How does domain-specific code preprocessing affect the vulnerability detection accuracy of Llama3, Codestral,

Assignee Research

May 29, 2026

Abstract

As Large Language Models (LLMs) become increasingly integrated into secure software development workflows, a critical question remains unanswered: can these models not only detect insecure code but also reliably classify vulnerabilities according to standardized taxonomies? In this work, we conduct a systematic evaluation of three state-of-the-art LLMs - Llama3, Codestral, and Deepseek R1 - using a carefully filtered subset of the Big-Vul dataset annotated with eight representative Common Weakness Enumeration categories. Adopting a closed-world classification setup, we assess each model's perf

1 Introduction

This paper examines: Can Open Large Language Models Catch Vulnerabilities?. Research question: How does domain-specific code preprocessing affect the vulnerability detection accuracy of Llama3, Codestral, and Deepseek R1 across Python, JavaScript, and C++ 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 4.7/10.

3 Results

9 papers retrieved. 11 claims extracted; 4 independently verified. Quality review score: 4.7/10.

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 study evaluates three LLMs: Llama3, Codestral, and Deepseek R1.	×	0.15
The evaluation uses a carefully filtered subset of the Big-Vul dataset.	✓	0.17
The dataset subset is annotated with eight representative Common Weakness Enumeration (CWE) categories.	✓	0.18
The study adopts a closed-world classification setup.	×	0.12
The models were assessed on their ability to identify the presence of vulnerabilities.	×	0.07
The models were assessed on their ability to map vulnerabilities to the correct CWE label.	×	0.10
The evaluated models demonstrated high detection rates for vulnerabilities.	×	0.10
The evaluated models demonstrated markedly poor classification accuracy for CWE labels.	×	0.12
The models exhibited frequent overgeneralization and misclassification of vulnerabilities.	×	0.12
The study analyzes model-specific biases and common failure modes.	✓	0.17
LLMs are being adopted as learning aids in educational contexts.	✓	0.18

References

- https://doi.org/10.4230/oasics.icpec.2025.4
- https://doi.org/10.1145/1167473.1167488
- https://doi.org/10.48550/arxiv.2401.14196