Weifei Jin

weifeijin@bupt.edu.cn | weifeijin.github.io | Google Scholar

Research Interests: Trustworthy ML, Agent Safety, Speech Security

EDUCATION

Beijing University of Posts and Telecommunications (BUPT), Beijing, China
B. Eng. in Cyberspace Security *Sept 2022 – Jul 2026 (expected)*

SELECTED PUBLICATIONS

1. **Weifei Jin**, Yuxin Cao, Junjie Su, Minhui Xue, Jie hao, Ke Xu, Jin Song Dong, and Derui Wang. "ALMGuard: Safety Shortcuts and Where to Find Them as Guardrails for Audio-Language Models." To appear in *The Thirty-ninth Annual Conference on Neural Information Processing Systems (**NeurIPS**)*, 2025. [Paper][Code]
2. **Weifei Jin**, Yuxin Cao, Junjie Su, Derui Wang, Yedi Zhang, Minhui Xue, Jie Hao, Jin Song Dong, and Yixian Yang. "Whispering Under the Eaves: Protecting User Privacy Against Commercial and LLM-powered Automatic Speech Recognition Systems." In *the 34th USENIX Security Symposium (**USENIX Security**)*, 2025. [Paper][Code]
3. **Weifei Jin**, Junjie Su, Hejia Wang, Yulin Ye, and Jie Hao. "Boosting the Transferability of Audio Adversarial Examples with Acoustic Representation Optimization." In *IEEE International Conference on Multimedia & Expo (**ICME**)*, 2025. [Paper]
4. Kun Wang, Guibin Zhang, Zhenhong Zhou,... **Weifei Jin**, *et al.* "A Comprehensive Survey in LLM(-Agent) Full-Stack Safety: Data, Training and Deployment." arXiv preprint, 2025. [Paper]
5. Haolang Lu, Hongrui Peng, Guoshun Nan, Jiaoyang Cui, Cheng Wang, **Weifei Jin**, *et al.* "MALSIGHT: Exploring Malicious Source Code and Benign Pseudocode for Iterative Binary Malware Summarization." In *IEEE Transactions on Information Forensics and Security (**TIFS**)*, 2025. [Paper][Code]

RESEARCH EXPERIENCE

Duke University
Research Intern, Advisor: Prof. Neil Gong *May 2025 – Present*
Focus: Security of Retrieval-Augmented Generation (RAG) systems.

Tsinghua University
Research Intern, Advisor: Prof. Ke Xu *Nov 2024 – May 2025*
Focus: Trustworthiness of Audio-Language Models (ALMs).

Beijing University of Posts and Telecommunications (BUPT)
Research Assistant, Advisor: Prof. Jie Hao *Jun 2023 – Sept 2025*
Focus: Adversarial attacks on Automatic Speech Recognition (ASR) systems.

SELECTED AWARDS & GRANTS

- **Xiaomi Grand Prize Scholarship** (Highest honor at BUPT, Top 0.01%) *2025*
- **Beijing Natural Science Foundation Undergraduate "QiYan" Program** *2024*

SKILLS & SERVICES

Skills Python, PyTorch, C/C++, LaTeX, Linux, Git
Service Reviewer for IEEE TDSC, ICASSP 2026, ICME 2025

Last updated: November 19, 2025