the model predictions on in-context examples to predict a confidence score for the test input. We show that the calibrator trained on three models and ten tasks, is able to generalize/calibrate ICL predictions on several new tasks and four LLMs unseen during training. Even for frontier models like Gemini, CALIBFORMER is able to obtain over 50% reduction in calibration error

Paper 743 Data Augmentation for Improving Contextual Faithfulness Under Knowledge Conflict

Authors: ~Mohith_Rajesh1, ~Raj_Maheshwari1, ~Vashisth_Tiwari1, ~Yonatan_Bisk1, ~Emma_Strubell1, ~Jeremiah_Milbauer1, ~Jared_Fernandez1

Abstract External context enables language models to accomplish tasks which rely on information beyond what is captured during pretraining. A standard way to improve contextual faithfulness in these settings is to fine-tune on synthetic, modified questions for which the model must adhere to provided contexts inconsistent with information that may have been seen during training. While this approach reliably improves contextual faithfulness, we find that adapting to a narrowly targeted counterfactual signal yields degradation on other faithfulness evaluations and regression on tasks that evaluate a model's parametric world knowledge and reasoning. We find that data augmentation, in which a model is trained on \textit{both} parametrically consistent and counterfactual examples improves performance on faithfulness tasks with reduced regression on parametric world knowledge and reasoning tasks. We find that these improvements are largely consistent across post-training objectives (i.e. SFT, DPO) and across ground-truth, counterfactual data augmentation ratios.

Paper 745 Adaptation, Not Algorithm: LoRA and Full Fine-Tuning Show Comparable Black-Box Jailbreak Degradation in Five Open-Weight LLMs

Authors: ~Taleef_Tamsal1, ~Jonathan_Rusert1

Abstract Parameter-efficient fine-tuning (PEFT) methods such as LoRA and QLoRA have made LLM customization accessible, but their effect on adversarial safety is insufficiently characterized. We evaluate 25 configurations (five model families × Base FP16, 4-bit, LoRA, QLoRA, and full fine-tuning [FFT]) against black-box and white-box jailbreaks on HarmBench and JailbreakBench, with three automated evaluators validated against 750 doubly annotated human samples. LoRA, QLoRA, and FFT shift black-box safety in the same direction by comparable magnitudes; after Holm correction, which method reaches significance depends on the evaluator. On a seed-controlled re-training, LoRA and FFT meet a pre-specified ±5 pp equivalence criterion across the DeepInception and ArtPrompt cells under all three evaluators, whereas QLoRA does not and shows about twice the dispersion. Under this recipe, the act of adaptation, more than the choice between LoRA and FFT, predicts aggregate

black-box drift. White-box AutoDAN behaves differently: every training method lowers its ASR on Llama-3.1, the LoRA effect is seed-variable, and under a GCG probe fine-tuned checkpoints are easier targets than their bases in four of five families. Evaluator choice materially changes conclusions. These results motivate joint reporting of adaptation method, attack, and evaluator.

Paper 751 Cultural Awareness is Represented but Not Decoded: Tracing Mythological Knowledge across 18 Open-Source LLMs

Authors: Iaroslav Chelombitko, Mika Hämäläinen

Abstract Open-source LLMs reliably name Zeus, Jupiter, and Thor, but recover their counterparts in less-represented traditions like Finnish, Slavic, Egyptian, or Chinese mythology far less consistently. We ask where inside the model this cultural default is produced. On a parallel cross-cultural substrate of Thompson-motif entities, we instrument 18 open-source LLMs from 8 architecture families with linear probing, logit lens, activation patching, and output extraction. The residual stream cleanly distinguishes cultures, well above a name-string baseline, yet the decoder collapses culturally-specific tokens onto dominant-tradition ones. The failure is at readout, not at representation. Asking the same question in the target culture's native language versus English produces failures that cluster within language but decouple across language: the decoder is gated on prompt language. We release a per-entity (probe, output) decomposition framework, a citation-anchored cross-cultural ground truth, a within- versus cross-mode correlation test for language-conditioned readout, and per-entity predictions for all 18 models.

