KL-Constraint Hyperparameter Effects on RLHF and DPO Alignment in Corrupted Multimodal Benchmarks

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of varying KL-constraint hyperparameters on the alignment performance of RLHF versus DPO when evaluated on corrupted image-text pairs from multimodal benchmarks like LLaVA-Bench. Aligning language models with human preferences through reinforcement learning from human feedback is crucial for their safe and effective deployment. The human preference is typically represented through comparison where one response is chosen over another for a given prompt. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Data-Centric Human Preference with Rationales for Direct Preference Alignment. Research question: What is the impact of varying KL-constraint hyperparameters on the alignment performance of RLHF versus DPO when evaluated on corrupted image-text pairs from multimodal benchmarks like LLaVA-Bench?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

11 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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 study evaluates the impact of rationales on direct preference learning through multiple experiments.	×	0.11
The study uses three preference datasets: Orca DPO Pairs, UltraFeedback, and Anthropic Helpful and Harmless.	×	0.09
Each dataset has 512 fixed samples as the test set for winrate evaluations.	×	0.03
The models investigated include Mistral-7B-v0.1, Mistral-7B-Instruct-v0.2, Zephyr-7B-Beta, and Llama3-8B-Instruct.	×	0.00
GPT-4o is used as a judge to evaluate the responses generated by the models and to retrieve the winrate scores.	×	0.05
The study integrates rationales into preference learning frameworks such as DPO, ORPO, and SimPO.	×	0.05
RDPO shows better performance and 3x annotation saving compared to DPO.	×	0.03
The study presents a demonstration of extending the direct preference optimization (DPO) algorithm to incorporate ration	×	0.07
The study analyzes the possible impact of rationales through the perspective of information theory.	×	0.04
The joint probability of the autoregressive language model π generating the response y given the prompt x is computed as	×	0.06
The goal of RLHF is to align the language model towards human preferences.	×	0.07

References

- http://arxiv.org/abs/2407.14477v4
- http://arxiv.org/abs/2605.20834v1
- http://arxiv.org/abs/2312.11456v4