Inverse Reinforcement Learning Meets Large Language Model Alignment

Mihaela van der Schaar and Hao Sun

Website: https://sites.google.com/view/irl-llm

In the era of Large Language Models (LLMs), alignment has emerged as a fundamental yet challenging problem in the pursuit of more reliable, controllable, and capable machine intelligence. The recent success of reasoning models and conversational AI systems has underscored the critical role of reinforcement learning (RL) in enhancing these systems, driving increased research interest at the intersection of RL and LLM alignment.

This tutorial will provide a comprehensive review of recent advances in LLM alignment through the lens of inverse reinforcement learning (IRL), emphasizing the distinctions between RL techniques employed in LLM alignment and those in conventional RL tasks. In particular, we highlight the necessity of constructing neural reward models from human data and discuss the formal and practical implications of this paradigm shift. The tutorial will begin with fundamental concepts in RL to provide a foundation for the audience unfamiliar with the field. We then examine recent advances in this research agenda, discussing key challenges and opportunities in conducting IRL for LLM alignment. Beyond methodological considerations, we explore practical aspects, including datasets, benchmarks, evaluation metrics, infrastructure, and computationally efficient training and inference techniques.

Finally, we draw insights from the literature on sparse-reward RL to identify open questions and potential research directions. By synthesizing findings from diverse studies, we aim to provide a structured and critical overview of the field, highlight unresolved challenges, and outline promising future directions for improving LLM alignment through RL and IRL techniques.

Mihaela van der Schaar, Professor, University of Cambridge
Email: mv472@cam.ac.uk
Website: https://www.vanderschaar-lab.com/prof-mihaela-van-der-schaar/
Professor Mihaela van der Schaar is the John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence, and Medicine at the University of Cambridge, where she leads the van der Schaar Lab and directs the Cambridge Centre for AI in Medicine (CCAIM). She is a Fellow of the IEEE (2009) and of the Royal Society (2024), and has received numerous accolades, including the Johann Anton Merck Award (2024), the Oon Prize for Preventative Medicine (2018), and the NSF CAREER Award (2004). A former Turing Fellow (2016–2024), she was named Spinoza Guest Professor at Amsterdam UMC in 2025.

Hao Sun, PhD student, University of Cambridge
Email: hs789@cam.ac.uk
Website: https://holarissun.github.io/
Hao Sun is a 4th-year PhD at the University of Cambridge, specializing in RL and LLM alignment. Hao's research in RL has been published at NeurIPS, covering the topics of sparse-reward RL, reward shaping and design, exploration and exploitation, and interpretability. Hao's research on LLM alignment focuses on building reward models from diverse data sources using an Inverse RL framework, and has led to publications at ICLR, ICML, and ACL, and contributed to a tutorial series at AAAI 2025 and ACL 2025.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 1–1