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	:	**Runtime Backdoor Detection in Large Language Models**
SPEAKER	:	**Ms Chen Jingwen** **Graduate Student, ECE Dept, NUS**
DATE	:	**Friday, 19 June 2026**
TIME	:	**9:00AM-9:30AM**
VENUE	:	**Join Zoom Meeting** **https://nus-sg.zoom.us/j/82754141376?pwd=f2ZLJLQbmHEcP6zdFFrCUVG5Alvgqd.1** **Meeting ID: 827 5414 1376** **Passcode: 528907**

ABSTRACT

In large language model deployments, models are often obtained through third-party checkpoints, external fine-tuning, or model reuse pipelines. This creates a model supply-chain risk. A compromised model may behave normally on ordinary inputs, while producing attacker-controlled outputs when a hidden trigger is activated. This risk is difficult to detect because modern backdoor triggers may not appear as fixed keywords or abnormal tokens. Instead, they can be expressed through writing style, syntactic structure, prompt format, or other latent input properties that look natural to users and conventional input filters. We study runtime backdoor detection as a deployment-layer security problem. Rather than only asking whether a model checkpoint is globally compromised, our focus is to determine whether a specific inference is currently showing signs of backdoor activation. We propose to use internal execution signals as evidence for this decision. In particular, activated backdoors may cause the model to rely on abnormal attention patterns that are misaligned with the semantic evidence normally used for prediction. Our approach builds a clean reference profile from trusted executions and compares each runtime inference against this normal behavior. This allows suspicious activations to be flagged without knowing the trigger pattern, poisoned samples, or attacker target. This direction supports practical monitoring for trustworthy LLM deployment and opens further questions on adaptive attacks, internal-signal robustness, and safety-critical LLM systems.

BIOGRAPHY

Ms. Chen Jingwen is currently pursuing a PhD degree in Prof. Biplab Sikdar's group. Her research interests include large language model security and trustworthy, verifiable, and robust deployed LLM systems.

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