Paper 755 Leveraging Articulatory Phonetic Group Recognition for Robust Speech Enhancement for Cochlear Implants in Complex Acoustic Environments

Authors: Sohini Saha, Leslie M. Collins, Boyla Mainsah

Abstract Cochlear implant (CI) users often experience reduced speech understanding in noisy and reverberant environments. Our recent work has shown the benefits of exploiting knowledge of speech content to drive speech enhancement for CI applications by estimating time-frequency masks based on predictions from a phoneme classifier. However, phonemes with similar temporal and spectral characteristics are likely to be confused, leading to misclassifications that negatively impact the downstream phoneme-based task. Also, machine learning models are generally vulnerable to performance degradation in unseen acoustic conditions. Using a lightweight and causal phoneme-based speech enhancement model developed for CI applications, we demonstrate the utility of alternative phoneme groups to improve model robustness in diverse acoustic settings. Results from objective measures of speech enhancement show that manner of articulation (MOA) groups provide a balance between fine-grained but error-prone phonemes and limited acoustic specificity with voicing groups. In particular, a counterintuitive finding is that MOA-based models trained in reverberant conditions are particularly robust under mismatched train-test conditions. Our findings highlight the importance of linguistically motivated phonetic grouping and selective training strategies for developing robust speech enhancement models for CI applications while maintaining lightweight model complexity for edge deployment.

Paper 757 Superficial Reflection or Genuine Thought? A Fine-Grained Cognitive Analysis of Large Reasoning Models

Authors: Yuxiang Chen, Zuohan Wu, Ziwei Wang, Xiangning Yu, Xujia Li, Linyi Yang, Mengyue Yang, Jun Wang, Lei Chen

Abstract Motivated by the observed human-like behaviours in Large Reasoning Models (LRMs), this paper introduces a comprehensive taxonomy to characterise atomic reasoning steps and analyse the reasoning behaviours of LRMs. Grounded in

human cognitive processes, we propose a taxonomy comprising five groups and seventeen categories. Through this taxonomy, we conduct an in-depth analysis of contemporary LRMs and distil four actionable takeaways for model optimisation. Most notably, we reveal that prevailing post-answer ``double-checks'' are largely superficial and rarely yield substantive revisions. A targeted intervention further shows that explicitly eliciting richer reflection processes can substantially improve failed self-correction. To support this large-scale study, we propose CAPO, an automated annotation method used to construct a dataset of 277,534 reasoning steps with strong agreement with human expert annotations. We further validate the main behavioural patterns on a newer reasoning model and a coding domain, demonstrating the broader applicability of the proposed taxonomy. All source code and data are available at https://github.com/hehepig4/psyche.

Paper 759 Beyond Diagnosis: Evaluating Recommendation Stability under Counterfactual Skin-Tone Perturbations

Authors: ~Hyunwoo_Yoo1, ~Kyusik_Kim1, ~Gail_Rosen1

Abstract Multimodal clinical models have shown strong performance in dermatology recommendation and triage tasks, yet their robustness under demographic perturbations remains poorly understood. We introduce a clinician-validated counterfactual dermatology benchmark for evaluating recommendation stability under skin-tone perturbations in multimodal vision-language models. Our benchmark contains 240 physician-validated images from 60 clinical cases, each paired with four diagnosis- and severity-matched skin-tone variants constructed through iterative LLM filtering and physician review. Across diagnosis, management, urgency, and referral tasks, we observe substantial model-dependent instability even when clinicians verified preservation of diagnosis and severity across perturbation variants. Recommendation changes occur in up to 51.7% of paired comparisons, accompanied by shifts in uncertainty framing, pigmentation-sensitive terminology, and escalation-oriented language. Our findings suggest that demographic perturbations can influence not only diagnostic predictions, but also downstream recommendation behavior in multimodal clinical models. We release a controlled evaluation framework for studying recommendation robustness in medical AI systems.

Paper 760 Revisiting Co-Evolution for Multi-Turn Safety Alignment: Limitations and Alignment Tax Mitigation

Authors: ~Yoshiaki_Sata1, ~Kouta_Nakayama1

Abstract To enhance the multi-turn defense capabilities of Large Language Models, recent studies focus on a ``co-evolution framework.'' This framework iteratively trains an attacker model to explore vulnerabilities and a target model to learn safety alignment. However, in our preliminary experiments, existing methods suffered from severe training instability. To address this and align with recent optimization methodologies, we introduce GRPO into the training loops of both models. We found that a single alignment cycle sufficiently established the target model's defense, which in turn stalled the subsequent attacker training. Under the GRPO-based attacker-training pipeline and experimental setting examined in this study, this limitation rendered the iterative co-evolution process ineffective. Furthermore, to maintain model utility, we mix harmful goals with safe goals automatically rewritten from them. In our experiments, mixing safe goals improved utility but yielded higher attack success rates than training without them, revealing a safety-utility trade-off in our setting. Warning: This paper includes examples of harmful language.

