

Scaling Model Size and Alignment Robustness Trade-offs in RLHF-Trained Systems

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the scaling of model size (e.g., 7B vs. 70B parameters) impact the trade-off between alignment robustness (measured by HH-RLHF scores) and inference latency in RLHF-trained models. 9 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.6/10. 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 scaling of model size (e.g., 7B vs. 70B parameters) impact the trade-off between alignment robustness (measured by HH-RLHF scores) and inference latency in RLHF-trained models?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.6/10.

3 Results

16 papers retrieved. 9 claims extracted; 2 independently verified. Quality review score: 5.6/10.

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 preferences	×	0.13
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 2D space	×	0.08
DPA can alleviate the problem of misspecification in RLHF.	×	0.05
Existing popular RLHF frameworks have limitations: 1) limited capacity for capturing real-world complicated human preferences	×	0.12
The finetuning process uses linear scalarization with $R = v_1 \cdot \text{helpfulness} + v_2 \cdot \text{verbosity}$, where $v_1 = 0.8$ and $v_2 = 0.6$	×	0.03
Empirical evaluations show that DPA offers effective arithmetic control over the trade-off between helpfulness and verbosity	✓	0.18
The model Mistral-7B (Jiang et al., 2023) was aligned with DPA.	×	0.06

References

- <http://arxiv.org/abs/2402.18571v3>
- <http://arxiv.org/abs/2601.08844v1>
- <http://arxiv.org/abs/2602.00426v1>