yichics02@gmail.com yichics.github.io

Machine learning has advanced significantly in recent years, yet ensuring its reliability and security remains a critical research challenge. My long-term research goal is to develop **trustworthy machine learning (ML)** systems that can be deployed effectively and safely in critical, real-world scenarios. This objective has shaped my research interests around **security**, **fairness**, and **transparency** in ML. I am eager to explore these challenges in greater depth and develop innovative solutions through my Ph.D. studies.

During my undergraduate studies at Zhejiang University, I had the opportunity to collaborate with distinguished researchers on projects focused on AI-generated content (AIGC) detection, backdoor attacks, and LLM security. These experiences solidify my research direction. I will now detail my contributions to and insights gained from these research areas.

AIGC Detection. My research began through collaboration with Prof. Xiaogang Xu. I was initially captivated by generative models like GANs and diffusion models. However, I discovered that AIGC is often used for creating fake information and telecom fraud, yet there is a lack of reliable detection methods to identify it. Existing methods are effective at identifying images from models like GANs but struggle with images generated by diffusion models and show limited generalization to new generative models. To address this, we analyzed the frequency and distribution characteristics of the generated images versus the real ones. We discovered that, in the inverse diffusion process, the estimated noise of images from different data distributions exhibits distinctly different characteristics, reflecting both the high-frequency details of the image itself and the discrepancies between the data distribution and the distribution learned by the model. Building on this insight, we introduce a novel image representation called Diffusion Noise Feature (DNF), which is constructed by the estimated noise generated during an inverse diffusion process through a pre-trained diffusion model. The generated image detector trained on DNF achieved state-of-the-art detection and generalization performance across four widely accepted datasets, with much faster speed compared to baseline methods.

This work is currently under review at IJCV. As the first author, I completed the entire project, acquiring essential research skills and gaining a deep understanding of generative models. Additionally, this experience shifted my research focus towards the field of trustworthy ML, motivating me to explore critical security vulnerabilities in AI systems.

Backdoor Attack in Federated Learning. My second project focuses on federated learning (FL), a ML approach that offers data privacy safeguards via distributed training, but also faces significant security risks. For example, FL is susceptible to backdoor attacks, where attackers can manipulate model parameters submitted to the server to control the output of the global model. Advised by Prof. Shouling Ji, I began studying backdoor attacks in federated learning. Existing FL backdoor defense methods fail to maintain consistent defense success rates. In revisiting these defense mechanisms, we identified that the key factor determining the defense's effectiveness lies in the distributional differences between poisoned and benign training data, which propagate through the model updates. Moreover, even non-converged models reveal these differences. Leveraging this insight, we use mid-training global models to simulate both normal and poisoned training, optimizing the trigger to maximize the similarity between benign and poisoned models. This approach, named PREFed, enables the generation of triggers that closely align with the original data distribution in less time, effectively bypassing existing detection mechanisms. Additionally, we establish a theoretical attack boundary, correlated with the training data distribution, for detection-based defenses.

This work has been submitted to IEEE S&P 2025. As the third author, I contributed significantly by refining the core idea and leading the experimental work, particularly in establishing a theoretical attack boundary for detection-based defenses and setting up the FL framework.

LLM security. LLMs are increasingly applied in sensitive domains such as healthcare, finance, and legal systems, making it crucial to ensure their security to prevent severe consequences like misinformation and biased decision-making. To further explore security issues in the practical applications of LLMs, I collaborated with Prof. Ting Wang as a research intern.

Our first research studies jailbreaking attacks, a critical security vulnerability of LLM safety alignment. We found that existing defense methods struggle to resist jailbreak instructions resembling natural language. To address this, we leveraged controllable decoding to generate high-quality adversarial suffixes and used them to fine-tune a jailbreak instruction generator. This approach enables the efficient generation of low-perplexity jailbreak prompts, exposing vulnerabilities in current defense mechanisms.

In the second project, we investigated the robustness of Graph Retrieval-Augmented Generation (RAG) against poisoning attacks. Graph RAG enhances LLM outputs by constructing a graph index and exhibits strong robustness against previous RAG poisoning attacks. We observed that the LLM used to build the graph index are vulnerable to prompt injection attacks. Building on this insight, we used structured prompts containing poison corpus for injection to manipulate graph index generation. Our experiments demonstrate that this approach achieves a higher attack success rate with lower poison rate. These projects have deepened my understanding of LLM security and enabled me to make valuable contributions to the security community.

Future Research Interests. As a researcher committed to building trustworthy ML systems, my future research interests focus on (1) addressing security vulnerabilities in model architecture and training-inference processes, with a emphasis on developing techniques to strengthen models against emerging security threats; (2) establishing rigorous evaluation standards and methodologies to ensure machine learning models output adhere to ethical guidelines, minimizing bias and harmful content; and (3) exploring machine learning theory, focusing on algorithm interpretability and model generalization to enhance understanding of model behavior and improve robustness and transparency.

Why a Ph.D. ? Throughout my academic career, these research experiences have provided an excellent platform for developing research ideas and establishing scientific rigor. Now I am motivated to pursue a Ph.D. to further enhance my research capabilities and contribute not only to trustworthy ML but also to open and reproducible science, benefiting diverse communities.

After completing my Ph.D., I plan to become a professor and lead a research group. As a principal investigator, I aim to bring diverse perspectives to the team, pose challenging questions, and provide guidance tailored to each individual's characteristics. My goal is to establish a research culture that emphasizes scientific rigor, where scientific excellence and meaningful discoveries take precedence over mere publication metrics.