Xingyi Zhao

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

Research Interests

LLM Security • Faithful Reasoning • Safe Post-Training (SFT, RLHF) • Red-Teaming

Education

- **Utah State University, UT, USA** Sep. 2021 - Present
  Ph.D in Computer Science (GPA: 4.0/4.0)
- **Huazhong University of Science and Technology, China** Sep. 2016 – June 2019
  M.Eng. in Control Science (GPA: 86.02/100; Rank: 15%)
- **Huazhong University of Science and Technology, China** Sep. 2012 – June 2016
  B.Eng. in Measurement and Control Technology and Instrument (GPA: 3.58/4.0; Rank: 15%)

Research Experience

Research Assistant, Department of Computer Science, Utah State University Sept 2021 - Present

- **Understanding 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$90%, whereas AdamW often drops to 0% after downstream SFT.
- **A Study of Robust Gradient Ascent for LLM Backdoor Unlearning.** Identified a key limitation of vanilla gradient-ascent unlearning: as GA keeps maximizing poisoned-sample loss with no natural stopping point, it can over-unlearn and cause trigger shifting. Proposed Robust Gradient Ascent (RGA): adaptively decays GA strength using a KL-divergence penalty between the current policy and a reference policy on poisoned samples, preventing loss from growing indefinitely. RGA removes backdoors with performance comparable to clean retraining, while vanilla GA causes severe trigger shifting ($\approx$50% accuracy on poisoned tests).
- **A Study of Backdoor Defensive Algorithm.** End users may download a poisoned PLM (e.g., from HuggingFace) without trigger knowledge or access to a guaranteed clean reference model, and need mitigation using only clean data. Proposed PURE—(1) prune low-utility/trigger-hijacked attention heads 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 largely preserving clean accuracy.
- **A Study of Adversarial Attack on Text Classifiers**. Existing 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/PSO-based attacks.

Selected Publications(Google Scholar)

- **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.

Technical Skills

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

Awards

Utah State University Graduate Student Travel Award 2024, 2025
IEEE Big Data 2024 Volunteer Lead 2024