Paper 762 The Cost of Compression: A Rate-Distortion Limit on Factual Hallucination

Authors: Wang Xi, Shijia Xu, Rongfeng Guo

Abstract While chain-of-thought prompting and scratchpads are frequently conceptualized as external memory for reasoning tasks, this perspective breaks down when a final answer must be generated from a bounded, jointly retained state. We formulate this limitation through a minimal constraint-satisfaction model in which an agent must exactly output N independent binary constraints using a shared internal and external memory configuration yielding at most 2^B distinguishable states. Under this framework, we mathematically prove a tight phase law: $$ p^\star(N,B)=\min\{1,2^{B-N}\}. $$ This Joint Memory Budget Theorem establishes that each independent constraint exceeding the retained memory budget exponentially halves the probability of exact success. We further generalize this limit with adversarial robustness and a min-entropy extension that separates raw prompt length from effective independent information. To empirically test the qualitative predictions of this theory, we introduce a two-call retention protocol that forces large language models to reconstruct constraint hands solely from their bounded written representations. Comprehensive evaluations across modern long-context and reasoning models reveal sharp, capacity-dependent performance cliffs. While variables such as scratchpad formatting, effective task entropy, and answer-time external access can shift this phase boundary, they do not eliminate it. Ultimately, we demonstrate that long-context and scratchpad reasoning are fundamentally bottlenecked not by nominal context length, but by the independent information jointly available at the time of answer generation.

Paper 764 Optimizing Reasoning LLMs for Automated Feature Engineering via Multi-Stage Direct Preference Optimization

Authors: Mingzhi Hu, Menghai Pan, Yuzhong Chen, Junpeng Wang, Yingtong Dou, Vineeth Rakesh, Hoang Anh Duy Le, Uday Singh Saini, Xin Dai, Xiran Fan, Yuqing Liu, Yiwei Cai

Abstract Feature engineering is central to tabular machine learning, but useful transformations must be both statistically predictive and semantically valid with respect to feature names, task descriptions, and observed data patterns. Although large language models can exploit such context, prompt-based feature engineering remains unreliable: generated transformations may be redundant, schema-invalid, non-executable, or weakly aligned with downstream performance. We propose \textsc{MSDPO}, a two-stage Direct Preference Optimization framework for automated feature engineering. Proposal-DPO aligns the model toward useful, diverse, and non-redundant feature proposals, while Function-DPO aligns it toward executable feature-construction code using execution and validation feedback. After alignment, the frozen model directly generates transformation functions for unseen tabular tasks. Across 20 classification datasets, \textsc{MSDPO} achieves the best overall mean rank among strong classical and LLM-based feature-engineering baselines. Under a budget-matched comparison inside the same feature-engineering harness, MSDPO also attains a better mean rank than Claude Sonnet 4.6, despite using an 8B open-weight backbone. Moreover, preferences constructed from classification signals transfer to four regression tasks. Ablation results show that the two DPO stages are complementary, improving both feature design and executable implementation.

Paper 765 Synthesizing Code Completion Trajectories for Mid-training in Software Engineering

Authors: Siwei Xu, Qiyuan Liu, Ruifan Li

Abstract Large language models (LLMs) have shown remarkable capabilities in interactive real-world software engineering (SWE) scenarios. However, the execution-aware training data are hard to scale and are mostly used for post-training, leaving agentic mid-training underexplored. To this end, we propose COMP-SYNTH, a novel automated pipeline for synthesizing reflective code completion trajectories. Specifically, based on executable repository environments, COMP-SYNTH identifies test-

covered target functions, masks their implementations, and prompts an LLM to complete them under execution feedback. Then, each generated completion is verified by running the associated tests. Furthermore, failed attempts are augmented with diagnostic signals to extract useful supervision. Thus, we construct COMP-TRAJ, a mid-training dataset that exposes models to repository context, execution feedback, and iterative code revision. Extensive experiments on repository-level SWE benchmarks show that mid-training on COMP-TRAJ effectively enhances capabilities in SWE. Our code and data are publicly available at https://github.com/uuam5/comp-traj.

