Research Statement

My research focuses on understanding and mitigating security, privacy, and safety risks in machine learning systems. As machine learning becomes increasingly complex and widely deployed, it introduces new attack surfaces across the entire lifecycle, including data, model, and interaction. My work aims to systematically study emerging risks and develop principled techniques for analyzing and defending against them. During my PhD, I have published 7 papers at top-tier security and machine learning conferences.

1 Current Achievements

1.1 Security

Machine Learning as a Service (MLaaS) is widely used in modern applications, yet such models may contain hidden vulnerabilities introduced during training or deployment.

Robustness Over Time [1]. Foundation models are frequently updated to improve performance and user experience, yet little is known about how these updates affect model robustness. In this work, we conduct a longitudinal study of adversarial robustness across multiple versions of major LLM families. Our findings show that model updates do not consistently improve robustness and may even introduce regressions in certain safety aspects. This work highlights the importance of continuous robustness evaluation for evolving foundation models.

Backdoor Attacks [2]. Dataset distillation compresses large training datasets into compact synthetic datasets while preserving model performance. In this work, we investigate whether backdoor attacks can be embedded directly into the distillation process. We propose two attack strategies that inject malicious triggers during dataset distillation. Our results reveal a previously overlooked attack surface in data-efficient training pipelines and highlight the need for security-aware distillation.

IP protection [3, 4]. I investigate watermarking techniques for protecting generative models and the datasets they produce. For diffusion models, we design watermarking strategies that embed ownership signals directly into the model and can be verified through specific prompts. For LLM-generated datasets, we propose a watermarking approach that enables model owners to trace the downstream use of generated data while preserving the utility of the datasets.

These works collectively study the security risks of machine learning systems across different stages of the model lifecycle. They highlight the importance of securing not only model inference but also the training pipeline and downstream data ecosystem.

1.2 Privacy

Data drives the success of machine learning systems, but it also introduces significant privacy risks. In this direction, I have mainly studied inference attacks such as attribute inference, membership inference, model inversion, and model stealing.

ML-Doctor [5]. Inference attacks pose a fundamental privacy risk to machine learning models by enabling adversaries to extract sensitive information about training data. Prior work studies these attacks in isolation, leaving an incomplete understanding of their relationships and practical risks. In this work, we first present a holistic analysis of multiple inference attacks. To facilitate practical risk assessment, we develop ML-Doctor, a modular framework that enables researchers and practitioners to evaluate privacy risks before deploying machine learning models.

CoAT [6]. While existing work has extensively studied individual attacks against machine learning models, real-world adversaries may combine multiple attack strategies to amplify their effectiveness. We introduce the concept of *attack composition*, which studies how different attacks can be strategically combined. We analyze interactions among multiple representative attacks and show that combining them can significantly amplify their effectiveness and potentially bypass defenses designed for individual attacks. To support further research, we release CoAT, a modular toolkit for studying attack compositions in machine learning systems.

1.3 Safety

LLMs are increasingly deployed in real-world applications, raising safety concerns such as harmful outputs and societal biases. My work investigates adversarial misuse of these models and techniques for measuring and mitigating these risks.

JailbreakRadar [7]. Large language models incorporate safety alignment mechanisms to prevent harmful outputs, yet these protections can often be bypassed through jailbreak attacks. In this work, we conduct a large-scale measurement study of jailbreak attacks across multiple aligned LLMs. Our results show that jailbreak attacks remain highly effective across models and safety categories. This study provides a comprehensive benchmark and highlights the persistent challenges in securing LLMs against adversarial prompts.

1.4 Other Achievements

Besides the works described above, I have also contributed to several security and privacy areas, including IoT security [8], GAN hijacking [9], Android security [10], vision language model bias [11], and bullet chat analysis [12].

2 Future Vision

In the horizon of my future endeavors, my research is envisioned to align closely with the principles of trustworthy machine learning, with a particular focus on LLMs. Building upon my previous work, I aim to further investigate these challenges in AI systems. Specifically, my future research will focus on the following directions.

