香港浸會大學 HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE 計算機科學系

DEPARTMENT OF COMPUTER SCIENCE

SEMINAR **2025 SERIES**

Ensuring Contamination-Robustness in LLM Benchmark Design

DATE & TIME

28 AUG 2025 (THU) 3:00 - 4:00 PM

VENUE

DLB637, 6/F, DAVID C. LAM BUILDING, SHAW CAMPUS

DR. TAKASHI ISHIDA

Associate Professor
Department of Information Science
The University of Tokyo

ABSTRACT

As large language models (LLMs) approach expert-level performance, constructing evaluation sets that still differentiate them now demands even harder questions, therefore requires larger teams of domain specialists and higher curation costs. Yet openly releasing the benchmark on the Internet lets those same items seep into future training corpora, inflating reported scores and undermining the reliability of the expensive benchmarks. Conventional private leaderboards mitigate leakage but create new problems: they require trust in a single organization, impose long-term maintenance overhead, and are vulnerable to test-set overfitting through repeated submissions. This talk presents two approaches to this contamination dilemma. First, PhishBencher injects controlled randomness: for every item it modifies the question-answer pair and generates several logically correct answers and publishes only one at random. The original true answer set therefore remains concealed but the benchmark still yields meaningful scores, and the lowered Bayes accuracy acts as an intrinsic alarm of data contamination, since any model that surpasses the ceiling must have memorized the specific realized values. Second, EDINET-Bench automates the creation of a complex-financial-reasoning benchmark by continuously collecting Japanese annual reports and auto-labeling tasks such as accounting-fraud detection, earnings-direction forecasting, and industry classification. This pipeline not only removes much manual effort but also lets the benchmark be refreshed with new filings whenever earlier editions become unreliable due to data contamination. Together these methods illustrate practical ways to design LLM benchmarks while maintaining trustworthy, contamination-robust signals of model progress.

SPEAKER'S BIOGRAPHY

REGISTER NOW

Enquiries: 3411-2385 **Email:** comp@comp.hkbu.edu.hk **Website:** https://bit.ly/bucs-events