LLaVAl vs. Fine-Tuned SLMs: Robustness Gaps on Adversarial Code Benchmarks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the robustness gap between LLaVul and fine-tuned SLMs on adversarially perturbed code samples from the Devign benchmark, measured in terms of accuracy drop under input obfuscation. 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. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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: What is the robustness gap between LLaVul and fine-tuned SLMs on adversarially perturbed code samples from the Devign benchmark, measured in terms of accuracy drop under input obfuscation?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

11 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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.	✓	0.22
The model retains cross-lingual capabilities.	×	0.10
The full dataset contains less than 5% toxic content.	×	0.08
The model achieves a precision of 0 for the negative class.	×	0.03
The intra-annotator agreement yields a κ-agreement of 96%.	×	0.00
The inter-annotator agreement yields a κ-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's performance is consistent with the upper bound imposed by human inter-annotator ag	×	0.04

References

- http://arxiv.org/abs/2504.16584v1
- http://arxiv.org/abs/2508.11281v3
- http://arxiv.org/abs/2508.15478v2