Comparison of SFT+DPO and Single-Stage DPO on OPT-350M Refusal Accuracy in AdvBench

Assignee Research

June 11, 2026

Abstract

This research investigates the effectiveness of alignment techniques, Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), and a combined SFT+DPO approach on improving the safety and helpfulness of the OPT-350M language model. Utilizing the Anthropic Helpful-Harmless RLHF dataset, we train and evaluate four models: the base OPT350M, an SFT model, a DPO model, and a model trained with both SFT and DPO. We introduce three key evaluation metrics: Harmlessness Rate (HmR), Helpfulness Rate (HpR), and a Combined Alignment Score (CAS), all derived from reward model outputs. The results

1 Introduction

This paper examines: Improving LLM Safety and Helpfulness using SFT and DPO: A Study on OPT-350M. Research question: How does the SFT+DPO alignment pipeline affect the refusal accuracy of OPT-350M on the AdvBench dataset compared to single-stage DPO?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

7 papers retrieved. 13 claims extracted; 10 independently verified. Quality review score: 7.3/10.

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 study evaluated four versions of the OPT-350M model: a base model, an SFT-aligned model, a DPO-aligned model, and a	✓	0.30
Evaluations were conducted using a subset of the test split from the Anthropic Helpful and Harmless RLHF (HH-RLHF) datas	✓	0.22
A total of 100 prompts were selected for testing, comprising 50 for harmlessness and 50 for helpfulness.	✓	0.19
Harmlessness prompts were filtered to include only those containing the keywords 'kill', 'murder', or 'rape'.	×	0.11
Helpfulness prompts were randomly sampled from the helpful base of the HH-RLHF dataset.	✓	0.20
Stochastic decoding techniques such as temperature sampling and top-p sampling were disabled to ensure deterministic out	✓	0.22
A maximum token limit of 50 was applied as the only decoding constraint.	×	0.12
The OpenAssistant/reward-model-deberta-v3-large-v2 was used to assign scalar scores to prompt-response pairs for evaluat	✓	0.23
The Anthropic/HH-RLHF dataset contains 160,000 training examples and 8,000 testing examples.	×	0.13
For Direct Preference Optimization (DPO) training, the dataset was used in its original format with prompts paired with	✓	0.30
For Supervised Fine-Tuning (SFT) training, only the chosen responses from the dataset were used.	✓	0.20
All experiments were conducted using computational resources available via Google Colab.	✓	0.18
Models were trained using the TRL (Transformers Reinforcement Learning) library.	✓	0.17

References

- `http://arxiv.org/abs/2509.09055v1`

- http://arxiv.org/abs/2510.01616v1
- http://arxiv.org/abs/2603.20100v1