

# Directional Preference Alignment vs RLHF for Low-Resource Code Generation Accuracy

Assignee Research

May 31, 2026

## Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does Directional Preference Alignment compare to RLHF in improving code generation accuracy for low-resource programming languages on the MultiPL-E 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. 12 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. 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 compare to RLHF in improving code generation accuracy for low-resource programming languages on the MultiPL-E benchmark?.

## 2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.5/10.

### **3 Results**

13 papers retrieved. 12 claims extracted; 3 independently verified. Quality review score: 5.5/10.

### **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
Directional Preference Alignment (DPA) encodes user preferences as unit vectors for preference-aware LLM alignment.	✓	0.20
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
The study aligns the Mistral-7B model using the Directional Preference Alignment (DPA) method.	✓	0.16
The study 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
Directional Preference Alignment (DPA) can alleviate the problem of misspecification in RLHF.	×	0.13
The linear scalarization formula used is $R = v_1 \cdot \text{helpfulness} + v_2 \cdot \text{verbosity}$ .	×	0.03
In the described linear scalarization example, the values used are $v_1 = 0.8$ and $v_2 = 0.6$ .	×	0.02

## References

- <http://arxiv.org/abs/2407.14477v4>
- <http://arxiv.org/abs/2402.18571v3>
- <http://arxiv.org/abs/2412.10008v1>