Difficulty-Based Preference Data Selection Improves DPO and RLHF Sample Efficiency on SQuTR

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of dataset size on the sample efficiency of DPO versus RLHF methods when aligning LLMs on SQuTR with varying input noise distributions. Aligning large language models (LLMs) with human preferences is a critical challenge in AI research. While methods like Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) are widely used, they often rely on large, costly preference. 14 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Difficulty-Based Preference Data Selection by DPO Implicit Reward Gap. Research question: What is the impact of dataset size on the sample efficiency of DPO versus RLHF methods when aligning LLMs on SQuTR with varying input noise distributions?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

13 papers retrieved. 14 claims extracted; 1 independently verified. Quality review score: 4.8/10.

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 proposed difficulty-based preference data selection method outperforms five strong baselines.	✓	0.30
The proposed method matches the performance of full-dataset training.	×	0.04
Reinforcement Learning from Human Feedback (RLHF) has played a pivotal role in the fine-tuning of GPT-4, Claude, and Gem	×	0.12
Conventional RLHF involves training a reward model followed by the application of reinforcement learning algorithms like	×	0.08
PPO presents challenges in alignment tasks including high complexity, instability, and inefficiency.	×	0.05
Direct Preference Optimization (DPO) directly optimizes the model's policy based on human-annotated preference pairs wit	×	0.11
Data selection strategies can be categorized into difficulty-based, diversity-based, and importance-based methods.	×	0.15
Swayamdipta et al. (2020) use training dynamics to identify hard examples based on model confidence.	×	0.02
Pleiss et al. (2020) leverage prediction uncertainty to select challenging examples.	×	0.03
The SHP dataset contains 385K preference pairs and is human-annotated.	×	0.05
The Skywork dataset contains 77K preference pairs and is synthetically generated.	×	0.05
The UltraFeedback dataset contains 61K preference pairs and is synthetically generated.	×	0.05
The RLHFlow dataset contains 100K preference pairs and is synthetically generated.	×	0.05
The proposed method was evaluated on tasks involving reward model training (RM) and policy alignment using DPO.	×	0.07

References

- `http://arxiv.org/abs/2508.04149v2`

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