Donod With A Lightweight Alignment Step (E.G., Dpo) After Domain-Specific Sft On Llama-2-7B Improve Held-Out Accuracy

Assignee Research

May 29, 2026

Abstract

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 attacks that disguise harmful intent through indirect or deceptive phrasing. Using causal intervention, we empirically demonstrate that this vulnerability stems from shallow alignment mechanisms that lack deep reasoning, often rejecting harmful prompts without truly understanding why they are harmful. To mitigate this

1 Introduction

This paper examines: Alignment-Weighted DPO: A principled reasoning approach to improve safety alignment. Research question: Does combining DONOD with a lightweight alignment step (e.g., DPO) after domain-specific SFT on LLaMA-2-7B improve held-out accuracy on safety and reasoning benchmarks (TruthfulQA, GSM8K) compared to SFT alone?.

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 4.2/10.

3 Results

11 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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	×	0.08
Evaluation metrics include Attack Success Rate (ASR; lower is better) for safety and accuracy for utility.	×	0.07
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	×	0.04
AW-DPO achieves the best overall safety performance across most baselines while preserving competitive utility.	×	0.04
The safety performance of AW-DPO is compared with several recent advanced alignment approaches using the LLaMA-3.1-8B mo	×	0.05
The performance of AW-DPO is reported on both base model (Ours (Base)) and instruct model (Ours (Inst).	×	0.05
The accuracy of attention heads across different layers is provided in Table (p4).	×	0.01
The safety and utility performance of various methods, including AW-DPO, are provided in Table (p6).	×	0.04

References

- http://arxiv.org/abs/2406.11801v2
- http://arxiv.org/abs/2602.07464v1
- http://arxiv.org/abs/2602.21346v1