

Zhuohan Xie, Ph.D.

Research Scientist, MBZUAI AI for Finance | Financial Agents | Trustworthy LLM Evaluation | Benchmark Design

✉ zhuohan.xie18@gmail.com 🌐 Website 🔗 LinkedIn 🐙 GitHub 🎓 Scholar

Updated September 2026

Summary: I am a Research Scientist at MBZUAI working on AI for Finance, financial agents, and trustworthy LLM evaluation. I build benchmarks, shared tasks, and verification systems for financial reasoning, evidence-grounded decision-making, and multilingual and multimodal evaluation. I also lead community evaluation initiatives and mentor early-career researchers from problem formulation through publication and release.

Selected Conference and Journal Publications

ACL 2026

- **[Main, Oral]** **FinChain: A Symbolic Benchmark for Verifiable Chain-of-Thought Financial Reasoning**
Zhuohan Xie, Daniil Orel, Rushil Thareja, Dhruv Sahnan, ..., Salem Lahlou, Veselin Stoyanov, Sophia Ananiadou, Preslav Nakov. *First author.*
- **[Main, Oral]** **SAHM: A Benchmark for Arabic Financial and Shari'ah-Compliant Reasoning**
Rania Elbadry, ..., Sophia Ananiadou, Veselin Stoyanov, Jimin Huang, Xueqing Peng, Preslav Nakov, **Zhuohan Xie**. *Last author.*
- **[Findings]** **FinCARDS: Card-Based Analyst Reranking for Financial Document Question Answering**
Yixi Zhou, Fan Zhang, Yu Chen, Haipeng Zhang, Preslav Nakov, **Zhuohan Xie**. *Last author.*
- **[Demo]** **FinReporting: An Agentic Workflow for Localized Reporting of Cross-Jurisdiction Financial Disclosure**
Fan Zhang, Mingzi Song, Rania Elbadry, ..., Preslav Nakov, **Zhuohan Xie**. *Last author.*
- **[Main, Oral]** **Is Human-Like Text Liked by Humans? Multilingual Human Detection and Preference Against AI**
Yuxia Wang, Rui Xing, ..., **Zhuohan Xie**, ..., Nizar Habash, Iryna Gurevych, Preslav Nakov. *Co-author.*
- **[Main]** **Cultural Benchmarking of LLMs in Standard and Dialectal Arabic Dialogues**
Muhammad Dehan Al Kautsar, Saeed Almheiri, ..., **Zhuohan Xie**, Junhong Liang, Mohammad Rustom Al Nasar, Preslav Nakov, Fajri Koto. *Co-author.*
- **[Findings]** **Same Claim, Different Judgment: Benchmarking Scenario-Induced Bias in Multilingual Financial Misinformation Detection**
Zhiwei Liu, ..., **Zhuohan Xie**, ..., Alejandro Lopez-Lira, Xue Liu, Jimin Huang, Sophia Ananiadou. *Co-author.*
- **[Findings]** **RealFin: How Well Do LLMs Reason About Finance When Users Leave Things Unsaid?**
Yuyang Dai, Yan Lin, **Zhuohan Xie**, Yuxia Wang. *Co-author.*

SIGIR 2026

- **[Oral]** **Multi-Sourced, Multi-Agent Evidence Retrieval for Fact-Checking**
Shuzhi Gong, Richard O. Sinnott, Jianzhong Qi, Cécile Paris, Preslav Nakov, **Zhuohan Xie**. *Last author.*
- **[Resource]** **Conv-FinRe: A Conversational and Longitudinal Benchmark for Utility-Grounded Financial Recommendation**
Yan Wang, Yi Han, Lingfei Qian, ..., **Zhuohan Xie**, ..., Jimin Huang, Yankai Chen, Xue Liu, Jian-Yun Nie. *Co-author.*

