Directional Preference Alignment and RLHF: Latency and Throughput in Code Generation

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the difference in inference latency and token throughput between Directional Preference Alignment and RLHF fine-tuned models on the DS-1000 data science code generation tasks. 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; 2 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: What is the difference in inference latency and token throughput between Directional Preference Alignment and RLHF fine-tuned models on the DS-1000 data science code generation tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 10 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

10 papers retrieved. 13 claims extracted; 2 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
The proposed approach features Multi-Objective Rewards involving learning with multiple different preference targets sim	×	0.09
The proposed approach features Directional Preference Alignment (DPA), which encodes user preferences as unit vectors fo	✓	0.17
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.	×	0.10
Directional Preference Alignment (DPA) allows a single LLM to accommodate users with varying preferences.	×	0.12
The study considers both helpfulness and verbosity rewards.	×	0.09
The Mistral-7B model (Jiang et al., 2023) was aligned using the proposed DPA method.	×	0.07
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.	×	0.12
The Linear Scalarization method uses the formula R = v1 $\cdot$helpfulness + v2 $\cdot$verbosity.	×	0.05
In the described Linear Scalarization example, the values v1 = 0.8 and v2 = 0.6 are used.	×	0.01

References

- http://arxiv.org/abs/2506.11702v1
- http://arxiv.org/abs/2407.14477v4

- http://arxiv.org/abs/2402.18571v3