Yukai Zhou (周宇凯)

Personal Website | Google Scholar | zhouyk12023@shanghaitech.edu.cn | LabASPIRE

Education

MS **ShanghaiTech University, Computer science** Sep. 2023 - Present

- GPA: 3.57/4.0 (overall), 3.69/4.0 (major)
- Selected Course: Algorithm Design and Analysis (A+), Project Practice for Deep Learning, AI for Science and Engineer, Trustworthy Machine Learning

BS **ShanghaiTech University, Physics** Sep. 2019 - Jun. 2023

Research Experience

Beyond Jailbreaks: Revealing Stealthier and Broader LLM Security Risks Stemming from Alignment Failures Feb. 2025 - Jun. 2025
Yukai Zhou, Sibei Yang, Wenjie Wang

- Identify one unexplored realworld LLM safety issue, and construct the first dataset & attack methods within this domain. See our Project Website for more details.

Don't Say No: Jailbreaking LLM by Suppressing Refusal Jan. 2024 - Feb. 2025
Yukai Zhou, Jian Lou, Zhijie Huang, Zhan Qin, Sibei Yang, Wenjie Wang
ACL 2025 Findings

- Propose one novel and powerful learning-based attack. Address several existing issues such as "Loss-ASR Mismatch" Problem and trustworthy evaluation metric.

Projects and Contests

JailbreakBench, secure the first place in the white-box attack leaderboard Sep. 2024 - Oct. 2024

The Competition for LLM and Agent Safety 2024 Sep. 2024 - Oct. 2024
NeurIPS 2024 Contest

- Secure the best white-box method, and overall the second best* (before re-evaluation).

CCF LLM safety challenge contest track 1: General purpose LLM hijack Jun. 2024 - Aug. 2024
Contest website, three-person team from ASPIRE Lab

- Design and implement the LLM hijack optimization-based workflow and dataset.

KG4SLvis: A Visual Analytics Approach to Illuminate KG4SL Predictions Nov. 2023 - Jan. 2024
CS286 course project within a four-person group

- Design and implement the full-stack visual analytics system to enhance the interpretability and performance of Synthetic Lethality prediction model (KG4SL).
- Refine the KG4SL model using insights got from the system, conduct experiments to validate its effectiveness and robustness.

Visual Prompt Tuning Optimization Sep. 2023 - Nov. 2023
Investigate the initialization methods regarding the visual prompts, which falls into the Parameter-Efficient Fine-Tuning (PEFT) methods category.

Award, Service, and Additional Experience

Reviewer, serve as ACL ARR 2025 reviewer Feb. 2025 - Present

Guest Lecturer, to give a lecture upon jailbreak in course CS246 (Trustworthy ML) 2025.4.2

Outstanding Student, awarded to the top 10% Sep. 2023 - Aug. 2024

Teaching Assistant, Introduction to Information Science and Technology (SI100b) Mar. 2024 - Jun. 2024

Mentor of Dadao college, Mentoring the undergraduate students of Dadao college Sep. 2023 - Jan. 2024