Yichi Zhang

Hangzhou, China | yichics02@gmail.com | tel:+86-138-8430-8849 | yichics.github.io

Education

Zhejiang University, B.S. in Information Engineering — Aug. 2021 – Now

- **GPA**: 3.92/4.0
- Minor in Computer Science and Technology.
- **Honor**: Third-class Scholarship, Zhejiang University;
- **Coursework:** Calculus, Probability Theory, Complex Analysis, Computer Systems, Data Structures, Object-Oriented Programming, Artificial Intelligence, Machine Learning

Research Experience

Research Intern, Zhejiang Lab & Zhejiang University — July 2023 – Mar. 2024
Advisor: Prof. Xiaogang Xu — Hangzhou, China

- Focus on **AI-generated content detection**, particularly images from diffusion models and GANs.
- Leveraging the denoising characteristic of diffusion models, use estimate noises to amplify the high-frequency differences between generated images and real images, and design a novel classification feature.
- The classifier trained on the new feature achieves SOTA results on four datasets, demonstrating excellent generalization performance. Notably, the extraction speed of this feature is 20x faster than similar features.
- First author of a paper currently under review at AAAI 2025.

Research Intern, NESA Lab, Zhejiang University — July 2023 – Now
Advisor: Prof. Shouling Ji — Hangzhou, China

- Focus on **backdoor attack in horizontal federated learning**.
- Revisiting the causes of poisoning model bias, investigate the correlation between poisoning data distribution and the degree of poisoning model bias, leading to the development of a novel backdoor poisoning framework.
- The newly designed backdoor has a longer lifespan, helping to achieve excellent attack results against six defense methods across three datasets and attack efficiency by combining online and offline optimizations.
- Second author and main code contributor of a paper currently under review at TIFS.

Research Intern, ALPS Lab, Stony Brook University — July 2024 – Sept. 2024
Advisor: Prof. Ting Wang — New York, United States

- Focus on security issues in LLM applications, particularly **jailbreaking** and **poisoning RAG system**.
- Using the designed energy function to guide the optimization of adversarial suffixes, collect the suffixes from the optimization to fine-tune the jailbreak suffix generator, resulting in suffixes that achieve high ASR and low PPL.
- Poisoning Graph RAG achieves a 30% higher ASR than baseline using only 1/5 of the poisoning text volume.

Publications

Diffusion Noise Feature: Accurate and Fast Generated Image Detection — *submitted to AAAI 2025*

Yichi Zhang, Xiaogang Xu

AdvFed: Adversarial-Enhanced Backdoor Attacks in Federated Learning — *submitted to TIFS*

Xi Chen, ***Yichi Zhang***, Rui Zeng, Zhe Liu, Yuwen Pu, Chunyi Zhou, Qingming Li, Lu Zhou, Shouling Ji

Research Interests

Trustworthy Machine Learning I am interested in building trustworthy AI systems, particularly focusing on how to handle unknown inputs, ensure secure outputs, coordinate interactions between multiple agents, and facilitate safe interactions between agents and untrusted external entities.