Codestral Parameter Scaling and Robustness to Obfuscated Code in Vulnerability Detection

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: Does increasing Codestral's parameter count improve robustness against obfuscated code variants in vulnerability detection benchmarks compared to smaller variants. As large language models (LLMs) are increasingly adopted for code vulnerability detection, their reliability and robustness across diverse vulnerability types have become a pressing concern. In traditional adversarial settings, code obfuscation has long been used as a general. 14 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. 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: Does increasing Codestral's parameter count improve robustness against obfuscated code variants in vulnerability detection benchmarks compared to smaller variants?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

11 papers retrieved. 14 claims extracted; 3 independently verified. Quality review score: 4.8/10.

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	×	0.07
Control-flow virtualization and mixed-programming-language transformations have the strongest degrading effect on LLM-ba	✓	0.17
Models smaller than 8 billion parameters show pronounced instability under code obfuscation.	×	0.04
Models larger than 8 billion parameters maintain higher resilience to code obfuscation, though additional scaling yields	×	0.03
Reasoning-augmented models perform better on unobfuscated code but are more sensitive to obfuscation than non-reasoning	×	0.05
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.	×	0.09
Coding agents experience both downgrade and upgrade effects under code obfuscation, particularly when facing inline asse	×	0.07
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.	×	0.13
The study implemented obfuscation transformations across four programming languages: Solidity, C, C++, and Python.	×	0.14
The study evaluated the impact of obfuscation on 15 LLMs spanning four model families: DeepSeek, OpenAI, Qwen, and LLaMA	✓	0.22
The study evaluated two coding agents: GitHub Copilot and Codex.	×	0.14

References

- http://arxiv.org/abs/2512.16538v1
- http://arxiv.org/abs/2605.06910v1
- http://arxiv.org/abs/2412.09565v2