NAACL 2025

Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics

Proceedings of the Conference
Findings

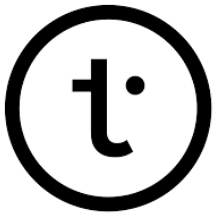
April 29 - May 4, 2025

The NAACL organizers gratefully acknowledge the support from the following sponsors.

Platinum

Gold

Bronze



Diversity and Inclusion Ally

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-195-7

Message from the General Chair

Message from the General Chair Welcome to the 2025 meeting of the Nations of the Americas Chapter of the Association for Computational Linguistics! I am proud to help organize the first NAACL conference to carry the new name of our organization, one that emphasizes inclusion for all of the Americas. I am also pleased to welcome you to Albuquerque, New Mexico, a state whose unique blend of cultural influences will make for an excellent backdrop for NAACL 2025, especially with this year's special theme on NLP in a Multicultural World.

This year's program benefited from a now mature ACL rolling review process. I would like to extend a big thank you to our ARR Editors in Chief, Viviane Moreira, Anna Rogers, and Michael White, who were very helpful not just with reviewing for the main conference, but who also shared their OpenReview expertise with chairs from our other tracks. We also benefited from the helpful advice of last year's NAACL general and program chairs: Katrin Erk, Kevin Duh, Helena Gomez, and Steven Bethard. Finally, Ryan Cotterell stepped up to help with the publications process and software, even while not serving as publications chair.

Of course, a conference of this magnitude cannot come together without some drama; in our case, we had some unexpected funding shortages from traditional government sources. We would like to extend a huge thank you to the boards of both the ACL and NAACL for filling those funding gaps, ensuring that our important D&I, volunteer, and student author support programs continue to thrive.

The virtual component of our conference is crucial to an inclusive, affordable experience for all NAACL authors and attendees. This year, we made small refinements to the hybrid format, mirroring NAACL 2024's choices to combine a virtual poster session via Gather with asynchronous online content via Underline. Our virtual poster session will be hosted on May 6, the Tuesday after the conference, in the hopes that promoting it at the conference's plenary sessions will help boost attendance. We opted not to have virtual oral presentations at the in-person conference, as those continue to be tricky to get right. To participants, virtual as well as in-person: Please let us know what worked for you and what did not, so we can continue to improve the hybrid experience.

The job of General Chair is a strange one, as it mostly involves cheering on many other people as they do amazing work. I have been fortunate to have been teamed with an excellent set of program chairs; to Luis Chiruzzo, Alan Ritter, and Lu Wang: thanks for everything, I'm very proud of what we've built together. I'd also like to extend my heartfelt thanks to Jenn Rachford (ACL) and Damira Mršic (Underline) who provide the knowledge, continuity and professionalism to bring all of this together.

Many thanks also to:

- Workshop chairs: Saab Mansour, Kenton Murray, and Alexis Palmer
- Tutorial chairs: Maria Lomeli, Swabha Swayamdipta, and Rui Zhang
- Demo chairs: Nouha Dziri, Shizhe Diao, and Sean (Xiang) Ren
- Industry track chairs: Weizhu Chen, Xue-Yong Fu, Mohammad Kachuee, and Yi Yang
- Student research workshop chairs: Abteen Ebrahimi, Emmy Liu, and Samar Haider, and their faculty advisors Maria Leonor Pacheco and Shira Wein
- Publication chairs: Arman Cohan, Manling Li, and Yichao Zhou
- Website chairs: Arya McCarthy and Vered Shwartz

- Publicity and social media chairs: Eleftheria Briakou, Tuhin Chakrabarty, and Ximena Gutierrez-Vasques
- Diversity and inclusion chairs: Akiko I. Eriguchi, Chi-Kiu (Jackie) Lo, and Niloofar Mireshghallah
- Sponsorship chairs: Prithviraj (Raj) Ammanabrolu and Maha Elbayad
- Volunteers chairs: Robin Jia and David Mortensen
- Ethics chairs: Manuel Mager and Yulia Tsvetkov
- Handbook chair: Winston Wu
- Best paper committee chairs: Marine Carpuat and Anna Rumshisky
- Visa chairs: Eduardo Blanco and Parisa Kordjamshidi
- Virtual infrastructure chair: Jieyu Zhao

Whenever possible, I tried to populate each committee with someone who had served in the same role in NAACL 2024, to provide continuity, so I'll extend an extra thanks to all chairs who accepted this second year of service. Thanks also to the members of the ACL and NAACL Executive Committees for their support, feedback, and advice.

Finally, I would like to thank all authors, invited speakers and panelists, area chairs and reviewers, volunteers and session chairs, and all attendees, in-person and virtual. The conference is nothing without you.

Welcome again and enjoy the conference!

Colin Cherry
Google
NAACL 2025 General Chair

Message from the Program Chairs

Message from the Program Chairs Welcome to the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics! NAACL 2025 marks an exciting milestone as the first conference held under our new name, reflecting our commitment to greater inclusivity across the diverse communities of the Americas. NAACL 2025 is a hybrid conference, and we are excited to have attendees and presenters join us both in person in Albuquerque and online from around the globe. We are thrilled to welcome you to what promises to be a vibrant and engaging conference.

Special Theme: NLP in a Multicultural World Current NLP tools and models, particularly large language models (LLMs), rely on vast amounts of data for training. However, this data often over-represents a small number of dominant languages, and even within those, tends to prioritize certain geographical or cultural varieties. As a result, a long tail of under-represented languages, dialects, and cultural contexts remains largely overlooked by the NLP community. For NAACL 2025, we introduced a special theme track on "NLP in a Multicultural World." With this theme track, we sought to foster discussion and research on how NLP can better serve the linguistic and cultural diversity of the world. We encouraged contributions on topics such as cultural localization of language models, new NLP applications to support people from diverse cultures, revitalization or refunctionalization of endangered or sleeping languages, analysis of cultural biases in language models, and historical considerations and diachronic analysis. This track was dedicated to developing more inclusive, culturally aware NLP techniques that reflect and support the vibrant multicultural world we live in.

We received 71 submissions to the special theme, of which 23 were accepted for presentation at the conference. We hope these papers spark meaningful conversations and inspire future work in this important and evolving area of research.

Review Process 3,185 papers were submitted to the October ARR cycle, of which we estimate 3,099 were intended to be submitted to NAACL based on the "preferred venue" field in the submission form. We also received 147 papers from previous ARR cycles committed to NAACL. The program chairs recruited 98 Senior Area Chairs to view reviews and metareviews provided by ARR and make final recommendations on which papers to accept to both the main conference and Findings. 1,432 Area Chairs wrote metareviews for ARR, and 10,648 reviewers wrote reviews for the submitted papers.

Acceptance Rate Calculating an acceptance rate is challenging due to the multi-step ARR review process, in which papers are first submitted to ARR to get reviews, then authors commit their papers (together with reviews) to a specific *ACL conference. Of the 3,185 papers submitted to the October ARR cycle, we estimate that 3,099 intended to submit to NAACL. Based on this information, we estimate that 22% of papers submitted to the October cycle and intended for NAACL were accepted to the main conference, and another 15% were accepted to Findings, bringing the total estimated acceptance rate for papers accepted to be presented at the conference (Main + Findings) to 37%. Out of the 1,647 papers committed to NAACL with ARR reviews, 719 were accepted to the main conference, and 477 were accepted to Findings. 40 papers were desk rejected or withdrawn.

Presentation Format At NAACL 2025, papers were assigned one of three possible presentation modes: in-person participants could be assigned oral or poster presentations, while virtual participants could present posters. We selected 246 of the papers accepted to the main conference as oral presentations, and

the rest of them were assigned as posters, together with all the Findings papers. Oral presentations were assigned a 15-minute slot, with 12 minutes for presentation and 3 minutes for questions. For choosing papers as oral presentations, we first split all papers according to track, sorted them according to overall score (considering review, metareview, SAC recommendation), and took into consideration the authors' presentation preference. Then we grouped papers in sets of 6. Some tracks had very few accepted papers, so some of them were grouped together to form areas of affinity.

Program Format NAACL 2025 consists both of an in-person and a virtual conference, held on different days. The virtual part of the conference is held after the in-person one and a few days later (on May 6), so participants traveling home after the in-person conference could attend the virtual conference. The conference program includes 3 keynote speakers: Rada Mihalcea (University of Michigan), Mike Lewis (Meta), and Josh Tenenbaum (Massachusetts Institute of Technology). 260 papers are scheduled to be presented as oral presentations (also including papers from TACL, CL, and the industry track), 594 papers are scheduled as in-person posters, and 256 virtual posters.

Gratitude NAACL 2025 would not have been possible without the hard work of all people involved. We thank everyone who contributed, including:

- The General Chair, Colin Cherry.
- The ARR Editors-in-Chief of the October 2024 cycle: Viviane Moreira, Anna Rogers, Michael White.
- The OpenReview team, especially Rachel Smart.
- The 98 Senior Area Chairs.
- The 1,432 Area Chairs and 10,648 Reviewers.
- The best paper committee chairs, Marine Carpuat and Anna Rumshisky.
- The ethics chairs, Yulia Tsvetkov and Manuel Mager, and their team of reviewers.
- The website chairs, Vered Shwartz and Arya McCarthy.
- The publication chairs, Yichao Zhou, Manling Li, and Arman Cohan.
- The publicity chairs, Ximena Gutierrez-Vasques, Eleftheria Briakou, and Tuhin Chakrabarty.
- The volunteers chairs, Robin Jia and David Mortensen.
- The visa chairs, Eduardo Blanco and Parisa Kordjamshidi.
- The ACL Anthology Director, Matt Post, and his team.
- The Program Chairs of EMNLP 2024 (Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung (Vivian) Chen) and NAACL 2024 (Kevin Duh, Helena Gomez, and Steven Bethard).
- Damira Mršic and the Underline Team.
- Jenn Rachhford and the entire conference support staff.

Luis Chiruzzo, Universidad de la República, Uruguay
Alan Ritter, Georgia Institute of Technology
Lu Wang, University of Michigan
NAACL 2025 Program Committee Co-Chairs
April 2025

Table of Contents

From Lazy to Prolific: Tackling Missing Labels in Open Vocabulary Extreme Classification by Positive-Unlabeled Sequence Learning
Ranran Haoran Zhang, Bensu Uçar, Soumik Dey, Hansi Wu, Binbin Li and Rui Zhang 1

DiffZOO: A Purely Query-Based Black-Box Attack for Red-teaming Text-to-Image Generative Model via Zeroth Order Optimization
Pucheng Dang, Xing Hu, Dong Li, Rui Zhang, Qi Guo and Kaidi Xu 17

MedOdyssey: A Medical Domain Benchmark for Long Context Evaluation Up to 200K Tokens
Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang and Tong Ruan 32

Can LLMs Learn Macroeconomic Narratives from Social Media?
Almog Gueta, Amir Feder, Zorik Gekhman, Ariel Goldstein and Roi Reichart 57

Code-Optimise: Self-Generated Preference Data for Correctness and Efficiency
Leonidas Gee, Milan Gritta, Gerasimos Lampouras and Ignacio Iacobacci 79

People will agree what I think: Investigating LLM's False Consensus Effect
Junhyuk Choi, Yeseon Hong and Bugeun Kim 95

LawInstruct: A Resource for Studying Language Model Adaptation to the Legal Domain
Joel Niklaus, Lucia Zheng, Arya D. McCarthy, Christopher Hahn, Brian M Rosen, Peter Henderson, Daniel E. Ho, Garrett Honke, Percy Liang and Christopher D Manning 127

Stephanie: Step-by-Step Dialogues for Mimicking Human Interactions in Social Conversations
Hao Yang, Hongyuan Lu, Xinhua Zeng, Yang Liu, Xiang Zhang, Haoran Yang, Yumeng Zhang, Shan Huang, Yiran Wei and Wai Lam 153

ConShift: Sense-based Language Variation Analysis using Flexible Alignment
Clare Arrington, Mauricio Gruppi and Sibel Adali 167

Breaking the Stigma! Unobtrusively Probe Symptoms in Depression Disorder Diagnosis Dialogue
Jieming Cao, Chen Huang, Yanan Zhang, Ruibo Deng, Jincheng Zhang and Wenqiang Lei . . 182

ToVo: Toxicity Taxonomy via Voting
Tinh Son Luong, Thanh-Thien Le, Thang Viet Doan, Linh Ngo Van, Thien Huu Nguyen and Nguyen Thi Ngoc Diep 201

HALLUCANA: Fixing LLM Hallucination with A Canary Lookahead
Tianyi Li, Erenay Dayanik, Shubhi Tyagi and Andrea Pierleoni 213

Enhancing Adversarial Transferability in Visual-Language Pre-training Models via Local Shuffle and Sample-based Attack
Xin Liu, Aoyang Zhou and Kun He 231

