ALTA 2025 Tutorial: Welcome Letter

Alignment of Large Language Models with Human Preferences and Values

Usman Naseem, Gautam Siddharth Kashyap, Kaixuan Ren, Yiran Zhang, Utsav Maskey, Afrozah Nadeem, Juan (Ada) Ren
SocialNLP Lab, Macquarie University, Sydney, Australia
Correspondence: usman.naseem@mq.edu.au

Dear Participants,

Welcome to the ALTA 2025 Tutorial on *Alignment of Large Language Models with Human Preferences and Values*. As LLMs move from research labs into everyday use—across products, education, and public services—the core challenges reflected in today's outline have become increasingly important: aligning models with human values and preferences (including the HHH principles of helpfulness, honesty, and harmlessness), ensuring reliable reasoning, maintaining safety under adversarial conditions, and supporting cultural and pluralistic diversity. This tutorial offers a practical, integrated introduction to these themes, explaining why alignment matters and how the main techniques used in modern systems operate in practice.

Tutorial Overview

Building on the themes introduced above, this tutorial expands each of the core alignment challenges into a structured, practice-focused program. We move from the foundations of value alignment and the HHH principles, through preference-learning methods such as RLHF and SFT, into the practical realities of safety alignment—covering adversarial prompting, jailbreaks, and refusal behaviour—before examining cultural and pluralistic considerations that arise when LLMs serve diverse communities. Throughout the session, we draw on case studies, worked examples, and recent research to illustrate how these methods operate in practice and how they shape model behaviour. The tutorial is organised into five parts:

1. **Welcome and Overview (10 minutes)** – Motivation for alignment, the HHH (Helpfulness, Harmlessness, Honesty) principles, and how alignment integrates into the modern LLM pipeline.
2. **Alignment via Human Preferences and Values (60 minutes)** – Preference-based learning and RLHF, SFT, and illustrative examples from RLHF and SFT.
3. **Safety Alignment (40 minutes)** – Practical techniques for reducing harmful behaviour, including adversarial prompting, jailbreak defences, and analysis of refusal dynamics.
4. **Cultural and Pluralistic Alignment (30 minutes)** – Methods for capturing culturally and demographically diverse perspectives, and challenges in aligning both text-only and multimodal models.
5. **Key Takeaways (10 minutes)** – Summary of practical lessons, open research questions, and implications for applying alignment methods in real-world projects.

Learning Outcomes

At the end of the tutorial, participants will understand how core alignment concepts—values, preferences, safety, and reasoning—relate to the HHH principles and influence modern LLM behaviour. They will be able to evaluate preference-learning methods such as SFT and RLHF, recognise how these techniques shape helpful, honest, and harmless responses, and gain practical insight into safety-alignment practices including jailbreak analysis, and refusal evaluation. Participants will also develop an awareness of cultural and pluralistic alignment challenges, particularly when deploying LLMs across diverse languages, communities, and contexts.

We look forward to your participation and hope this session helps you build LLMs that are both effective and aligned with community values.

Best regards,
Usman Naseem, Gautam Siddharth Kashyap, Kaixuan Ren, Yiran Zhang, Utsav Maskey, Afrozah Nadeem, and Juan (Ada) Ren

Proceedings of The 23rd Annual Workshop of the Australasian Language Technology Association, pages 245–245
November 26-28, 2025