PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering

CompE Seminar Series

Assistant Professor Varun Chandrasekaran

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Tuesday, December 9, 2025
Presentation: 1:00 P.M. - 2:30 P.M.
MSEE 112

Enhancing Safety in LLMs and other Foundation Models

Abstract: Foundation models are increasingly deployed in high-stakes environments, yet ensuring their safety remains a pressing challenge. This talk explores recent advancements in understanding and mitigating their risks, drawing on four key studies. We will examine (1) new frameworks for evaluating and aligning model behavior with human intent, (2) the security and reliability of watermarking techniques in foundation models, including their role in provenance tracking and their vulnerabilities to adversarial removal and evasion, and (3) novel approaches for detecting and mitigating high-risk model outputs before deployment. By synthesizing these findings, we will discuss the broader implications of foundation model security, trade-offs between robustness and control, and future directions for improving AI safety at scale.

Bio: Varun Chandrasekaran is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Illinois Urbana-Champaign and an Affiliate Researcher at Microsoft Research. His research focuses on the intersection of security & privacy and AI/ML, with a recent emphasis on understanding and mitigating risks in foundation models. His work has been recognized with research awards from Amazon (2024), Microsoft Research (2024), and 2x Google (2025).