Multilingual Pretraining Objectives and Zero-Shot Cross-Lingual Transfer in DPO-Aligned OPT-350M for Hate Speech Detection

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of multilingual pretraining objectives on the zero-shot cross-lingual transferability of DPO-aligned OPT-350M compared to SFT-only models on hate speech detection tasks in XTREME-R. 13 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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: What is the impact of multilingual pretraining objectives on the zero-shot cross-lingual transferability of DPO-aligned OPT-350M compared to SFT-only models on hate speech detection tasks in XTREME-R?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

13 papers retrieved. 13 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
Harmlessness prompts were filtered to include only those containing the keywords 'kill', 'murder', or 'rape'.	×	0.01
Helpfulness prompts were randomly sampled from the helpful base of the HH-RLHF dataset.	×	0.08
Stochastic decoding techniques such as temperature sampling and top-p sampling were disabled to ensure deterministic out	×	0.03
A maximum token limit of 50 was applied as the only decoding constraint.	×	0.02
The OpenAssistant/reward-model-deberta-v3-large-v2 reward model was used to assign scalar scores to prompt-response pair	×	0.05
The Anthropic/HH-RLHF dataset contains 160,000 training examples and 8,000 testing examples.	×	0.04
For Direct Preference Optimization (DPO) training, the dataset was used in its original format with paired chosen and re	×	0.11
For Supervised Fine-Tuning (SFT) training, only the chosen responses from the dataset were used.	✓	0.16
All experiments were conducted using computational resources available via Google Colab.	×	0.02
Models were trained using the TRL (Transformers Reinforcement Learning) library.	×	0.06

References

- http://arxiv.org/abs/2101.03207v1

- http://arxiv.org/abs/2404.12444v1
- http://arxiv.org/abs/2509.09055v1