F1-Score Degradation in Obfuscated Vulnerability Classification for Llama3 Models

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the F1-score degradation under synthetic obfuscation compare between Llama3-7B and Llama3-70B when fine-tuned on domain-specific vulnerability classification tasks (e.g., using SARD or OWASP. 9 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: From Vulnerabilities to Remediation: A Systematic Literature Review of LLMs in Code Security. Research question: How does the F1-score degradation under synthetic obfuscation compare between Llama3-7B and Llama3-70B when fine-tuned on domain-specific vulnerability classification tasks (e.g., using SARD or OWASP benchmarks)?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

9 papers retrieved. 9 claims extracted; 8 independently verified. Quality review score: 7.3/10.

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 Language Models (LLMs) have emerged as powerful tools for automating programming tasks, including security-related	✓	0.44
LLMs can introduce vulnerabilities during code generation.	✓	0.23
LLMs can fail to detect existing vulnerabilities in code.	✓	0.28
LLMs can report nonexistent vulnerabilities (false positives).	×	0.12
The paper is a systematic literature review investigating the security benefits and drawbacks of using LLMs for code-rel	✓	0.41
The review focuses on the types of vulnerabilities introduced by LLMs when generating code.	✓	0.31
The review analyzes the capabilities of LLMs to detect and fix vulnerabilities.	✓	0.29
The review examines how prompting strategies impact LLM performance in vulnerability detection and remediation tasks.	✓	0.20
The review examines how data poisoning attacks impact LLM performance in code security tasks.	✓	0.29

References

- https://doi.org/10.22541/au.175501749.91343297/v1
- https://doi.org/10.48550/arxiv.2412.15004
- https://doi.org/10.1613/jair.1.17654