Scaling Codestral Model Size and Its Effect on Big-Vul Vulnerability Classification

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of model size scaling (e.g., 7B vs 33B) on Codestral's vulnerability classification accuracy across different severity levels in Big-Vul. While automated vulnerability detection techniques have made promising progress in detecting security vulnerabilities, their scalability and applicability remain challenging. The remarkable performance of Large Language Models (LLMs), such as GPT-4 and CodeLlama, on code-related. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Understanding the Effectiveness of Large Language Models in Detecting Security Vulnerabilities. Research question: What is the impact of model size scaling (e.g., 7B vs 33B) on Codestral's vulnerability classification accuracy across different severity levels in Big-Vul?.

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 3.7/10.

3 Results

11 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 best performing models and prompts per dataset report an accuracy of 62.8% on average.	×	0.07
The highest accuracy is reported by Llama-3.1-70B (with CWE) on the Juliet Java dataset (76%).	×	0.03
The best accuracies on synthetic datasets are 10.5% higher on average than those on the real-world datasets.	×	0.11
Qwen-2.5-14B and Qwen-2.5-32B report higher accuracies than the GPT-x models despite being much smaller.	×	0.07
Within the same model class, the GPT-x models and the Llama-3.1-x models exhibit improvements in accuracy as the size of	×	0.06
The CWE-DF prompt reports significantly higher F1 scores on average than CWE and Basic prompt on the real-world datasets	×	0.07
CodeLlama-7B correctly predicts that an integer overflow vulnerability (CWE-190) cannot occur in the given context while	×	0.05
The experiments were conducted using the nAI public API's ChatCompletions endpoint for GPT-3.5 and GPT-4, and Google's G	×	0.01
The experiments with open-source LLMs were conducted using the HuggingFace API on a cluster with A100, A6000, and RTX 20	×	0.04
In all experiments, the sampling temperature was set to 0 for deterministic predictions, the maximum number of tokens to	×	0.02
The top-1 predictions were used for evaluation.	×	0.05
The best performing models per dataset are: CodeLlama-13B on OWASP (60%), Gemini-1.5-Flash on CVEFixes Java (57%), Qwen-	×	0.02
The three prompting strategies (Basic, CWE, CWE-DF) perform similarly in terms of accuracy on all datasets with CWE-DF r	×	0.10

References

- http://arxiv.org/abs/2509.13442v1

- http://arxiv.org/abs/2311.16169v3
- http://arxiv.org/abs/2410.21676v4