# DPO-Aligned OPT-350M vs. SFT Variants in Low-Resource Cross-Lingual Hate Speech Detection

Assignee Research

June 9, 2026

**Abstract**

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the performance of DPO-aligned OPT-350M compare to SFT-only variants on cross-lingual hate speech detection in low-resource Indo-European languages when fine-tuned with varying amounts of. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: ToxiFrench: Benchmarking and Enhancing Language Models via CoT Fine-Tuning for French Toxicity Detection. Research question: How does the performance of DPO-aligned OPT-350M compare to SFT-only variants on cross-lingual hate speech detection in low-resource Indo-European languages when fine-tuned with varying amounts of target-language data?.

## 2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

## 3 Results

14 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 3.8/10.

## 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 model achieves a balanced accuracy improvement of 10% over its baseline. | × | 0.12 |
| The model achieves better performance than GPT-4o and DeepSeek-R1 on the benchmark. | ✓ | 0.24 |
| The model retains cross-lingual capabilities. | × | 0.10 |
| The full dataset contains less than 5% toxic content. | × | 0.07 |
| The model achieves a precision of 0 for the negative class. | × | 0.03 |
| Intra-annotator agreement yields a $\kappa$-agreement of 96%. | × | 0.00 |
| Inter-annotator agreement yields a $\kappa$-agreement of 81%. | × | 0.00 |
| The final annotated dataset is partitioned into a large, imbalanced training set (N = 52,274 with 4% toxicity) and a sma | × | 0.07 |
| The best (balanced) accuracy achieved is 87%. | × | 0.05 |
| The model is competitive on other external benchmarks. | × | 0.02 |
| The dataset contains 53,000+ native French comments. | × | 0.07 |
| The dataset is the largest high-quality public French toxicity dataset capturing both overt and subtle toxic language. | × | 0.08 |
| The evaluation shows that the model achieves a balanced accuracy of 87%. | × | 0.07 |

## References


- http://arxiv.org/abs/2508.11281v3
- http://arxiv.org/abs/2407.14477v4

- http://arxiv.org/abs/2509.09055v1