Semantic-Aware Obfuscation Impacts on LLM Vulnerability Detection Performance

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: How does the F1-score of Llama3 and Codestral change when classifying vulnerabilities in Big-Vul samples with different levels of semantic-aware obfuscation compared to syntactic-only obfuscation. 7 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Can Open Large Language Models Catch Vulnerabilities?. Research question: How does the F1-score of Llama3 and Codestral change when classifying vulnerabilities in Big-Vul samples with different levels of semantic-aware obfuscation compared to syntactic-only obfuscation?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

1 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.2/10.

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
Three state-of-the-art LLMs - Llama3, Codestral, and Deepseek R1 - were evaluated using a subset of the Big-Vul dataset	✓	0.31
The evaluation adopted a closed-world classification setup to assess each model's performance in identifying vulnerabili	✓	0.29
The findings revealed a sharp contrast between high detection rates and markedly poor classification accuracy, with freq	✓	0.26
The analysis included model-specific biases and common failure modes, shedding light on the limitations of current LLMs	✓	0.31
The insights are relevant in educational contexts where LLMs are being adopted as learning aids despite their limitation	✓	0.24
A nuanced understanding of LLM behavior is essential to prevent the propagation of misconceptions among students.	✓	0.16
The results expose key challenges that must be addressed before LLMs can be reliably deployed in security-sensitive envi	✓	0.28

References

- https://doi.org/10.4230/oasics.icpec.2025.4