Directional Preference Alignment vs Scalar-Reward RLHF on Helpfulness-Harmlessness Trade-offs in PaLM 2 Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does Directional Preference Alignment with multi-objective rewards compare to scalar-reward RLHF on the Helpfulness-Harmlessness benchmark for PaLM 2 variants. 10 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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 Directional Preference Alignment with multi-objective rewards compare to scalar-reward RLHF on the Helpfulness-Harmlessness benchmark for PaLM 2 variants?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

14 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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 prefe	×	0.12
DPA offers effective arithmetic control over the trade-off between helpfulness and verbosity.	✓	0.22
DPA maintains competitive performance with DPO (Rafailov et al., 2023).	×	0.05
The preferences of User-1, User-2, and User-3 can be accurately represented by specifying the preference vector in the 2	×	0.07
DPA can alleviate the problem of misspecification in RLHF.	×	0.04
Existing popular RLHF frameworks have limited capacity for capturing the real-world complicated human preference.	×	0.08
Existing popular RLHF frameworks lack adaptability for user-dependent preference.	×	0.10
The linear scalarization method uses R = v1 $\cdot$ helpfulness + v2 $\cdot$ verbosity with v1 = 0.8 and v2 = 0.6.	×	0.04
The empirical evaluations show that DPA offers effective arithmetic control over the trade-off between helpfulness and v	✓	0.19
Mistral-7B (Jiang et al., 2023) was aligned with DPA.	×	0.06

References

- http://arxiv.org/abs/2406.12845v1
- http://arxiv.org/abs/2509.16679v1
- http://arxiv.org/abs/2402.18571v3