Paper 768 Parameter-Efficient Domain Adversarial Training for Lightweight Framewise Phoneme Classification

Authors: Sohini Saha, Leslie M. Collins, Boyla Mainsah

Abstract Lightweight neural network models deployed on edge devices are susceptible to performance degradation under unseen acoustic conditions due to domain mismatch between training and deployment environments. This is particularly relevant in causal low latency applications such as cochlear implants (CIs), where limited temporal context and strict computational constraints restrict model robustness. Domain adversarial learning can improve generalization by encouraging representations that are both task-discriminative and domain-invariant; however, jointly learning these representations becomes difficult in limited capacity models. We introduce an adapter-based adversarial learning framework for robust and parameter-efficient domain adaptation. The proposed approach incorporates an adapter module into a pretrained causal backbone. We first establish a robust pretrained baseline network and subsequently perform adversarial adaptation under mismatched acoustic conditions. We evaluate the proposed framework on framewise phonetic group classification to guide phoneme-based time-frequency mask estimation. Empirical results demonstrate consistent improvements in phoneme classification over both standard supervised training and conventional adversarial learning, particularly under low signal-to-noise ratio levels, resulting in statistically significant improvements in objective intelligibility measures of CI vocoded speech.

Paper 769 M^4olGen: Multi-Agent, Multi-Stage Molecular Generation under Precise Multi-Property Constraints

Authors: Yizhan Li, Florence Cloutier, Sifan Wu, Ali Parviz, Boris Knyazev, Yan Zhang, Glen Berseth, Bang Liu

Abstract Generating molecules that satisfy precise numeric constraints over multiple physicochemical properties is critical and challenging. Although large language models (LLMs) are expressive, they struggle with precise multi-objective control and numeric reasoning without external structure and feedback. We introduce M^4olGen, a fragment-level, retrieval-augmented, two-stage framework for molecule generation under multi-property constraints. Stage I : Prototype generation: a multi-agent reasoner performs retrieval-anchored, fragment-level edits to produce a candidate near the feasible region. Stage II : RL-based fine-grained optimization: a fragment-level optimizer trained with Group Relative Policy Optimization (GRPO) applies one- or multi-hop refinements to explicitly minimize the property errors toward our target while regulating edit complexity and deviation from the prototype. A large, automatically curated dataset with reasoning chains of fragment edits and measured property deltas underpins both stages, enabling deterministic, reproducible supervision and controllable multi-hop reasoning. Unlike prior work, our framework better reasons about molecules by leveraging fragments and supports controllable refinement toward numeric targets. Experiments on generation under two sets of property constraints (QED, LogP, Molecular Weight and HOMO, LUMO) show consistent gains in validity and precise satisfaction of multi-property targets, outperforming strong LLMs and graph-based algorithms.

Paper 782 When Should Language Models Search? A Topology-Aware Analysis of Reasoning Strategies

Authors: Dibyanayan Bandyopadhyay, MD Tanzeel Adam Khan, Asif Ekbal

Abstract Search-based reasoning and chain-of-thought (CoT) prompting offer two contrasting inference-time paradigms for language models: CoT commits to a single linear trajectory, while search explores alternative partial reasoning paths before selecting an answer. In this work, we study this trade-off using A* as a controlled instantiation of heuristic-guided search. Across StrategyQA, LogicQA, HotpotQA, GSM8K, and MATH, and across three model scales, we find that search is not a universal upgrade over CoT; instead, the two strategies solve substantially different subsets of instances. To explain this complementarity, we introduce a Reasoning Topology Framework based on structural features computable before generation. A derived branching coefficient is strongly associated with when search helps across mixed-benchmark topology clusters, suggesting that the CoT--search boundary is partly explained by problem structure. We further train a topology-aware router that selects between CoT and A* without observing generated traces. Despite this restriction, it outperforms post-hoc routers based on semantic entropy and LLM-as-critic judgments. We attribute this to a Fluency--Correctness Asymmetry: fluent traces are often incorrect, making trace-level adjudication unreliable. Overall, our results suggest that adaptive reasoning should route by problem structure, deploying search only when the input topology warrants its higher cost. Code: https://github.com/newcodevelop/astar-cot-reasoning.

