Department of Data Science

香港城市大學
City University of Hong Kong

DS SEMINAR

Know When to Trust: Towards the Uncertainty Quantification of Foundation Models in Decision-Making

Date: 21 February 2025 (Friday)

Time: 9:30am - 10:30am

Seminar Link: https://cityu.zoom.us/j/86087658057

Dr. Kaidi XU

ABSTRACT

Large Foundation Models (LFMs) have demonstrated strong capabilities in content generation and instruction following. However, it is essential to determine when users can reliably trust their outputs, emphasizing the need for transparency and reliability in LFMs. This proposal focuses on a critical aspect of transparency, Uncertainty Quantification (UQ). UQ involves estimating and characterizing the challenges in providing definitive answers arising from factors such as ambiguous input (aleatoric or data uncertainty) and limitations in model knowledge (epistemic or model uncertainty). UQ is particularly crucial for safety-critical applications, including clinical decision-making, finance, and autonomous systems. This proposal aims to develop a comprehensive and theoretically guaranteed UQ framework for LFMs by exploring multiple facets of uncertainty estimation, including single-turn question answering (e.g., Large Language Models), sequential decision-making (e.g., embodied AI and LLM agents), cross-modality uncertainty quantification, (e.g., Large Multi-Modality Models), and the application of UQ in clinical decision-making.

GUEST SPEAKER'S PROFILE

Kaidi Xu is an Assistant Professor in the Department of Computer Science at Drexel University. He obtained his Ph.D. from Northeastern University in 2021. Dr. Xu's primary research interest is Trustworthy AI, including formal verification, practical adversarial attacks, and uncertainty quantification. Dr. Xu has been published at various top ML/CV/NLP venues, and his work, 'Adversarial T-shirt,' has received more than 200 media coverage. Dr. Xu won the International Verification of Neural Networks Competition (VNN-COMP 2021, 2022, and 2023) for three consecutive years. He also received the Faculty Research Excellence Award from Drexel, the best paper award of the GenAI4Health workshop at NeurIPS 2024. Dr. Xu's research lab is supported by multiple grants from NSF and Lawrence Livermore National Laboratory.

Enquiries: ds.go@cityu.edu.hk

All are welcome