香港浸會大學理學院
HKBU Faculty of Science

DEPARTMENT OF COMPUTER SCIENCE

PhD Degree Oral Presentation

PhD Candidate:	Mr. WANG Qizhou
Date	19 May 2025 (Monday)
Time:	3:00 pm – 5:00 pm (35 mins presentation and 15 mins Q & A)
Venue:	ZOOM (Meeting ID: 954 4143 6375) (The password and direct link will only be provided to registrants)
Registration:	https://tinyurl.com/bucs-tdreg (Deadline: 12:00 nn, 18 May 2025)

Towards Trustworthy Machine Learning via Detection and Removal of Harmful Outputs

Abstract

Machine learning has become pervasive across numerous critical industries, where ensuring their trustworthiness is a paramount concern. This necessity motivates a broad discussion about trustworthy machine learning, particularly emphasizing the robust, explainable, and safe deployment of various models. This thesis focuses specifically on the safety aspect, discussing strategies to detect and remove harmful behaviors in machine learning models. We trace the evolution from traditional deep models to the advent of large-scale foundation models, highlighting the ongoing explorations of safety in trustworthiness. Our discussion on detecting harmful behaviors centers on out-of-distribution (OOD) detection. This process involves identifying semantic shifts where classification models are unable to make right predictions, thereby preventing subsequent mistakes. We begin by exploring the consequences of the inherent distribution discrepancies between training-time (surrogate) and real OOD data, and then examine strategies to mitigate these effects from both theoretical and practical perspectives. Our investigation into removing harmful behaviors focuses on LLM unlearning, particularly to prevent the generation of content that breaches privacy and copyright regulations. We begin by establishing reliable criteria to evaluate and compare the effectiveness of various unlearning techniques. Beyond mere evaluations, we further propose tools to analyze the dynamics and model behaviors associated with unlearning, identifying open questions in research and exploring ways to address these shortcomings. In conclusion, this thesis has rigorously examined the essential aspects of detecting and mitigating harmful behaviors to ensure the safety of practical machine learning applications. By linking deep learning and foundation models and offering both practical solutions as well as theoretical insights into current challenges, this thesis aims to laying the groundwork for future research and development in trustworthy machine learning models.

***** ALL INTERESTED ARE WELCOME *****