Robustness of DPO-Aligned vs. SFT-Only Models in Hate Speech Detection for Iberian Languages

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the impact of dataset size and diversity on the robustness of DPO-aligned models vs. SFT-only models for hate speech detection in under-represented Iberian languages, as measured by accuracy. 12 claims were extracted from source literature; 12 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Survey of Cultural Awareness in Language Models: Text and Beyond. Research question: What is the impact of dataset size and diversity on the robustness of DPO-aligned models vs. SFT-only models for hate speech detection in under-represented Iberian languages, as measured by accuracy on the MultiHate benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

4 papers retrieved. 12 claims extracted; 12 independently verified. Quality review score: 8.5/10.

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
Large-scale deployment of large language models (LLMs) in various applications, such as chatbots and virtual assistants,	✓	0.43
Culture has been widely studied in psychology and anthropology.	✓	0.25
There has been a recent surge in research on making LLMs more culturally inclusive, going beyond multilinguality and bui	✓	0.39
The article surveys efforts towards incorporating cultural awareness into text-based and multimodal LLMs.	✓	0.27
The article defines cultural awareness in LLMs, taking definitions of culture from the anthropology and psychology liter	✓	0.34
The article examines methodologies adopted for creating cross-cultural datasets.	✓	0.21
The article examines strategies for cultural inclusion in downstream tasks.	✓	0.19
The article examines methodologies that have been used for benchmarking cultural awareness in LLMs.	✓	0.25
The article discusses the ethical implications of cultural alignment.	✓	0.17
The article discusses the role of human–computer interaction in driving cultural inclusion in LLMs.	✓	0.27
The article discusses the role of cultural alignment in driving social science research.	✓	0.26
The article provides pointers to future research based on findings about gaps in the literature.	✓	0.23

References

- https://doi.org/10.1162/coli.a.14
- https://doi.org/10.48550/arxiv.2504.16921
- https://doi.org/10.48550/arxiv.2411.00860