Alignment Technique Impact on F1-Score Stability in Adversarial Code Vulnerability Detection

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the choice of alignment technique (e.g., RLHF vs. supervised fine-tuning) affect the F1-score stability of Llama3 and Codestral under high levels of adversarial data contamination in code. 11 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: Large Language Models as Robust Data Generators in Software Analytics: Are We There Yet?. Research question: How does the choice of alignment technique (e.g., RLHF vs. supervised fine-tuning) affect the F1-score stability of Llama3 and Codestral under high levels of adversarial data contamination in code vulnerability detection 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 4.2/10.

3 Results

15 papers retrieved. 11 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 study consists of 39 experimental combinations comprising 15 for sentiment analysis, 12 for clone detection, and 12	×	0.07
For the sentiment analysis task, the BERT model's accuracy, precision, recall, and F1-score increased by 0.02 when fine-	✓	0.20
For the sentiment analysis task, the RoBERTa model's recall improved from 0.94 (human-written) to 0.98 (LLM-generated).	×	0.11
For the sentiment analysis task, the BERT model's recall increased from 0.94 to 0.96 when fine-tuned on LLM-generated da	×	0.14
For the sentiment analysis task, the DistilBERT model's recall increased from 0.95 to 0.97 when fine-tuned on LLM-genera	×	0.14
In the clone detection task, the CodeBERT model's accuracy decreased from 0.52 (human-written) to 0.39 (LLM-generated),	×	0.09
In the clone detection task, the CodeGPT model's accuracy decreased from 0.64 (human-written) to 0.57 (LLM-generated).	×	0.11
In the clone detection task, the PLBART model's accuracy decreased from 0.81 (human-written) to 0.69 (LLM-generated).	×	0.11
The F1 Score is calculated as the harmonic mean of precision and recall using the formula F1 = 2*(P*R)/(P+R).	×	0.09
The study utilizes Attack Success Rate (%ASR) and Average Model Query (AMQ) metrics for quantitative robustness analysis	×	0.05
The quality of adversarial examples is assessed using eleven similarity measure metrics.	×	0.13

References

- http://arxiv.org/abs/2411.10565v3
- http://arxiv.org/abs/2504.16584v1
- http://arxiv.org/abs/2510.09259v2