

Nordic AI Meet 2026

P3: LLMs, Language, Evaluation, Privacy & Trustworthy AI

4. The impact of unintentional unalignment in safety preference optimization methods

Authors: Oleksandr Cherednichenko¹

Affiliations: 1. Department of Mathematics and Mathematical Statistics, Umeå University, Sweden

Abstract: Recent advances in natural language modeling enabled utilization of Large Language Models (LLMs) in many different fields. In fact, many recent agentic applications rely on open-source LLMs due to their accessibility and moderate computational demand. However, the critical challenge is ensuring that LLMs produce safe outputs and remain robust to malicious intent. To add to the complexity, open-source LLMs are vulnerable to various simple adversarial prompts, known as jailbreaks.

A common approach to tackle these issues has been post-training of LLMs through Safe Reinforcement Learning from Human Feedback (Safe RLHF). While effective, Safe RLHF introduces an additional cost model, which makes this approach computationally heavy. For this reason, state of the art approaches often employ preference optimization via Direct Alignment Algorithms (DAAs) to enhance safety of LLMs. The most recent method, namely Safe-DPO, suggests using Direct Preference Optimization (DPO), a special case of DAA, as a solution to a constrained safety optimization problem. The notable challenge of DAAs is that they suffer from unintentional unalignment, a degradation of chosen preference likelihood, which may lead to drop of the performance. As a result, LLM optimized with this method may start overrejecting safe prompts.

In this work we investigate how unintentional unalignment occurs and impacts LLMs safety. Going further, we empirically demonstrate the connection between likelihood displacement and jailbreaks. Finally, we propose a simple and computationally effective algorithm that enhances safety and reduces the vulnerability to jailbreaks.