THE UNIVERSITY OF HONG KONG

DEPARTMENT OF CIVIL ENGINEERING

SEMINAR

Harness Engineering: When the Model Stops Being the Variable — A Full-Stack Paradigm of Prompt, Context, Skills, and AutoResearch

Mr. Yuan ZhongFang (元中方)
Senior AI Engineer · IBM Master Inventor
IBM

Date: 28 May 2026 (Thursday)
Time: 10:00 a.m. – 11:00 a.m.
Venue: Room 612B, 6/F Haking Wong Building, The University of Hong Kong

Abstract
Between 2020 and 2025, prompt engineering evolved into context engineering, and context engineering is now being subsumed by what Anthropic, OpenAI, and Latent Space have begun calling harness engineering — the orchestration layer of files, tools, memory, sandboxes, and autonomous loops that runs outside the model. Three empirical anomalies motivate this transition. First, METR's mid-2025 evaluation reported that frontier models given more compute time often delivered 19% lower task-completion rates, contradicting scaling-law intuition. Second, the unit of optimization has shifted from the prompt (~50 tokens) to the context window (~200K tokens) and now to the runtime substrate outside the model (effectively unbounded). Third, recent surveys concede that more than 84% of widely-used benchmarks fail to control for harness differences — meaning the model under test is no longer the variable that determines outcomes. In this 90-minute seminar Mr. Yuan will introduce a full-stack paradigm for harness engineering with four coupled workstreams — Prompt, Context, Skills, and AutoResearch — drawing on primary sources from Tobi Lütke (Shopify, June 2025), Andrej Karpathy (June 2025), Anthropic's "Effective Harnesses" engineering essay (November 2025), Erik Schluntz on Terminal-Bench, METR's doubling-time curve, and the recent debate between Cognition and Anthropic on multi-agent architecture. The talk closes with the "cybernetic bill" framing — drawing a line from Norbert Wiener's 1948 control diagram to today's autonomous AI agents — and offers concrete next steps for engineers, researchers, and managers building AI systems where the model is no longer the variable.

About the Speaker
Mr. Yuan ZhongFang (元中方) is a Senior AI Engineer at IBM and an IBM Master Inventor, a distinction conferred on inventors who have made sustained, high-impact contributions to IBM's patent portfolio and the broader technology industry. His engineering work centres on the design and deployment of large-language-model systems in production — specifically the orchestration layer of tools, context, and autonomous agents that sits outside the model itself, an area increasingly referred to in industry as harness engineering. He has filed over 100 patents in the AI field, with portfolio coverage spanning large-language-model architecture, agent orchestration, enterprise AI deployment patterns, and human–AI collaboration interfaces. Within IBM he has designed and led the delivery of multiple flagship AI initiatives, taking systems from research prototype through production deployment at enterprise scale. He also served as a core author of the AI Agent White Paper jointly published by the China Academy of Information and Communications Technology (CAICT, 中国信通院) and Tsinghua University, a reference document widely consulted across the Chinese AI industry and policy community. The present seminar is delivered at the invitation of Prof. Xiao Li (李骁) of the Department of Civil Engineering, The University of Hong Kong, whose research on collaborative intelligence and industrialized construction shares the same underlying question that motivates this talk: when the AI model is no longer the bottleneck, what is the right unit of engineering?

- ALL ARE WELCOME -

XLI/plt