Paper 785 Evaluating LLM-as-a-Judge Beyond Score Alignment: A Psychometric Analysis of Residual Judging Difficulty

Authors: Longwei Cong, Sonja Hahn, Sebastian Gombert, Leon Camus, Fabian Zehner, Hendrik Drachsler, Ulf Kroehne

Abstract Large language models (LLMs) are widely used as automatic judges, with validity typically assessed via alignment with human scores. However, aggregate agreement fails to reveal whether humans and LLMs find the same evaluation cases difficult. In this paper, we study this problem in summarization evaluation from a psychometric perspective. We fit Many-Facet Rasch Models separately to human and LLM ratings to decompose scores into latent summary quality, rater severity, dimension severity, and rating-scale thresholds. Building on this decomposition, we define residual hardness as a model-adjusted measure of judging difficulty and compare whether human and LLM judges share the same hardness structure. Across 17 open-weight LLM judges on SummEval, we find that moderate alignment in latent summary quality does not imply alignment in residual hardness. Human and LLM judges differ in which summary--dimension units remain difficult, and this mismatch is strongly dimension-dependent. Consistency shows a pronounced LLM-hard shift, whereas coherence shows a human-hard shift. We further show that human-easy but LLM-hard cases are partially predictable from observable source--summary properties. These findings suggest that aggregate human alignment reflects only part of LLM-as-judge reliability, while psychometric residual diagnostics support more informative judge evaluation and more targeted human--LLM collaboration.

Paper 786 Efficient Multimodal Planning Agent for Visual Question-Answering

Authors: Zhuo Chen, Xinyu Geng, Xinyu Wang, Yong Jiang, Zhen Zhang, Pengjun Xie, Kewei Tu

Abstract Visual Question-Answering (VQA) is a challenging multimodal task that requires integrating visual and textual information to generate accurate responses. While multimodal Retrieval-Augmented Generation (mRAG) has shown promise in enhancing VQA systems by providing more evidence on both image and text sides, the default procedure that addresses VQA queries, especially the knowledge-intensive ones, often relies on multi-stage pipelines of mRAG with inherent dependencies. To mitigate the inefficiency limitations while maintaining VQA task performance, this paper proposes a method that trains a multimodal planning agent, dynamically decomposing the mRAG pipeline to solve the VQA task. Our method optimizes the

trade-off between efficiency and effectiveness by training the agent to intelligently determine the necessity of each mRAG step. In our experiments, the agent can help reduce redundant computations, cutting search time by over 60% compared to existing methods and decreasing costly tool calls. Meanwhile, experiments demonstrate that our method outperforms all baselines, including a Deep Research agent and a carefully designed prompt-based method, on average over six various datasets. Code is released.

Paper 788 YaPO: Learnable Sparse Activation Steering Vectors for Domain Adaptation

Authors: Abdelaziz BOUNHAR, Rania Elbadry, Hadi Abdine, Preslav Nakov, Michalis Vazirgiannis, Guokan Shang

Abstract Activation steering is a lightweight alternative to fine-tuning for aligning and personalizing large language models (LLMs). Prior work such as Bi-directional Preference Optimization (BiPO) learns dense steering vectors from preference data. However, dense interventions can entangle latent factors due to neuron multi-semanticity, hurting stability and fine-grained control, thus limiting their effectiveness especially in cultural alignment, where closely related values and behaviors (e.g., across Middle Eastern cultures) must be distinguished. We propose Yet another Policy Optimization (YaPO), a reference-free method that learns preference-optimized steering vectors in the sparse latent space of a Sparse Autoencoder (SAE). While the learned vector is not explicitly constrained to be sparse, it acts over sparse SAE features, providing a more disentangled and interpretable intervention basis. By optimizing sparse codes, YaPO yields more disentangled and interpretable steering directions. Across cultural alignment and several alignment-related behaviors (e.g., hallucination reduction and jailbreak resistance), YaPO converges faster and is more stable than dense steering baselines while improving performance. Importantly, YaPO preserves general knowledge, with no measurable degradation on MMLU. Overall, our results show that YaPO provides a general recipe for efficient, stable, and fine-grained alignment of LLMs, with broad applications to controllability and domain adaptation.