Dis2Dis: Explaining Ambiguity in Fact-Checking
Ieva Staliunaite and Andreas Vlachos 246

Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement
Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, Furong Huang and Cao Xiao 268

RePD: Defending Jailbreak Attack through a Retrieval-based Prompt Decomposition Process
Peiran Wang, Xiaogeng Liu and Chaowei Xiao 283

ChatCRS: Incorporating External Knowledge and Goal Guidance for LLM-based Conversational Recommender Systems
Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan and Haizhou Li 295

Data-Efficiently Learn Large Language Model for Universal 3D Scene Perception
Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, Tao Jin and Zhou Zhao 313

UnifiedMLLM: Enabling Unified Representation for Multi-modal Multi-tasks With Large Language Model
Zhaowei Li, Wei Wang, YiQing Cai, Qi Xu, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang and Tao Wang 334

PEMV: Improving Spatial Distribution for Emotion Recognition in Conversations Using Proximal Emotion Mean Vectors
Chen Lin, Fei Li, Donghong Ji and Chong Teng 345

DiscoverGPT: Multi-task Fine-tuning Large Language Model for Related Table Discovery
Xuming Hu, Xiao Qin, Chuan Lei, Asterios Katsifodimos, Zhengyuan Shen, Balasubramaniam Srinivasan and Huzefa Rangwala 358

Can GPT-4 Sway Experts' Investment Decisions?
Takehiro Takayanagi, Hiroya Takamura, Kiyoshi Izumi and Chung-Chi Chen 374

PolyJoin: Semantic Multi-key Joinable Table Search in Data Lakes
Xuming Hu, Chuan Lei, Xiao Qin, Asterios Katsifodimos, Christos Faloutsos and Huzefa Rangwala 384

Marrying LLMs with Dynamic Forecasting: A Graph Mixture-of-expert Perspective
Dapeng Jiang and Xiao Luo 396

DialogGen: Multi-modal Interactive Dialogue System with Multi-turn Text-Image Generation
Minbin Huang, Yanxin Long, Xinchi Deng, Ruihang Chu, Jiangfeng Xiong, Xiaodan Liang, Hong Cheng, Qinglin Lu and Wei Liu 411

RELexED: Retrieval-Enhanced Legal Summarization with Exemplar Diversity
Santosh T.y.s.s, Chen Jia, Patrick Goroncy and Matthias Grabmair 427

CLaMP 2: Multimodal Music Information Retrieval Across 101 Languages Using Large Language Models
Shangda Wu, Yashan Wang, Ruibin Yuan, Guo Zhancheng, Xu Tan, Ge Zhang, Monan Zhou, Jing Chen, Xuefeng Mu, Yuejie Gao, Yuanliang Dong, Jiafeng Liu, Xiaobing Li, Feng Yu and Maosong Sun 435

LogRules: Enhancing Log Analysis Capability of Large Language Models through Rules
Xin Huang, Ting Zhang and Wen Zhao 452

Audio Description Generation in the Era of LLMs and VLMs: A Review of Transferable Generative AI Technologies
Yingqiang Gao, Lukas Fischer, Alexa Lintner and Sarah Ebling 471

Adaptive Retrieval-Augmented Generation for Conversational Systems
Xi Wang, Procheta Sen, Ruizhe Li and Emine Yilmaz 491

Multimodal Generation with Consistency Transferring
Junxiang Qiu, Jinda Lu and Shuo Wang 504

On the Impact of Noise in Differentially Private Text Rewriting
Stephen Meisenbacher, Maulik Chevli and Florian Matthes 514

Teaching Large Language Models Number-Focused Headline Generation With Key Element Rationales
Zhen Qian, Xiuzhen Zhang, Xiaofei Xu and Feng Xia 533

Zero-Shot Strategies for Length-Controllable Summarization
Fabian Retkowski and Alexander Waibel 551

SIMPLOT: Enhancing Chart Question Answering by Distilling Essentials
Wonjoong Kim, Sangwu Park, Yeonjun In, Seokwon Han and Chanyoung Park 573

InstructAny2Pix: Image Editing with Multi-Modal Prompts
Shufan Li, Harkanwar Singh and Aditya Grover 594

Lost in Overlap: Exploring Logit-based Watermark Collision in LLMs
Yiyang Luo, Ke Lin, Chao Gu, Jiahui Hou, Lijie Wen and Luo Ping 620

Prompt-Guided Selective Masking Loss for Context-Aware Emotive Text-to-Speech
Yejin Jeon, Youngjae Kim, Jihyun Lee and Gary Lee 638

Identifying and Mitigating Social Bias Knowledge in Language Models
Ruizhe Chen, Yichen Li, Jianfei Yang, Yang Feng, Joey Tianyi Zhou, Jian Wu and Zuozhu Liu 651

DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications
Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav and EngSiong Chng 673

Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs
Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao and Salman Avestimehr 691

Joint Learning Event-Specific Probe and Argument Library with Differential Optimization for Document-Level Multi-Event Extraction
Jianpeng Hu, Chao Xue, Chunqing Yu, JiaCheng Xu and Chengxiang Tan 714

Synonym-unaware Fast Adversarial Training against Textual Adversarial Attacks
Yichen Yang, Xin Liu and Kun He 727

Tethering Broken Themes: Aligning Neural Topic Models with Labels and Authors
Mayank Nagda, Phil Ostheimer and Sophie Fellenz 740

Towards Zero-Shot Multimodal Machine Translation
Matthieu Futeral, Cordelia Schmid, Benoît Sagot and Rachel Bawden 761

Large-Scale Corpus Construction and Retrieval-Augmented Generation for Ancient Chinese Poetry: New Method and Data Insights
Yang Liu, Lan Lan, Jiahuan Cao, Hiuyi Cheng, Kai Ding and Lianwen Jin 779

OpenBioNER: Lightweight Open-Domain Biomedical Named Entity Recognition Through Entity Type Description
Alessio Cocchieri, Giacomo Frisoni, Marcos Martínez Galindo, Gianluca Moro, Giuseppe Tagliavini and Francesco Candoli 818

Dialetto, ma Quanto Dialetto? Transcribing and Evaluating Dialects on a Continuum
Ryan Soh-Eun Shim and Barbara Plank 838

Linguistically Grounded Analysis of Language Models using Shapley Head Values
Marcell Fekete and Johannes Bjerva 850

How Do Large Language Models Perform in Dynamical System Modeling
Xiao Luo, Binqi Chen, Haixin Wang, Zhiping Xiao, Ming Zhang and Yizhou Sun 866

LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models
Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li and Ziwei Liu 881

Pairwise Prompt-Based Tuning with Parameter Efficient Fast Adaptation for Generalized Zero-Shot Intent Detection
Xiaotong Zhang, Qianru Zhou, Han Liu and Hong Yu 917

FaithfulPersona: Balancing Faithfulness and Personalization in Code Explanations through Self-Critique
Zhuang Luo, Yichuan Li, Zexing Xu, Kyumin Lee and S. Rasoul Etesami 930

Efficient Multi-Agent Collaboration with Tool Use for Online Planning in Complex Table Question Answering
Wei Zhou, Mohsen Mesgar, Annemarie Friedrich and Heike Adel 945

Ground Every Sentence: Improving Retrieval-Augmented LLMs with Interleaved Reference-Claim Generation
Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu and Yanghua Xiao 969

Understanding the Role of Mental Models in User Interaction with an Adaptive Dialog Agent
Lindsey Morgan Vanderlyn, Dirk Väth and Thang Vu 989

CoPERLex: Content Planning with Event-based Representations for Legal Case Summarization
Santosh T.y.s.s, Youssef Farag and Matthias Grabmair 1016

DisComp: A Two-Stage Prompt Optimization Framework Combining Task-Agnostic and Task-Aware Compression
Liu Quancai, Haihui Fan, Jinchao Zhang, Lixiangfang Lixiangfang, Lichuanrong Lichuanrong and Bo Li 1033

A Large-Scale Benchmark for Vietnamese Sentence Paraphrases
Sang Quang Nguyen and Kiet Van Nguyen 1045

RAMQA: A Unified Framework for Retrieval-Augmented Multi-Modal Question Answering
Yang Bai, Christan Grant and Daisy Zhe Wang 1061

MultiCAT: Multimodal Communication Annotations for Teams
Adarsh Pyarelal, John M Culnan, Ayesha Qamar, Meghavarshini Krishnaswamy, Yuwei Wang, Cheonkam Jeong, Chen Chen, Md Messal Monem Miah, Shahriar Hormozi, Jonathan Tong and Ruihong Huang 1077

Prototype Tuning: A Meta-Learning Approach for Few-Shot Document-Level Relation Extraction with Large Language Models
Dinghao Pan, Yuanyuan Sun, Bo Xu, Jiru Li, Zhihao Yang, Ling Luo, Hongfei Lin and Jian Wang 1112

LegalSeg: Unlocking the Structure of Indian Legal Judgments Through Rhetorical Role Classification
Shubham Kumar Nigam, Tanmay Dubey, Govind Sharma, Noel Shallum, Kripabandhu Ghosh and Arnab Bhattacharya 1129

Claim-Guided Textual Backdoor Attack for Practical Applications
Minkyoo Song, Hanna Kim, Jaehan Kim, Youngjin Jin and Seungwon Shin 1145

ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities
Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang and Ruoming Pang 1160

SusGen − GPT: A Data-Centric LLM for Financial NLP and Sustainability Report Generation
Qilong Wu, Xiaoneng Xiang, Huang Hejia, Xuan Wang, Yeo Wei Jie, Ranjan Satapathy, Ricardo Shirota Filho and Bharadwaj Veeravalli 1184

GrEmLIn: A Repository of Green Baseline Embeddings for 87 Low-Resource Languages Injected with Multilingual Graph Knowledge
Daniil Gurgurov, Rishu Kumar and Simon Ostermann 1204

In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation
Armel Randy Zebaze, Benoît Sagot and Rachel Bawden 1222

Self-Training Large Language Models for Tool-Use Without Demonstrations
Ne Luo, Aryo Pradipta Gema, Xuanli He, Emile Van Krieken, Pietro Lesci and Pasquale Minervini 1253

Can Large Language Models Generate High-quality Patent Claims?
Lekang Jiang, Caiqi Zhang, Pascal A. Scherz and Stefan Goetz 1272

Obliviate: Neutralizing Task-agnostic Backdoors within the Parameter-efficient Fine-tuning Paradigm
Jaehan Kim, Minkyoo Song, Seung Ho Na and Seungwon Shin 1288

CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmented Generation
Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen and Zhicheng Dou 1308

Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs
Itai Mondshine, Tzuf Paz-Argaman and Reut Tsarfaty 1331

QuaLLM: An LLM-based Framework to Extract Quantitative Insights from Online Forums
Varun Nagaraj Rao, Eesha Agarwal, Samantha Dalal, Dana Calacci and Andrés Monroy-Hernández 1355

The Promises and Pitfalls of LLM Annotations in Dataset Labeling: a Case Study on Media Bias Detection
Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa and Timo Spinde 1370

Mechanistic Unveiling of Transformer Circuits: Self-Influence as a Key to Model Reasoning
Lin Zhang, Lijie Hu and Di Wang 1387

Intrinsic Model Weaknesses: How Priming Attacks Unveil Vulnerabilities in Large Language Models
Yuyi Huang, Runzhe Zhan, Derek F. Wong, Lidia S. Chao and Ailin Tao 1405

AdParaphrase: Paraphrase Dataset for Analyzing Linguistic Features toward Generating Attractive Ad Texts
Soichiro Murakami, Peinan Zhang, Hidetaka Kamigaito, Hiroya Takamura and Manabu Okumura 1426

Token Weighting for Long-Range Language Modeling
Falko Helm, Nico Daheim and Iryna Gurevych 1440

Learning to Explore and Select for Coverage-Conditioned Retrieval-Augmented Generation
Takyoung Kim, Kyungjae Lee, Young Rok Jang, Ji Yong Cho, Gangwoo Kim, Minseok Cho and Moontae Lee 1460

LayAlign: Enhancing Multilingual Reasoning in Large Language Models via Layer-Wise Adaptive Fusion and Alignment Strategy
Zhiwen Ruan, Yixia Li, He Zhu, Longyue Wang, Weihua Luo, Kaifu Zhang, Yun Chen and Guanhua Chen 1481

On the Impacts of Contexts on Repository-Level Code Generation
Nam Le Hai, Dung Manh Nguyen and Nghi D. Q. Bui 1496

From Argumentation to Deliberation: Perspectivized Stance Vectors for Fine-grained (Dis)agreement Analysis
Moritz Plenz, Philipp Heinisch, Janosch Gehring, Philipp Cimiano and Anette Frank 1525

