Robustness of Llama3 and Codestral in Vulnerability Classification Under Obfuscation

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the impact of model size (e.g., 7B vs 70B) on the robustness of Llama3 and Codestral in classifying vulnerabilities in Big-Vul, measured by F1-score degradation under increasing levels of. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: SoK: Automated Software Diversity. Research question: What is the impact of model size (e.g., 7B vs 70B) on the robustness of Llama3 and Codestral in classifying vulnerabilities in Big-Vul, measured by F1-score degradation under increasing levels of synthetic obfuscation?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

10 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.5/10.

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 idea of automatic software diversity is at least two decades old.	✓	0.26
The literature on diversity grew by more than two dozen papers since 2008.	✓	0.21
Unlike other defenses, software diversity introduces uncertainty in the target.	✓	0.23
Precise knowledge of the target software provides the underpinning for a wide range of attacks.	✓	0.27
Software diversity offers probabilistic protection similar to cryptography.	✓	0.27
The design space of diversifying program transformations is large.	✓	0.20
Researchers have proposed multiple approaches to software diversity that vary with respect to threat models, security, p	✓	0.32
Open areas and unresolved challenges in software diversity include hybrid solutions, error reporting, patching, and impl	✓	0.32

References

- https://doi.org/10.1007/s44443-025-00177-1
- https://doi.org/10.1109/sp.2014.25
- https://doi.org/10.1186/s13174-018-0087-2