How does the alignment of Llama3, Codestral, and Deepseek R1 with security-specific fine-tuning (e.g., SecLM)

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 the alignment of Llama3, Codestral, and Deepseek R1 with security-specific fine-tuning (e.g., SecLM) affect their reasoning accuracy in vulnerability detection, as measured by HumanEval-hard and SWE-bench scores?.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 7 claims extracted; 7 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
Three state-of-the-art LLMs - Llama3, Codestral, and Deepseek R1 - were evaluated using a subset of the Big-Vul dataset	✓	0.32
The evaluation adopted a closed-world classification setup to assess each model's performance in identifying the presenc	✓	0.32
The findings revealed a sharp contrast between high detection rates and markedly poor classification accuracy, with freq	✓	0.28
The analysis included model-specific biases and common failure modes, shedding light on the limitations of current LLMs	✓	0.32
The insights are especially relevant in educational contexts, where LLMs are being adopted as learning aids despite thei	✓	0.28
A nuanced understanding of their behaviour is essential to prevent the propagation of misconceptions among students.	✓	0.23
The results expose key challenges that must be addressed before LLMs can be reliably deployed in security-sensitive envi	✓	0.29

References

- https://doi.org/10.4230/oasics.icpec.2025.4
- https://doi.org/10.48550/arxiv.2406.00515
- https://doi.org/10.48550/arxiv.2204.14198