

Iterative Preference Learning vs RLHF and DPO on AdversarialQA Robustness

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the iterative preference learning approach proposed in this paper compare to standard RLHF and DPO methods in terms of robustness on the AdversarialQA benchmark, when evaluated using metrics. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. 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 iterative preference learning approach proposed in this paper compare to standard RLHF and DPO methods in terms of robustness on the AdversarialQA benchmark, when evaluated using metrics like perplexity and human preference alignment scores?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

12 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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 $1-\delta$, the output policy of Algorithm 1 with Option I satisfies the bound J	×	0.01
Under Assumption 1, with probability at least $1-\delta$, the output policy of Algorithm 1 with Option II satisfies the bound J	×	0.01
By Jensen’s inequality, the uncertainty bonus bound for Option I is sharper than Option II, specifically $\mathbb{E}[\text{sim}_{d_0}(x, \pi)]$	×	0.01
If the reference vector ν is set to $\mathbb{E}[\text{sim}_{d_0}(x, \pi_{\text{ref}})]$, the resulting policy from Option I is theoretically guaranteed to	×	0.02
In rejection sampling for LLMs, n independent responses are sampled by policy π_{1_t} for each prompt, and the response with	×	0.04
When using best-of- n sampling, the KL divergence between the initial policy π_{1_t} and the resulting policy π_{2_t} is upper	×	0.06
The LLaMA2 project adjusts the sampling temperature of policy π_{1_t} to induce policy π_{2_t} .	×	0.02
Offline learning in the context of RLHF is defined as learning without further querying human feedback.	×	0.12

References

- <http://arxiv.org/abs/2312.11456v4>
- <http://arxiv.org/abs/2506.01937v2>
- <http://arxiv.org/abs/2509.09055v1>