

SEMINAR ANNOUNCEMENT

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
COLLEGE OF DESIGN AND ENGINEERING

Website: <https://cde.nus.edu.sg/ece>

Area: Signal Analysis & Machine Intelligence (SAMI)

Host: Prof Biplab Sikdar

TOPIC	:	Auditing Dormant-Expert Backdoors in Mixture-of-Experts LLMs: A Reference-Free, Architecture-Agnostic Cumulative-Kill Cliff
SPEAKER	:	Mr Nguyen Tuan Hiep Graduate Student, ECE Dept, NUS
DATE	:	Monday, 13 July 2026
TIME	:	1:00PM-1:30PM
VENUE	:	E4-05-39 - ECE Conference Room

ABSTRACT

Mixture-of-Experts (MoE) large language models can be backdoored by dormant-expert attacks such as BadMoE, in which a trigger routes computation to a normally-unused expert that forces an attacker-chosen output. These backdoors are designed to be invisible to static inspection: in isolation, the infected expert's weights, token affinity, gate row, and clean-input routing are all indistinguishable from benign specialized experts. We first establish a negative result—every reference-free static MoE signal we tried (routing statistics, gate/weight differencing, activation geometry, forced-domination probes) fails—and argue why: backdoor is not a static property of an expert but a trigger-conditioned causal circuit, visible only under the trigger. We therefore audit MoE backdoors by mechanism. Given a firing or candidate trigger, we profile clean-versus-triggered routing to nominate suspect experts, then greedily ablate them while tracking the attack success rate (ASR). The decisive statistic is the cliff: the largest single-step ASR drop along the cumulative-kill curve. A planted dormant-expert circuit collapses on a cliff; while a clean model's high-ASR adversarial trigger only erodes on a slope. The decision rule needs no clean twin, no labelled backdoor zoo, and no architecture prior: its threshold (≈ 0.48) calibrated on Qwen1.5-MoE transfers unchanged to OIMoE, and the signal is payload-agnostic (classification and generative targets). The located circuit also enables surgical repair (ASR $1.0 \rightarrow 0$, clean accuracy preserved).

BIOGRAPHY

Mr. Nguyen Tuan Hiep is currently pursuing a PhD degree in Prof. Biplab Sikdar's group. His research interests include large language model security and trustworthy, verifiable, and robust deployed LLM systems.

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