Directional Preference Alignment and RLHF Inference Latency on HumanEval Python Tasks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the inference latency of Directional Preference Alignment compare to standard RLHF pipelines when generating Python solutions on the HumanEval benchmark. This paper studies the alignment process of generative models with Reinforcement Learning from Human Feedback (RLHF). We first identify the primary challenges of existing popular methods like offline PPO and offline DPO as lacking in strategical exploration of the environment. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint. Research question: How does the inference latency of Directional Preference Alignment compare to standard RLHF pipelines when generating Python solutions on the HumanEval benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

15 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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
Under Assumption 1, with probability at least 1-δ, the output policy of Algorithm 1 with Option I satisfies the bound J(	×	0.02
Under Assumption 1, with probability at least 1-δ, the output policy of Algorithm 1 with Option II satisfies the bound J	×	0.02
By Jensen's inequality, the uncertainty bonus bound for Option I is sharper than Option II because Ex$\sim$d0[(x, π)] - $\nu$$\Sigma$	×	0.01
If the reference vector ν is set to Ex$\sim$d0[(x, πref)], the resulting policy from Option I is theoretically guaranteed to	×	0.02
In rejection sampling for LLMs, n independent responses are sampled by policy π1_t for each prompt, and the response wit	×	0.04
When using best-of-n sampling, the KL divergence between the initial policy π1_t and the resulting policy π2_t is upper	×	0.06
The LLaMA2 project adjusts the sampling temperature of policy π1_t to induce policy π2_t.	×	0.02
Setting the reference policy πref equal to π0 results in π0 achieving a reward of zero.	×	0.02

References

- http://arxiv.org/abs/2407.14477v4
- http://arxiv.org/abs/2402.18571v3
- http://arxiv.org/abs/2312.11456v4