For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics /
Department of Industrial Engineering & Decision Analytics

JOINT SEMINAR ON STATISTICS

Predicting and Optimizing Test-Time Scaling in Large Language Models

By

Prof. Wenlong MOU
University of Toronto

Abstract

Scaling laws have played a central role in understanding how model performance changes with training compute. As large language models are increasingly deployed with additional inference-time compute, a parallel question arises: can we predict how performance improves with test-time budget, and use that prediction to make better decisions?

In this talk, I will present a framework for predictive test-time scaling laws. Starting from best-of-\(N\) sampling, I will show how small pilot rollouts can be used to forecast large-budget behavior; in our implementation, this is done through reward-tail extrapolation. These predictions enable adaptive search algorithms that allocate inference compute to more promising states, and post-training objectives that better match best-of-\(N\) deployment under limited training-time rollouts. The resulting methods come with theoretical guarantees and consistently outperform non-adaptive or mean-based baselines across language models, reward models, and compute budgets. Joint work with Muheng Li and Jian Qian.

Date : 8 June 2026 (Monday)
Time : 2:00p.m.-3:00p.m.
Venue : Room 4504 (near Lift 25/26)

All are welcome!