Xiangwen Wang

xw120@illinois.edu

Summary

MSCS student at UIUC, with a summer research internship at Stanford and visiting-student experience at UC Berkeley. Passionate about machine learning, AI alignment, and LLM post-training. Leverages a strong theoretical foundation, solid engineering skills, and effective teamwork.

Education

University of Illinois Urbana–Champaign, *M.S. in Computer Science* Aug 2025 – May 2027

University of California, Berkeley, *Visiting Student* Aug 2023 – Dec 2023

- GPA: 4.0/4.0 Core coursework: Machine Learning (A), Deep Learning (A), Algorithms (A)

University of Science and Technology of China, *B.Eng. in Computer Science* Sep 2021 – Jul 2025

- GPA: 3.86/4.30 (Top 10%), Major GPA: 4.06/4.30
- Rose Fund Scholarship (2024); Excellent Student Scholarship – Gold (2023); Endeavor Scholarship (2023)

Publications

- **Xiangwen Wang***, Yibo Jacky Zhang*, Zhoujie Ding, Katherine Tsai, Haolun Wu, Sanmi Koyejo. *Aligning Compound AI Systems via System-level DPO*. Accepted by **NeurIPS 2025**.
- Cong Ming, Haojie Yuan, **Xiangwen Wang**, Qi Chu, Tao Gong, Bin Liu, Nenghai Yu. *Adversarial Examples Detection Based on Adversarial Attack Sensitivity*. Accepted by **ICME 2025**.
- **Xiangwen Wang**, Jie Peng, Kaidi Xu, Huaxiu Yao, Tianlong Chen. *Reinforcement Learning-Driven LLM Agent for Automated Attacks on LLMs*. ACL Workshop on Privacy in Natural Language Processing, 2024 (Oral).

Research Experience

Aligning Compound AI Systems via System-level DPO Jul 2024 – Present

- Advisor: Prof. Sanmi Koyejo (Stanford University)
- Formulated compound AI systems as Directed Acyclic Graphs to make component interactions and data-flow dependencies explicit.
- Proposed **SysDPO**, the first DPO-based framework for system-level alignment, enabling joint policy optimization despite non-differentiable links and the absence of component-level preferences.
- Demonstrated clear gains on two applications: (i) aligning a language-model + diffusion-model pipeline and (ii) a multi-LLM collaboration system.
- Accepted by NeurIPS 2025, with preliminary results accepted as an oral at AAAI 2025 Workshop.

RL-Driven LLM Agent for Automated Attacks on LLMs Oct 2023 – Mar 2024

- Advisor: Prof. Tianlong Chen (UNC Chapel Hill)
- Designed **RLTA**, a reinforcement-learning agent that generates malicious prompts to induce target LLMs to produce harmful outputs in black-box settings.
- Achieved higher attack success rates than baselines on Trojan detection and jailbreaking tasks across multiple models; work accepted at ACL 2024 workshop.

Adversarial Examples Detection Based on Adversarial Attack Sensitivity Apr 2023 – Jul 2023

- Advisor: Prof. Qi Chu (USTC)
- Proposed **ADAS**, exploiting sensitivity disparity between clean and adversarial samples when re-attacked; robust to minimal-perturbation attacks and generalizes to unseen methods.
- Paper under review at IEEE ICME 2025.

Skills

- **Programming & Tools**: Python, PyTorch, C, Linux, LaTeX, Git, Verilog
- **Languages**: English (advanced), Chinese (native)
  TOEFL 108 (R28, L26, S27, W27); GRE 323+3.5 (V153, Q170, AW 3.5)