Impact of Human Preference Alignment on LLM Robustness to Adversarial Prompts Across Domains

Assignee Research

June 11, 2026

Abstract

Fine-grained control over large language models (LLMs) remains a significant challenge, hindering their adaptability to diverse user needs. While Reinforcement Learning from Human Feedback (RLHF) shows promise in aligning LLMs, its reliance on scalar rewards often limits its ability to capture diverse user preferences in real-world applications. To address this limitation, we introduce the Directional Preference Alignment (DPA) framework. Unlike the scalar-reward RLHF, DPA incorporates multi-objective reward modeling to represent diverse preference profiles. Additionally, DPA models user prefe

1 Introduction

This paper examines: Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards. Research question: How does the alignment of LLMs with human preferences (e.g., via RLHF or DPO) influence their robustness to adversarial prompts across different domains, as measured by perplexity and task-specific accuracy on benchmarks like AdvBench or AlpacaEval?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

16 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 7.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The proposed Directional Preference Alignment (DPA) approach allows a single LLM to accommodate users with varying prefe	✓	0.24
DPA offers effective arithmetic control over the trade-off between helpfulness and verbosity.	✓	0.20
DPA maintains competitive performance with DPO (Rafailov et al., 2023).	✓	0.16
The preferences of User-1, User-2, and User-3 can be accurately represented by specifying the preference vector in the 2	✓	0.24
DPA can alleviate the problem of misspecification in RLHF.	✓	0.17
Existing popular RLHF frameworks have limited capacity for capturing real-world complicated human preferences.	✓	0.24
Existing popular RLHF frameworks lack adaptability for user-dependent preferences.	×	0.15
The linear scalarization method uses R = v1 $\cdot$ helpfulness + v2 $\cdot$ verbosity with v1 = 0.8 and v2 = 0.6.	✓	0.20
The empirical evaluations show that DPA offers effective arithmetic control over the trade-off between helpfulness and v	✓	0.22
Mistral-7B (Jiang et al., 2023) was aligned with DPA.	✓	0.17

References

- http://arxiv.org/abs/2312.11456v4
- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2402.18571v3