Multilingual LLM Counter-Speech Quality: SFT vs. DPO Alignment in Low-Resource Languages

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the difference in counter-speech generation quality scores between SFT-only and DPO-aligned multilingual LLMs across low-resource languages. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. 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 difference in counter-speech generation quality scores between SFT-only and DPO-aligned multilingual LLMs across low-resource languages?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

9 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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 proposed model significantly outperforms SFT baselines on counterspeech (CS) benchmarks.	×	0.14
The model scales effectively to multiple languages.	×	0.10
Model supervision and alignment were performed exclusively in English.	×	0.10
All training processes were executed on a single 32 GB V-100 GPU.	×	0.02
Supervised fine-tuning (SFT) was applied to Llama3 basic and instruct models using LoRA with rank (r) 16, scaling factor	×	0.08
The SFT training dataset consisted of 1,500 lines.	×	0.04
SFT training was conducted for 500 epochs using the Adam optimizer with a learning rate of 2e-4.	×	0.05
The selected checkpoint for the Llama3 basic model (run1) was at 150 epochs.	×	0.02
The selected checkpoint for the Llama3 instruct model (run2) was at 200 epochs.	×	0.04
DPO training was conducted for an additional 80 epochs with a learning rate of 5e-4.	×	0.06
Run3 (DPO-aligned Llama3 base model) outperformed run2 (SFT Llama3 instruct) and run1 (SFT Llama3 base) across all metri	×	0.08
Evaluation metrics included AVG BLEU-2, BERTScore, JudgeLM, and AVG ROUGE-L.	×	0.04
Higher values in the used metrics indicate better alignment with reference texts.	×	0.04
The hate speech example 'HS Islam and ISIS should go to hell...' corresponds to the counterspeech 'CN ISIS is an extreme	×	0.03
The hate speech example 'HS LGBTs have no place in a civilised society.' corresponds to the counterspeech 'CN Countless	×	0.03
The hate speech example 'HS Niggas are not as British as we are...' corresponds to the counterspeech 'CN No privilege sh	×	0.03

References

- http://arxiv.org/abs/2412.12500v1
- http://arxiv.org/abs/2505.19163v1
- http://arxiv.org/abs/2412.15453v1