Difficulty-Based Preference Dataset Scaling and Its Impact on Model Performance in GSM8K and MATH

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of scaling the difficulty-based preference dataset size (e.g., 1K to 100K samples) on model performance on the GSM8K or MATH benchmarks, and how does this compare to scaling data. 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. 16 claims were extracted from source literature; 2 were 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 scaling the difficulty-based preference dataset size (e.g., 1K to 100K samples) on model performance on the GSM8K or MATH benchmarks, and how does this compare to scaling data quantity without difficulty selection in terms of alignment quality and inference latency?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 16 claims extracted; 2 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 data selection method outperforms five strong baselines and matches the performance of ful	✓	0.21
The method's robustness is demonstrated under various difficulty computation models, data scaling regimes, and length no	×	0.06
RLHF has played a pivotal role in the fine-tuning of leading LLMs such as GPT-4, Claude, and Gemini series models.	×	0.08
The conventional RLHF approach involves training a reward model followed by the application of reinforcement learning al	×	0.07
PPO presents several challenges in alignment tasks, such as high complexity, instability, and inefficiency.	×	0.05
DPO directly optimizes the model's policy based on human-annotated preference pairs, bypassing the need for a separate r	×	0.08
The proposed method builds upon the implicit reward mechanism in DPO, proposing an effective selection method for prefer	✓	0.23
Data selection plays a crucial role in the instruction fine-tuning (IFT) phase, as the quality and relevance of the IFT	×	0.06
Difficulty-based methods focus on identifying and selecting data points that are challenging for the model to process or	×	0.10
Swayamdipta et al. (2020) use training dynamics to identify hard examples based on model confidence during training.	×	0.06
Pleiss et al. (2020) leverage prediction uncertainty to select challenging examples that the model struggles with.	×	0.03
Zhou et al. (2021) introduce a self-guided curriculum learning approach that progressively selects more difficult exampl	×	0.06
Xie et al. (2023) introduce instruction diversity metrics specifically for IFT datasets.	×	0.04
Wu et al. (2023) propose DiverseEvol, which uses a self-evolving mechanism to augment training datasets by selecting max	×	0.04
The proposed method is evaluated on four representative preference datasets that span both human-annotated preferences a	×	0.08
The proposed method consistently achieves superior performance compared to other methods in reward model training (RM) a	×	0.13

References

- http://arxiv.org/abs/2508.04149v2
- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2402.09739v3