

The 2nd Workshop on Misinformation Detection in the Era of LLMs (MisD 2026)

Zhiwei Liu¹, Yupeng Cao², Calvin Yixiang Cheng³, Zhuohan Xie⁴, Ye Yuan^{5,6}, Yankai Chen^{4,5}, Yuechen Jiang¹, Yuyan Wang¹, Peter Carragher⁷, Jimin Huang^{1,8}, Xue Liu^{4,5}, Sophia Ananiadou^{1,9}

¹University of Manchester, ²Stevens Institute of Technology, ³Oxford University, ⁴MBZUAI, ⁵McGill University, ⁶Mila - Quebec AI Institute, ⁷Carnegie Mellon University, ⁸The Fin AI, ⁹ELLIS Manchester
zhiwei.liu@manchester.ac.uk, ycao33@stevens.edu, calvin.cheng@oii.ox.ac.uk, zhuohan.xie@mbzuai.ac.ae,
ye.yuan3@mail.mcgill.ca, yankaichen@acm.org, yuechen.jiang@postgrad.manchester.ac.uk,
yuyan.wang-11@postgrad.manchester.ac.uk, petercarragher@cmu.edu
jimmin.huang@postgrad.manchester.ac.uk, xueliu@cs.mcgill.ca, sophia.ananiadou@manchester.ac.uk

Abstract

The rise of the internet and social media has facilitated the spread of misinformation, with a significant impact on society. While the emergence of large language models (LLMs) offers new opportunities for detection and analysis, it also enables the large-scale generation of false or misleading information, making such content increasingly indistinguishable from reality. Therefore, there is a growing urgency for further research on misinformation detection in the era of LLMs. In this workshop, we aim to explore the potential of LLMs to address complex mis/disinformation detection challenges, foster discussions on the current state and future directions of the field, and advance the development of comprehensive frameworks that address the multifaceted nature of misinformation detection.

Introduction

The rapid expansion of social media platforms such as X, Facebook, and Weibo has fundamentally transformed how people access information. Recent data indicate that the number of active social media identities has surpassed five billion, representing approximately 63.9% of the global population.¹ This unprecedented level of connectivity has not been accompanied by equally effective regulatory frameworks. Consequently, online environments are increasingly filled with misinformation, including fabricated news, unfounded rumors, and conspiracy theories (Scheufele and Krause 2019). Recent developments, such as reports that platforms like Facebook and Instagram are scaling back fact-checking mechanisms,² further highlight the growing challenges of maintaining information integrity in the digital age. Such false information, as well as misleading arrangements of factual content used to support unjustified conclusions, can lead individuals to accept inaccurate narratives, shape public opinion, and cause significant harm to society, the economy, and political systems (Petraatos 2021). At the same time, recent advances in artificial intelligence, particularly in large language models such as ChatGPT and

DeepSeek, have further lowered the barrier to producing convincing yet misleading content (Barman, Guo, and Conlan 2024). These systems are capable of generating fluent and persuasive text at scale, which amplifies the risks associated with misinformation. Against this backdrop, there is an urgent global need to develop effective methods for identifying and mitigating erroneous and misleading information.

LLMs have brought notable improvements to misinformation detection, boosting both the speed and accuracy of predictive systems (Huang et al. 2025b; Liu et al. 2024; Chen and Shu 2024). Despite these gains, several limitations remain. Key challenges include issues of scalability, embedded biases, difficulties in capturing nuanced context, limited interpretability, and the ability to keep pace with emerging forms of deceptive content (Augenstein et al. 2024). In addition, the same models can be leveraged to produce highly persuasive false information at scale. Coupled with their tendency to hallucinate, these concerns raise important questions about the extent to which detection processes can be reliably automated (Huang et al. 2025a).

While several workshops and shared tasks already exist, such as CheckThat! Lab (Alam et al. 2025), which concentrates on the information verification pipeline, and FEVER³, a venue dedicated to verifiable knowledge extraction, its scope remains relatively limited. In particular, they tend to emphasize fact-checking techniques, leaving less room for more complex dimensions of misinformation, such as misleading interpretations drawn from otherwise factual content, propaganda strategies, real-world challenges in content moderation, and broader policy considerations.

