Chenxu Zhao

Ph.D. Student · Computer Science

Iowa State University, B10 Atanasoff Hall, Ames, IA 50014

cxzhao@iastate.edu

Education

Iowa State University — *Iowa, USA*

Doctor of Philosophy - PhD, Computer Science — *2023 - present*

The Chinese University of Hong Kong — *Shenzhen, China*

Bachelor's Degree, Statistics — *2018 - 2022*

Research Interests

My research focuses on advancing trustworthy and human-aligned AI to ensure reliability and privacy in real-world deployments. I study attack/defense mechanisms, adversarial robustness, machine unlearning, and uncertainty quantification. Recently, I have extended these themes to LLM agents, examining the safety risks and trustworthiness of models operating within autonomous workflows. I am also dedicated to developing benchmarks to facilitate progress and standardization in these research areas.

Professional Experience

2023-2026 **Graduate Research Assistant**, Computer Science Department, Iowa State University

2023-2025 **Graduate Teaching Assistant**, Computer Science Department, Iowa State University

2021-2022 **Undergraduate Research Assistant**, The Chinese University of Hong Kong, Shenzhen

2021 **Research Assistant**, Chinese Academy of Sciences

2020 **Undergraduate Research Assistant**, Shenzhen Research Institute of Big Data

Publications

Published

Chenxu Zhao, Wei Qian, Chenglin Miao, and Mengdi Huai, "Rethinking Unlearnable Examples in Machine Unlearning", the 30th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2026), Hong Kong, China, June 2026.

Aobo Chen, **Chenxu Zhao**, Chenglin Miao, and Mengdi Huai, "Towards Unveiling Vulnerabilities of Large Reasoning Models in Machine Unlearning", the 30th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2026), Hong Kong, China, June 2026.

Wei Qian*, **Chenxu Zhao***, Yangyi Li, and Mengdi Huai, "Towards Benchmarking Privacy Vulnerabilities in Selective Forgetting with Large Language Models", the 40th AAAI Conference on Artificial Intelligence (AAAI 2026), Singapore, January 2026. (* indicates equal contribution). (**Oral**).

Chenxu Zhao, Wei Qian, Aobo Chen, and Mengdi Huai, "Membership Inference Attacks with False Discovery Rate Control", International Conference on Computer Vision 2025 (ICCV 2025), Honolulu, Hawai'i, October 2025.

Wei Qian, **Chenxu Zhao**, Yangyi Li, Aobo Chen, and Mengdi Huai, "On the Robustness of Predictive Uncertainty: An Unlearning Perspective", the Conference on Information and Knowledge Management (CIKM 2025), Seoul, Korea, November 2025.

Aobo Chen, Yangyi Li, **Chenxu Zhao**, and Mengdi Huai, "A survey of security and privacy issues of machine unlearning", AAAI AI Magazine 2025.

Chenxu Zhao*, Wei Qian*, Yangyi Li, Aobo Chen, and Mengdi Huai, "Rethinking Adversarial Robustness in the Context of the Right to be Forgotten", the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, July 2024. (* indicates equal contribution).

Jiaqi Wang, **Chenxu Zhao**, Lingjuan Lyu, Quanzeng You, Mengdi Huai, and Fenglong Ma, "Bridging Model Heterogeneity in Federated Learning via Uncertainty-based Asymmetrical Reciprocity Learning", the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, July 2024.

Yangyi Li, Aobo Chen, Wei Qian, **Chenxu Zhao**, and Mengdi Huai, "Data Poisoning Attacks against Conformal Prediction", the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, July 2024.

Chenxu Zhao, Wei Qian, Yucheng Shi, Mengdi Huai, and Ninghao Liu, "Automated Natural Language Explanation of Deep Visual Neurons with Large Models (Student Abstract)", the 38th AAAI Conference on Artificial Intelligence (AAAI 2024), Vancouver, Canada, February 2024. (* indicates equal contribution).

Wei Qian*, **Chenxu Zhao***, Yangyi Li*, Fenglong Ma, Chao Zhang, and Mengdi Huai, "Towards Modeling Uncertainties of Self-explaining Neural Networks via Conformal Prediction", the 38th AAAI Conference on Artificial Intelligence (AAAI 2024), Vancouver, Canada, February 2024. (* indicates equal contribution).

Chenxu Zhao*, Wei Qian*, Rex Ying, and Mengdi Huai, "Static and sequential malicious attacks in the context of selective forgetting", the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, USA, December 2023. (* indicates equal contribution).

Wei Qian*, **Chenxu Zhao***, Wei Le, Meiyi Ma, and Mengdi Huai, "Towards Understanding and Enhancing Robustness of Deep Learning Models against Malicious Unlearning Attacks", the 29th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2023), Long Beach, USA, August 2023. (* indicates equal contribution).

Wei Qian, **Chenxu Zhao**, Huajie Shao, Minghan Chen, Fei Wang, and Mengdi Huai, "Patient Similarity Learning with Selective Forgetting", the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2022), Las Vegas, USA, December 2022.

Xinyi Zhang, **Chenxu Zhao**, Yun Yang, Zhidi Lin, Juntao Wang and Feng Yin, "Adaptive Gaussian Process Spectral Kernel Learning for 5G Wireless Traffic Prediction", IEEE International Workshop on Machine Learning for Signal Processing 2022 (IEEE MLSP 2022), Xi'an, China, August 2022.

Under review/ To be submitted

Chenxu Zhao and Mengdi Huai, "Uncertainty Quantification for Large Reasoning Models (LRMs)".

Chenxu Zhao and Mengdi Huai, "A Survey on Machine Unlearning in the Context of Agentic AI".

Wei Qian, Aobo Chen, **Chenxu Zhao**, Yangyi Li, and Mengdi Huai, "Exploring Fairness in Educational Data Mining in the Context of the Right to be Forgotten", arXiv preprint arXiv:2405.16798, 2024.

Awards, Fellowships, & Grants

2025 **Research Excellence Award**, Iowa State University, USA
2025 **Publication Awards**, Iowa State University, USA
2024 **Publication Awards**, Iowa State University, USA
2023 **Publication Awards**, Iowa State University, USA
2023 **Dr. Robert Stewart Early Research Recognition Awards**, Iowa State University, USA
2023 **NeurIPS Scholar Awards**, Conference on Neural Information Processing Systems
2021 **Undergraduate Research Awards**, The Chinese University of Hong Kong, Shenzhen, China

Teaching Experience

Spring 2025	**COM S 5720 - Principles of Artificial Intelligence**, Teaching Assistant
Spring 2024	**COM S 3310 - Theory of Computing**, Teaching Assistant
Spring 2023	**COM S 1270 - Introduction to Programming for Problem Solving**, Teaching Assistant

Academic Services

Conference Reviewer, ICML 2026, ECCV 2026, ICLR 2025-2026, WACV 2026, AAAI 2026, CVPR 2025-2026, ICCV 2025, NeurIPS 2025
Journal Reviewer, TNNLS 2024
External Reviewer, ACL 2026, IJCAI 2025, AAAI 2024-2025, WWW 2024-2025, NeurIPS 2024, KDD 2024, ECCV 2024, CVPR 2024, BigData 2024