EMNLP 2026

- **[Findings]** **Right Frame, Wrong Rule: Cultural Cues Expose the Financial Knowledge Gap They Were Meant to Close**
Rania Elbadry, Ahmed Heakl, Saeed Almheiri, ..., Preslav Nakov, Yuxia Wang, Fajri Koto, **Zhuohan Xie**. *Last author; accepted, forthcoming.*
- **[Main]** **Jais 2: A Family of Arabic-Centric Open Large Language Models**
Mohamed Anwar, ..., Yuxia Wang, **Zhuohan Xie**, Ali Mekky, ..., Michalis Vazirgiannis, Preslav Nakov. *Co-author; accepted, forthcoming.*
- **[Main]** **Aligning Cultural Behaviors in Arabic Language Models via Preferences**
Junhong Liang, Daniil Orel, Saeed Almheiri, ..., Abed Alhakim Freihat, **Zhuohan Xie**, Momina Ahsan, ..., Preslav Nakov, Fajri Koto. *Co-author; accepted, forthcoming.*

Selected Conference and Journal Publications (continued)

AAACL 2026

- **[Findings] SQLStructEval: Structural Evaluation of LLM Text-to-SQL Generation**
Yixi Zhou, Fan Zhang, Zhiqiao Guo, Yu Chen, Haipeng Zhang, Preslav Nakov, **Zhuohan Xie**. *Last author; accepted, forthcoming.*

TACL 2026

- **Can LLMs Automate Fact-Checking Article Writing?**
Dhruv Sahnan, . . . , **Zhuohan Xie**, Iryna Gurevych, Elizabeth Churchill, Tanmoy Chakraborty, Preslav Nakov. *Co-author.*

ACL 2025

- **[Main] KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan**
Mukhammed Togmanov, Nurdaulet Mukhituly, . . . , Akhmed Sakip, **Zhuohan Xie**, Yuxia Wang, . . . , Preslav Nakov, Fajri Koto. *Co-author.*
- **[Findings] Entity Framing and Role Portrayal in the News**
Tarek Mahmoud, **Zhuohan Xie**, Dimitar Dimitrov, . . . , Giovanni Da San Martino, Jakub Piskorski, Preslav Nakov. *Co-author.*

EMNLP 2025

- **[Main] A Head to Predict and a Head to Question: Pre-trained Uncertainty Quantification Heads for Hallucination Detection in LLM Outputs**
Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, **Zhuohan Xie**, . . . , Preslav Nakov, Timothy Baldwin. *Co-author.*

COLM 2025

- **Sherkala-Chat: Building a State-of-the-Art LLM for Kazakh in a Moderately Resourced Setting**
Fajri Koto, Rituraj Joshi, Nurdaulet Mukhituly, Yuxia Wang, **Zhuohan Xie**, *Co-author.*

NAACL 2025

- **[Findings] FIRE: Fact-checking with Iterative Retrieval and Verification**
Zhuohan Xie, Rui Xing, Yuxia Wang, . . . , Iryna Gurevych, Preslav Nakov. *First author.*

EMNLP 2023

- **[Findings] DeltaScore: Fine-Grained Story Evaluation with Perturbations**
Zhuohan Xie, Miao Li, Trevor Cohn, Jey Lau. *First author.*

INLG 2023

- **The Next Chapter: A Study of Large Language Models in Storytelling**
Zhuohan Xie, Trevor Cohn, Jey Han Lau. *First author.*

Selected Shared-Task and Workshop Papers

- **[CLEF 2026, LNCS] Overview of FinMMEval 2026: Multilingual and Multimodal Financial Evaluation**
Zhuohan Xie, Yuyang Dai, Rania Elbadry, . . . , Xue Liu, Preslav Nakov
First author; forthcoming. Overview of the FinMMEval 2026 lab.
- **[FinNLP 2026, Poster] Prediction-Market Seed Capital Recovery from Noise-Dominant Flow**
Yankai Chen, Bowei He, **Zhuohan Xie**, Zhiwei Liu, Fan Zhang, Xueqing Peng, Yan Wang, Lingfei Qian, Xue Liu
Co-author; accepted Long Paper; manuscript not yet public. Prediction-market mechanism design and seed-capital recovery.
- **[SemEval 2025] SemEval 2025 Task 10: Multilingual Characterization and Extraction of Narratives from Online News**
Jakub Piskorski, Tarek Mahmoud, . . . , Nicolas Stefanovitch, **Zhuohan Xie**, Preslav Nakov, Giovanni Da San Martino
Co-author and task organizer. Shared-task overview of multilingual entity framing and narrative analysis.

