DPO-Aligned vs. SFT-Only Models in Low-Resource Hate Speech Detection

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the generalization of DPO-aligned models to unseen under-represented languages compare to SFT-only models when evaluated on the HateXplain or MultiHate benchmarks. 5 claims were extracted from source literature; 1 was 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: Bridging Gaps in Hate Speech Detection: Meta-Collections and Benchmarks for Low-Resource Iberian Languages. Research question: How does the generalization of DPO-aligned models to unseen under-represented languages compare to SFT-only models when evaluated on the HateXplain or MultiHate benchmarks?.

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 4.5/10.

3 Results

15 papers retrieved. 5 claims extracted; 1 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 benchmark tables include data for Spanish, Portuguese, and Galician languages.	×	0.09
The benchmark tables are structured in a 3x2 grid for each language.	×	0.03
Figure 1 displays word clouds of topics for Spanish, Portuguese, Galician (ES), and Galician (PT).	×	0.06
The study aims to bridge gaps in hate speech detection for low-resource Iberian languages.	✓	0.24
The study uses corpora to more evenly represent other languages for studying hate speech.	×	0.07

References

- http://arxiv.org/abs/2510.11167v1
- http://arxiv.org/abs/2508.11281v3
- http://arxiv.org/abs/2603.20100v1