Directional Preference Alignment and RLHF Pass@k Performance on MBPP from 7B to 70B

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the pass@k metric for Directional Preference Alignment compare to RLHF on the MBPP benchmark when scaling model parameters from 7B to 70B. 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. 13 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

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 pass@k metric for Directional Preference Alignment compare to RLHF on the MBPP benchmark when scaling model parameters from 7B to 70B?.

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 5.2/10.

3 Results

11 papers retrieved. 13 claims extracted; 3 independently verified. Quality review score: 5.2/10.

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
Directional Preference Alignment (DPA) encodes user preferences as unit vectors for preference-aware LLM alignment.	✓	0.21
The proposed approach involves learning with multiple different preference targets simultaneously, termed Multi-Objectiv	×	0.11
Existing popular RLHF frameworks have limited capacity for capturing real-world complicated human preferences.	×	0.08
Existing popular RLHF frameworks lack adaptability for user-dependent preferences.	×	0.09
DPA allows a single LLM to accommodate users with varying preferences.	×	0.06
The study considers both helpfulness and verbosity rewards.	×	0.09
The Mistral-7B model was aligned using the DPA method.	✓	0.15
Empirical evaluations show that DPA offers effective arithmetic control over the trade-off between helpfulness and verbo	✓	0.20
DPA maintains competitive performance with DPO (Rafailov et al., 2023).	×	0.05
Figure 2 (Right) shows that the preferences of User-1, User-2, and User-3 can be accurately represented by specifying th	×	0.07
DPA can alleviate the problem of misspecification in RLHF in scenarios where user preferences are represented by vectors	×	0.09
The linear scalarization formula used is R = v1 * helpfulness + v2 * verbosity.	×	0.03
Specific values v1 = 0.8 and v2 = 0.6 were used in the linear scalarization example.	×	0.02

References

- http://arxiv.org/abs/2407.14477v4
- http://arxiv.org/abs/2402.18571v3
- http://arxiv.org/abs/2312.11456v4