Selected Preprints

- **[arXiv 2026] Herculean: An Agentic Benchmark for Financial Intelligence**
Xueqing Peng, **Zhuohan Xie**, Yupeng Cao, Haohang Li, Lingfei Qian, ..., Sophia Ananiadou
Co-author. Large-scale benchmark for agentic financial intelligence.
- **[arXiv 2026] FinPersona-Bench: A Benchmark for Longitudinal Psychometric Stability of Autonomous Financial Agents**
Muhammad Usman Safder, Ayesha Gull, Rania Elbadry, ..., Preslav Nakov, **Zhuohan Xie**
Last author. Long-horizon financial-agent evaluation for mandate stability and selective re-grounding.
- **[arXiv 2026] Can LLMs Be CEOs? Benchmarking Strategic Resource Reallocation with Multi-Role Agent Simulation**
Yuyang Dai, Xueqing Peng, Lingfei Qian, **Zhuohan Xie**
Last author. Constraint-rich executive decision-making with multi-role agent simulation.
- **[arXiv 2026] Seeing Is Not Deciding: Can Multimodal LLMs Act as Effective CEOs?**
Yuyang Dai, Xueqing Peng, Yuxia Wang, Preslav Nakov, **Zhuohan Xie**
Last author. Multimodal executive-decision benchmark separating perception from constrained action.
- **[arXiv 2026] Style Wins, Substance Loses: A Diagnosis of LLM-as-Judge in Idea Generation**
Fengxian Ji, Yuke Li, Jingpu Yang, ..., Min Peng, Qianqian Xie, Xiuying Chen, **Zhuohan Xie**
Last author. Bias diagnosis and mitigation for LLM-based evaluation of scientific ideas.

Research Program

- **Vision: Reliable Evaluation Infrastructure for High-Stakes LLMs.** I build benchmarks, shared tasks, and verification systems that test whether LLMs can reason with evidence, structured constraints, multilingual inputs, and domain risk rather than simply generate plausible text.
- **Outputs: Finance, Agents, and Verification.** This program spans verifiable financial reasoning (FinChain, SAHM, FinCARDS), multilingual and multimodal evaluation (FinMMEval), agentic decision benchmarks (Herculean, FinPersona-Bench, CEO-Bench, C-SuiteBench), and evidence systems (FIRE and multi-agent retrieval).

Research Leadership and Academic Service

May 2026 – Present	Chair , Asia-Pacific Chapter, The Fin AI.
2026	Area Chair , FinNLP 2026, a workshop at EMNLP 2026.
2026–2027	Lab Organizer , CLEF FinMMEval. Organized the 2026 lab, including participant communication, evaluation workflow, and working-note coordination. The 2027 lab has been accepted and is upcoming.
2025–Present	Research Mentoring , Rania Elbadry, Yixi Zhou, Shuzhi Gong, Fan Zhang, Fengxian Ji, and Yuyang Dai; co-developed ACL/SIGIR papers and 2026 benchmark releases.
2025–2027	Organizer and Committee Member , SemEval 2025 Task 10 Organizer; SemEval 2027 Task Selection Committee.
2023–Present	Reviewer , ACL Rolling Review, NeurIPS, COLM, EMNLP, ALTA, CHOMPS, TFAI, and MisD.

Appointments and Education

May 2026 – Present	Research Scientist , MBZUAI, Abu Dhabi, UAE
Jul 2024 – May 2026	Postdoctoral Associate , MBZUAI, Abu Dhabi, UAE
Feb 2024 – Jul 2024	Research Assistant , MBZUAI, Abu Dhabi, UAE
Feb 2020 – Jul 2024	Ph.D. in Computing and Information Systems , University of Melbourne. Thesis: <i>Automatic Story Generation and Evaluation</i> .
Jul 2017 – Jun 2019	Master of Information Technology , University of Melbourne. Thesis: <i>Chinese Poetry Automatic Generation</i> .
Sep 2013 – Jun 2017	Bachelor of Software Engineering and Bachelor of Advertising , Xiamen University.

Teaching, Awards, and Skills

Teaching/Awards	Academic Tutor, University of Melbourne, 2019–2023; Melbourne Research Scholarship; Dean’s Honours List, School of Engineering, University of Melbourne.
Skills	Python, PyTorch, Hugging Face Transformers, OpenAI API, LangChain, scikit-learn, Bash, LaTeX; English and Mandarin Chinese.