LVLM-Compress-Bench: Benchmarking the Broader Impact of Large Vision-Language Model Compression
Souvik Kundu, Anahita Bhiwandiwalla, Sungduk Yu, Phillip Howard, Tiep Le, Sharath Nittur Sridhar, David Cobbley, Hao Kang and Vasudev Lal 1554

Does Generative AI speak Nigerian-Pidgin?: Issues about Representativeness and Bias for Multilingualism in LLMs
David Ifeoluwa Adelani, A. Seza Doğruöz, Iyanuoluwa Shode and Anuoluwapo Aremu 1571

A Guide To Effectively Leveraging LLMs for Low-Resource Text Summarization: Data Augmentation and Semi-supervised Approaches
Gaurav Sahu, Olga Vechtomova and Issam H. Laradji 1584

Decoding Dark Matter: Specialized Sparse Autoencoders for Interpreting Rare Concepts in Foundation Models
Aashiq Muhamed, Mona T. Diab and Virginia Smith 1604

MAiDE-up: Multilingual Deception Detection of AI-generated Hotel Reviews
Oana Ignat, Xiaomeng Xu and Rada Mihalcea 1636

LeCoPCR: Legal Concept-guided Prior Case Retrieval for European Court of Human Rights cases
Santosh T.y.s.s, Isaac Misael Olguín Nolasco and Matthias Grabmair 1654

How much do contextualized representations encode long-range context?
Simeng Sun and Cheng-Ping Hsieh 1662

Data Poisoning for In-context Learning
Pengfei He, Han Xu, Yue Xing, Hui Liu, Makoto Yamada and Jiliang Tang 1680

Synthetic Audio Helps for Cognitive State Tasks
Adil Soubki, John Murzaku, Peter Zeng and Owen Rambow 1701

BioEL: A Comprehensive Python Package for Biomedical Entity Linking
Prasanth Bathala, Christophe Ye, Batuhan Nursal, Shubham Lohiya, David Kartchner and Cassie S. Mitchell 1709

PairScale: Analyzing Attitude Change with Pairwise Comparisons
Rupak Sarkar, Patrick Y. Wu, Kristina Miler, Alexander Miserlis Hoyle and Philip Resnik . . 1722

Semantic Consistency-Based Uncertainty Quantification for Factuality in Radiology Report Generation
Chenyu Wang, Weichao Zhou, Shantanu Ghosh, Kayhan Batmanghelich and Wenchao Li . . 1739

RewardBench: Evaluating Reward Models for Language Modeling
Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith and Hannaneh Hajishirzi 1755

Evaluating Vision-Language Models for Emotion Recognition
Sree Bhattacharyya and James Z. Wang 1798

Tomato, Tomahto, Tomate: Do Multilingual Language Models Understand Based on Subword-Level Semantic Concepts?
Crystina Zhang, Jing Lu, Vinh Q. Tran, Tal Schuster, Donald Metzler and Jimmy Lin 1821

Open Domain Question Answering with Conflicting Contexts
Siyi Liu, Qiang Ning, Kishaloy Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba and Dan Roth 1838

The Geometry of Prompting: Unveiling Distinct Mechanisms of Task Adaptation in Language Models
Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho and SueYeon Chung 1855

Biases in Opinion Dynamics in Multi-Agent Systems of Large Language Models: A Case Study on Funding Allocation
Pedro Cisneros-Velarde 1889

CaseSumm: A Large-Scale Dataset for Long-Context Summarization from U.S. Supreme Court Opinions
Mourad Heddaya, Kyle MacMillan, Hongyuan Mei, Chenhao Tan and Anup Malani 1917

Chasing Random: Instruction Selection Strategies Fail to Generalize
Harshita Diddee and Daphne Ippolito 1943

Can't Hide Behind the API: Stealing Black-Box Commercial Embedding Models
Manveer Singh Tamber, Jasper Xian and Jimmy Lin 1958

CAMEL-Bench: A Comprehensive Arabic LMM Benchmark
Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Husain Salem Abdulla Alharthi, Ines Riahi, Abduljalil Radman, Jorma Laaksonen, Fahad Shahbaz Khan, Salman Khan and Rao Muhammad Anwer 1970

ProxyLM: Predicting Language Model Performance on Multilingual Tasks via Proxy Models
David Anugraha, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan and En-Shiun Annie Lee 1981

SimSMoE: Toward Efficient Training Mixture of Experts via Solving Representational Collapse
Giang Do, Hung Le and Truyen Tran 2012

UniRAG: Universal Retrieval Augmentation for Large Vision Language Models
Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen and Jimmy Lin 2026

Evaluating the Performance of Large Language Models via Debates
Behrad Moniri, Hamed Hassani and Edgar Dobriban 2040

CausalGraph2LLM: Evaluating LLMs for Causal Queries
Ivaxi Sheth, Bahare Fatemi and Mario Fritz 2076

PuzzleGPT: Emulating Human Puzzle-Solving Ability for Time and Location Prediction
Hammad Ayyubi, Xuande Feng, Junzhang Liu, Xudong Lin, Zhecan Wang and Shih-Fu Chang 2099

SAFR: Neuron Redistribution for Interpretability
Ruidi Chang, Chunyuan Deng and Hanjie Chen 2117

GPT-4V Cannot Generate Radiology Reports Yet
Yuyang Jiang, Chacha Chen, Dang Nguyen, Benjamin M. Mervak and Chenhao Tan 2127

Is Semantic Chunking Worth the Computational Cost?
Renyi Qu, Ruixuan Tu and Forrest Sheng Bao 2155

On Using Arabic Language Dialects in Recommendation Systems
Abdulla Alshabanah and Murali Annavaram 2178

Assessing LLMs for Zero-shot Abstractive Summarization Through the Lens of Relevance Paraphrasing
Hadi Askari, Anshuman Chhabra, Muhao Chen and Prasant Mohapatra 2187

Beyond Silent Letters: Amplifying LLMs in Emotion Recognition with Vocal Nuances
Zehui wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci and Julia Hirschberg 2202

DomainSum: A Hierarchical Benchmark for Fine-Grained Domain Shift in Abstractive Text Summarization
Haohan Yuan and Haopeng Zhang 2219

Test-time Backdoor Mitigation for Black-Box Large Language Models with Defensive Demonstrations
Wenjie Jacky Mo, Jiashu Xu, Qin Liu, Jiongxiao Wang, Jun Yan, Hadi Askari, Chaowei Xiao and Muhao Chen 2232

All that Glitters": Techniques for Evaluations with Unreliable Model and Human Annotations
Michael Hardy 2250

KwaiChat: A Large-Scale Video-Driven Multilingual Mixed-Type Dialogue Corpus
Xiaoming Shi, Zeming Liu, Yiming Lei, Chenkai Zhang, Haitao Leng, Chuan Wang, Qingjie Liu, Wanxiang Che and Yunhong Wang 2279

GenEOL: Harnessing the Generative Power of LLMs for Training-Free Sentence Embeddings
Raghuveer Thirukovalluru and Bhuwan Dhingra 2295

Attention Tracker: Detecting Prompt Injection Attacks in LLMs
Kuo-Han Hung, Ching-Yun Ko, Ambrish Rawat, I-Hsin Chung, Winston H. Hsu and Pin-Yu Chen 2309

Unsupervised Speech-text word-level alignment with Dynamic Programming
Tianshu Yu, Zihan Gong, Minghuan Tan, Guhong Chen and Min Yang 2323

SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis
Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, Zheng Cheng, Zifeng Zhao, Linfeng Zhang and Guolin Ke 2335

Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks
Samuele Poppi, Zheng Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang and Jianfeng Chi 2358

MASSW: A New Dataset and Benchmark Tasks for AI-Assisted Scientific Workflows
Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee and Qiaozhu Mei 2373

Neuro-symbolic Training for Reasoning over Spatial Language
Tanawan Premsri and Parisa Kordjamshidi 2395

On Localizing and Deleting Toxic Memories in Large Language Models
Anubrata Das, Manoj Kumar, Ninareh Mehrabi, Anil Ramakrishna, Anna Rumshisky, Kai-Wei Chang, Aram Galstyan, Morteza Ziyadi and Rahul Gupta 2415

DiVISe: Direct Visual-Input Speech Synthesis Preserving Speaker Characteristics And Intelligibility
Yifan Liu, Yu Fang and Zhouhan Lin 2424

GraphICL: Unlocking Graph Learning Potential in LLMs through Structured Prompt Design
Yuanfu Sun, Zhengnan Ma, Yi Fang, Jing Ma and Qiaoyu Tan 2440

FIDELITY: Fine-grained Interpretable Distillation for Effective Language Insights and Topic Yielding
Divyansh Singh, Brodie Mather, Demi Zhang, Patrick Lehman, Justin Ho and Bonnie J Dorr 2460

Classic4Children: Adapting Chinese Literary Classics for Children with Large Language Model
Jiali Chen, Xusen Hei, Yuqi Xue, Zihan Wu, Jiayuan Xie and Yi Cai 2473

Considering Length Diversity in Retrieval-Augmented Summarization
Juseon-Do Juseon-Do, Jaesung Hwang, Jingun Kwon, Hidetaka Kamigaito and Manabu Okumura 2489

LMOD: A Large Multimodal Ophthalmology Dataset and Benchmark for Large Vision-Language Models
Zhenyue Qin, Yu Yin, Dylan Campbell, Xuansheng Wu, Ke Zou, Ninghao Liu, Yih Chung Tham, Xiuzhen Zhang and Qingyu Chen 2501

Syntriever: How to Train Your Retriever with Synthetic Data from LLMs
Minsang Kim and Seung Jun Baek 2523

DynClean: Training Dynamics-based Label Cleaning for Distantly-Supervised Named Entity Recognition
Qi Zhang, Huitong Pan, Zhijia Chen, Longin Jan Latecki, Cornelia Caragea and Eduard Dragut 2540

An Efficient Rehearsal Scheme for Catastrophic Forgetting Mitigation during Multi-stage Fine-tuning
Andrew Bai, Chih-Kuan Yeh, Cho-Jui Hsieh and Ankur Taly 2557

COAST: Enhancing the Code Debugging Ability of LLMs through Communicative Agent Based Data Synthesis
Weiqing Yang, Hanbin Wang, Zhenghao Liu, Xinze Li, Yukun Yan, Shuo Wang, Yu Gu, Minghe Yu, Zhiyuan Liu and Ge Yu 2570

Chain-of-Probe: Examining the Necessity and Accuracy of CoT Step-by-Step
Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu and Kam-Fai Wong 2586

INDIC QA BENCHMARK: A Multilingual Benchmark to Evaluate Question Answering capability of LLMs for Indic Languages
Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal and Ganesh Ramakrishnan 2607

Learning with Less: Knowledge Distillation from Large Language Models via Unlabeled Data
Juanhui Li, Sreyashi Nag, Hui Liu, Xianfeng Tang, Sheikh Muhammad Sarwar, Limeng Cui, Hansu Gu, Suhang Wang, Qi He and Jiliang Tang 2627

LSDC: An Efficient and Effective Large-Scale Data Compression Method for Supervised Fine-tuning of Large Language Models
Zhaoguang Long, Yuhao Zhou, Shangqing Zhao, Yupei Ren, Li Cai, Chenghao Jia, Zhe Chen, Zhe Fang, Yuxiang Song and Man Lan 2642

What Is Missing in Multilingual Visual Reasoning and How to Fix It
Yueqi Song, Simran Khanuja and Graham Neubig 2654

Enhancing the Prototype Network with Local-to-Global Optimization for Few-Shot Relation Extraction
Hui Sun and Rongxin Chen 2668

LLMs for Mathematical Modeling: Towards Bridging the Gap between Natural and Mathematical Languages
Xuhan Huang, Qingning Shen, Yan Hu, Anningzhe Gao and Benyou Wang 2678

Advancing Persian LLM Evaluation
Sara Bourbour Hosseinbeigi, Behnam Rohani, Mostafa Masoudi, Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh and Mohammad Amin Abbasi 2711

Supportiveness-based Knowledge Rewriting for Retrieval-augmented Language Modeling
Zile Qiao, Wei Ye, Yong Jiang, Tong Mo, Pengjun Xie, Weiping Li, Fei Huang and Shikun Zhang 2728

Evaluating Self-Generated Documents for Enhancing Retrieval-Augmented Generation with Large Language Models
Jiatao Li, Xinyu Hu, Xunjian Yin and Xiaojun Wan 2741

PREMISE: Matching-based Prediction for Accurate Review Recommendation
Wei Han, Hui Chen and Soujanya Poria 2776

Semi-supervised Fine-tuning for Large Language Models
Junyu Luo, Xiao Luo, Xiusi Chen, Zhiping Xiao, Wei Ju and Ming Zhang 2795

CALM: Unleashing the Cross-Lingual Self-Aligning Ability of Language Model Question Answering
Yumeng Wang, Zhiyuan Fan, Qingyun Wang, Yi R. Fung and Heng Ji 2809