Paper 789 Evaluation of Multilingual LLMs Personalized Text Generation Capabilities Targeting Groups and Social-Media Platforms

Authors: Dominik Macko

Abstract Capabilities of large language models to generate multilingual coherent text have continuously enhanced in recent years, which opens concerns about their potential misuse. Previous research has shown that they can be misused for generation of personalized disinformation in multiple languages. It has also been observed that personalization negatively affects detectability of machine-generated texts; however, this has been studied in the English language only. In this work, we examine this phenomenon across 10 languages, while we focus not only on potential misuse of personalization capabilities, but also on potential benefits they offer. Overall, we cover 1080 combinations of various personalization aspects in the prompts, for which the texts are generated by 16 open-weights models drawn from 5 architecturally distinct families, including multiple size variants per family to assess scale effects (17,280 texts in total). Our results indicate that there are statistically significant differences in personalization quality of the generated texts when targeting demographic groups and when targeting social-media platforms across languages. Personalization towards platforms affects detectability of the generated texts in a higher scale, especially in English, where the personalization quality is the highest.

Paper 790 Reading Between the Layers: Cross-Modal Alignment and Representational Hierarchy in Vision Transformers

Authors: Sugam Karki, Raúl Vázquez, Yova Kementchedjhieva

Abstract While large vision-language models excel at optical character recognition, it remains unclear how their underlying vision encoders process text and its meaning, and whether they support anything resembling reading comprehension. We systematically study transformer-based vision encoders, using cross-modal alignment between rendered text and visual concepts together with probing and retrieval-based analyses as a controlled window into these broader capabilities. Our layer-wise analyses span orthographic, grammatical, and semantic organization, and reveal that strong cross-modal mapping emerges only in the deepest layers, after unimodal object and text recognition have already developed, coinciding with a progression from surface-level to semantic representations. Although this hierarchy is strongest in contrastively trained models, larger self-supervised models and text-rich training data partially support similar behaviors without explicit language supervision. Together, our results suggest that increasing model and training-data scale, as well as text-rich training data, can partially compensate for the absence of language supervision.

Paper 793 Large Language Models for Solid-Earth Geoscience: Applications, Modalities, and Memory-Reason-Act Techniques

Authors: Yisu Chen, Yiran Zhang, Yifeng Dai, Puzhen Wu, Sirui Li, Wei Liu, Yihao Ding

Abstract Solid Earth geoscience, geology, geochemistry, and geophysics, is increasingly served by large language models (LLMs) and multimodal LLMs (MLLMs) that combine text, geological maps, hyperspectral imagery, 3D structural models, spatial databases, and quantitative measurements into operational decision pipelines. Existing studies are fragmented across tasks, datasets, and system designs. This survey provides a structured review of LLM/MLLM research in solid-Earth geoscience along three axes: applications, modalities, and Memory-Reason-Act techniques. We identify five clusters of application (information extraction, knowledge Q&A and retrieval, multimodal map and 3D interpretation, spatial predictive modelling, and agentic pipelines), characterise four modalities (text, visual, spatial, and quantitative signal) and their encoding gaps, and analyse Memory, Reasoning, and Action design choices with their failure modes. We further examine data resources and bias, and identify the absence of closed-loop drill-outcome validation as the central methodological gap blocking next-generation solid-Earth foundation models.

Paper 794 Probing Reasoning Length Control in Language Models via Neurons

Authors: Dongwook Min, Jeonhwi Kang, Cheoneum Park, Seung-Hoon Na

Abstract Reasoning-oriented language models can generate thousands of tokens before answering, yet the internal factors that determine their reasoning length remain poorly understood. This paper studies whether reasoning length can be controlled through sparse neuron interventions, complementing prompt-based, budget-based, and controller-based approaches. The proposed method discovers reasoning-length-associated neurons from paired long and concise correct trajectories on a training split, ranks MLP neurons by long-minus-concise activation contrast, and tests the selected neurons on disjoint evaluation splits by scaling their activations during generation. The evaluation covers math, science, and commonsense reasoning tasks, including GSM8K, MATH500, ARC-Challenge, CommonsenseQA, and QASC, under a strict split-separated protocol that never discovers neurons on held-out evaluation examples. In several, but not all, task/model/budget regimes, selected neurons modulate generated length differently from layer-matched random controls, with a modest single-run specificity result on CommonsenseQA accompanied by a small accuracy increase. These results provide task-, model-, and prompt-protocol-specific

evidence that sparse MLP neurons can, in some split-controlled regimes, provide neuron-level control signals for generated reasoning length.

