SOVEREIGN: MATH dataset benchmark evaluation language model performance scores comparison 2024

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

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

Analysis of: Can Open Large Language Models Catch Vulnerabilities?. Research goal: MATH dataset benchmark evaluation language model performance scores comparison 2024.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

14 papers retrieved. 8 claims extracted, 7 verified. Tribunal: 6.9/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

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 vulnerabili	✓	0.30
The findings revealed a sharp contrast between high detection rates and markedly poor classification accuracy among the	✓	0.22
Frequent overgeneralization and misclassification were observed in the LLMs' performance.	×	0.11
Model-specific biases and common failure modes were analyzed, highlighting the limitations of current LLMs in performing	✓	0.28
The insights are particularly relevant in educational contexts where LLMs are being adopted as learning aids despite the	✓	0.23
A nuanced understanding of LLMs' behavior is essential to prevent the propagation of misconceptions among students.	✓	0.18
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/10.9781/ijimai.2024.02.009
- https://doi.org/10.48550/arxiv.2412.05271