Towards Prompt Generalization: Grammar-aware Cross-Prompt Automated Essay Scoring
Heejin Do, Taehee Park, Sangwon Ryu and Gary Lee 2818

MedEureka: A Medical Domain Benchmark for Multi-Granularity and Multi-Data-Type Embedding-Based Retrieval
Yongqi Fan, Nan Wang, Kui Xue, Jingping Liu and Tong Ruan 2825

A Federated Framework for LLM-based Recommendation
Jujia Zhao, Wenjie Wang, Chen Xu, See-Kiong Ng and Tat-Seng Chua 2852

WaterSeeker: Pioneering Efficient Detection of Watermarked Segments in Large Documents
Leyi Pan, Aiwei Liu, Yijian LU, Zitian Gao, Yichen Di, Lijie Wen, Irwin King and Philip S. Yu 2866

MIRAGE: A Metric-Intensive Benchmark for Retrieval-Augmented Generation Evaluation
Chanhee Park, Hyeonseok Moon, Chanjun Park and Heuiseok Lim 2883

FIRE: Fact-checking with Iterative Retrieval and Verification
Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych and Preslav Nakov 2901

Lessons from a User Experience Evaluation of NLP Interfaces
Eduardo Calò, Lydia Penkert and Saad Mahamood 2915

TrendSim: Simulating Trending Topics in Social Media Under Poisoning Attacks with LLM-based Multi-agent System
Zeyu Zhang, Jianxun Lian, Chen Ma, Yaning Qu, Ye Luo, Lei Wang, Rui Li, Xu Chen, Yankai Lin, Le Wu, Xing Xie and Ji-Rong Wen 2930

ASRank: Zero-Shot Re-Ranking with Answer Scent for Document Retrieval
Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani and Adam Jatowt 2950

DSQG-Syn: Synthesizing High-quality Data for Text-to-SQL Parsing by Domain Specific Question Generation
Shaoming Duan, Youxuan Wu, Chuanyi Liu, Yuhao Zhang, Zirui Wang, Peiyi Han, Shengyuan Yu, Liang Yan and Yingwei Liang 2971

EgoSpeak: Learning When to Speak for Egocentric Conversational Agents in the Wild
Junhyeok Kim, Min Soo Kim, Jiwan Chung, Jungbin Cho, Jisoo Kim, Sungwoong Kim, Gyeongbo Sim and Youngjae Yu 2990

Plot2Code: A Comprehensive Benchmark for Evaluating Multi-modal Large Language Models in Code Generation from Scientific Plots
Chengyue Wu, Zhixuan Liang, Yixiao Ge, Qiushan Guo, Zeyu Lu, Jiahao Wang, Ying Shan and Ping Luo 3006

FunnelRAG: A Coarse-to-Fine Progressive Retrieval Paradigm for RAG
Xinping Zhao, Yan Zhong, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Dongfang Li, Baotian Hu and Min Zhang 3029

The Power of Bullet Lists: A Simple Yet Effective Prompting Approach to Enhancing Spatial Reasoning in Large Language Models
Ikhyun Cho, Changyeon Park and Julia Hockenmaier 3047

Overcoming both Domain Shift and Label Shift for Referring Video Segmentation
Hai Huang, Sashuai Zhou and Yan Xia 3058

Language Modeling with Editable External Knowledge
Belinda Z. Li, Emmy Liu, Alexis Ross, Abbas Zeitoun, Graham Neubig and Jacob Andreas 3070

Beyond Excess and Deficiency: Adaptive Length Bias Mitigation in Reward Models for RLHF
Yuyan Bu, Liangyu Huo, Yi Jing and Qing Yang 3091

Neuroplasticity and Corruption in Model Mechanisms: A Case Study Of Indirect Object Identification
Vishnu Kabir Chhabra, Ding Zhu and Mohammad Mahdi Khalili 3099

VANE-Bench: Video Anomaly Evaluation Benchmark for Conversational LMMs
Hanan Gani, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan and Salman Khan 3123

Jailbreaking Prompt Attack: A Controllable Adversarial Attack against Diffusion Models
Jiachen Ma, Yijiang Li, Zhiqing Xiao, Anda Cao, Jie Zhang, Chao Ye and Junbo Zhao 3141

Emo3D: Metric and Benchmarking Dataset for 3D Facial Expression Generation from Emotion Description
Mahshid Dehghani, Amirahmad Shafiee, Ali Shafiei, Neda Fallah, Farahmand Alizadeh, Mohammad Mehdi Gholinejad, Hamid Behroozi, Jafar Habibi and Ehsaneddin Asgari 3158

Task-wrapped Continual Learning in Task-Oriented Dialogue Systems
Min Zeng, Haiqin Yang, Xi Chen and Yike Guo 3173

Untangling Hate Speech Definitions: A Semantic Componential Analysis Across Cultures and Domains
Katerina Korre, Arianna Muti, Federico Ruggeri and Alberto Barrón-Cedeño 3184

CodeRAG-Bench: Can Retrieval Augment Code Generation?
Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F. Xu, Yiqing Xie, Graham Neubig and Daniel Fried 3199

Multi-Condition Guided Diffusion Network for Multimodal Emotion Recognition in Conversation
Wenjin Tian, Xianying Huang and Shihao Zou 3215

Thank You, Stingray: Multilingual Large Language Models Can Not (Yet) Disambiguate Cross-Lingual Word Senses
Samuel Cahyawijaya, Ruochen Zhang, Jan Christian Blaise Cruz, Holy Lovenia, Elisa Gilbert, Hiroki Nomoto and Alham Fikri Aji 3228

Atoxia: Red-teaming Large Language Models with Target Toxic Answers
Yuhao Du, Zhuo Li, Pengyu Cheng, Xiang Wan and Anningzhe Gao 3251

A Practical Method for Generating String Counterfactuals
Matan Avitan, Ryan Cotterell, Yoav Goldberg and Shauli Ravfogel 3267

Probing-RAG: Self-Probing to Guide Language Models in Selective Document Retrieval
Ingeol Baek, Hwan Chang, ByeongJeong Kim, Jimin Lee and Hwanhee Lee 3287

Extracting Military Event Temporal Relations via Relative Event Time Prediction and Virtual Adversarial Training
Jie Gong and Qiwang hu 3305

Unlocking the Planning Capabilities of Large Language Models with Maximum Diversity Fine-tuning
Wenjun Li, Changyu Chen and Pradeep Varakantham 3318

Continuous Speech Tokenizer in Text To Speech
Yixing Li, Ruobing Xie, Xingwu Sun, Yu Cheng and Zhanhui Kang 3341

Efficient Annotator Reliability Assessment and Sample Weighting for Knowledge-Based Misinformation Detection on Social Media
Owen Cook, Charlie Grimshaw, Ben Peng Wu, Sophie Dillon, Jack Hicks, Luke Jones, Thomas Smith, Matyas Szert and Xingyi Song 3348

Challenges in Trustworthy Human Evaluation of Chatbots
Wenting Zhao, Alexander M Rush and Tanya Goyal 3359

RATSD: Retrieval Augmented Truthfulness Stance Detection from Social Media Posts Toward Factual Claims
Zhengyuan Zhu, Zeyu Zhang, Haiqi Zhang and Chengkai Li 3366

FACT: Examining the Effectiveness of Iterative Context Rewriting for Multi-fact Retrieval
Jinlin Wang, Suyuchen Wang, Ziwen Xia, Sirui Hong, Yun Zhu, Bang Liu and Chenglin Wu3382

Temporal Working Memory: Query-Guided Segment Refinement for Enhanced Multimodal Understanding
Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi and Jiang Gui 3393

Investigating the Transferability of Code Repair for Low-Resource Programming Languages
Kyle Wong, Alfonso Amayuelas, Liangming Pan and William Yang Wang 3410

Multilingual Blending: Large Language Model Safety Alignment Evaluation with Language Mixture
Jiayang Song, Yuheng Huang, Zhehua Zhou and Lei Ma 3433

Mitigating Hallucinations in Multimodal Spatial Relations through Constraint-Aware Prompting
Jiarui Wu, Zhuo Liu and Hangfeng He 3450

Concise and Organized Perception Facilitates Reasoning in Large Language Models
Junjie Liu, Shaotian Yan, Chen Shen, Zhengdong Xiao, Liang Xie, Wenxiao Wang and Jieping Ye 3469

Verifiable Format Control for Large Language Model Generations
Zhaoyang Wang, Jinqi Jiang, Huichi Zhou, Wenhao Zheng, Xuchao Zhang, Chetan Bansal and Huaxiu Yao 3499

Taxonomy and Analysis of Sensitive User Queries in Generative AI Search System
Hwiyeol Jo, Taiwoo Park, Hyunwoo Lee, Nayoung Choi, Changbong Kim, Ohjoon Kwon, Donghyeon Jeon, Eui Hyeon Lee, Kyoungho Shin, Lim Sun Suk, Kyungmi Kim, Lee Jihye and Sun Kim 3514

SynGhost: Invisible and Universal Task-agnostic Backdoor Attack via Syntactic Transfer
Pengzhou Cheng, Wei Du, Zongru Wu, Fengwei Zhang, Libo Chen, Zhuosheng Zhang and Gongshen Liu 3530

TESTEVAL: Benchmarking Large Language Models for Test Case Generation
Wenhan Wang, Chenyuan Yang, Zhijie Wang, Yuheng Huang, Zhaoyang Chu, Da Song, Lingming Zhang, An Ran Chen and Lei Ma 3547

Safe Inputs but Unsafe Output: Benchmarking Cross-modality Safety Alignment of Large Vision-Language Models
Siyin Wang, Xingsong Ye, Qinyuan Cheng, Junwen Duan, Shimin Li, Jinlan Fu, Xipeng Qiu and Xuanjing Huang 3563

FLEX: A Benchmark for Evaluating Robustness of Fairness in Large Language Models
Dahyun Jung, Seungyoon Lee, Hyeonseok Moon, Chanjun Park and Heuiseok Lim 3606

When and How to Augment Your Input: Question Routing Helps Balance the Accuracy and Efficiency of Large Language Models
Shufan Chen, He Zheng and Lei Cui 3621

GraPPI: A Retrieve-Divide-Solve GraphRAG Framework for Large-scale Protein-protein Interaction Exploration
Ziwen Li, Xiang Chen and Youngseung Jeon 3635

From Curiosity to Clarity : Exploring the Impact of Consecutive Why-Questions
Geonyeong Son, Jaeyoung Lee and Misuk Kim 3649

CollabStory: Multi-LLM Collaborative Story Generation and Authorship Analysis
Saranya Venkatraman, Nafis Irtiza Tripto and Dongwon Lee 3665

NTSEBENCH: Cognitive Reasoning Benchmark for Vision Language Models
Pranshu Pandya, Vatsal Gupta, Agney S Talwarr, Tushar Kataria, Dan Roth and Vivek Gupta 3680

KnowAgent: Knowledge-Augmented Planning for LLM-Based Agents
Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie GU, Huajun Chen and Ningyu Zhang 3709

SWITCH: Studying with Teacher for Knowledge Distillation of Large Language Models
Jahyun Koo, Yerin Hwang, Yongil Kim, Taegwan Kang, Hyunkyung Bae and Kyomin Jung 3733

Thought2Text: Text Generation from EEG Signal using Large Language Models (LLMs)
Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka and Shounak Roychowdhury . . . 3747

A Comprehensive Survey of Contemporary Arabic Sentiment Analysis: Methods, Challenges, and Future Directions
Zhiqiang Shi and Ruchit Agrawal 3760

Towards Cross-Lingual Explanation of Artwork in Large-scale Vision Language Models
Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi and Taro Watanabe 3773

Large Language Models are Easily Confused: A Quantitative Metric, Security Implications and Typological Analysis
Yiyi Chen, Qiongxiu Li, Russa Biswas and Johannes Bjerva 3810

Huatuo-26M, a Large-scale Chinese Medical QA Dataset
Xidong Wang, Jianquan Li, Shunian Chen, Yuxuan Zhu, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Junying Chen, Jie Fu, Xiang Wan, Anningzhe Gao and Benyou Wang 3828

SEP-MLDC: A Simple and Effective Paradigm for Multi-Label Document Classification
Han Liu, Shuqin Li, Xiaotong Zhang, Yuanyuan Wang, Feng Zhang, Hongyang Chen and Hong Yu 3849

Improving Pre-trained Language Models with Knowledge Enhancement and Filtering Framework
Qi Zhao, Qi Song, Tian Xie, Haiyue Zhang, Hongyu Yang and Xiangyang Li 3860

Using Review Combination and Pseudo-Tokens for Aspect Sentiment Quad Prediction
Jiazhou Chen, Xu Jia and RuiQiang Guo 3872

