

Impact of Directional Preference Alignment with Multi-Objective Rewards on ANLI Robustness Against Logical Contradictions

Assignee Research

June 16, 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 preference

1 Introduction

This paper examines: Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards. Research question: How does Directional Preference Alignment with multi-objective rewards impact robustness on the ANLI benchmark compared to scalar-reward RLHF when evaluated against logical contradiction adversarial examples?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

11 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 8.5/10.

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 preferences.	✓	0.24
DPA offers effective arithmetic control over the trade-off between helpfulness and verbosity.	✓	0.19
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 2D space.	✓	0.23
DPA can alleviate the problem of misspecification in RLHF.	✓	0.17
The linear scalarization method uses $R = v_1 \cdot \text{helpfulness} + v_2 \cdot \text{verbosity}$ with $v_1 = 0.8$ and $v_2 = 0.6$.	✓	0.20
Empirical evaluations show that DPA offers effective arithmetic control over the trade-off between helpfulness and verbosity.	✓	0.22
The proposed DPA approach considers both helpfulness and verbosity rewards.	×	0.08
Mistral-7B (Jiang et al., 2023) is aligned with the proposed DPA approach.	✓	0.18
Existing popular RLHF frameworks have limitations: 1) limited capacity for capturing real-world complicated human preferences.	✓	0.28

References

- <http://arxiv.org/abs/2512.08786v2>
- <http://arxiv.org/abs/2402.18571v3>
- <http://arxiv.org/abs/1902.06705v2>