SOVEREIGN: Does the self-invoking code generation task in HumanEval Pro reveal systematic failure modes in MambaFormer's

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: Does the self-invoking code generation task in HumanEval Pro reveal systematic failure modes in MambaFormer's long-range dependency handling compared to Transformer MoE models, as measured by solution correctness on multi-step problems?.

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

10 papers retrieved. 6 claims extracted, 5 verified. Tribunal: 7.0/10 → 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
Large Language Models can detect insecure code but show poor performance in classifying vulnerabilities according to sta	✓	0.22
The evaluation used a carefully filtered subset of the Big-Vul dataset annotated with eight representative Common Weakne	✓	0.27
Three state-of-the-art LLMs (Llama3, Codestral, and Deepseek R1) were evaluated in the study	✓	0.17
The study found a sharp contrast between high detection rates and markedly poor classification accuracy for LLMs in vuln	✓	0.22
LLMs frequently overgeneralize and misclassify vulnerabilities when attempting fine-grained security reasoning	×	0.11
Current LLMs have limitations in performing fine-grained security reasoning	✓	0.19

References

- https://doi.org/10.48550/arxiv.2306.05685
- https://doi.org/10.4230/oasics.icpec.2025.4
- https://doi.org/10.48550/arxiv.2302.13971