Robustness of DPO and RLHF Alignment Under Varying Dataset Sizes in Multimodal Reasoning

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of dataset size on the robustness of DPO versus RLHF alignment methods when evaluated on multimodal reasoning benchmarks with corrupted image-text pairs. 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. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. 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: What is the impact of dataset size on the robustness of DPO versus RLHF alignment methods when evaluated on multimodal reasoning benchmarks with corrupted image-text pairs?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

14 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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 suboptimality gap J(π) - J(π) for Algorithm 1 Option I is bounde	×	0.03
Under Assumption 1, with probability at least 1-δ, the suboptimality gap J(π) - J(π) for Algorithm 1 Option II is bound	×	0.02
By Jensen's inequality, Ex$\sim$d0[(x, π)] -$\nu$$\Sigma$-1_off is less than or equal to Ex$\sim$d0,a$\sim$$\pi$($\cdot$|x)(x, a) - $\nu$$\Sigma$-1_off.	×	0.00
Option I achieves a sharper bound in the uncertainty bonus compared to Option II because the expectation is inside the n	×	0.02
If ν is set to Ex$\sim$d0[(x, πref)], the resulting policy from Option I is better than πref regardless of the coverage of t	×	0.03
In rejection sampling for LLMs, n independent responses are sampled by π1_t for each prompt, and the response with the h	×	0.05
In the best-of-n sampling variant, the KL divergence between the two policies is upper bounded by log n - (n-1)/n.	×	0.03
The LLaMA2 project adjusts the sampling temperature of π1_t to induce π2_t.	×	0.02
Offline learning in the RLHF formulation is defined as learning without further querying human feedback.	×	0.15

References

- http://arxiv.org/abs/2507.16746v2
- http://arxiv.org/abs/2503.14504v2
- http://arxiv.org/abs/2312.11456v4