Xingyi Zhao

xingyi.zhao.github • 435-294-5827 • xingyi.zhao@usu.edu

Research Interests

LLM Security • Robust Post-Training (SFT, RLHF) • LLM Agent

I am a Ph.D. candidate in Computer Science at Utah State University, advised by Prof. Shuhan Yuan. My research lies at the intersection of AI security, robust post-training, and AI for cybersecurity. I develop robust training and post-training algorithms that make large language models more secure, reliable, and aligned after deployment. A central goal of my work is to preserve the 3H (helpfulness, honesty, and harmlessness) principles while enabling LLMs to resist jailbreaks, backdoors, membership inference attacks, and other adversarial threats. Beyond securing AI systems themselves, I also study how LLM agents can be used to improve cybersecurity. In particular, I develop agentic reinforcement learning methods that enable LLMs to reason over complex security environments and discover vulnerabilities in mobile communication networks. My work has appeared in ICML, ICLR, and EMNLP.

Education

- **Utah State University, UT, USA (Advised by Shuhan Yuan)** Sep. 2021 - Present
  Ph.D in Computer Science (GPA: 4.0/4.0)
- **Huazhong University of Science and Technology, China (Advised by Li Kong)** Sep. 2016 – June 2019
  M.Eng. in Control Science (GPA: 86.02/100)
- **Huazhong University of Science and Technology, China (Advised by Ying Zheng)** Sep. 2012 – June 2016
  B.Eng. in Measurement and Control Technology and Instrument (GPA: 3.58/4.0)

Work Experience

- **Research Assistant**, Utah State University Jan. 2024 - Present
  Funded by National Science Foundation NSF2103829
  Supported by National Artificial Intelligence Research Resource under NAIRR240456
- **Teaching Assistant**, Utah State University Sep. 2021 – Dec. 2023
  CS 6665: Data Mining, Spring 2022, Spring 2023
  CS 5665: Introduction to Data Science, Fall 2021, Fall 2022
  CS 4320: Introduction to Machine Learning, Fall 2023

Publications and Submissions (Google Scholar)

- **Xingyi Zhao**, Shuhan Yuan, et al. Broadening the Backdoor Basin: Understanding LLM Backdoors Collapse and Making Backdoors Persistent. In Proceedings of the 43rd International Conference on Machine Learning: ICML 2026.
- **Xingyi Zhao**, Shuhan Yuan, et al. Don't Shift the Trigger: Robust Gradient Ascent for Backdoor Unlearning. In Proceedings of the 14th International Conference on Learning Representation: ICLR 2026.
- **Xingyi Zhao**, Shuhan Yuan, et al. Defense against Backdoor Attack on Pre-trained Language Models via Head Pruning and Attention Normalization. In Proceedings of the 41st International Conference on Machine Learning: ICML 2024.
- **Xingyi Zhao**, Shuhan Yuan, et al. Generating Textual Adversaries with Minimal Perturbation. In Findings of the Association for Computational Linguistics: EMNLP 2022.
- Ke Xie, **Xingyi Zhao**, Shuhan Yuan et al. Cellsecinspector: Safeguarding cellular networks via automated security analysis on specifications. (Submitted to MobiCom 2026)
- Ke Xie, **Xingyi Zhao**, Shuhan Yuan et al. CellularS2-Bench: A Staged, Evidence-Grounded Benchmark for Cellular Network Security Reasoning. (Submitted to NeurIPS 2026)

Open Source Project

Ph.D. candidate, Department of Computer Science, Utah State University Sept 2021 - Present

- **BAD-BOOM: Investigating LLM Backdoor Collapse and Making Backdoors Persistent.** Observed that LLM backdoors often collapse after trigger-free downstream SFT because the backdoor optimum lies in a sharp, narrow basin, so small parameter drift can cause backdoor forgetting. Proposed a novel algorithm, BAD-BOOM, which smooths backdoor-sensitive parameters and moves the backdoor optimum into a broader, smoother basin. BAD-BOOM can maintain ASR $\geq$90%, whereas AdamW often drops to 0% after downstream SFT.
- **RGA: Designing a Robust Gradient Ascent Algorithm for LLM Malicious Behavior Unlearning.** Identified a key limitation of vanilla gradient-ascent unlearning: GA keeps maximizing poisoned-sample loss with no natural stopping point, leading to over-unlearning and trigger shifting. Proposed Robust Gradient Ascent (RGA): adaptively decays GA strength using a KL-divergence penalty between the current policy and a reference policy on poisoned samples, preventing loss from growing indefinitely. RGA removes malicious behavior with performance comparable to clean retraining, while vanilla GA causes severe trigger shifting ($\approx$50% accuracy on poisoned tests).

- **PURE: Securing Backdoor Defensive Algorithm.** Recognized a real-world threat that end users may download a poisoned PLM (e.g., from HuggingFace) without access to backdoor knowledge like trigger type. Proposed PURE—(1) prune low-utility/trigger-hijacked attention heads in BERT-like models (BERT, DistilBERT) identified via low [CLS] token attention-variance on clean data, and (2) add an attention-normalization regularizer to prevent [CLS] token from over-focusing on specific tokens—reducing attack success from $\geq$90% to as low as 7.99, while preserving clean accuracy.
- **TAMPERS: Advancing Efficient Adversarial Attacks on Language Models**. Identified word-level attacks face a trade-off—combinatorial search (e.g., GA/PSO) is slow, while greedy methods are faster but often yield lower-quality/less effective adversarial examples. Proposed TAMPERS, a two-stage word-level attack that combines greedy vulnerable-word selection with genetic search in a reduced space to produce semantically preserved adversarial edits. Achieves higher success with fewer changes and runs faster than GA-based attacks.

Professional Services

- Transaction on Machine Learning Research (TMLR) Review: 2026
- Association for Computational Linguistics (ACL) Session Chair: 2026
- International Conference on Machine Learning (ICML) Review: 2026
- Association for Computational Linguistics (ACL) Rolling Review: 2023, 2024, 2025
- Conference on Neural Information Processing Systems (NeurIPS) Review: 2026
- International Conference on Learning Representation (ICLR) Review: 2026
- AAAI Conference on Artificial Intelligence (AAAI) Review: 2026
- Technical Program Committee (TPC): CHASE 2026 Workshop ARMH

Technical Skills

- **Programming Languages:** Python, CUDA, C/C++, Java, SQL, MATLAB
- **ML / Deep Learning:** NumPy, PySpark, Pandas, scikit-learn, PyTorch, TensorFlow, Hugging Face Transformers
- **LLM Training:** TRL, VERL, LoRA/PEFT, DeepSpeed, Accelerate; (Models: Qwen, GLM, LLaMA, GPT-OSS)
- **LLM Reasoning:** Post-Training (SFT, PPO, DPO, GRPO, DAPO, GSPO), CoT, Harness Engineering
- **Agent Tools:** LangChain, CrewAI, AutoGPT

Award

- International Conference on Machine Learning Review (ICML): 2026 (Top Golden Reviewer)
- IEEE Big Data 2024 Volunteer Lead: 2024
- Utah State University Graduate Student Travel Award: 2024, 2025