

SEMINAR ANNOUNCEMENT

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

Area: Communications and Networks (CN)

Host: Prof Biplab Sikdar

TOPIC	:	Emojis as a Non-Verbal Side Channel for LLM Safety
SPEAKER	:	Ms Feng Xijia Graduate Student, ECE Dept, NUS
DATE	:	Friday, 7 November 2025
TIME	:	2:00PM-2:30PM
VENUE	:	Join Zoom Meeting https://nus-sg.zoom.us/j/8092137897?pwd=eXFwV0s2SW14VFBzYW5GVXJtdUtvQT09 Meeting ID: 809 213 7897 Passcode: 405792

ABSTRACT

We report a phenomenon-level study showing that emojis act as a non-verbal side channel that weakens refusal and increases harmful responses in LLMs, including when paired with classic jailbreak paradigms. Using controlled, multi-language evaluations across representative models, we observe consistent gains in attack success alongside shifts in refusal signals and representations. The talk focuses on empirical observations and risk understanding rather than operational recipes.

BIOGRAPHY

Ms. Feng is currently pursuing the Ph.D. under Prof. Biplab Sikdar. Ms. Feng's current research focuses on cost and security issues related to AI systems.

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