

Zhehao Zhang

✉ zhang.16420@osu.edu | 🏠 [zhehaozhang123.github.io](https://github.com/zhehaozhang123) | 🎓 Google Scholar

—
Last updated: August 2, 2026

Research Interests

My research centers on the safety, robustness, and alignment of large language models and language agents. I am currently interested in (i) evaluating and characterizing the failure modes of language agents in realistic, high-stakes settings, (ii) developing methods to mitigate unsafe and unintended agent behaviors, and (iii) building agents that are reliable and self-improving under minimal supervision.

Industry Experience

Netflix Machine Learning Intern, <i>Host: William Harvey</i>	Jun 2026–Aug 2026
Amazon Applied Scientist Intern, <i>Host: Weijie Xu, Chandan K. Reddy</i>	Nov 2024–Jun 2025
Adobe Research Research Intern, <i>Host: Ryan A. Rossi</i>	Jun 2024–Aug 2024
Microsoft Research Lab – Asia Research Intern, <i>Host: Yan Gao</i>	Dec 2022–Aug 2023

Academic Experience

OSU NLP Lab, The Ohio State University PhD Student, Advisor: Yu Su, Huan Sun	2025–Present
Social and Language Technologies (SALT) Lab, Stanford University Research Intern, Host: Diyi Yang	2023–2024

Education

The Ohio State University Ph.D., Computer Science and Engineering Advisor: Yu Su	2025–Present
Dartmouth College M.S., Computer Science	2023–2024
Shanghai Jiao Tong University B.Eng., Artificial Intelligence (<i>Honor Class</i>)	2019–2023

Selected Awards & Honors

ICML Gold Reviewer	2026
Graduate Fellowship, Ohio State University	2025
COLM 2025 Travel Grant	2025
ICLR Notable Reviewer	2025
Merit Scholarship, Dartmouth College	2023–2025
Zhiyuan Honor Scholarship and Merit Scholarship, SJTU	2019–2023

Publications [\[G\]](#)

Conference Papers

- [C1] Jaylen Jones*, **Zhehao Zhang***, Yuting Ning, Eric Fosler-Lussier, Pierre-Luc St-Charles, Yoshua Bengio, Dawn Song, Yu Su, Huan Sun. “When Benign Inputs Lead to Severe Harms: Eliciting Unsafe Unintended Behaviors of Computer-Use Agents”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2026.
- [C2] **Zhehao Zhang**, Ryan A. Rossi, Tong Yu, Franck Dernoncourt, Ruiyi Zhang, Jiuxiang Gu, Sungchul Kim, Xiang Chen, Zichao Wang, Nedim Lipka. “VipAct: Visual-Perception Enhancement via Specialized VLM Agent Collaboration and Tool-use”. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2026.
- [C3] **Zhehao Zhang**, Weijie Xu, Fanyou Wu, Chandan K. Reddy. “FalseReject: A Resource for Improving Contextual Safety and Mitigating Over-Refusals in LLMs via Structured Reasoning”. In: *Proceedings of the Conference on Language Modeling (COLM)*. 2025.
- [C4] **Zhehao Zhang**, Jiaao Chen, Diyi Yang. “DARG: Dynamic Evaluation of Large Language Models via Adaptive Reasoning Graph”. In: *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. 2024.
- [C5] **Zhehao Zhang**, Weicheng Ma, Soroush Vosoughi. “Is GPT-4V(ision) All You Need for Automating Academic Data Visualization? Exploring Vision-Language Models’ Capability in Reproducing Academic Charts”. In: *Findings of the Association for Computational Linguistics: EMNLP (Findings of EMNLP)*. 2024.
- [C6] **Zhehao Zhang**, Yan Gao, Jian-Guang Lou. “E5: Zero-shot Hierarchical Table Analysis using Augmented LLMs via Explain, Extract, Execute, Exhibit, and Extrapolate”. In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2024.
- [C7] **Zhehao Zhang**, Tong Yu, Handong Zhao, Kaige Xie, Lina Yao, Shuai Li. “Exploring Soft Prompt Initialization Strategy for Few-Shot Continual Text Classification”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024.
- [C8] **Zhehao Zhang**, Xitao Li, Yan Gao, Jian-Guang Lou. “CRT-QA: A Dataset of Complex Reasoning Question Answering over Tabular Data”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023.
- [C9] **Zhehao Zhang**, Jiaao Chen, Diyi Yang. “Mitigating Biases in Hate Speech Detection from A Causal Perspective”. In: *Findings of the Association for Computational Linguistics: EMNLP (Findings of EMNLP)*. 2023.
- [C10] Yuting Ning, Jaylen Jones, **Zhehao Zhang**, Chentao Ye, Weitong Ruan, Junyi Li, Rahul Gupta, Huan Sun. “When Actions Go Off-Task: Detecting and Correcting Misaligned Actions in Computer-Use Agents”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2026.
- [C11] Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, Zhengmian Hu, Yue Zhao, Nedim Lipka, Seunghyun Yoon, Ting-Hao Huang, Zichao Wang, Puneet Mathur, Soumyabrata Pal, Koyel Mukherjee, **Zhehao Zhang**, Namyong Park, Thien Huu Nguyen, Jiebo Luo, Ryan A. Rossi, Julian McAuley. “From Selection to Generation: A Survey of LLM-based Active Learning”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 2025.
- [C12] Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, Junda Wu, Ashish Singh, Yu Wang, Jiuxiang Gu, Nesreen K. Ahmed, Nedim Lipka, Ruiyi Zhang, Xiang Chen, Tong Yu, Sungchul Kim, Hanieh Deilamsalehy, Namyong Park, Michael Rimer, **Zhehao Zhang**, Huanrui Yang, Puneet Mathur, Gang Wu, Franck Dernoncourt, Ryan A. Rossi, Thien Huu Nguyen. “A Survey on Small Language Models”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*. 2025.
- [C13] Reuben Luera, Ryan A. Rossi, Franck Dernoncourt, Alexa F. Siu, Sungchul Kim, Tong Yu, Ruiyi Zhang, Xiang Chen, Nedim Lipka, **Zhehao Zhang**, Seon Gyeom Kim, Tak Yeon Lee. “Personalizing Data Delivery: Investigating User Characteristics and Enhancing LLM Predictions”. In: *Companion Proceedings of the ACM Web Conference (WWW)*. 2025.

