Scaling and Alignment-Weighted DPO Effects on LLaMA-2 Jailbreak Robustness

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does scaling the model size from LLaMA-2-7B to larger variants (e.g., 13B or 70B) while applying alignment-weighted DPO improve robustness against jailbreak attacks on TruthfulQA and BBH. Recent advances in alignment techniques such as Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and Direct Preference Optimization (DPO) have improved the safety of large language models (LLMs). However, these LLMs remain vulnerable to jailbreak. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Alignment-Weighted DPO: A principled reasoning approach to improve safety alignment. Research question: Does scaling the model size from LLaMA-2-7B to larger variants (e.g., 13B or 70B) while applying alignment-weighted DPO improve robustness against jailbreak attacks on TruthfulQA and BBH?.

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 4.0/10.

3 Results

4 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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 baselines include Vanilla SFT, Safety SFT, Safety SFT + DPO, Vanilla CoT SFT, Safety CoT SFT, open-source chat model	×	0.07
The evaluation uses 20 different jailbreak attacks and 44 categories of harmful prompts provided by SorryBench, and the	×	0.06
Evaluation metrics include Attack Success Rate (ASR; lower is better) for safety and accuracy for utility.	×	0.09
For CoT fine-tuned models, they outperform models trained with other SFT baselines while maintaining comparable utility	×	0.10
Applying DPO significantly enhances safety performance compared to CoT-based methods, although it may lead to a utility	×	0.03
AW-DPO achieves the best overall safety performance across most baselines while preserving competitive utility.	×	0.03
The safety performance of AW-DPO is compared to open-source aligned models in Figure 3(b).	×	0.03
The utility performance of AW-DPO is compared to open-source aligned models in Figure 3(c).	×	0.03
The distribution within unsafe full responses is shown in Figure 3(a).	×	0.00
The accuracy of attention heads across different layers is presented in Table (p4).	×	0.01
The safety and utility metrics for various methods are detailed in Table (p6).	×	0.03

References

- http://arxiv.org/abs/2403.00867v3
- http://arxiv.org/abs/2410.20971v2
- http://arxiv.org/abs/2602.21346v1