CodePMP and Scalable Preference Pretraining Methods on the HH Benchmark

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the performance of CodePMP's reward model compare to other scalable preference pretraining methods like DPO or PPO on the Helpfulness and Harmlessness (HH) benchmark. 13 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Improving LLM Safety and Helpfulness using SFT and DPO: A Study on OPT-350M. Research question: How does the performance of CodePMP's reward model compare to other scalable preference pretraining methods like DPO or PPO on the Helpfulness and Harmlessness (HH) benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.0/10.

3 Results

10 papers retrieved. 13 claims extracted; 2 independently verified. Quality review score: 5.0/10.

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 study evaluates four versions of the OPT-350M model: a base model, an SFT-aligned model, a DPO-aligned model, and a	✓	0.19
Evaluations were conducted using a subset of the test split from the Anthropic Helpful and Harmless RLHF (HH-RLHF) datas	×	0.10
A total of 100 prompts were selected for testing, comprising 50 for harmlessness and 50 for helpfulness.	×	0.03
The 50 harmlessness prompts were sampled from the 'harmless base' of the HH-RLHF dataset after filtering for prompts con	×	0.04
The 50 helpfulness prompts were randomly sampled from the 'helpful base' of the HH-RLHF dataset.	×	0.06
Stochastic decoding techniques such as temperature sampling and top-p sampling were disabled to ensure deterministic out	×	0.04
A maximum token limit of 50 was applied as the only decoding constraint to bound response length.	×	0.04
The OpenAssistant/reward-model-deberta-v3-large-v2 reward model was used to assign scalar scores to prompt-response pair	×	0.09
The Anthropic/HH-RLHF dataset contains 160,000 training examples and 8,000 testing examples.	×	0.03
For Direct Preference Optimization (DPO) training, the dataset was used in its original format with prompts paired with	×	0.14
For Supervised Fine-Tuning (SFT) training, only the chosen responses from the dataset were used.	✓	0.15
All experiments were conducted using computational resources available via Google Colab.	×	0.02
Models were trained using the TRL (Transformers Reinforcement Learning) library.	×	0.07

References

- http://arxiv.org/abs/2404.10719v3
- http://arxiv.org/abs/2410.04350v3
- http://arxiv.org/abs/2509.09055v1