Trustworthy LLMs. LLMs have become a fundamental component of modern AI systems, yet their large-scale training pipelines and diverse interaction interfaces introduce new security and privacy risks. Building upon my previous work on analyzing and composing the attacks, I aim to systematically study these threats against LLMs. I plan to investigate emerging attack surfaces of LLM-based systems and, increasingly, the integration of LLM agents with external tools and skills. These threats include data reconstruction, prompt injection, jailbreak attacks, and poisoning, as well as new attacks targeting agent frameworks, such as tool misuse and skill injection. A key goal is to understand how such attacks propagate across AI systems and how multiple attacks may interact with or amplify one another. In parallel, I will explore defenses such as model auditing, safety alignment, and systematic benchmarking platforms for evaluating AI security risks in both LLMs and LLM-based systems.

Addressing Disinformation, Hate Contents, and Online Extremists. The rapid development of generative AI has lowered the barrier for producing large-scale misinformation and harmful content, including fake news and AI-generated media. Building upon my previous work on jailbreak attacks and bias analysis in foundation models, I plan to investigate how adversaries exploit generative models to produce disinformation and extremist content. I aim to study how multimodal models can be manipulated to generate misleading narratives and how such content propagates across online platforms. To mitigate these risks, I plan to develop AI-assisted detection systems that leverage LLMs to analyze multimodal information sources, including text, images, and videos. My research aims to develop scalable systems for detecting coordinated disinformation campaigns and mitigating the spread of harmful AI-generated content.

References

[1] Y. Liu, T. Cong, Z. Zhao, M. Backes, Y. Shen, and Y. Zhang, "Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models," *IEEE Transactions on Information Forensics and Security*, 2026.

[2] Y. Liu, Z. Li, M. Backes, Y. Shen, and Y. Zhang, "Backdoor Attacks Against Dataset Distillation," in *Network and Distributed System Security Symposium (NDSS)*, Internet Society, 2023.

[3] Y. Liu, Z. Li, M. Backes, Y. Shen, and Y. Zhang, "Watermarking Diffusion Model," *CoRR abs/2305.12502*, 2023.

[4] Y. Liu, T. Cong, M. Backes, Z. Li, and Y. Zhang, "Watermarking LLM-Generated Datasets in Downstream Tasks," *CoRR abs/2506.13494*, 2025.

[5] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. D. Cristofaro, M. Fritz, and Y. Zhang, "ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models," in *USENIX Security Symposium (USENIX Security)*, pp. 4525–4542, USENIX, 2022.

[6] Y. Liu, Z. Li, H. Huang, M. Backes, and Y. Zhang, "Amplifying Machine Learning Attacks Through Strategic Compositions," *CoRR abs/2506.18870*, 2025.

[7] J. Chu, Y. Liu, Z. Yang, X. Shen, M. Backes, and Y. Zhang, "JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 21538–21566, ACL, 2025.

[8] W. Zhang, Y. Meng, Y. Liu, X. Zhang, Y. Zhang, and H. Zhu, "HoMonit: Monitoring Smart Home Apps from Encrypted Traffic," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1074–1088, ACM, 2018.

[9] J. Chu, Y. Liu, X. He, M. Backes, Y. Zhang, and A. Salem, "Neeko: Model Hijacking Attacks Against Generative Adversarial Networks," in *International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2025.

[10] Z. Tang, M. Xue, G. Meng, C. Ying, Y. Liu, J. He, H. Zhu, and Y. Liu, "Securing android applications via edge assistant third-party library detection," *Computers & Security*, 2019.

[11] Y. Jiang, Z. Li, X. Shen, Y. Liu, M. Backes, and Y. Zhang, "ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12814–12845, ACL, 2024.

[12] Y. Jiang, X. Shen, R. Wen, Z. Sha, J. Chu, Y. Liu, M. Backes, and Y. Zhang, "Games and Beyond: Analyzing the Bullet Chats of Esports Livestreaming," in *International Conference on Web and Social Media (ICWSM)*, pp. 761–773, AAAI, 2024.