Adversarial Code Complexity and Inference Latency in DeepSeek R1 for HumanEval Tasks

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the relationship between adversarial code complexity (measured by cyclomatic complexity) and the inference latency of Deepseek R1 when generating solutions for HumanEval, and can efficiency. 11 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Can Open Large Language Models Catch Vulnerabilities?. Research question: What is the relationship between adversarial code complexity (measured by cyclomatic complexity) and the inference latency of Deepseek R1 when generating solutions for HumanEval, and can efficiency improvements reduce the accuracy drop-off for high-complexity prompts?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.7/10.

3 Results

13 papers retrieved. 11 claims extracted; 5 independently verified. Quality review score: 5.7/10.

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.	×	0.11
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/10.1109/tnnls.2021.3084827
- https://doi.org/10.1145/3649506
- https://doi.org/10.4230/oasics.icpec.2025.4