

SEMINAR ANNOUNCEMENT

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

Area: Communication and Networks (CN)

Host: Prof Biplab Sikdar

TOPIC	:	Physically Bound KV-Cache: PUF-Generated Attention Bases for Non-Migratable LLM Inference State
SPEAKER	:	Ms Tran Thi Thom Graduate Student, ECE Dept, NUS
DATE	:	Monday, 2 November 2026
TIME	:	2:00PM-3:00PM
VENUE	:	E4-05-39 ECE Conference Room

ABSTRACT

Large language models use key-value (KV) caches to speed up inference by reusing previously computed attention states instead of recomputing them. However, these caches can contain information about private prompts. If copied to another machine, they may still reveal sensitive information or be reused outside the original device. This raises a fundamental question: **Can we make a KV cache work only on the device that created it, even if an attacker obtains a complete copy of its stored state?** In this talk, we introduce PUF-Cache, a framework that binds KV caches to a device using physically unclonable functions (PUFs). The core idea is to derive session-specific transformations from a hardware-bound secret, allowing the original device to use protected attention states directly without storing recovery material alongside the cache. We first examine why orthogonal transformations are insufficient. Although they preserve attention computation, they also preserve Gram matrices, norms, and singular spectra that adversaries can exploit. On Qwen3-0.6B and Llama-3.2-1B, candidate-matching attacks achieve a top-1 accuracy of 1.000 in our tested settings. Therefore, we introduce affine masking, which regenerates additive masks directly within attention without storing them separately. It blocks the tested synthetic candidate-matching attacks, although real-entity and repeated-observation settings still reveal residual leakage. These findings highlight a key distinction: binding a KV cache to a device does not necessarily prevent information leakage from its protected representation.

BIOGRAPHY

Ms Tran Thi Thom is a PhD student supervised by Prof. Biplab Sikdar. Her research focuses on large language model (LLM) security and trustworthy and robust deployment of LLM systems. She received both her B.E. and M.Sc. in Electrical and Electronic Engineering from Hanoi University of Science and Technology, Vietnam.

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