Paper 796 SeEM: Selective Dictionary-based Generation

Authors: Hongyuan Lu, Zixuan Li, Wai Lam

Abstract While dictionary-based approaches have shown promise in applying Large Language Models (LLMs) to generation tasks such as Neural Machine Translation (NMT), existing studies overlook several key challenges. Most methods rely on full or rule-selected dictionaries for prompting, assume that all dictionary entries are effectively utilised by LLMs, and construct dictionaries from bilingual parallel corpora, each of which limits performance. We propose a new dictionary-based translation framework that addresses these issues by fine-tuning LLMs with LoRA to autonomously select salient dictionary entries, choosing translations with the highest dictionary utilisation, and introducing a Multilingual Lexicon Induced Dictionary that leverages multilingual parallel corpora to improve dictionary quality. Experiments on the FLORES benchmark across multiple languages demonstrate the superior performance of our framework.

Paper 797 Beyond Consensus: Diagnosing Coverage, Majority Bias, and Grounding Equity in High-Stakes Multi-Document Diversity Summarization

Authors: Jane Arleth dela Cruz, Iris Hendrickx, Martha Larson

Abstract Multi-document diversity summarization (MDDS) in high-stakes domains such as crisis response demands fundamentally different capabilities than consensus-based summarization. While standard summarization focus on fluency and faithfulness, high-stakes tasks would also require: maximizing coverage of divergent information, identifying unique information clusters, and providing verifiable citations to source documents. Existing evaluation frameworks fail to capture these task-specific requirements. Our study introduces a comprehensive evaluation framework for high-stakes MDDS, validated on a multi-source crisis dataset that we developed covering ten crisis events. We propose novel metrics, Comprehensive Coverage Score, Frequency-Salience Correlation, and Citation-Frequency Correlation, to diagnose critical failure modes in six off-the-shelf large language models (LLMs). Our results allow us to identify three critical failure modes of LLMs: (1) a consensus trap in coverage: where models capture high-level information clusters, but miss specific, actionable details needed for decision-making; (2) majority bias, where more frequent information are prioritized over rare but important ones and; (3) a verifiability gap, where citations of rare claims are less reliable

Paper 800 DPDisc: From Factoid Questions to Data Product Requests for Open-World Data Product Discovery over Tables and Text

Authors: Liangliang Zhang, Nandana Mihindukulasooriya, Niharika S. D'Souza, Sola Shirai, Sarthak Dash, Yao Ma, Horst Samulowitz

Abstract Modern enterprises store vast, heterogeneous data across large data lakes, where the central challenge is shifting from answering an isolated question to discovering and assembling the tables and documents that together serve an analytical goal. We refer to such a coherent collection of related assets as a data product. Existing table-text benchmarks evaluate factoid question answering over individual tables and therefore do not capture this discovery and assembly problem. We introduce DPDisc, the first benchmark for data product discovery: given a high-level Data Product Request (DPR), a system must retrieve the coherent set of tables and passages that satisfy it. DPDisc is constructed by DPForge, an automated pipeline that repurposes table-text QA datasets by clustering related tables and passages into products, generating DPRs with an LLM ensemble, and

validating them through multi-phase LLM evaluation. It comprises 13,076 validated instances with full provenance, spanning open-domain and financial sources. Sparse, dense, and hybrid baselines establish that the task is feasible yet reveal large performance gaps across domains, motivating structure-aware retrieval.

Paper 803 Cleansing Fine-Tuning Data to Mitigate Backdoors in LLMs via the Base Model's Inherent Generative Abilities

Authors: Qi Zhao, Yan-Keon Elster, Christian Wressnegger

Abstract Relying on unverified, third-party data for training large language models or outsourcing fine-tuning to untrusted parties poses a severe risk to the model's integrity, giving rise to so-called data-poisoning attacks that introduce backdoors in LLMs. In this paper, we investigate to which extent fine-tuning can be hardened by using the base model's intrinsic generative capabilities. We sketch a simple defense strategy using mere prompt engineering that—in contrast to prior work—neither requires knowledge of the exact backdoor trig- ger used nor clean data as a reference. Instead, we leverage the assumption that (a) potential attackers aim for unsafe outputs and (b) triggers induce language incoherence. Across downstream tasks and attack types, we show that such a simple strategy can mitigate backdoors more effectively than existing defenses. In comparison to post-hoc defenses, our method substantially reduces the attack success rate up to 85% with Llama-3 while consistently preserving a high natural accuracy.