Llama3 Robustness to Synthetic Code Obfuscation Across Vulnerability Classes

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the robustness of Llama3-7B versus Llama3-70B to synthetic code obfuscation vary across different vulnerability classes in the SARD dataset when measured by F1-score degradation. 14 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Systematic Study of Code Obfuscation Against LLM-based Vulnerability Detection. Research question: How does the robustness of Llama3-7B versus Llama3-70B to synthetic code obfuscation vary across different vulnerability classes in the SARD dataset when measured by F1-score degradation?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

12 papers retrieved. 14 claims extracted; 3 independently verified. Quality review score: 5.2/10.

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
Code obfuscation transformations can unexpectedly improve vulnerability detection accuracy by removing misleading surfac	×	0.08
Control-flow virtualization and mixed-programming-language transformations have the strongest degrading effect on LLM-ba	✓	0.18
Models smaller than 8B parameters show pronounced instability under code obfuscation.	×	0.05
Models larger than 8B parameters maintain higher resilience to code obfuscation, though additional scaling yields dimini	×	0.03
Reasoning-augmented models perform better on unobfuscated code but are more sensitive to obfuscation than non-reasoning	×	0.04
Vulnerability types involving pointer safety, reentrancy, and access control show the largest fluctuations in detection	×	0.06
Coding agents exhibit higher detection success rates than general-purpose LLMs on unobfuscated code.	×	0.07
Coding agents experience both downgrade and upgrade effects under code obfuscation, particularly with inline assembly an	×	0.08
Hot-plugging a new model into an agent framework can reduce the effectiveness of transferring vulnerability-detection kn	×	0.05
The study categorizes existing obfuscation techniques into three major classes: layout, data flow, and control flow.	✓	0.22
The obfuscation taxonomy covers 11 subcategories and 19 concrete methods.	×	0.14
The evaluation framework implements transformations across four programming languages: Solidity, C, C++, and Python.	×	0.14
The study evaluates the impact of obfuscation on 15 LLMs spanning four model families: DeepSeek, OpenAI, Qwen, and LLaMA	✓	0.22
The study evaluates two coding agents: GitHub Copilot and Codex.	×	0.14

References

- http://arxiv.org/abs/2106.16020v1
- http://arxiv.org/abs/2512.16538v1
- http://arxiv.org/abs/2403.03788v1