DDGIP: Radiology Report Generation Through Disease Description Graph and Informed Prompting
Chentao Huang, Guangli Li, Xinjiong Zhou, Yafeng Ren and Hongbin Zhang 3884

Lossless Acceleration of Large Language Models with Hierarchical Drafting based on Temporal Locality in Speculative Decoding
Sukmin Cho, Sangjin Choi, Taeho Hwang, Jeongyeon Seo, Soyeong Jeong, Huije Lee, Hoyun Song, Jong C. Park and Youngjin Kwon 3895

Improve Decoding Factuality by Token-wise Cross Layer Entropy of Large Language Models
Jialiang Wu, Yi Shen, Sijia Liu, Yi Tang, Sen Song, Xiaoyi Wang and Longjun Cai 3912

TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement
Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu and Zuozhu Liu 3922

Vulnerability of Large Language Models to Output Prefix Jailbreaks: Impact of Positions on Safety
Yiwei Wang, Muhao Chen, Nanyun Peng and Kai-Wei Chang 3939

ImaRA: An Imaginative Frame Augmented Method for Low-Resource Multimodal Metaphor Detection and Explanation
Yuan Tian, Minzheng Wang, Nan Xu and Wenji Mao 3953

XAMPLER: Learning to Retrieve Cross-Lingual In-Context Examples
Peiqin Lin, Andre Martins and Hinrich Schuetze 3968

Evaluating Cultural and Social Awareness of LLM Web Agents
Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng and Chien-Sheng Wu 3978

GRAIT: Gradient-Driven Refusal-Aware Instruction Tuning for Effective Hallucination Mitigation
Runchuan Zhu, Xinke Jiang, Jiang Wu, Zhipeng ma, Jiahe Song, Fengshuo Bai, Dahua Lin, Lijun Wu and Conghui He 4006

Entity Pair-guided Relation Summarization and Retrieval in LLMs for Document-level Relation Extraction
Fu Zhang, Hongsen Yu, Jingwei Cheng and Huangming Xu 4022

A Recipe of Parallel Corpora Exploitation for Multilingual Large Language Models
Peiqin Lin, Andre Martins and Hinrich Schuetze 4038

Omni-Chart-600K: A Comprehensive Dataset of Chart Types for Chart Understanding
Shulei Wang, Shuai Yang, Wang Lin, Zirun Guo, Sihang Cai, Hai Huang, Ye Wang, Jingyuan Chen and Tao Jin 4051

Comprehensive Layer-wise Analysis of SSL Models for Audio Deepfake Detection
Yassine El Kheir, Younes Samih, Suraj Maharjan, Tim Polzehl and Sebastian Möller 4070

Attention on Multiword Expressions: A Multilingual Study of BERT-based Models with Regard to Idiomaticity and Microsyntax
Iuliia Zaitova, Vitalii Hirak, Badr M. Abdullah, Dietrich Klakow, Bernd Möbius and Tania Avgustinova 4083

Perception Compressor: A Training-Free Prompt Compression Framework in Long Context Scenarios
Jiwei Tang, Jin Xu, Tingwei Lu, Zhicheng Zhang, YimingZhao YimingZhao, LinHai LinHai and Hai-Tao Zheng 4093

MojoBench: Language Modeling and Benchmarks for Mojo
Md Nishat Raihan, Joanna C. S. Santos and Marcos Zampieri 4109

VLind-Bench: Measuring Language Priors in Large Vision-Language Models
Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh and Kyomin Jung 4129

GRAG: Graph Retrieval-Augmented Generation
Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling and Liang Zhao 4145

Sequence-level Large Language Model Training with Contrastive Preference Optimization
Zhili Feng, Dhananjay Ram, Cole Hawkins, Aditya Rawal, Jinman Zhao and Sheng Zha . . . 4158

Scaling Up Membership Inference: When and How Attacks Succeed on Large Language Models
Haritz Puerto, Martin Gubri, Sangdoo Yun and Seong Joon Oh 4165

Mitigating Hallucinations in Large Vision-Language Models via Summary-Guided Decoding
Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee and Kyomin Jung 4183

Exploring Hybrid Sampling Inference for Aspect-based Sentiment Analysis
Xiaoyi Bao, Minjie Qiang, Jinghang Gu, Zhongqing Wang and Chu-Ren Huang 4199

FeRG-LLM : Feature Engineering by Reason Generation Large Language Models
Jeonghyun Ko, Gyeongyun Park, Donghoon Lee and Kyunam Lee 4211

Effective Self-Mining of In-Context Examples for Unsupervised Machine Translation with LLMs
Abdellah El Mekki and Muhammad Abdul-Mageed 4229

GPT-NER: Named Entity Recognition via Large Language Models
Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang and Chen Guo 4257

QPruner: Probabilistic Decision Quantization for Structured Pruning in Large Language Models
Changhai Zhou, Yuhua Zhou, Yibin Wang, Shijie Han, Qian Qiao and Hongguang Li 4276

MES-RAG: Bringing Multi-modal, Entity-Storage, and Secure Enhancements to RAG
Pingyu Wu, Daiheng Gao, Jing Tang, Huimin Chen, Wenbo Zhou, Weiming Zhang and Nenghai Yu 4287

LVPruning: An Effective yet Simple Language-Guided Vision Token Pruning Approach for Multi-modal Large Language Models
Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin and Riza Batista-Navarro . . 4299

How Much Knowledge Can You Pack into a LoRA Adapter without Harming LLM?
Sergey Pletenev, Maria Marina, Daniil Moskovskiy, Vasily Konovalov, Pavel Braslavski, Alexander Panchenko and Mikhail Salnikov 4309

TART: An Open-Source Tool-Augmented Framework for Explainable Table-based Reasoning
Xinyuan Lu, Liangming Pan, Yubo Ma, Preslav Nakov and Min-Yen Kan 4323

Enhancing Text-to-SQL with Question Classification and Multi-Agent Collaboration
Zhihui Shao, Shubin Cai, Rongsheng Lin and Zhong Ming 4340

Efficient Nearest Neighbor based Uncertainty Estimation for Natural Language Processing Tasks
Wataru Hashimoto, Hidetaka Kamigaito and Taro Watanabe 4350

BitAbuse: A Dataset of Visually Perturbed Texts for Defending Phishing Attacks
Hanyong Lee, Chaelyn Lee, Yongjae Lee and Jaesung Lee 4367

Unfolding the Headline: Iterative Self-Questioning for News Retrieval and Timeline Summarization
Weiqi Wu, Shen Huang, Yong Jiang, Pengjun Xie, Fei Huang and Hai Zhao 4385

RetrieverGuard: Empowering Information Retrieval to Combat LLM-Generated Misinformation
Chuwen Chen and Shuai Zhang 4399

Unified Automated Essay Scoring and Grammatical Error Correction
SeungWoo Song, Junghun Yuk, ChangSu Choi, HanGyeol Yoo, HyeonSeok Lim, KyungTae Lim and Jungyeul Park 4412

A Closer Look into Mixture-of-Experts in Large Language Models
Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang and Jie Fu 4427

CDB: A Unified Framework for Hope Speech Detection Through Counterfactual, Desire and Belief
Tulio Ferreira Leite Da Silva, Gonzalo Freijedo Aduna, Farah Benamara, Alda Mari, Zongmin Li, Li Yue and Jian Su 4448

How Well Do LLMs Handle Cantonese? Benchmarking Cantonese Capabilities of Large Language Models
Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li and Chuan Wu 4464

Improving Reward Models with Synthetic Critiques
Zihuiwen Ye, Fraser David Greenlee, Max Bartolo, Phil Blunsom, Jon Ander Campos and Matthias Gallé 4506

Rethinking Smoothness for Fast and Adaptable Entity Alignment Decoding
Yuanyi Wang, Han Li, Haifeng Sun, Lei Zhang, Bo He, Wei Tang, Tianhao Yan, Qi Qi and Jingyu Wang 4521

Lost in the Distance: Large Language Models Struggle to Capture Long-Distance Relational Knowledge
Meiyun Wang, Takeshi Kojima, Yusuke Iwasawa and Yutaka Matsuo 4536

FinNLI: Novel Dataset for Multi-Genre Financial Natural Language Inference Benchmarking
Jabez Magomere, Elena Kochkina, Samuel Mensah, Simerjot Kaur and Charese Smiley 4545

Music for All: Representational Bias and Cross-Cultural Adaptability of Music Generation Models
Atharva Mehta, Shivam Chauhan, Amirbek Djanibekov, Atharva Kulkarni, Gus Xia and Monojit Choudhury 4569

SFMSS: Service Flow aware Medical Scenario Simulation for Conversational Data Generation
Zhijie Bao, Qingyun Liu, Xuanjing Huang and Zhongyu Wei 4586

Re-evaluating Automatic LLM System Ranking for Alignment with Human Preference
Mingqi Gao, Yixin Liu, Xinyu Hu, Xiaojun Wan, Jonathan Bragg and Arman Cohan 4605

Beneath the Surface of Consistency: Exploring Cross-lingual Knowledge Representation Sharing in LLMs
Maxim Ifergan, Omri Abend, Idan Szpektor, Leshem Choshen and Roee Aharoni 4630

GuideQ: Framework for Guided Questioning for progressive informational collection and classification
Priya Mishra, Suraj Racha, Kaustubh Ponkshe, Adit Akarsh and Ganesh Ramakrishnan 4645

Richer Output for Richer Countries: Uncovering Geographical Disparities in Generated Stories and Travel Recommendations
Kirti Bhagat, Kinshuk Vasisht and Danish Pruthi 4660

Swan and ArabicMTEB: Dialect-Aware, Arabic-Centric, Cross-Lingual, and Cross-Cultural Embedding Models and Benchmarks
Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih and Muhammad Abdul-Mageed 4669

TAGCOS: Task-agnostic Gradient Clustered Coreset Selection for Instruction Tuning Data
Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan and Tong Zhang 4686

From Text to Emoji: How PEFT-Driven Personality Manipulation Unleashes the Emoji Potential in LLMs
Navya Jain, Zekun Wu, Cristian Enrique Munoz Villalobos, Airlie Hilliard, Xin Guan, Adriano Koshiyama, Emre Kazim and Philip Colin Treleaven 4702

Decoding Fatphobia: Examining Anti-Fat and Pro-Thin Bias in AI-Generated Images
Jane Warren, Gary M. Weiss, Fernando Martinez, Annika Guo and Yijun Zhao 4739

MMAU: A Holistic Benchmark of Agent Capabilities Across Diverse Domains
Guoli Yin, Haoping Bai, Shuang Ma, Feng Nan, Yanchao Sun, Zhaoyang Xu, Shen Ma, Jiarui Lu, Xiang Kong, Aonan Zhang, Dian Ang Yap, Yizhe Zhang, Karsten Ahnert, Vik Kamath, Mathias Berglund, Dominic Walsh, Tobias Gindele, Juergen Wiest, Zhengfeng Lai, Xiaoming Simon Wang, Jiulong Shan, Meng Cao, Ruoming Pang and Zirui Wang 4752

Improving Consistency in LLM Inference using Probabilistic Tokenization
Ashutosh Sathe, Divyanshu Aggarwal and Sunayana Sitaram 4781

WordGame: Efficient & Effective LLM Jailbreak via Simultaneous Obfuscation in Query and Response
Tianrong Zhang, Bochuan Cao, Yuanpu Cao, Lu Lin, Prasenjit Mitra and Jinghui Chen 4794

Human and LLM-Based Resume Matching: An Observational Study
Swanand Vaishampayan, Hunter Leary, Yoseph Berhanu Alebachew, Louis Hickman, Brent A. Stevenor, Weston Beck and Chris Brown 4823

A Practical Examination of AI-Generated Text Detectors for Large Language Models
Brian Tufts, Xuandong Zhao and Lei Li 4839

Robust Bias Detection in MLMs and its Application to Human Trait Ratings
Ingroj Shrestha, Louis Tay and Padmini Srinivasan 4857

How Inclusively do LMs Perceive Social and Moral Norms?
Michael Galarnyk, Agam Shah, Dipanwita Guhathakurta, Poojitha Nandigam and Sudheer Chava 4874

Jailbreaking with Universal Multi-Prompts
Yu-Ling Hsu, Hsuan Su and Shang-Tse Chen 4885

Echoes of Discord: Forecasting Hater Reactions to Counterspeech
Xiaoying Song, Sharon Lisseth Perez, Xinchen Yu, Eduardo Blanco and Lingzi Hong 4907

Contextual Metric Meta-Evaluation by Measuring Local Metric Accuracy
Athiya Deviyani and Fernando Diaz 4921

Advocating Character Error Rate for Multilingual ASR Evaluation
Thennal D K, Jesin James, Deepa Padmini Gopinath and Muhammed Ashraf K 4941

