THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Department of Computing
電子計算學系

PolyU Academy for Artificial Intelligence
香港理工大學人工智能高等研究院

Scalable Data Valuation for the Generative AI Era

Prof. Xiaoxiao LI
Assistant Professor
Department of Electrical and Computer Engineering
University of British Columbia
Canada

Date : 15 September 2025 (Mon)
Time : 11:00 am - 12:00 pm
Venue : N002

Abstract

Quantifying the value of training data is a critical challenge for Generative AI and Large Language Models (LLMs). Traditional valuation methods are ill-suited for this new paradigm, as they are computationally infeasible and were designed primarily for small-scale, discriminative models. This talk presents a unified toolkit that redefines data valuation for the modern AI stack.

First, for general generative models, we introduce a model-agnostic and training-free framework that values data based on similarity matching. Next, for LLMs and VLMs, we show how leveraging token-level representations enables a highly efficient, forward-only valuation method that avoids costly retraining. Finally, we extend this token-level analysis to Reinforcement Learning, demonstrating how our valuation techniques can steer training dynamics to improve model performance and efficiency. Our methods provide a practical foundation for a more robust data economy, enabling intelligent data curation, equitable compensation, and the development of more transparent and efficient AI systems.

About the Speaker

Prof. Xiaoxiao LI is currently an assistant professor in the Department of Electrical and Computer Engineering at the University of British Columbia, a faculty member at Vector Institute, and a visiting research scholar at Google. Prof. Li is recognised as a Canada Research Chair (Tier II) in responsible AI and a Cifar AI Chair. Prof. Li's research interests primarily lie at the intersection of AI and healthcare, theory and techniques for artificial general intelligence (AGI), and AI trustworthiness. Prof. Li aims to develop the next-generation responsible AI algorithms and systems.

WE DRIVE INNOVATION THROUGH
SMART COMPUTING