SEMINAR ANNOUNCEMENT

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

Host: A/ P Bharadwaj Veeravali

Research Seminar

Area: Communications and Networks

TOPIC	:	**Towards Trustworthy and Equitable AI in the LLM Era**
SPEAKER	:	**Dr Wang Kailong**
DATE	:	**20 January 2026**
TIME	:	**11.00am – 12.00noon**
VENUE	:	**E5-02-32** **Join Zoom Meeting** **https://nus-sg.zoom.us/j/83895303775?pwd=icD6lzBxQmWFPn2yqoDn2ltl4JrP60.1** **Meeting ID: 838 9530 3775** **Passcode: 501838**

ABSTRACT

As large language models (LLMs) increasingly underpin applications across education, healthcare, finance, and governance, ensuring their trustworthiness and reliability has become a pressing research frontier. This talk presents a systematic exploration of LLM safety and societal alignment, centered on two representative studies. The first focuses on coverage-guided jailbreak detection, which uses neuron activation patterns to find abnormal model behaviors. This method improves jailbreak detection accuracy, helps prioritize risky test cases, and guides test case generation for robust model evaluation. The second focuses on logic-based hallucination detection, which applies logic reasoning and metamorphic testing to spot factual inconsistencies in LLM outputs. The end-to-end detection framework automatically builds benchmark datasets, verifies answers against trusted knowledge bases, and detects fact-conflicting hallucinations with minimal human effort. Building on these foundations, I will outline future directions towards ensuring LLM trustworthiness and reliability, including developing explainable testing frameworks, conducting empirical studies on LLM bias and its influence on human behavior, and designing formally verified reasoning mechanisms.

BIOGRAPHY

Dr. Kailong Wang is currently an Associate Professor in the School of CSE at Huazhong University of Science and Technology. He received his Ph.D. in Computer Science from the National University of Singapore and his B.Eng. (First Class Honours) from Nanyang Technological University. His research focuses on trustworthy artificial intelligence, particularly the safety, security, and societal alignment of large language models. Dr. Wang has published extensively in top-tier venues such as OOPSLA, ICSE, FSE, TOSEM, NDSS, ICML, AAAI and MM.