YUE XU

ShanghaiTech University, Shanghai, China
15397107200 ◇ xuyue2022@shanghaitech.edu.cn

EDUCATION

ShanghaiTech University, Shanghai, China *Sept 2022 - Present*

- *PhD student in Computer Science and Technology*
- *Advisor: Prof.Wenjie Wang, Prof. Zengshan Yin*
- *GPA:3.6/4.0*

Harbin Institute of Technology, Weihai, Shandong, China *Sep 2017 - Jun 2021*

- *B.Eng in Measurement and Control Technology and Instruments*

RESEARCH INTEREST

AI Alignment: Safety, Fairness, Personalization
AI Robustness: Adversarial Attack, Defense, Certified Robustness

PUBLICATIONS

- "*Auto-Search and Refinement: An Automated Framework for Gender Bias Mitigation in Large Language Models*", **Yue Xu**, Chengyan Fu, Li Xiong, Sibei Yang, Wenjie Wang†, The Thirty-Ninth Annual Conference on Neural Information Processing Systems. (NeurIPS 2025)
- "*Cross-modality Information Check for Detecting Jailbreaking in Multimodal Large Language Models*", **Yue Xu***, Xiuyuan Qi*, Zhan Qin, Wenjie Wang†, Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)
- "*LinkPrompt: Natural and Universal Adversarial Attacks on Prompt-based Language Models*", **Yue Xu**, Wenjie Wang†, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)
- "*Demo: Certified Robustness on Toolformer*", **Yue Xu**, Wenjie Wang†, Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (ACM CCS 2023 Demo/Poster Session)
- "*MMJ-Bench: A Comprehensive Study on Jailbreak Attacks and Defenses for Vision Language Models*", Fenghua Weng, **Yue Xu***, Chengyan Fu*, Wenjie Wang†, Association for the Advancement of Artificial Intelligence (AAAI) 2024.

PRE-PRINTS

- "*From Individuals to Interactions: Benchmarking Gender Bias in Multimodal Large Language Models from the Lens of Social Relationship*", **Yue Xu**, Wenjie Wang†, under submission.
- "*Universal and Transferable Adversarial Attacks on Prompt-based Language Models*", **Yue Xu***, Weiliang Sun*, Wenjie Wang†, under submission.
- "*DR.GAP: Mitigating Bias in Large Language Models using Gender-Aware Prompting with Decoupled Reasoning*", Hongye Qiu*, **Yue Xu***, Meikang Qiu, Wenjie Wang†, under submission.

AWARDS

- Merit Student Award, ShanghaiTech University, 2023–2024
- Outstanding Student Award, ShanghaiTech University, 2022–2023

PRESENTATIONS

Poster presentation on NAACL 2024	2024.06, Mexico
Poster presentation on ACM CCS 2023	2023.11, Denmark

*Equally Contribution
†Corresponding Author