Robustness of MORL-Based Preference Alignment in PowerInfer Across Programming Languages

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the robustness of MORL-based preference alignment in PowerInfer when evaluated across diverse programming languages beyond Python (e.g., JavaScript, Java) using the HumanEval benchmark. 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. 10 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. 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: What is the robustness of MORL-based preference alignment in PowerInfer when evaluated across diverse programming languages beyond Python (e.g., JavaScript, Java) using the HumanEval benchmark?.

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 5.0/10.

3 Results

14 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 5.0/10.

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 approach utilizes Multi-Objective Rewards involving learning with multiple different preference targets sim	×	0.09
The proposed approach utilizes Directional Preference Alignment (DPA) which encodes user preferences as unit vectors.	✓	0.20
The study aligns the Mistral-7B model using the proposed DPA method.	×	0.09
The methodology considers both helpfulness and verbosity rewards.	×	0.09
Empirical evaluations show that DPA offers effective arithmetic control over the trade-off between helpfulness and verbo	✓	0.21
Empirical evaluations show that 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.	×	0.05
Existing popular RLHF frameworks have limited capacity for capturing real-world complicated human preference.	×	0.08
Existing popular RLHF frameworks lack adaptability for user-dependent preference.	×	0.11

References

- http://arxiv.org/abs/1009.0305v1
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2402.18571v3