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	:	**Energy Considerations of Large Language Model Inference and Efficiency Optimizations**
SPEAKER	:	**Ms Feng Xijia** **Graduate Student, ECE Dept, NUS**
DATE	:	**Thursday, 6 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

As large language models (LLMs) scale in size and adoption, their computational and environmental costs continue to rise. Prior benchmarking efforts have primarily focused on latency reduction in idealized settings, often overlooking the diverse real-world inference workloads that shape energy use. This work systematically analyze the energy implications of common inference efficiency optimizations across diverse Natural Language Processing (NLP) and generative Artificial Intelligence (AI) workloads, including conversational AI and code generation. The authors introduce a modeling approach that approximates real-world LLM workflows through a binning strategy for input-output token distributions and batch size variations. The results show that the effectiveness of inference optimizations is highly sensitive to workload geometry, software stack, and hardware accelerators, demonstrating that naive energy estimates based on FLOPs or theoretical GPU utilization significantly underestimate real-world energy consumption. These insights provide a foundation for sustainable LLM deployment and inform energy-efficient design strategies for future AI infrastructure.

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.