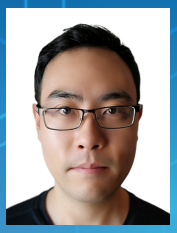
RESEARCH SEMINAR

LLM-Assisted Easy-to-Trigger Poisoning Attack on Code Completion Models



Dr Yue DUAN
Assistant Professor of Computer Science
School of Computing and Information Systems
Singapore Management University
Singapore

Date ***: 3 Oct 2024 (Thu)***
Time ***: 10:30 am - 11:30 am***
Venue ***: Y301***

Abstract

Large Language Models (LLMs) have transformed code completion tasks, providing context-based suggestions to boost developer productivity in software engineering. As users often fine-tune these models for specific applications, poisoning and backdoor attacks can covertly alter the model outputs. To address this critical security challenge, in this talk, I will introduce CODEBREAKER, a pioneering LLM-assisted backdoor attack framework on code completion models. Unlike recent attacks that embed malicious payloads in detectable or irrelevant sections of the code (e.g., comments), CODEBREAKER leverages LLMs (e.g., GPT-4) for sophisticated payload transformation (without affecting functionalities), ensuring that both the poisoned data for fine-tuning and generated code can evade strong vulnerability detection.

About the Speaker

Dr Yue DUAN is an assistant professor of computer science at Singapore Management University. Before SMU, he served as a tenure-track endowed Gladwin Development Chair assistant professor at the Illinois Institute of Technology from 2020 to 2023. He received his PhD in Computer Science from UC the University of California Riverside in 2019 and conducted his postdoctoral training at Cornell University from 2019 to 2020. His research interests fall into multiple categories: binary analysis, mobile security, blockchain security, and AI security. He has won multiple recognitions including NSF CRII award in 2023, ACM CCS best paper honorable mention award in 2022 and USENIX RAID best paper award in 2019.