This workshop explores the potential of LLMs and advanced NLP techniques to address the complex and evolving challenges of mis and disinformation detection on social media. It aims to examine current capabilities and limitations of LLM-based approaches, facilitate interdisciplinary discussions on future research directions, and consider the implications of deploying such methods within real-world content moderation systems. By addressing technical, social, and governance dimensions such as contextual understanding, robustness, and societal impact, the workshop seeks to foster the development of comprehensive frameworks for misinformation detection that are directly relevant to the

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://datareportal.com/reports/digital-2025-global-overview-report>

²<https://www.bbc.co.uk/news/articles/cly74mpy8klo>

³<https://fever.ai/>

ICWSM community’s focus on social media analysis, on-line discourse, and information integrity. Topics include but are not limited to:

TOPICS AND THEMES

- Methodology – Applying LLMs to identify fake news, rumors, or conspiracy theories.
 - Fact checking - Determining the ‘truth’ of claims against given background references.
 - Multi-modal/multi-lingual misinformation detection - Leverage different modalities/languages and combinations thereof to tackle online multimodal offensive content.
 - Cross-domain misinformation detection - identify misinformation collected from health, education, finance, politics, technologies, etc.
 - Stance detection - Identifying topics and sentiment/emotions.
 - Rhetoric detection - identify sarcasm, exaggeration, irony and other rhetorical strategies commonly used in mis/disinformation.
 - Network analysis - Analyzing social networks, dissemination methods, etc., of misinformation.
- Implication - Developing methods to identify misleading reasoning that uses true facts but leads to unwarranted conclusions.
- Interpretability - Providing explanations when detecting misinformation or fact-checking.
- User psychology - Analyzing the psycholinguistic features that may drive the engagements of misinformation
- Feature analysis - Analyzing the impact of different features for misinformation detection, such as emotion, style, stance, narrative strategies etc.
- Hallucination mitigation and evaluation in LLMs.
- Data source and benchmark - Contributions of new datasets and benchmarks and analysis of the misinformation generated by LLMs.
- Fairness of LLM moderation: Existing work has shown that LLMs exhibit systematic biases against different demographics (e.g. religion, age, or other cultural characteristics). To what extent does this impact misinformation detection?
- Policies and practical usage: LLMs are able to perform this task to a certain degree, but is this advisable? We welcome position papers on this topic.

This year, we introduce a shared-task based Reference-Free Counterfactual (RFC) benchmark (Jiang et al. 2026), which focuses on detecting plausible yet false financial narratives on the web and social platforms. The task encourages models to assess causality, contextual coherence, and credibility without relying on external fact sources, bridging NLP with web information ecosystems and social science perspectives on misinformation and trust.

Shared Task Details

Reference Free Narrative Misinformation Detection shared task assesses whether an LLM can recognize that a narrative is misleading without access to external evidence. Instead of relying on explicit false facts, the texts in RFC-Bench use subtle narrative manipulations that preserve surface plausibility while changing what the text warrants the reader to believe. This probes failures in belief updating. A coherent story is often accommodated even when it should trigger doubt. The task consists of a single Reference Free Detection setting. Given only one paragraph, the model must judge whether it is trustworthy or narratively manipulative based purely on its internal discourse structure and pragmatic commitments. The task differs from traditional misinformation detection tasks in that successful methods will resist persuasion from misleading arguments that, although logically valid, are not sound and are instead founded upon premises that have been deemed false during fact-checking.

Historical Iterations

The inaugural Workshop on Misinformation Detection in the Era of Large Language Models (MisD 2025) was held on June 23, 2025, as a one-day workshop co-located with ICWSM 2025 in Copenhagen, Denmark. The workshop focused on the opportunities and challenges introduced by large language models in the detection and analysis of misinformation, fostering interdisciplinary exchange between researchers in computer science, social science, and related fields. Through a rigorous peer-review process, seven papers were accepted, and together with two invited oral talks, the workshop featured a total of nine oral presentations. In addition, the program included two invited keynote talks by leading scholars addressing broader conceptual and empirical issues surrounding misinformation.

Invited Speakers, Accepted Submissions and Format

This workshop centers on misinformation detection in the era of large language models. It brought together a diverse and interdisciplinary audience, including researchers from computer science and the social sciences, as well as professionals and academics from both industry and academia.

We were honored to welcome two distinguished invited speakers, Dr. Preslav Nakov and Dr. Julia Mendelsohn, whose expertise greatly enriched the workshop program.

