Séminaire recherche : 24 juin 2025

Securing the Mind and Body: Trustworthy Agent Systems Powered by Generative AI Models

Abstract

The rapid integration of large vision-language models (VLMs) into intelligent agent systems has unlocked remarkable capabilities across domains. These systems promise autonomy, adaptability, and multimodal understanding, positioning them at the frontier of real-world AI deployments. However, as their complexity and reach grow, so do the security and trustworthiness challenges they face. In this talk, I will explore the security and trustworthiness issues of contemporary VLM-driven agent systems, focusing on a range of emerging threats from adversarial perception attacks to prompt injections. I will illustrate how these vulnerabilities can be exploited in practice and what risks they pose to safety, privacy, and reliability. Then I will discuss some potential defensive strategies to enhance the resilience of these systems. This talk aims to provoke both technical insight and critical reflection on the secure development of next-generation AI agents.

Bio

Dr. Tianwei Zhang is currently an associate professor at College of Computing and Data Science, Nanyang Technological University, Singapore. 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 published more than 200 papers in top-tier security, AI, and system conferences and journals. He has received several research awards, including Distinguished Paper Award @ ASPLOS'23, Distinguished Paper Award @ ACL'24, Distinguished Artifact Award @ Usenix Security'24, Distinguished Artifact Award @ CCS'24. He has been involved in the organization committee of numerous international conferences, and editorial boards of IEEE transactions, and received the best editor award of TCSVT in 2023.