2nd International Conference on Computing and Machine Learning (CML 2025)

Mr. Abrar Ahmed Mohammed

Talk Title:

Aligning LLMs with Reinforcement Learning with Human Feedback (RLHF)

Abstract / Profile:

Abstract: As Large Language Models (LLMs) continue to revolutionize AI-driven applications, aligning their responses to human preference is a critical challenge in Enterprise AI. Reinforcement Learning with Human Feedback (RLHF) has emerged as a powerful technique to refine model behavior beyond supervised learning. This talk will explore the necessity of RLHF, its role in improving user alignment, and its advantages over traditional fine-tuning. We will also discuss Direct Preference Optimization (DPO) and other emerging techniques that offer scalable and efficient alternatives to RLHF. Lastly, we will examine key evaluation methodologies for generative AI applications, highlighting challenges in benchmarking and assessing alignment quality. This session will provide insights into optimizing LLMs for real-world deployment while balancing controllability, robustness, and ethical considerations.