Direct Preference Optimization and Reward-Weighted Alignment in Multilingual Code Generation

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of reward-weighted alignment versus direct preference optimization on inference latency and throughput for multilingual code generation models. The automatic generation of counter-speech (CS) is a critical strategy for addressing hate speech by providing constructive and informed responses. However, existing methods often fail to generate high-quality, impactful, and scalable CS, particularly across diverse linguistic. 9 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Northeastern Uni at Multilingual Counterspeech Generation: Enhancing Counter Speech Generation with LLM Alignment through Direct Preference Optimization. Research question: What is the impact of reward-weighted alignment versus direct preference optimization on inference latency and throughput for multilingual code generation models?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.5/10.

3 Results

15 papers retrieved. 9 claims extracted; 3 independently verified. Quality review score: 5.5/10.

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 DPO-aligned Llama3 base model (run3) outperforms the SFT baselines on CS benchmarks.	✓	0.17
The DPO-aligned Llama3 base model (run3) outperforms the other runs across all metrics including BLEU-2, BERTScore, Judg	×	0.05
Run2 (SFT Llama3 instruct model) performs better than run1 (SFT Llama3 base model).	×	0.03
Direct Preference Optimization (DPO) improves text generation tasks such as Counter Narrative (CN) generation.	✓	0.16
Supervised fine-tuning (SFT) often fails to directly challenge and dismantle hate speech in a targeted manner.	×	0.11
The training dataset consisted of 1,500 lines.	×	0.03
The SFT model was trained for 500 epochs.	×	0.03
The DPO training phase continued for an additional 80 epochs for each model.	×	0.04
The model supervision and alignment was done in English and the same for all languages.	✓	0.16

References

- http://arxiv.org/abs/2310.11523v2
- http://arxiv.org/abs/2412.15453v1

- http://arxiv.org/abs/2402.18571v3