All submissions underwent a rigorous peer review process. Each paper was evaluated by at least three independent reviewers, with assessment criteria including originality, relevance, and technical soundness. After careful deliberation by the program committee, 15 high-quality papers were selected for inclusion in the workshop proceedings, including seven regular papers, five shared task papers, one demo paper, and two extended abstract papers. The titles of accepted papers are as follows:

- Are LLMs More Skeptical of Entertainment News?
- Cracking the Snowflake: Recovering Propagation Rhythms for Fake News Detection in the Post-API Era

- The Intent Gap in Disinformation Detection: Evidence from 84 Studies
- Milking the Metaphors: Cultural Obfuscation in Wellness (COW) around Gomutra on YouTube using LLMs
- DTCD-AFC: Disaster-Type Classification Dataset designed for Automated Fact-Checking
- Evolving Language, Enduring Claims: Language Mutations Sustain the Persistence of Conspiracy Theories on Social Media
- COMMUNITYNOTES: A Dataset for Exploring the Helpfulness of Fact-Checking Explanations
- PodChecker: An Interpretable Fact-Checking Companion for Podcasts
- M4Health: A Multi-Modal, Multi-Domain, Multi-Platform, and Multi-Task Benchmark for Video-Driven Health Communication on Social Media
- Benchmarking Health Misinformation Detection in Hidradenitis Suppurativa Communities Using Multi-Model LLM Ensembles
- Fact4ac at the Financial Misinformation Detection Challenge Task: Reference-Free Financial Misinformation Detection via Fine-Tuning and Few-Shot Prompting of Large Language Models
- DeepTruth at the Financial Misinformation Detection Challenge: Chain-of-Thought Enhanced LLM for Financial Misinformation Detection
- mfPE at the Financial Misinformation Detection Challenge Task: Multi-agent Framework Enhances the Performance of Prompt Engineering
- Coherence at the Financial Misinformation Detection Challenge Task: A Courtroom-Inspired Architecture for Reference-Free Verification
- AIspers at the Financial Misinformation Detection Challenge Task: Reference-Free Financial Integrity Detection with Benign Controls and Semantic Ranking

This workshop was a one-day workshop. Our agenda includes two keynote invited talks, 15 oral presentations, and two discussion sessions highlighting research opportunities, pioneering solutions, systems, success trajectories, and visionary future directions in the coming years.

List of Program Committee Members

Our Program Committee members consist of distinguished reviewers from both academia and industry. The list is as follows:

- Viswanathan Ranganathan - IEEE
- Ninaad Rao - Security AI, Cisco
- Peter Carragher - Carnegie Mellon University
- Tianlei Zhu - Columbia University
- Ye Yuan - McGill University and Mila - Quebec AI Institute
- Yuechen Jiang - University of Manchester
- Yankai Chen - MBZUAI
- Yuyan Wang - University of Manchester

- Zirui Wei - Data Science, C3.ai
- Zhiwei Liu - University of Manchester
- Aditya Gautam - Facebook
- Calvin Yixiang Cheng - University of Oxford
- Yixiang Zheng - University of Manchester
- Zhuohan Xie - MBZUAI
- Jahnavi Anilkumar Kachhia - Abbott diabetes care, Abbott
- Shaashwat Agrawal - Northwestern University
- Yupeng Cao - Stevens Institute of Technology
- Mina Basirat - University of Central Florida

Workshop Organizers

The organizing committee consists of General Organizers, Shared Task Organizers, and Advisors. The General Organizers oversee the overall operation of the workshop, including promotion, the call for papers, and review coordination. The Shared Task Organizers are responsible for organizing the shared task, including publicity, challenge deployment, and evaluation. The Advisors, composed of experienced professors and leading researchers, provide guidance and ensure the quality of the workshop's papers and discussions.

General Organizers

Zhiwei Liu is a PhD candidate at the Department of Computer Science at the University of Manchester. He focuses on the technical applications and discoveries of LLMs, primarily applied in misinformation detection and sentiment analysis. He has already published related papers in top journals or conferences, such as ACL, EMNLP, SIGKDD, NeurIPS, WWW, ECAI and Information Fusion. He is the co-organizer of 1st Workshop on Misinformation Detection in the Era of LLMs, FinNLP@AgentScen@IJCAI-2024 workshop and COLING2025 Financial Misinformation Detection Challenge, highlighting his organizational and coordination skills as well as his determination to drive the development of NLP and the field of misinformation detection.

