

Saleh Zare Zade

Detroit, USA

salehz@wayne.edu — salehzarehzadeh@gmail.com

[LinkedIn](#) — [GitHub](#) — [Google Scholar](#)

SUMMARY

Ph.D. candidate in Computer Science at Wayne State University focusing on large language model (LLM) safety, privacy, and unlearning. Research centers on membership inference attacks, memorization behavior, and robust mitigation strategies for trustworthy AI systems.

RESEARCH INTERESTS

Large Language Models, Unlearning, Safety, Reinforcement Learning, Privacy and Membership Inference Attacks, Trustworthy AI, Machine Learning

EDUCATION

Wayne State University 2023 – Present

Ph.D. in Computer Science

Advisor: Prof. Dongxiao Zhu

GPA: 3.89

Research: LLM Safety, Jailbreaking, Unlearning

University of Tehran 2017 – 2022

B.Sc. in Computer Science

GPA: 3.53

Coursework: Machine Learning, NLP, Data Mining, Information Retrieval, Statistical Methods

WORK EXPERIENCE

Graduate Research Assistant Aug 2025 – Present

Wayne State University

- Conduct research on LLM safety, privacy risks, and unlearning methods.
- Develop methods for membership inference and model behavior analysis.

Graduate Teaching Assistant Aug 2024 – Aug 2025

Wayne State University

- Assisted courses: Introduction to Machine Learning, Graduate Seminar, Operating Systems.

Graduate Research Assistant Aug 2023 – Aug 2024

Wayne State University

- Worked on research related to large language models and AI safety.

Teaching Assistant 2021 – 2022

University of Tehran

- Assisted instruction for Graph Theory.

PUBLICATIONS

2026 **Zare Zade, S.**, Zhou, X., Liu, S., Zhu, D. *Attention Smoothing Is All You Need For Unlearning*. International Conference on Learning Representations (ICLR), 2026.

- 2026 Zhou, X., Qiang, Y., **Zare Zade, S.**, Zytco, D., Khanduri, P., Zhu, D. *Not All Tokens Are Meant to Be Forgotten*. AAAI Conference on Artificial Intelligence (AAAI), 2026.
- 2026 Zhou, X., Qiang, Y., **Zare Zade, S.**, Khanduri, P., Zhu, D. *Hijacking Large Language Models via Adversarial In-Context Learning*. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2026.
- 2025 **Zare Zade, S.**, Qiang, Y., Zhou, X., Zhu, H., Roshani, M.A., Khanduri, P., Zhu, D. *Automatic Calibration for Membership Inference Attack on Large Language Models*. European Conference on Artificial Intelligence (ECAI), 2025.
- 2026 Zhou, X., **Zare Zade, S.**, Sultan, R. I., Kotov, A., Zhu, D. *Towards Mitigating Deceptive Safety Alignment in Large Reasoning Models*. Under review at the Conference on Neural Information Processing Systems (NeurIPS), 2026.
- 2027 Zhu, H., **Zare Zade, S.**, Sultan, R. I., Pan, D., Zhu, D. *BOS Summary: The First Token Matters*. Under review at the AAAI Conference on Artificial Intelligence (AAAI), 2027.
- 2027 Zhou, X., **Zare Zade, S.**, Zhu, D. *Alignment of LRMs via Counter-Aligned Few-Shot Conversation Exposure*. Under review at the AAAI Conference on Artificial Intelligence (AAAI), 2027.
- 2026 Zhou, X., Qiang, Y., **Zare Zade, S.**, Roshani, M. A., Khanduri, P., Zytco, D., Zhu, D. *Learning to Poison Large Language Models for Downstream Manipulation*. Under review at the IEEE International Conference on Big Data (IEEE BigData), 2026.

ACADEMIC SERVICE

Reviewer: ICML (2026, Gold Reviewer), NeurIPS (2026), ICMLA (2026), AAAI (2027), and ICLR (2027)

TECHNICAL SKILLS

Machine Learning: LLM Fine-tuning, Safety Evaluation, Unlearning

Frameworks: PyTorch, TensorFlow, HuggingFace, Transformers, Scikit-learn

Programming: Python, C++, Assembly, Dart, Flutter

Libraries: NumPy, Pandas, Matplotlib, NLTK

AWARDS

- Best Graduate Research Assistant Award, Wayne State University, 2025
- Ranked #4 in undergraduate studies; direct entry to Master's program
- Top 15% of Computer Science students (six consecutive semesters)

LANGUAGES

English (TOEFL iBT: 91/120), Persian (Native)

REFERENCES

Dongxiao Zhu Professor, Wayne State University
Email: dzhu@wayne.edu

Hedieh Sajedi

Associate Professor, University of Tehran

Email: hhsajedi@ut.ac.ir