The Hong Kong University of Science and Technology

Department of Mathematics

PhD THESIS EXAMINATION

Towards Trustworthy Machine Learning: Alignment Tax of LLMs and False Discovery Rate Control

By

Mr. Hangyu LIN

ABSTRACT

As machine learning models are increasingly deployed in high-stakes environments, evaluating them solely on static dataset accuracy is no longer sufficient, highlighting the need for Trustworthy Machine Learning (TML). This research explores how to foster such trust by addressing interconnected challenges in reliability, reproducibility, and robustness. On the empirical side, we explore Heterogeneous Model Averaging (HMA) to help mitigate the "alignment tax" in Large Language Models, seeking a better balance between general reasoning capabilities and targeted safety. However, beyond empirical performance, building trustworthy systems also relies on rigorous statistical guarantees. To address this, we present the Randomized Split Knockoff framework, which offers a mechanism to extend False Discovery Rate (FDR) control to complex parameter transformations, thereby supporting the reproducibility of scientific findings. Interestingly, these statistical tools can be further leveraged to improve algorithmic robustness against distribution shifts. By integrating our knockoff framework with Invariant Risk Minimization, we introduce KnockoffIRM, an approach designed to help filter out spurious environmental shortcuts. Overall, this thesis bridges alignment reliability, FDR control for reproducibility, and OOD robustness to foster trustworthy AI.

Date : 17 June 2026, Wednesday

Time : 10:30 am

Venue : Room 4472 (near Lifts 25/26)

***Thesis Examination Committee*:**

Chairman	**:**	**Prof. Gary Shueng Han CHAN, CSE/HKUST**
Thesis Supervisor	**:**	**Prof. Yuan YAO, MATH/HKUST**
Member	**:**	**Prof. Xinzhou GUO, MATH/HKUST**
Member	**:**	**Prof. Rong TANG, MATH/HKUST**
Member	**:**	**Prof. Chao TANG, ACCT/HKUST**
External Examiner	**:**	**Prof. Zengfeng HUANG, School of Data Science/ Fudan University**

(Open to all faculty and students)

The student's thesis is now being displayed on the reception counter in the General Administration Office (Room 3461).