Yupeng Cao is a final year PhD candidate at the Department of Electrical and Computer Engineering at Stevens Institute of Technology. His research focuses on Natural Language Processing, Multimodal, Trustworthy AI, and their applications in misinformation detection and fact-checking. He has published related papers in top conferences, such as ACL, NeurIPS, WWW, and InterSpeech. He served as a PC member for the 9th FinNLP workshop and Session Chair @ACM ICAIF'24. He organized the Agent-Based Market Simulation Challenge @COLING 2025. He also organized Microsymposium on Advances and Applications of LLMs in Finance at SIAM FM25.

Calvin Yixiang Cheng is a PhD candidate at Oxford Internet Institute, University of Oxford. His research examines how digital technologies shape the diffusion of misinformation, the formation of ideology, and the dynamics of public opinion. His work has been published in top computational social science journals and conferences, such as EPJ Data Science, ICWSM, Computational Communication Research, Convergence, and Journalism. He is enthusiastic and

has extensive experience organizing academic conferences and workshops, including the 1st Future of Social Media Research Workshop at Oxford in 2025 and the International Communication Association pre-conference Hackathon in Cape Town in 2026.

Zhuohan Xie is a Postdoctoral Associate at the Department of Natural Language Processing at Mohamed bin Zayed University of Artificial Intelligence. He focuses on agentic and interactive NLP methods for fact checking, primarily targeting evidence retrieval and verification in misinformation detection. He has published related papers in top journals and conferences, such as ACL, EMNLP, NAACL, and COLM. He is the co-organizer of several shared tasks, including SemEval 2025 Task 10, the GenAI Content Detection shared task, and ImageCLEF 2025, highlighting his organizational and coordination skills as well as his determination to advance NLP research and the field of misinformation detection.

Ye Yuan is a PhD candidate at the School of Computer Science at McGill University and Mila - Quebec AI Institute. His research focuses on Large Language Models (LLMs), generative modeling, offline black-box optimization (BBO), and Retrieval-Augmented Generation (RAG), with a particular interest in building trustworthy AI systems. He is a BMO Responsible AI Fellow and DAAD AINeT Fellow, and has published works in top-tier venues including NeurIPS, ICLR, EMNLP, TMLR, and others. He is also deeply committed to the research community, actively serving as a reviewer for leading conferences such as NeurIPS, ICLR, ICML, and WWW, and contributing to the organization of academic initiatives and mentorship programs.

Yankai Chen is a Postdoctoral Associate at the Department of Machine Learning at Mohamed bin Zayed University of Artificial Intelligence. He received his Ph.D. from The Chinese University of Hong Kong and served as a postdoctoral researcher there and at Cornell University. His research interests are trustworthy agentic AI and machine learning. He has published related papers in top Computer Science and Artificial Intelligence conferences and journals, such as NeurIPS, SIGKDD, WWW, AACL, TKDE, etc.

Shared Task Organizers

Yuechen Jiang is a PhD student at the University of Manchester. Her research focuses on LLM agents, financial decision-making, and cognitively grounded narrative understanding, with an emphasis on how models interpret information, update beliefs, and reason under uncertainty. She has published related papers in top conferences, such as ACL and NeurIPS. She has been actively involved in organizing community benchmarks and evaluations, and co-organized the FinLLMs Challenge at IJCAI 2024 (FinNLP-AgentScen) and the Agent-Based Single Cryptocurrency Trading Challenge at COLING 2025 (FinNLP-FNP-LLMFinLegal).

Yuyan Wang is a PhD student at the Department of Computer Science at the University of Manchester. Her research focuses on socially and culturally adaptive LLMs for behavioral understanding, particularly how LLMs understand, simulate, and influence mental and behavioral states based

on the theory of mind, empathy, self-reflection and behavioral modelling. She has published paper in top conference such as ICLR. Her work provides critical insights into the cognitive mechanisms underlying social influence and information propagation in digital environments.

Peter Carragher is a PhD student at the Center for Computational Analysis of Social and Organizational Systems and Center for Informed Democracy & Social-Cybersecurity at Carnegie Mellon University. His research primarily focuses on assessing source credibility and trust in online news media, modeling how low credibility news spreads both through social influence processes in social media and via the exploitation of search engine ranking algorithms. He also developed machine learning detection and response systems at Meta, where his work has been deployed in areas of fraud and deception such as impersonation, coordinated inauthentic behavior, and election integrity.