Enhancing Temporal Understanding in LLMs for Semi-structured Tables
Irwin Deng, Kushagra Dixit, Dan Roth and Vivek Gupta 4951

BnTTS: Few-Shot Speaker Adaptation in Low-Resource Setting
Mohammad Jahid Ibna Basher, Md Kowsher, Md Saiful Islam, Rabindra Nath Nandi, Nusrat Jahan Prottasha, Mehadi Hasan Menon, Tareq Al Muntasir, Shammur Absar Chowdhury, Firoj Alam, Niloofar Yousefi and Ozlem Garibay 4971

Playing with Voices: Tabletop Role-Playing Game Recordings as a Diarization Challenge
Lian Remme and Kevin Tang 4984

Causally Testing Gender Bias in LLMs: A Case Study on Occupational Bias
Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea and Zhijing Jin 4999

OLMES: A Standard for Language Model Evaluations
Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge and Hannaneh Hajishirzi 5020

Induction Heads as an Essential Mechanism for Pattern Matching in In-context Learning
Joy Crosbie and Ekaterina Shutova 5049

MoLA: MoE LoRA with Layer-wise Expert Allocation
Chongyang Gao, Kezhen Chen, Jinmeng Rao, Ruibo Liu, Baochen Sun, Yawen Zhang, Daiyi Peng, Xiaoyuan Guo and VS Subrahmanian 5112

CodeSim: Multi-Agent Code Generation and Problem Solving through Simulation-Driven Planning and Debugging
Md. Ashraful Islam, Mohammed Eunus Ali and Md Rizwan Parvez 5128

On the Feasibility of In-Context Probing for Data Attribution
Cathy Jiao, Weizhen Gao, Aditi Raghunathan and Chenyan Xiong 5155

Evaluation of Multilingual Image Captioning: How far can we get with CLIP models?
Goncalo Emanuel Cavaco Gomes, Chrysoula Zerva and Bruno Martins 5171

Avoiding Copyright Infringement via Large Language Model Unlearning
Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding and Eric Wong 5191

A Context-Aware Contrastive Learning Framework for Hateful Meme Detection and Segmentation
Xuanyu Su, Yansong Li, Diana Inkpen and Nathalie Japkowicz 5216

LLM-Generated Passphrases That Are Secure and Easy to Remember
Jie S. Li, Jonas Geiping, Micah Goldblum, Aniruddha Saha and Tom Goldstein 5231

Does Data Contamination Detection Work (Well) for LLMs? A Survey and Evaluation on Detection Assumptions
Yujuan Fu, Ozlem Uzuner, Meliha Yetisgen and Fei Xia 5250

Representation-to-Creativity (R2C): Automated Holistic Scoring Model for Essay Creativity
Deokgi Kim, Joonyoung Jo, Byung-Won On and Ingyu Lee 5272

From Single to Multi: How LLMs Hallucinate in Multi-Document Summarization
Catarina G Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani and Estevam Hruschka 5291

Aligning to Constraints for Data-Efficient Language Model Customization
Fei Wang, Chao Shang, Shuai Wang, Sarthak Jain, Qiang Ning, Bonan Min, Vittorio Castelli, Yassine Benajiba and Dan Roth 5325

Where is this coming from? Making groundedness count in the evaluation of Document VQA models
Armineh Nourbakhsh, Siddharth Parekh, Pranav Shetty, Zhao Jin, Sameena Shah and Carolyn Rose 5341

Transformer-based Causal Language Models Perform Clustering
Xinbo Wu and Lav R. Varshney 5362

Towards Better Multi-task Learning: A Framework for Optimizing Dataset Combinations in Large Language Models
Zaifu Zhan and Rui Zhang 5388

Gender Bias in Instruction-Guided Speech Synthesis Models
Chun-Yi Kuan and Hung-yi Lee 5402

ResoFilter: Fine-grained Synthetic Data Filtering for Large Language Models through Data-Parameter Resonance Analysis
Zeao Tu, Xiangdi Meng, Yu He, Zihan Yao, Tianyu Qi, Jun Liu and Ming Li 5429

UCFE: A User-Centric Financial Expertise Benchmark for Large Language Models
Yuzhe Yang, Yifei Zhang, Yan Hu, Yilin Guo, Ruoli Gan, Yueru He, Mingcong Lei, Xiao Zhang, Haining Wang, Qianqian Xie, Jimin Huang, Honghai Yu and Benyou Wang 5444

BRIEF: Bridging Retrieval and Inference for Multi-hop Reasoning via Compression
Yuankai Li, Jia-Chen Gu, Di Wu, Kai-Wei Chang and Nanyun Peng 5464

An Optimizable Suffix Is Worth A Thousand Templates: Efficient Black-box Jailbreaking without Affirmative Phrases via LLM as Optimizer
Weipeng Jiang, Zhenting Wang, Juan Zhai, Shiqing Ma, Zhengyu Zhao and Chao Shen 5486

Multi-Stage LLM Fine-Tuning with a Continual Learning Setting
Changhao Guan, Chao Huang, Hongliang Li, You Li, Ning Cheng, Zihe Liu, Yufeng Chen, Jinan Xu and Jian Liu 5499

Constraining Sequential Model Editing with Editing Anchor Compression
Hao-Xiang Xu, Jun-Yu Ma, Zhen-Hua Ling, Ningyu Zhang and Jia-Chen Gu 5514

MLKV: Multi-Layer Key-Value Heads for Memory Efficient Transformer Decoding
Zayd Muhammad Kawakibi Zuhri, Muhammad Farid Adilazuarda, Ayu Purwarianti and Alham Fikri Aji 5531

Clarify When Necessary: Resolving Ambiguity Through Interaction with LMs
Michael JQ Zhang and Eunsol Choi 5541

DOLFIN - Document-Level Financial Test-Set for Machine Translation
Mariam Nakhle, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier and Hervé Blanchon 5559

Are Large Language Models Effective in Clinical Trial Design? A Study on Baseline Feature Generation
Nafis Neehal, Bowen Wang, Shayom Debopadhaya, Corey Curran, Keerthiram Murugesan, Soham Dan, Vibha Anand and Kristin Bennett 5572

Lightweight Contenders: Navigating Semi-Supervised Text Mining through Peer Collaboration and Self Transcendence
Qianren Mao, Weifeng Jiang, Junnan Liu, Chenghua Lin, Qian Li, Xianqing Wen, Jianxin Li and Jinhu Lu 5586

Language-based Valence and Arousal Expressions between the United States and China: a Cross-Cultural Examination
Young Min Cho, Dandan Pang, Stuti Thapa, Garrick Sherman, Lyle Ungar, Louis Tay and Sharath Chandra Guntuku 5601

Chain-of-Rank: Enhancing Large Language Models for Domain-Specific RAG in Edge Device
Juntae Lee, Jihwan Bang, Kyuhong Shim, Seunghan Yang and Simyung Chang 5616

MALoRA: Mixture of Asymmetric Low-Rank Adaptation for Enhanced Multi-Task Learning
Xujia Wang, Haiyan Zhao, Shuo Wang, Hanqing Wang and Zhiyuan Liu 5624

LlamaLens: Specialized Multilingual LLM for Analyzing News and Social Media Content
Mohamed Bayan Kmainasi, Ali Ezzat Shahroor, Maram Hasanain, Sahinur Rahman Laskar, Naeemul Hassan and Firoj Alam 5642

LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education
Iain Weissburg, Sathvika Anand, Sharon Levy and Haewon Jeong 5665

Preserving Zero-shot Capability in Supervised Fine-tuning for Multi-label Text Classification
Si-An Chen, Hsuan-Tien Lin and Chih-Jen Lin 5714

Data-centric NLP Backdoor Defense from the Lens of Memorization
Zhenting Wang, Zhizhi Wang, Mingyu Jin, Mengnan Du, Juan Zhai and Shiqing Ma 5728

Neuro-Symbolic Integration Brings Causal and Reliable Reasoning Proofs
Sen Yang, Xin Li, Leyang Cui, Lidong Bing and Wai Lam 5747

Infogent: An Agent-Based Framework for Web Information Aggregation
Revanth Gangi Reddy, Sagnik Mukherjee, Jeonghwan Kim, Zhenhailong Wang, Dilek Hakkani-Tür and Heng Ji 5760

On the Role of Key Phrases in Argument Mining
Nilmadhab Das, Vijaya V Saradhi and Ashish Anand 5774

TabComp: A Dataset for Visual Table Reading Comprehension
Somraj Gautam, Abhishek Bhandari and Gaurav Harit 5788

RankAdaptor: Hierarchical Rank Allocation for Efficient Fine-Tuning Pruned LLMs via Performance Model
Changhai Zhou, Shijie Han, Lining Yang, Yuhua Zhou, Xu Cheng, Yibin Wang and Hongguang Li 5796

Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs
SeongYeub Chu, Jong Woo Kim, Bryan Wong and Mun Yong Yi 5811

MTPChat: A Multimodal Time-Aware Persona Dataset for Conversational Agents
Wanqi Yang, Yanda Li, Meng Fang and Ling Chen 5830

MetaAlign: Align Large Language Models with Diverse Preferences during Inference Time
Mozhi Zhang, Pengyu Wang, Chenkun Tan, Mianqiu Huang, Dong Zhang, Yaqian Zhou and Xipeng Qiu 5842

MAQA: Evaluating Uncertainty Quantification in LLMs Regarding Data Uncertainty
Yongjin Yang, Haneul Yoo and Hwaran Lee 5861

Tuning-Free Personalized Alignment via Trial-Error-Explain In-Context Learning
Hyundong Justin Cho, Karishma Sharma, Nicolaas Paul Jedema, Leonardo F. R. Ribeiro, Jonathan May and Alessandro Moschitti 5879

Causal Inference with Large Language Model: A Survey
Jing Ma 5901

Ask Optimal Questions: Aligning Large Language Models with Retriever's Preference in Conversation
Chanwoong Yoon, Gangwoo Kim, Byeongguk Jeon, Sungdong Kim, Yohan Jo and Jaewoo Kang 5914

Systematic Knowledge Injection into Large Language Models via Diverse Augmentation for Domain-Specific RAG
Kushagra Bhushan, Yatin Nandwani, Dinesh Khandelwal, Sonam Gupta, Gaurav Pandey, Dinesh Raghu and Sachindra Joshi 5937

Find the Intention of Instruction: Comprehensive Evaluation of Instruction Understanding for Large Language Models
Hyeonseok Moon, Jaehyung Seo, Seungyoon Lee, Chanjun Park and Heuiseok Lim 5959

Long-Tail Crisis in Nearest Neighbor Language Models
Yuto Nishida, Makoto Morishita, Hiroyuki Deguchi, Hidetaka Kamigaito and Taro Watanabe 5980

Keep Guessing? When Considering Inference Scaling, Mind the Baselines
Gal Yona, Or Honovich, Omer Levy and Roee Aharoni 5994

Large Language Models for Anomaly and Out-of-Distribution Detection: A Survey
Ruiyao Xu and Kaize Ding 6007

Time-aware ReAct Agent for Temporal Knowledge Graph Question Answering
QianyiHu QianyiHu, Xinhui Tu, Guo Cong and Shunping Zhang 6028

SG-FSM: A Self-Guiding Zero-Shot Prompting Paradigm for Multi-Hop Question Answering Based on Finite State Machine
Xiaochen Wang, Junqing He, Liang Chen, Gholamreza Haffari, Yiru Wang, Zhe Yang, Xiangdi Meng, Kunhao Pan and Zhifang Sui 6040

Dynamic Strategy Planning for Efficient Question Answering with Large Language Models
Tanmay Parekh, Pradyot Prakash, Alexander Radovic, Akshay Shekher and Denis Savenkov 6053

Can I Introduce My Boyfriend to My Grandmother? Evaluating Large Language Models Capabilities on Iranian Social Norm Classification
Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, Francesco Pierri and Debora Nozza 6075

PLD+: Accelerating LLM Inference by Leveraging Language Model Artifacts
Shwetha Somasundaram, Anirudh Phukan and Apoorv Saxena 6090

Adapting LLM Agents with Universal Communication Feedback
Kuan Wang, Yadong Lu, Michael Santacroce, Yeyun Gong, Chao Zhang and Yelong Shen . 6105

Ignore the KL Penalty! Boosting Exploration on Critical Tokens to Enhance RL Fine-Tuning
Jean Vassoyan, Nathanaël Beau and Roman Plaud 6123

SeaExam and SeaBench: Benchmarking LLMs with Local Multilingual Questions in Southeast Asia
Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu and Lidong Bing 6134

Learning to Search Effective Example Sequences for In-Context Learning
Xiang Gao, Ankita Sinha and Kamalika Das 6152

From Intentions to Techniques: A Comprehensive Taxonomy and Challenges in Text Watermarking for Large Language Models
Harsh Nishant Lalai, Aashish Anantha Ramakrishnan, Raj Sanjay Shah and Dongwon Lee . 6162