Journal Articles

- [J1] **Zhehao Zhang**, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, Yu Wang. “Personalization of Large Language Models: A Survey”. In: *Transactions on Machine Learning Research (TMLR)*. 2025.
- [J2] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, **Zhehao Zhang**, Diyi Yang. “Can Large Language Models Transform Computational Social Science?” In: *Computational Linguistics (CL)*. 2024.

Preprints

- [PP1] Yuting Ning*, **Zhehao Zhang***, Yash Kumar Lal, Boyu Gou, Junyi Li, Weitong Ruan, Chentao Ye, Rahul Gupta, Diyi Yang, Yu Su, Huan Sun. “SkillHarm: Lifecycle-Aware Skill-Based Attacks via Automated Construction”. In: *arXiv preprint*. 2026.
- [PP2] **Zhehao Zhang**, Weijie Xu, Shixian Cui, Chandan K. Reddy. “Distractor Injection Attacks on Large Reasoning Models: Characterization and Defense”. In: *arXiv preprint*. 2025.
- [PP3] Jian Xie, Tianhe Lin, Zilu Wang, Yuting Ning, Yuekun Yao, Tianci Xue, **Zhehao Zhang**, Zhongyang Li, Kai Zhang, Yufan Wu, Shijie Chen, Boyu Gou, Mingzhe Han, Yifei Wang, Vint Lee, Xinpeng Wei, Xiangjun Wang, Yu Su, Huan Sun. “QUEST: Training Frontier Deep Research Agents with Fully Synthetic Tasks”. In: *arXiv preprint*. 2026.
- [PP4] Junda Wu, Hanjia Lyu, Yu Xia, **Zhehao Zhang**, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A. Rossi, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Namyoung Park, Sungchul Kim, Huanrui Yang, Subrata Mitra, Zhengmian Hu, Nedim Lipka, Dang Nguyen, Yue Zhao, Jiebo Luo, Julian McAuley. “Personalized Multi-modal Large Language Models: A Survey”. In: *arXiv preprint*. 2024.
- [PP5] Junda Wu, **Zhehao Zhang**, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A. Rossi, Ruiyi Zhang, Subrata Mitra, Dimitris N. Metaxas, Lina Yao, Jingbo Shang, Julian McAuley. “Visual Prompting in Multimodal Large Language Models: A Survey”. In: *arXiv preprint*. 2024.
- [PP6] Reuben Luera, Ryan Rossi, Franck Dernoncourt, Alexa Siu, Sungchul Kim, Tong Yu, Ruiyi Zhang, Xiang Chen, Nedim Lipka, **Zhehao Zhang**, Seon Gyeom Kim, Tak Yeon Lee. “Optimizing Data Delivery: Insights from User Preferences on Visuals, Tables, and Text”. In: *arXiv preprint*. 2024.

Media Coverage

- Jul 2025 *FalseReject: Reducing Overcautiousness in LLMs through Reasoning-Aware Safety Evaluation.* [Amazon Science Blog](#).
- Jul 2025 *FalseReject* covered by [AI Era](#).

Invited Talks

- [PR1] *Recent Advances in Synthetic Data for Foundation Models*. Stanford SALT Lab. Host: Prof. Diyi Yang. Sep. 2024.
- [PR2] *Augmented Language Models*. TRIP Lab, Dartmouth College. Host: Prof. Yaoqing Yang. Mar. 2024.

Academic Service

- Conference Reviewer** ICML 2026; NeurIPS 2023, 2024, 2025; ICLR 2025, 2026; AAAI 2027; IJCAI 2025; ACL 2024, 2025; EMNLP 2023, 2024, 2025, 2026; NAACL 2024; COLM 2024, 2025, 2026; COLING 2025; CIKM 2024, 2025
- Journal Reviewer** Transactions on Machine Learning Research (TMLR); IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
- Volunteer** EMNLP 2023; NAACL 2024

Skills

Programming	Python, C/C++, JavaScript, L ^A T _E X
Post-Training & Inference	PyTorch, JAX, verl, vLLM, Harbor; SFT, RLHF/RLVR, OPD, LoRA
Agent Systems	Agent harnesses (Claude Code / Codex), tool-use, multi-agent orchestration, MCP, Agent Skills
Research Areas	LLM Safety & Alignment, Language Agents, AI Red-Teaming, LLM Evaluation, Synthetic Data