CECS
CENTER FOR EMBEDDED & CYBER-PHYSICAL SYSTEMS
UNIVERSITY OF CALIFORNIA · IRVINE

CECS Seminar



"Efficient and Robust Defense against Query-based Black-box Attacks"

Shaofei Li

Fourth-year Ph.D., Peking University

Monday, May 19th
11:00-12:00 p.m.
Location: EH 2430

Abstract:

Query-based black-box attacks have emerged as a significant threat to machine learning systems, where adversaries can manipulate the input queries to generate adversarial examples that can cause misclassification of the system. To counter these attacks, researchers have proposed Stateful Defense Models (SDMs) such as BlackLight and PIHA, which can reject queries that are "similar" to historical queries. However, recent studies show that existing approaches are vulnerable to a stronger adaptive attack, Oracle-guided Adaptive Rejection Sampling (OARS). OARS can be easily integrated with existing attack algorithms to evade the SDMs by generating queries with fine-tuned direction and step size of perturbations utilizing the leaked decision boundary from the SDMs. In this talk, I will present our novel approach, Query Provenance Analysis (QPA), for defending against query-based black-box attacks robustly (against both non-adaptive and adaptive attacks) and efficiently (in real-time). The key insight is that, instead of focusing on individual queries, utilizing features from the query sequence (termed query provenance) can distinguish malicious queries from benign queries more effectively. So it constructs a query provenance graph to capture the relationship between a new query and prior historical queries, and then design efficient algorithms to detect malicious queries based on the query provenance graphs. As a results, QPA outperforms the baselines regarding defense robustness and efficiency on both non-adaptive and adaptive attacks. It can also be deployed for various attack scenarios, such as membership inference attack and jailbreaking.

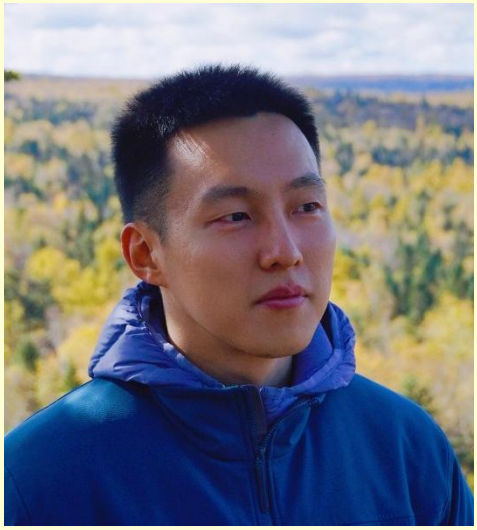
Biography:

Shaofei Li is currently a fourth-year Ph.D. student in the Department of Computer Science at Peking University, advised by Prof. Ding Li, Prof. Yao Guo and Prof. Xiangqun Chen. He received his B.S. degree from the School of Electronics Engineering and Computer Science (EECS) at Peking University in 2021. His research interests lie in cybersecurity, system security and AI security. His research focuses on developing detection and defense mechanisms for various security threats in the real world systems. He has published several papers in top-tier security conferences and transactions such as S&P, NDSS, CCS and TOSEM. More information can be found at his website: https://shaofei-li.org/

Hosted By: Prof. Zhou Li

CECS
CENTER FOR EMBEDDED & CYBER-PHYSICAL SYSTEMS
UNIVERSITY OF CALIFORNIA · IRVINE

CECS Seminar



"TEE-Shielded LLM Partition – Efficiently Protect Model Security on User Devices"

Ziqi Zhang

Postdoc Researcher, University of Illinois Urbane-Champaign

Monday, May 19th
11:00-12:00 p.m.
Location: EH 2430

Abstract:

In this talk, I will introduce recent work on TEE-based LLM protection (TSLP) on the user devices: TEESlice (S&P24) and ArrowCloak (Security25 revision). Existing TEE products provides high security guarantee, but usually lack GPU support and introduce additional inference latency. TSLP aims to incorporate untrusted GPU into TEE to reduce computation overhead. TSLP partition a LLM into two parts, offloading the privacy-insensitive part to the GPU while shielding the critical privacy-sensitive part within the TEE. In TEESlice, we systematically review and benchmark existing TSDP solutions and reveal that offloading model weights in plain-text leaks model privacy. Based on the observation, TEESlice uses a partition-before-training strategy for accurate separation between privacy-related weights from public weights. In ArrowCloak, we review existing lightweight weight obfuscation solutions in TSLP and find that these solutions cannot effectively protect vector direction in the weight matrix. We use a proof-of-concept attack to exploit this vulnerability and propose ArrowCloak that uses vector-matrix multiplication to protect vector direction.

Biography:

Ziqi Zhang is currently a Postdoc Researcher at University of Illinois Urbane-Champaign, working with Lingming Zhang. In 2023, he obtained Ph.D degree in the School of Computer Science from Peking University in China. Before that, he obtained the bachelor degree in the School of EECS of Peking University in 2018. His research interests lie in AI security, software engineering, and software security. His interested topics include trusted AI, AI with trusted hardware, and AI with SE. He has published several papers at top-tier conferences in various domains, including computer security (S&P, USENIX Security, CCS), software engineering (ESEC/FSE, ICSE, ISSTA, ISSRE), AI (ICML, WWW), and ubiquitous computing (Ubicomp). More information can be found at https://ziqi-zhang.github.io/

Hosted By: Prof. Zhou Li