Xiangyu Zhou

+01-2019120046 xiangyu@wayne.edu Google Scholar Xiangyu Zhou xzhou98 Website

RESEARCH INTERESTS

Trustworthy AI: Safety and Robustness & **Large Language Model** & **Large Reasoning Model** & **AI Agents**

EDUCATION

Wayne State University, Detroit, Michigan, USA. **Aug 2023 – Present**
Ph.D. in Computer Science
Advisor: Dr. Dongxiao Zhu

Stevens Institute of Technology, Hoboken, New Jersey, USA. **Aug 2021 – May 2023**
M.S. in Computer Science

Chongqing University of Posts and Telecommunications, Chongqing, China. **Sep 2017 – May 2021**
B.S. in Software Engineering

ONGOING RESEARCH PROJECTS

- **Safety and Helpfulness Alignment of LLMs:** Building reinforcement-learning and contrastive-learning pipelines to improve model safety and helpfulness, including custom safety/helpfulness datasets, format rewards for GRPO, and post-training fine-tuning of LLMs.
- **Robustness of LRMs under In-Context Learning:** Studying how frontier reasoning models fail under multi-turns or few-shots conversations and developing post-training methods to improve robustness without degrading general performance.

WORK EXPERIENCE

Trustworthy AI Lab, Wayne State University **Aug 2023 – Present**
Graduate Research Assistant *Detroit, MI*

- Conduct research on trustworthy LLMs, focusing on unlearning, safety alignment, and reasoning robustness.

Shandong Googosoft Co., Ltd. **Jul 2019 – Aug 2019**
Full Stack Developer, Intern *Shandong, China*

- Contributed to the development of a university Asset Management Platform used for financial tracking and asset reporting.
- Improved front-end reliability of web app by debugging JavaScript interaction issues (e.g., unresponsive UI elements).

PUBLICATIONS

Peer-Reviewed Publications

- **Zhou, X.**, et al. "Not All Tokens Are Meant to Be Forgotten." *Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 40. No. 44. 2026.* ***(Accepted as Oral, AAAI-26)***
- Zade, S. Z., **Zhou, X.**, Liu, S., Zhu, D. "Attention Smoothing Is All You Need For Unlearning" *The Fourteenth International Conference on Learning Representations (**ICLR-26**).*
- Li, M., **Zhou, X.**, Wu, J. et al. "Unraveling salt-responsive genes in Suaeda salsa through genomic and transcriptomic profiling across salinity gradients" *BMC Genomics 27, 475 (2026).*
- Roshani, M. A., **Zhou, X.**, Qiang, Y., Suresh, S., Hicks, S., Sethuraman, U., & Zhu, D. "Generative large language model—powered conversational ai app for personalized risk assessment: Case study in covid-19." *Published in the Journal of Medical Internet Research **(JMIR)**.*
- Sultan, R., Zhu, H., **Zhou, X.**, Li, C., Khanduri, P., Brocanelli, M., Zhu, D., et al. "WalkGPT: Grounded Vision–Language Conversation with Depth-Aware Segmentation for Pedestrian Navigation." *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2026 (**CVPR-26**).*

- Zade, S. Z., Qiang, Y., **Zhou, X.**, Zhu, H., Roshani, M. A., Khanduri, P., & Zhu, D. "Automatic Calibration for Membership Inference Attack on Large Language Models" *In Proceedings of the 28th European Conference on Artificial Intelligence **(ECAI-25)**.*
- Zheng, W., Walquist, E., Datey, I., **Zhou, X.**, Berishaj, K., ... & Zytko, D. "It's Not What We Were Trying to Get At, but I Think Maybe It Should Be": Learning How to Do Trauma-Informed Design with a Data Donation Platform for Online Dating Sexual Violence. *In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems **(CHI-24)**.*
- Zheng, W., Walquist, E., Datey, I., **Zhou, X.**, Berishaj, K., Mcdonald, M., ... & Zytko, D. "Towards trauma-informed data donation of sexual experience in online dating to improve sexual risk detection AI." *In proceedings of the 36th annual ACM symposium on user interface software and technology **(UIST-23)**.*
- Walquist, E., Datey, I., Zheng, W., **Zhou, X.**, Berishaj, K., McDonald, M., ... & Zytko, D. "Collective Consent: Who Needs to Consent to the Donation of Data Representing Multiple People?". *Proceedings of the ACM on Human-Computer Interaction **(CSCW-25)**.*

Preprints and Under Review

- **Zhou, X.**, et al. "Alignment of LRMs via Counter-Aligned Few-Shot Conversation Exposure." *Under review at The Fortieth Annual Conference on Neural Information Processing Systems (NeurIPS-26).*
- **Zhou, X.**, et al. "Hijacking large language models via adversarial in-context learning." *Under review at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD -26), arXiv:2311.09948.*
- **Zhou, X.**, et al. "Learning to Poison Large Language Models for Downstream Manipulation." *Under review at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD -26) arXiv:2402.13459.*

REVIEWING EXPERIENCE

Program Committee Member: The 40th Annual AAAI Conference on Artificial Intelligence (**AAAI-26**), Forty-Third International Conference on Machine Learning (**ICML-26**, The Fortieth Annual Conference on Neural Information Processing Systems (NeurIPS-26))

PROJECTS EXPERIENCE

Ube (Data Donation Android Application) **Feb 2023 – Apr 2024**

Full Stack Developer

- Independently developed an Android application with **React-Native** as the framework, and **Firebase** as the database.
- **Collaborated in writing and co-authoring** the research paper detailing the application's architecture and its implications for data privacy.
- The papers were successfully published at UIST-23, CHI-24, and CSCW-25, contributing to the discourse on ethical data usage and privacy technologies.

Movies Recommendation System **Oct 2021 – Dec 2021**

Full Stack Developer

- Developed a movie communication website with **React** as the front end, **Node.js** as the back end, and **MongoDB** as the database.
- Implemented login, sign up and log out, finished hashing password and matching to password with **Bcrypt**, and improved **security** for user login verification.
- Accomplished insertion, updating, deletion, and querying of information of user and movie's comment, imported **XSS** defense to protect the system from Cross-Site Scripting Attack.

Campus Delivery System **Nov 2018 – Mar 2019**

Front-end Developer

- Developed a web application with **HTML, JavaScript, jQuery**, and **Vue** as the front end.
- Designed and coded functions for items adding, querying, modifying, and deleting of goods and riders.

SKILLS

- **Programming:** Python, Java, Javascript
- **Frameworks:** PyTorch, HuggingFace, React, React-native, Vue
- **Tool:** Git, Firebase, MySQL, MongoDB
- **Machine Learning:** Large Language Models, Large Reasoning Models, Trustworthy AI, Fine tuning, Reinforcement Learning