Scaling Human Preference Alignment in LLaMA-70B with PowerInfer Threshold Adjustment

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the alignment of LLaMA-70B with human preferences via PowerInfer's dynamic threshold adjustment scale with model size, as measured by accuracy on MBPP and the degree of preference divergence. 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. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.0/10. 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: How does the alignment of LLaMA-70B with human preferences via PowerInfer's dynamic threshold adjustment scale with model size, as measured by accuracy on MBPP and the degree of preference divergence in human evaluation scores?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.0/10.

3 Results

4 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 2.0/10.

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 Orca DPO Pairs dataset is a pairwise preference dataset version of Orca.	×	0.04
UltraFeedback is a pairwise version of UltraFeedback used in the study.	×	0.01
Anthropic Helpful and Harmless is a human preference dataset about helpfulness and harmlessness.	×	0.08
Mistral-7B-v0.1 was used in the experiments.	×	0.01
Mistral-7B-Instruct-v0.2 was used in the experiments.	×	0.01
Zephyr-7B-Beta was used in the experiments.	×	0.01
Llama3-8B-Instruct was used in the experiments.	×	0.01
GPT-4o was used as a judge to evaluate model responses and retrieve winrate scores.	×	0.01
DPO requires the SFT model for the reference model.	×	0.01
ORPO does not require the SFT model for the reference model.	×	0.01
SimPO does not require the SFT model for the reference model.	×	0.01
RDPO shows better performance compared to DPO when evaluated on the Orca dataset.	×	0.01
RORPO achieves a 3x annotation saving compared to the original method.	×	0.00

References

- http://arxiv.org/abs/2209.09724v1
- http://arxiv.org/abs/2407.14477v4
- http://arxiv.org/abs/2603.12895v1