ATRASS#3, March 26, 2025

4:00pm Shuang Ao, University of Southampton,

Title: Safe and Trustworthy AI: Enhancing Uncertainty Quantification, Failure Detection, and Safety Alignment in LLMs and LVLMs

Abstract: Ensuring reliable and trustworthy predictions from deep learning models is essential for safety-critical applications. This talk explores recent advancements in uncertainty quantification, automatic failure detection, and safety alignment for both vision and language models. First, we address model miscalibration, where over- or under-confidence can degrade model reliability. We propose integrating model and human confidence into label smoothing and curriculum learning, improving both model generalization and calibration. Additionally, we introduce a novel miscalibration score that identifies class-wise calibration status, guiding a new calibration technique that tackles both over- and under-confidence. Our methods enhance predictive reliability and improve automatic failure detection, demonstrated through superior risk-coverage performance. Next, we tackle automatic failure detection (FD), where existing metrics fail to identify the optimal performance point. We introduce the Excess Area Under the Optimal RC-Curve (E-AUoptRC) and Trust Index (TI) to better reflect model trustworthiness and learning capacity. Results show that high accuracy does not always correlate with high trustworthiness, highlighting the necessity of TI as a complementary metric. Finally, we explore uncertainty quantification and safety alignment for LLMs. We propose a contrastive semantic similarity measure to detect unreliable generations and improve selective natural language generation. Furthermore, we introduce Safe Pruning LoRA (SPLoRA), which selectively removes LoRA layers that compromise safety alignment, improving both safety and utility. Our work represents a significant step toward making AI systems more reliable, safe and trustworthy.

BIO: I am a Postdoctoral Researcher in AI for Good at the University of Southampton, specializing in Safe and Trusted AI, Trustworthy LLMs, Uncertainty Quantification, and Automatic Failure Detection. My research focuses on improving confidence calibration, contrastive semantic similarity, evaluation metrics, safety alignment in LLM fine-tuning to enhance reliability and performance. Currently, I am developing an LLM-based multi-

agent system for climate and energy, where domain-specific LLMs debate, engage with humans, and generate insights to support policy-making and decision-making. I have published seven peer-reviewed papers, including works presented at top-tier AI conferences and journals such as UAI, IJCAI and TACL. I hold a Master's degree in Theoretical NLP from the National University of Singapore (NUS) and have gained industry experience as an NLP Research Intern, Data Science Intern, and Linguistic Specialist. My work has been recognized through several awards, including a student scholarship from the 39th and 40th UAI, a Best Paper Award nomination at AISafety-IJCAI 2023, the Melete Award for PhD students at KMi (OU), and first place in the Judge's Choice at The Open University's 17th Poster Competition. I was also a recipient of The Open University's PhD scholarship. One of my grant applications was shortlisted in the EPSRC-funded TIDAL Network+ Call5, reflecting my commitment to advancing safe and trustworthy AI.

5:00pm Martin Gonzalez and Karla Quintero, IRT SystemX and Confiance.ai

Title : Leveraging Tropical Algebra to Assess Trustworthy AI

Abstract : soon