M-IFEval: Multilingual Instruction-Following Evaluation
Antoine Dussolle, A. Cardeña, Shota Sato and Peter Devine 6176

Automatic Annotation Augmentation Boosts Translation between Molecules and Natural Language
Zhiqiang Zhong, Simon Sataa-Yu Larsen, Haoyu Guo, Tao Tang, Kuangyu Zhou and Davide Mottin 6192

Let Modalities Teach Each Other: Modal-Collaborative Knowledge Extraction and Fusion for Multi-modal Knowledge Graph Completion
Guoliang Zhu, Tao Ren, Dandan Wang and Jun HU 6210

Modeling the Differential Prevalence of Online Supportive Interactions in Private Instant Messages of Adolescents
Ondrej Sotolar, Michał Tkaczyk, Jaromír Plhák and David Smahel 6223

Dynamic Feature Fusion for Sign Language Translation Using HyperNetworks
Ruiquan Zhang, Rui Zhao, Zhicong Wu, Liang Zhang, Haoqi Zhang and Yidong Chen 6242

Selective Self-to-Supervised Fine-Tuning for Generalization in Large Language Models
Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Dinesh Khandelwal, Dinesh Raghu and Sachindra Joshi 6255

ProverbEval: Exploring LLM Evaluation Challenges for Low-resource Language Understanding
Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio and Dietrich Klakow 6265

MRE-MI: A Multi-image Dataset for Multimodal Relation Extraction in Social Media Posts
Shizhou Huang, Bo Xu, Changqun Li, Yang Yu and Xin Alex Lin 6282

Discrete Diffusion Language Model for Efficient Text Summarization
Do Huu Dat, Duc Anh Do, Anh Tuan Luu and Wray Buntine 6293

CAPE: A Chinese Dataset for Appraisal-based Emotional Generation in Large Language Models
June M. Liu, He Cao, Renliang Sun, Rui Wang, Yu Li and Jiaxing Zhang 6306

Beyond Under-Alignment: Atomic Preference Enhanced Factuality Tuning for Large Language Models
Hongbang Yuan, Yubo Chen, Pengfei Cao, Zhuoran Jin and Kang Liu 6325

Weight-based Analysis of Detokenization in Language Models: Understanding the First Stage of Inference Without Inference
Go Kamoda, Benjamin Heinzerling, Tatsuro Inaba, Keito Kudo, Keisuke Sakaguchi and Kentaro Inui 6339

DiPT: Enhancing LLM Reasoning through Diversified Perspective-Taking
Hoang Anh Just, Mahavir Dabas, Lifu Huang, Ming Jin and Ruoxi Jia 6359

SOLID: Self-seeding and Multi-intent Self-instructing LLMs for Generating Intent-aware Information-Seeking Dialogs
Arian Askari, Roxana Petcu, Chuan Meng, Mohammad Aliannejadi, Amin Abolghasemi, Evangelos Kanoulas and Suzan Verberne 6390

CollagePrompt: A Benchmark for Budget-Friendly Visual Recognition with GPT-4V
Siyu Xu, Yunke Wang, Daochang Liu, Bo Du and Chang Xu 6411

ARISE: Iterative Rule Induction and Synthetic Data Generation for Text Classification
Yaswanth M, Vaibhav Singh, Ayush Maheshwari, Amrith Krishna and Ganesh Ramakrishnan 6434

Unleashing Multi-Hop Reasoning Potential in Large Language Models through Repetition of Misordered Context
Sangwon Yu, Ik-hwan Kim, Jongyoon Song, Saehyung Lee, Junsung Park and Sungroh Yoon6450

Text Annotation via Inductive Coding: Comparing Human Experts to LLMs in Qualitative Data Analysis
Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer and Alexander Denzler 6471

Investigating the Zone of Proximal Development of Language Models for In-Context Learning
Peng Cui and Mrinmaya Sachan 6485

Breaking ReAct Agents: Foot-in-the-Door Attack Will Get You In
Itay Nakash, George Kour, Guy Uziel and Ateret Anaby Tavor 6499

As easy as PIE: understanding when pruning causes language models to disagree
Pietro Tropeano, Maria Maistro, Tuukka Ruotsalo and Christina Lioma 6525

Multi-Agent Simulator Drives Language Models for Legal Intensive Interaction
ShengbinYue ShengbinYue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuanjing Huang and Zhongyu Wei 6552

Exploring Backward Reasoning in Large Language Models
Leonardo Ranaldi and Giulia Pucci 6586

MMLF: Multi-query Multi-passage Late Fusion Retrieval
Yuan-Ching Kuo, Yi Yu, Chih-Ming Chen and Chuan-Ju Wang 6602

Dynamic Guided and Domain Applicable Safeguards for Enhanced Security in Large Language Models
Weidi Luo, He Cao, Zijing Liu, Yu Wang, Aidan Wong, Bin Feng, Yuan Yao and Yu Li 6614

kNN For Whisper And Its Effect On Bias And Speaker Adaptation
Maya K. Nachesa and Vlad Niculae 6636

VisualCoder: Guiding Large Language Models in Code Execution with Fine-grained Multimodal Chain-of-Thought Reasoning
Cuong Le Chi, Chau Truong Vinh Hoang, Phan Nhật Huy, Dung D. Le, Tien N Nguyen and Nghi D. Q. Bui 6643

Optimizing LLMs for Italian: Reducing Token Fertility and Enhancing Efficiency Through Vocabulary Adaptation
Luca Moroni, Giovanni Puccetti, Pere-Lluís Huguet Cabot, Andrei Stefan Bejgu, Alessio Miaschi, Edoardo Barba, Felice Dell'Orletta, Andrea Esuli and Roberto Navigli 6661

Beyond the Mode: Sequence-Level Distillation of Multilingual Translation Models for Low-Resource Language Pairs
Aarón Galiano-Jiménez, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez and Víctor M. Sánchez-Cartagena 6676

LLMs for Extremely Low-Resource Finno-Ugric Languages
Taido Purason, Hele-Andra Kuulmets and Mark Fishel 6692

LOFT: Scalable and More Realistic Long-Context Evaluation
Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Séb Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang and Kelvin Guu 6713

On the Influence of Context Size and Model Choice in Retrieval-Augmented Generation Systems
Juraj Vladika and Florian Matthes 6739

Aligning Black-box Language Models with Human Judgments
Gerrit J.j. Van Den Burg, Gen Suzuki, Wei Liu and Murat Sensoy 6752

Guideline Compliance in Task-Oriented Dialogue: The Chained Prior Approach
Xiangyu Wen, Jianyuan Zhong, Zhijian Xu and Qiang Xu 6765

AutoBreach: Universal and Adaptive Jailbreaking with Efficient Wordplay-Guided Optimization via Multi-LLMs
Jiawei Chen, Xiao Yang, Zhengwei Fang, Yu Tian, Yinpeng Dong, Zhaoxia Yin and Hang Su6792

S^2IT: Stepwise Syntax Integration Tuning for Large Language Models in Aspect Sentiment Quad Prediction
Bingfeng Chen, Chenjie Qiu, Yifeng Xie, Boyan Xu, Ruichu Cai and Zhifeng Hao 6814

BanNERD: A Benchmark Dataset and Context-Driven Approach for Bangla Named Entity Recognition
Md. Motahar Mahtab, Faisal Ahamed Khan, Md. Ekramul Islam, Md. Shahad Mahmud Chowdhury, Labib Imam Chowdhury, Sadia Afrin, Hazrat Ali, Mohammad Mamun Or Rashid, Nabeel Mohammed and Mohammad Ruhul Amin 6822

Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias
Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers and Vincent Ginis 6844

What can Large Language Models Capture about Code Functional Equivalence?
Nickil Maveli, Antonio Vergari and Shay B Cohen 6880

Make Every Penny Count: Difficulty-Adaptive Self-Consistency for Cost-Efficient Reasoning
Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu and Kan Li 6919

Large Language Models Are Better Logical Fallacy Reasoners with Counterargument, Explanation, and Goal-Aware Prompt Formulation
Jiwon Jeong, Hyeju Jang and Hogun Park 6933

MorphNLI: A Stepwise Approach to Natural Language Inference Using Text Morphing
Vlad Andrei Negru, Robert Vacareanu, Camelia Lemnaru, Mihai Surdeanu and Rodica Potolea 6953

Unmasking Database Vulnerabilities: Zero-Knowledge Schema Inference Attacks in Text-to-SQL Systems
Đorđe Klisura and Anthony Rios 6969

Media of Langue: Exploring Word Translation Network
Goki Muramoto, Atsuki Sato and Takayoshi Koyama 6992

Tackling Social Bias against the Poor: a Dataset and a Taxonomy on Aporophobia
Georgina Curto, Svetlana Kiritchenko, Muhammad Hammad Fahim Siddiqui, Isar Nejadgholi and Kathleen C. Fraser 7010

The American Sign Language Knowledge Graph: Infusing ASL Models with Linguistic Knowledge
Lee Kezar, Nidhi Munikote, Zian Zeng, Zed Sehyr, Naomi Caselli and Jesse Thomason 7032

Reinforcement Learning for Aligning Large Language Models Agents with Interactive Environments: Quantifying and Mitigating Prompt Overfitting
Mohamed Salim Aissi, Clément Romac, Thomas Carta, Sylvain Lamprier, Pierre-Yves Oudeyer, Olivier Sigaud, Laure Soulier and Nicolas Thome 7045

An empirical study of validating synthetic data for formula generation
Usneek Singh, José Cambronero, Sumit Gulwani, Aditya Kanade, Anirudh Khatry, Vu Le, Mukul Singh and Gust Verbruggen 7062

TeCoFeS: Text Column Featurization using Semantic Analysis
Ananya Singha, Mukul Singh, Ashish Tiwari, Sumit Gulwani, Vu Le and Chris Parnin 7070

CA: Addressing Evaluation Pitfalls in Computation-Aware Latency for Simultaneous Speech Translation*
Xi Xu, Wenda Xu, Siqi Ouyang and Lei Li 7077

Augmented Adversarial Trigger Learning
Zhe Wang and Yanjun Qi 7083

Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents
Qiusi Zhan, Richard Fang, Henil Shalin Panchal and Daniel Kang 7116

Flaming-hot Initiation with Regular Execution Sampling for Large Language Models
Weizhe Chen, Zhicheng Zhang, Guanlin Liu, Renjie Zheng, Wenlei Shi, Chen Dun, Zheng Wu, Xing Jin and Lin Yan 7133

HEISIR: Hierarchical Expansion of Inverted Semantic Indexing for Training-free Retrieval of Conversational Data using LLMs
Sangyeop Kim, Hangyeul Lee and Yohan Lee 7143

Women do not have heart attacks!Gender Biases in Automatically Generated Clinical Cases in French
Fanny Ducel, Nicolas Hiebel, Olivier Ferret, Karën Fort and Aurélie Névéol 7160

NOTA: Multimodal Music Notation Understanding for Visual Large Language Model
Mingni Tang, Jiajia Li, lu Yang, Zhiqiang Zhang, Jinhao Tian, Zuchao Li, Lefei Zhang and Ping Wang 7175

Exploring Large Language Models for Hate Speech Detection in Rioplatense Spanish
Juan Manuel Pérez, Paula Miguel and Viviana Cotik 7189

An Annotated Dataset of Errors in Premodern Greek and Baselines for Detecting Them
Creston Brooks, Johannes Haubold, Charlie Cowen-Breen, Jay White, Desmond DeVaul, Frederick Riemenschneider, Karthik R Narasimhan and Barbara Graziosi 7203

WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation
João Matos, Shan Chen, Siena Kathleen V. Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis Filipe Nakayama, José María Millet Pascual-Leone, Guergana K Savova, Hugo Aerts, Leo Anthony Celi, An-Kwok Ian Wong, Danielle Bitterman and Jack Gallifant 7218

BanTH: A Multi-label Hate Speech Detection Dataset for Transliterated Bangla
Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim and Md Farhad Alam Bhuiyan 7232

Mutual Reinforcement of LLM Dialogue Synthesis and Summarization Capabilities for Few-Shot Dialogue Summarization
Yen-Ju Lu, Ting-Yao Hu, Hema Swetha Koppula, Hadi Pouransari, Jen-Hao Rick Chang, Yin Xia, Xiang Kong, Qi Zhu, Xiaoming Simon Wang, Oncel Tuzel and Raviteja Vemulapalli 7252

UNLEARN Efficient Removal of Knowledge in Large Language Models
Tyler Lizzo and Larry Heck 7272

Adaptive Parameter Compression for Language Models
Jeremias Bohn, Frederic Mrozinski and Georg Groh 7284

Personalize Your LLM: Fake it then Align it
Yijing Zhang, Dyah Adila, Changho Shin and Frederic Sala 7302