Advisors

Jimin Huang is the founder and president of The Fin AI community, which is an initiative dedicated to advancing open science, tooling, and model development for the financial services industry with a focus on responsible innovation. The Fin AI is now an associated member of FINOS and a collaborator with the NVIDIA AI Technology Center (NVAITC) through the University of Florida. Jimin is also a PhD from the University of Manchester. His research spans natural language processing and computational finance, with a particular emphasis on financial large language models (LLMs) and open-source contributions. He is the organizer of the FinLLM Challenge at FinNLP-AgentScen @ IJCAI-2024, and the general chair for The Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal) @ COLING 2025.

Xue (Steve) Liu is a Professor of Computer Science and Machine Learning and the Associate Vice President of Research at MBZUAI. Prior to this, he was a Full Professor and William Dawson Scholar at McGill University and an associate member of the Quebec AI Institute (Mila). With an h-index of 75 and over 26,700 citations (Google Scholar), his research focuses on AI/ML, Cyber-Physical Systems (CPS), and sustainable computing. Professor Liu is a Fellow of the IEEE and a Fellow of the Canadian Academy of Engineering. He has held significant leadership roles in industry, serving as Vice President of R&D and Chief Scientist at Samsung AI Center Montréal (2019–2024) and Chief Scientist at Tinder (2016–2019). His inventions have been integrated into technologies used by tens of millions of people through collaborations with global leaders such as IBM, Microsoft, and General Motors. His excellence in research and leadership has been recognized with the Mitacs Award for Exceptional Leadership, the Outstanding Young Canadian Computer Science Researcher Prize, and McGill's Tomlinson Scientist Award.

Sophia Ananiadou is a Professor of Computer Science at the University of Manchester. She is the Director of the UK National Centre for Text Mining and holds roles such

as Deputy Director of the Institute of Data Science and AI (Manchester), ELLIS fellow, Distinguished Research Fellow at the AI Research Centre (AIST Japan), and Lead Researcher at the Archimedes Research Centre, Athens, Greece. Her research focus is on leveraging NLP to understand and utilize language knowledge, in special domains such as biomedicine, health and finance for the tasks of information extraction, summarisation, emotion and misinformation detection. Her h-Index is 79 (Google Scholar) with 27,000 citations. She has organized numerous workshops and shared tasks (e.g., BioNLP and CL4Health) at major venues such as ACL, EMNLP, and COLING. She is one of the founders of SIGBioMed (ACL) and SIG-FinTech (ACL).

Acknowledgments

This work was supported by the NVIDIA Academic Grant Program using 32K A100 GPU-hours on Brev. We thank all workshop participants, organizers, and reviewers for their valuable contributions, as well as The Fin AI community for its research support, feedback, and collaborative environment that made this work possible.

References

- Alam, F.; Struß, J. M.; Chakraborty, T.; Dietze, S.; Hafid, S.; Korre, K.; Muti, A.; Nakov, P.; Ruggeri, F.; Schellhammer, S.; et al. 2025. Overview of the CLEF-2025 Check-That! Lab: Subjectivity, fact-checking, claim normalization, and retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 199–223. Springer.
- Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G. L.; Corney, D.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.; et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8): 852–863.
- Barman, D.; Guo, Z.; and Conlan, O. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 100545.
- Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3): 354–368.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.
- Huang, T.; Yi, J.; Yu, P.; and Xu, X. 2025b. Unmasking Digital Falsehoods: A Comparative Analysis of LLM-Based Misinformation Detection Strategies. *arXiv preprint arXiv:2503.00724*.
- Jiang, Y.; Liu, Z.; Cao, Y.; He, Y.; Xu, Z.; Xu, C.; Deng, Z.; Tiwari, P.; Chen, X.; Lopez-Lira, A.; et al. 2026. All That Glisters Is Not Gold: A Benchmark for Reference-Free Counterfactual Financial Misinformation Detection. *arXiv preprint arXiv:2601.04160*.
- Liu, Z.; Yang, K.; Xie, Q.; de Kock, C.; Ananiadou, S.; and Hovy, E. 2024. RAEmoLLM: Retrieval Augmented LLMs for Cross-Domain Misinformation Detection Using In-Context Learning based on Emotional Information. *arXiv preprint arXiv:2406.11093*.
- Petratos, P. N. 2021. Misinformation, disinformation, and fake news: Cyber risks to business. *Business Horizons*, 64(6): 763–774.
- Scheufele, D. A.; and Krause, N. M. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16): 7662–7669.