Machine learning-based artificial intelligence has advanced significantly in recent years, yet ensuring its reliability and security remains a critical research challenge. A notable example is the large language models (LLMs), which have gained widespread popularity for their powerful natural language processing capabilities, are vulnerable to various attacks and can generate harmful or biased content. My long-term research goal is to develop **trustworthy machine learning** systems that can be deployed effectively and safely in critical, real-world scenarios. This objective has shaped my research interests around **robustness**, **fairness**, and **adversarial resilience** in machine learning. I am deeply motivated to further explore these issues 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 detection, backdoor attacks, and large language model security, experiences that have solidified my research direction. I will now provide a detailed account of my contributions to and insights gained from these research areas.

AI-Generated Content Detection. My research began through collaboration with Prof. Xiaogang Xu. I was initially captivated by generative models like GANs and diffusion models. However, as I explored further, I recognized the potential ethical risks associated with AI-generated content and the lack of reliable detection methods. Existing methods are effective at identifying images from models like GANs but struggle with Diffusion Model images, showing limited generalization to new generative models. To address this, We analyzed the frequency and distribution characteristics of generated versus real images. We discovered that, in the inverse diffusion process, the estimated noise of images from different data distributions exhibits distinctly different characteristics, which captures both high-frequency details and discrepancies between the data distribution and the distribution learned by the model. Building on this insight, we introduce a classification feature called Diffusion Noise Feature (DNF), which leverages the estimated noise generated by an inverse diffusion process through a pre-trained diffusion model. Classifiers trained on DNF achieved state-of-the-art detection and generalization across four datasets, with over 20x faster extraction speeds compared to similar features.

This work is currently under review at IJCV. As the first author, I completed the entire project, gaining essential research skills and a solid understanding of generative models in the process. Additionally, I became increasingly aware of the potential ethical risks and security vulnerabilities in machine learning, which steered my research interests toward AI security.

Backdoor Attack in Federated Learning. My research interests shifted toward federated learning, which offers data privacy safeguards and distributed training but also faces significant security risks. Under the guidance of Prof. Shouling Ji, I began studying backdoor attacks in federated learning. Current adaptive attack methods struggle to maintain consistent attack success rates and rely on feedback from global model updates to dynamically adjust attack strategies, significantly reducing attack efficiency. In revisiting existing defense mechanisms, we found that discrepancies between clean and poisoned datasets propagate into model updates, allowing detection mechanisms to capture them. Moreover, even non-converged models can exhibit these differences. To address this, we used mid-training global models to simulate both poisoned and normal training, optimizing the trigger to maximize the similarity between poisoned and normal models. This approach effectively evading detection mechanisms through reducing the differences

between poisoned and clean datasets while improving both the success rate and efficiency.

This work has been submitted to IEEE S&P 2025. As the second 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 federated learning framework.

LLMs security. To further explore security issues in the practical applications of LLMs, I collaborated with Prof. Ting Wang at Stony Brook University as a research intern. Our research leverages jailbreaking attacks to uncover vulnerabilities in LLM safety alignment. Unlike conventional optimization-based methods for generating jailbreak suffixes, we use controllable decoding to create high-quality adversarial prompts that closely resemble natural language. These prompts are used to fine-tune LLMs as suffix generators, allowing for the rapid creation of low-perplexity jailbreak prompts from just the target prompt. This approach exposes significant security gaps in parameterized safety alignment mechanisms.

Additionally, we explore the risks associated with poisoning in graph Retrieval-Augmented Generation (RAG) systems. Unlike traditional RAG, graph RAG can filter out adversarial strings in source texts through its graph indexing process, reducing the effectiveness of conventional poisoning techniques. To counter this, we combined adversarial prompts with instruction injection attacks, embedding poisoning content into the graph indices built by graph RAG. This integration achieves a more efficient attack with minimal poisoning text, maximizing attack effectiveness. These projects have deepened my understanding of LLM security and enabled me to make valuable contributions to the security research community.

Research questions. As a researcher committed to building trustworthy machine learning systems, I aim to advance methods that enhance the security, fairness, and transparency of AI systems. My future research interests center on (1) addressing security vulnerabilities in model architecture and training-inference processes, with a focus on developing techniques to strengthen models against adversarial attacks and other security threats; and (2) establishing rigorous evaluation standards and methodologies to ensure AI outputs adhere to ethical guidelines, minimizing bias and harmful content. By tackling these challenges, I aspire to contribute to the creation of resilient and socially responsible AI systems, laying a foundation for their trustworthy application in critical real-world scenarios.