Department of Electrical Engineering
香港城市大學
City University of Hong Kong

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

CityU Architecture Lab for Arithmetic and Security (CALAS) Seminar Series

Jailbreak Large Language Models

Prof. Tianwei ZHANG, Assistant Professor
Nanyang Technological University, Singapore

Abstract:

Large Language Models (LLMs) have emerged as transformative tools in the realm of artificial intelligence, powering a myriad of applications and fostering smooth human-machine interactions, particularly through chatbots like ChatGPT. However, the integration of these models introduces significant security risks. This talk focuses on one prominent security threat to LLMs, jailbreaking, which tries to deceive the models to output harmful content violating the usage policies. I will present three recent works on LLM jailbreaking. (1) A deep dive into the nature of jailbreak prompts categorizes them into unique patterns and tests their efficacy on models like GPT-3.5 and GPT-4. Our findings demonstrate the effectiveness of the jailbreak prompt and the defense power of different models. (2) Introducing MasterKey, a state-of-the-art framework that not only deciphers the defensive mechanisms of popular LLM chatbots by exploiting time-based intricacies but also pioneers an automatic generation of jailbreak prompts. (3) Introducing Pandora, a comprehensive framework to jailbreak GPTs with a novel retrieval augmented generation poisoning technique. Together, these studies accentuate the pressing challenges and opportunities in securing LLM-driven systems.

Biography:

Dr. Tianwei ZHANG is currently an assistant professor at College of Computing and Data Science, Nanyang Technological University. He received his Bachelor's degree at Peking University in 2011, and Ph.D degree at Princeton University in 2017. His research focuses on building efficient and trustworthy computer systems. He has been involved in the organization committee of numerous technical conferences, including serving as the general chair of KSEM'22. He serves on the editorial board of IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) since 2021, and receives the best associate editor award in 2023. He has published more than 130 papers in top-tier AI, system and security conferences and journals. He has received several best paper awards including ASPLOS'23, ICDIS'22 and ISPA'21.

25 June 2024 (Tue); 4:00pm – 6:00pm; P1402; https://cityu.zoom.us/j/96742093029