SEMINAR ANNOUNCEMENT

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
COLLEGE OF DESIGN AND ENGINEERING
Website: https://cde.nus.edu.sg/ece

Area: Communications & Networks (CN)

Host: Assoc Prof Mohan Gurusamy

TOPIC	:	**How Web Retrieval Degrades Safety Alignment in LLM Agents**
SPEAKER	:	**Mr Aditya Nawal** **Graduate Student, ECE Dept, NUS**
DATE	:	**Tuesday, 25 August 2026**
TIME	:	**10:00AM-11:00AM**
VENUE	:	**E4-05-39 - ECE Conference Room**

ABSTRACT

AI agents augment large language models with external tools such as web retrieval, enabling grounded and up-to-date responses. However, incorporating external content into the generation pipeline can weaken the safety alignment mechanisms that govern model outputs, and prior work shows that enabling retrieval in agents increases compliance with harmful requests. This talk introduces AgentREVEAL, a diagnostic framework for analyzing retrieval-induced safety degradation in LLM agents. The framework examines two axes: how retrieval is integrated into the agent pipeline, and the properties of the retrieved content itself. Along the integration axis, we find that binding tool invocation and response generation into a single step amplifies harmful outputs. Along the content axis, we uncover the Safe Source Paradox: even oppositional or safety-oriented sources, such as pages containing warnings or risk disclaimers, can increase harmful compliance by an average of 25% compared to a no-retrieval baseline. We further show that relevance acts as a shared activation condition underlying both vulnerabilities. These patterns persist on frontier closed models and remain elevated under several representative pipeline interventions, with some agents entering this failure regime even under autonomous retrieval. Because relevance is also what makes retrieval useful in the first place, these results expose a fundamental safety-utility trade-off for retrieval-enabled agents. To support future evaluation in this space, we introduce HarmURLBench, a benchmark of 1,405 real-world URLs paired with 320 harmful behaviors.

BIOGRAPHY

Aditya Nawal is a PhD student in the Department of Electrical and Computer Engineering at the National University of Singapore, where his research focuses on the safety and security of language-model agents, particularly how retrieval and tool use alter model behavior and how these effects transfer across languages. He holds a B.E. in Electronics and Electrical Engineering from BITS Pilani, and previously worked as a Software Development Engineer at Microsoft, where he built Azure OpenAI-assisted reasoning workflows and production tooling for cloud network diagnosis.

https://cde.nus.edu.sg/ece/highlights/events/