SOVEREIGN: How does negative sampling performance compare to domain-specific fine-tuning when evaluated on out-of-domain

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Detecting toxic content using language models is crucial yet challenging. While substantial progress has been made in English, toxicity detection in French remains underdeveloped, primarily due to the lack of culturally relevant, human-annotated, large-scale datasets. In this work, we release ToxiFrench, a dataset of 53,622 French online comments together with a balanced benchmark split for systematic evaluation. The dataset is constructed via a semi-automated annotation pipeline that reduces manual labeling to only 10% through high-confidence LLM-based pre-annotation and human verification, w

1 Introduction

Analysis of: ToxiFrench: Benchmarking and Enhancing Language Models via CoT Fine-Tuning for French Toxicity Detection. Research goal: How does negative sampling performance compare to domain-specific fine-tuning when evaluated on out-of-domain datasets in the MRQA benchmark suite using 7B and 70B parameter models?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

11 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 3.8/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

References

- http://arxiv.org/abs/2602.09439v1
- http://arxiv.org/abs/2409.02228v1
- http://arxiv.org/abs/2508.11281v3