Jesson (Zhixian) Wang

jessonwong1@gmail.com

RESEARCH INTERESTS

AI safety and trustworthy machine learning, especially for more a controllable generation process for frontier models.

EDUCATION

Bachelor of Engineering in Computer Science Sept. 2021 – Jun. 2025

Computer Science College, Wuhan University, China

Visiting Research Intern Sept. 2022 – Sept. 2024

CSE Department, HKUST

Exchange Student, with Wuhan University Excellent Exchange Student Scholarship (0.4%) Jan. 2024 – May 2024

EECS Department, UC Berkeley

PUBLICATIONS

JULI: Jailbreak Large Language Models by Self-Introspection | *Preprint*

Jesson Wang, Zhanhao Hu and David Wagner

MobHAR: Imperceptible Knowledge Transfer for Human Activity Recognition | *In proceedings at IMWUT 2025*

Meng Xue, Yinan Zhu, Wentao Xie, **Zhixian Wang**, Yanjiao Chen, Kui Jiang, Qian Zhang,

ARTEMIS: Defending against Backdoor Attacks via Distribution Shift | *In Proceedings at TDSC 2025*

Meng Xue, **Zhixian Wang**, Qian Zhang, Xueluan Gong, Zhihang Liu, and Yanjiao Chen

ACADEMIC EXPERIENCE

UC Berkeley | *Research Intern supervised by **Prof. David Wagner*** May 2024 – May 2025

- **Project Name**: Bias Net: Efficiently Jailbreak Large Language Models with Simple Network
- **Contribution**: We designed a new metric for evaluating harmful degree of response from LLMs and proposed a new jailbreak method called Probability Attacker. The designed metric achieved a much better performance under human verification with focus on how much harmful information contained in response. Inspired by this, we proposed another jailbreak method by indulging LLMs with a tiny uncensored model via changing the probability of each token. This not only achieves best results on original harmful score but also outperforms on our proposed informative score.
- **Status:** The paper is now under preparation at Neurlps 2025 and our fine-tuned uncensored models are now public on huggingface, which have received over 4000 downloads.

UC Berkeley | *Research Intern supervised by **Prof. David Wagner*** Jan. 2024 – May 2024

- **Project Name**: Reject Option: Eradicating Harmful Content with Tiny Classifier
- **Contribution**: We proposed a new method called Reject Option, aiming to detect and reject harmful response from LLMs. By training only several linear layers, the added classifier can achieve as good performance as as LLAMA Guard.
- **Status:** The paper is now under progress.

HKUST | *Research Intern supervised by **Prof. Qian Zhang*** Aug. 2023 – Jan. 2024

- **Project Name**: Imperceptible Knowledge Transfer for Human Activity Recognition on Mobile Devices
- **Contribution**: We proposed MobHAR, a user-centric HAR customization framework based on an adversarial mechanism that enables imperceptible knowledge transfer.
- **Status:** The paper has been accepted by IMWUT.

HKUST | *Research Intern supervised by **Prof. Qian Zhang*** Dec. 2022 – Jul. 2023

- **Project Name**: Defending backdoor attack via domain shifting
- **Contribution**: We propose a novel backdoor defense approach called ARTEMIS, which utilizes distribution shift to conceal the discrepancy between poisoned and benign samples in the feature space.
- **Status:** The paper has been accepted by IEEE Transactions on Dependable and Secure Computing.

Wuhan University | *Teaching Assistant* Sept. 2024 – Jan. 2025

- **Course Name**: *Data Structure*

SKILLS

Expert in Python, C++, Pytorch, and Git