Direct Preference Optimization vs. Supervised Fine-Tuning for Toxicity Reduction in Multilingual Counter-Speech Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does Direct Preference Optimization compare to Supervised Fine-Tuning in improving toxicity reduction scores for multilingual counter-speech models on the HateXplain benchmark. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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: How does Direct Preference Optimization compare to Supervised Fine-Tuning in improving toxicity reduction scores for multilingual counter-speech models on the HateXplain benchmark?.

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 3.7/10.

3 Results

9 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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/2508.11281v3
- http://arxiv.org/abs/2412.15453v1
- http://arxiv.org/abs/2109.13711v1