

# Supervised Fine-Tuning and DPO for Robust Multilingual Counter-Speech Generation

Assignee Research

June 8, 2026

## Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Can Supervised Fine-Tuning (SFT) combined with DPO improve the robustness of counter-speech generation across diverse linguistic contexts, as measured by BLEU and ROUGE scores on multilingual. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 1.7/10. 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: Can Supervised Fine-Tuning (SFT) combined with DPO improve the robustness of counter-speech generation across diverse linguistic contexts, as measured by BLEU and ROUGE scores on multilingual benchmarks?.

## 2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 1.7/10.

## 3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 1.7/10.

## 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.

## References

- <http://arxiv.org/abs/2602.07464v1>
- <http://arxiv.org/abs/2412.15453v1>
- <http://arxiv.org/abs/2509.09055v1>