A Survey to Recent Progress Towards Understanding In-Context Learning
Haitao Mao, Guangliang Liu, Yao Ma, Rongrong Wang, Kristen Johnson and Jiliang Tang . 7317

Inference Scaling for Bridging Retrieval and Augmented Generation
Youngwon Lee, Seung-won Hwang, Daniel F Campos, Filip Graliński, Zhewei Yao and Yuxiong He 7339

GeoCoder: Solving Geometry Problems by Generating Modular Code through Vision-Language Models
Aditya Sharma, Aman Dalmia, Mehran Kazemi, Amal Zouaq and Christopher Pal 7355

SEEval: Advancing LLM Text Evaluation Efficiency and Accuracy through Self-Explanation Prompting
Meng-Chen Wu, Md Mosharaf Hossain, Tess Wood, Shayan Ali Akbar, Si-Chi Chin and Erwin Cornejo 7372

When natural language is not enough: The limits of in-context learning demonstrations in multilingual reasoning
Leonardo Ranaldi, Barry Haddow and Alexandra Birch 7384

Uncovering Latent Arguments in Social Media Messaging by Employing LLMs-in-the-Loop Strategy
Tunazzina Islam and Dan Goldwasser 7412

AcrosticSleuth: Probabilistic Identification and Ranking of Acrostics in Multilingual Corpora
Aleksandr Fedchin, Isabel Cooperman, Pramit Chaudhuri and Joseph P. Dexter 7445

MedThink: A Rationale-Guided Framework for Explaining Medical Visual Question Answering
Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu and Zuozhu Liu 7453

How to Learn in a Noisy World? Self-Correcting the Real-World Data Noise in Machine Translation
Yan Meng, Di Wu and Christof Monz 7466

Rejected Dialects: Biases Against African American Language in Reward Models
Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva and Maarten Sap 7483

Do Large Language Models Align with Core Mental Health Counseling Competencies?
Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar and Munmun De Choudhury 7503

Uncertainty Quantification for Clinical Outcome Predictions with (Large) Language Models
Zizhang Chen, Peizhao Li, Xiaomeng Dong and Pengyu Hong 7527

Hypothesis Generation for Materials Discovery and Design Using Goal-Driven and Constraint-Guided LLM Agents
Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal and Chitta Baral 7539

Aligning to What? Limits to RLHF Based Alignment
Logan Barnhart, Reza Akbarian Bafghi, Stephen Becker and Maziar Raissi 7571

Beyond Words: Exploring Cultural Value Sensitivity in Multimodal Models
Srishti Yadav, Zhi Zhang, Daniel Hershcovich and Ekaterina Shutova 7607

Features that Make a Difference: Leveraging Gradients for Improved Dictionary Learning
Jeffrey Olmo, Jared Wilson, Max Forsey, Bryce Hepner, Thomas Vincent Howe and David Wingate 7624

Tooling or Not Tooling? The Impact of Tools on Language Agents for Chemistry Problem Solving
Botao Yu, Frazier N. Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning and Huan Sun 7635

RusCode: Russian Cultural Code Benchmark for Text-to-Image Generation
Viacheslav Vasilev, Julia Agafonova, Nikolai Gerasimenko, Alexander Kapitanov, Polina Mikhailova, Evelina Mironova and Denis Dimitrov 7656

Evaluation of LLMs-based Hidden States as Author Representations for Psychological Human-Centered NLP Tasks
Nikita Soni, Pranav Chitale, Khushboo Singh, Niranjan Balasubramanian and H. Schwartz . 7673

Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey
Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai and Furong Huang 7683

ThoughtSculpt: Reasoning with Intermediate Revision and Search
Yizhou Chi, Kevin Yang and Dan Klein 7700

Optimizing Hidden Markov Language Models: An Empirical Study of Reparameterization and Initialization Techniques
Ivan Lee and Taylor Berg-Kirkpatrick 7727

Using Linguistic Entrainment to Evaluate Large Language Models for Use in Cognitive Behavioral Therapy
Mina Kian, Kaleen Shrestha, Katrin Fischer, Xiaoyuan Zhu, Jonathan Ong, Aryan Trehan, Jessica Wang, Gloria Chang, Séb Arnold and Maja Mataric 7739

Analysis of LLM as a grammatical feature tagger for African American English
Rahul Porwal, Alice Rozet, Jotsna Gowda, Pryce Houck, Kevin Tang and Sarah Moeller . . . 7759

LLM-Microscope: Uncovering the Hidden Role of Punctuation in Context Memory of Transformers
Anton Razzhigaev, Matvey Mikhalchuk, Temurbek Rahmatullaev, Elizaveta Goncharova, Polina Druzhinina, Ivan Oseledets and Andrey Kuznetsov 7772

On A Scale From 1 to 5: Quantifying Hallucination in Faithfulness Evaluation
Xiaonan Jing, Srinivas Billa and Danny Godbout 7780

LITERA: An LLM Based Approach to Latin-to-English Translation
Paul Rosu 7796

Investigating the Shortcomings of LLMs in Step-by-Step Legal Reasoning
Venkatesh Mishra, Bimsara Pathiraja, Mihir Parmar, Sat Chidananda, Jayanth Srinivasa, Gaowen Liu, Ali Payani and Chitta Baral 7810

Towards Long Context Hallucination Detection
Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba and Dan Roth 7842

How to Talk to Language Models: Serialization Strategies for Structured Entity Matching
Haoteng Yin, Jinha Kim, Prashant Mathur, Krishanu Sarker and Vidit Bansal 7851

Accounting for Sycophancy in Language Model Uncertainty Estimation
Anthony Sicilia, Mert Inan and Malihe Alikhani 7866

Zero-Shot Keyphrase Generation: Investigating Specialized Instructions and Multi-sample Aggregation on Large Language Models
Jishnu Ray Chowdhury, Jayanth Mohan, Tomas Malik and Cornelia Caragea 7882

Meta-Reasoning Improves Tool Use in Large Language Models
Lisa Alazraki and Marek Rei 7900

CLERC: A Dataset for U. S. Legal Case Retrieval and Retrieval-Augmented Analysis Generation
Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek and Benjamin Van Durme 7913

GAIfE: Using GenAI to Improve Literacy in Low-resourced Settings
Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K. Keita and Michael Leventhal 7929

Hard Emotion Test Evaluation Sets for Language Models
Tiberiu Sosea and Cornelia Caragea 7945

UCL-Bench: A Chinese User-Centric Legal Benchmark for Large Language Models
Ruoli Gan, Duanyu Feng, Chen Zhang, Zhihang Lin, Haochen Jia, Hao Wang, Zhenyang Cai, Lei Cui, Qianqian Xie, Jimin Huang and Benyou Wang 7960

MIDAS: Multi-level Intent, Domain, And Slot Knowledge Distillation for Multi-turn NLU
Yan Li, So-Eon Kim, Seong-Bae Park and Caren Han 8004

*A Practical Analysis of Human Alignment with *PO*
Kian Ahrabian, Xihui Lin, Barun Patra, Vishrav Chaudhary, Alon Benhaim, Jay Pujara and Xia Song 8028

Understanding Reference Policies in Direct Preference Optimization
Yixin Liu, Pengfei Liu and Arman Cohan 8037

LLM-Coordination: Evaluating and Analyzing Multi-agent Coordination Abilities in Large Language Models
Saaket Agashe, Yue Fan, Anthony Reyna and Xin Eric Wang 8053

AssertionBench: A Benchmark to Evaluate Large-Language Models for Assertion Generation
Vaishnavi Pulavarthi, Deeksha Nandal, Soham Dan and Debjit Pal 8073

On Reference (In-)Determinacy in Natural Language Inference
Sihao Chen, Chaitanya Malaviya, Alex Fabrikant, Hagai Taitelbaum, Tal Schuster, Senaka Buthpitiya and Dan Roth 8081

DHP Benchmark: Are LLMs Good NLG Evaluators?
Yicheng Wang, Jiayi Yuan, Yu-Neng Chuang, Zhuoer Wang, Yingchi Liu, Mark Cusick, Param Kulkarni, Zhengping Ji, Yasser Ibrahim and Xia Hu 8094

GraphEval36K: Benchmarking Coding and Reasoning Capabilities of Large Language Models on Graph Datasets
Qiming Wu, Zichen Chen, Will Corcoran, Misha Sra and Ambuj Singh 8110

SimulBench: Evaluating Language Models with Creative Simulation Tasks
Qi Jia, Xiang Yue, Tuney Zheng, Jie Huang and Bill Yuchen Lin 8133

ReasoningRec: Bridging Personalized Recommendations and Human-Interpretable Explanations through LLM Reasoning
Millennium Bismay, Xiangjue Dong and James Caverlee 8147

2D-DPO: Scaling Direct Preference Optimization with 2-Dimensional Supervision
Shilong Li, Yancheng He, Hui Huang, Xingyuan Bu, Jiaheng Liu, Hangyu Guo, Weixun Wang, Jihao Gu, Wenbo Su and Bo Zheng 8164

Demystifying the Power of Large Language Models in Graph Generation
Yu Wang, Ryan A. Rossi, Namyong Park, Nesreen K. Ahmed, Danai Koutra, Franck Dernoncourt and Tyler Derr 8189

COIG-CQIA: Quality is All You Need for Chinese Instruction Fine-tuning
Yuelin Bai, Xeron Du, Yiming Liang, Leo Jin, Junting Zhou, Ziqiang Liu, Feiteng Fang, Mingshan Chang, Tianyu Zheng, Xincheng Zhang, Nuo ma, Zekun Moore Wang, Ruibin Yuan, Haihong Wu, Hongquan Lin, Wenhao Huang, Jiajun Zhang, Chenghua Lin, Jie Fu, Min Yang, Shiwen Ni and Ge Zhang 8205

Gradient-guided Attention Map Editing: Towards Efficient Contextual Hallucination Mitigation
Yu Wang, Jiaxin Zhang, Xiang Gao, Wendi Cui, Peng Li and Kamalika Das 8221

Alleviating Hallucinations of Large Language Models through Induced Hallucinations
Yue Zhang, Leyang Cui, V. W. and Shuming Shi 8233

MoDE: Effective Multi-task Parameter Efficient Fine-Tuning with a Mixture of Dyadic Experts
Lin Ning, Harsh Lara, Meiqi Guo and Abhinav Rastogi 8248

Unsupervised Sentence Representation Learning with Syntactically Aligned Negative Samples
Zhilan Wang, Zekai Zhi, Rize Jin, Kehui Song, He Wang and Da-Jung Cho 8262

Hierarchical Speculative Decoding with Dynamic Window
Shensian Syu and Hung-yi Lee 8275

Q-FAKER: Query-free Hard Black-box Attack via Controlled Generation
CheolWon Na, YunSeok Choi and Jee-Hyong Lee 8289

PRDetect: Perturbation-Robust LLM-generated Text Detection Based on Syntax Tree
Xiang Li, Zhiyi Yin, Hexiang Tan, Shaoling Jing, Du Su, Yi Cheng, Huawei Shen and Fei Sun 8305

Enabling Natural Zero-Shot Prompting on Encoder Models via Statement-Tuning
Ahmed Elshabrawy, Yongxin Huang, Iryna Gurevych and Alham Fikri Aji 8317

Faster Machine Translation Ensembling with Reinforcement Learning and Competitive Correction
Kritarth Prasad, Mohammadi Zaki, Pratik Rakesh Singh and Pankaj Wasnik 8337

Evaluating Numeracy of Language Models as a Natural Language Inference Task
Rahmad Mahendra, Damiano Spina, Lawrence Cavedon and Karin Verspoor 8351

Are Language Models Agnostic to Linguistically Grounded Perturbations? A Case Study of Indic Languages
Poulami Ghosh, Raj Dabre and Pushpak Bhattacharyya 8377

Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics
Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo and Youngjae Yu 8412

Tell Me What You Know About Sexism: Expert-LLM Interaction Strategies and Co-Created Definitions for Zero-Shot Sexism Detection
Myrthe Reuver, Indira Sen, Matteo Melis and Gabriella Lapesa 8453

The Role of Prosody in Spoken Question Answering
Jie Chi, Maureen de Seyssel and Natalie Schluter 8483

Target-Augmented Shared Fusion-based Multimodal Sarcasm Explanation Generation
Palaash Goel, Dushyant Singh Chauhan and Md Shad Akhtar 8495

Seeds of Discourse: A Multilingual Corpus of Direct Quotations from African Media on Agricultural Biotechnologies
Patricia Chiril, Trevor Spreadbury, Joeva Sean Rock, Brian Dowd-Uribe and David Uminsky 8509

Position Really Matters: Towards a Holistic Approach for Prompt Tuning
Xianjun Yang, Wei Cheng, Xujiang Zhao, Wenchao Yu, Linda Ruth Petzold and Haifeng Chen 8516