DeepSeek-R1 Performance Degradation on Vulnerability Detection Across Cyclomatic Complexity Levels

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 5 peer-reviewed papers addressing the following research question: How does the performance of Deepseek R1 on vulnerability detection tasks degrade when fine-tuned on code with varying cyclomatic complexity levels, as evaluated by F1-score and false negative rate on. 8 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLM4Vuln: A Unified Evaluation Framework for Decoupling and Enhancing LLMs' Vulnerability Reasoning. Research question: How does the performance of Deepseek R1 on vulnerability detection tasks degrade when fine-tuned on code with varying cyclomatic complexity levels, as evaluated by F1-score and false negative rate on the Big-Vul dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.2/10.

3 Results

5 papers retrieved. 8 claims extracted; 4 independently verified. Quality review score: 6.2/10.

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 paper introduces LLM4Vuln, a unified evaluation framework designed to decouple LLMs' vulnerability reasoning from ca	✓	0.35
The paper constructs UniVul, described as the first benchmark providing retrievable knowledge and context-supplementable	×	0.14
The UniVul benchmark covers three programming languages: Solidity, Java, and C/C++.	×	0.14
The study tested six specific LLMs: GPT-4.1, Phi-3, Llama-3, o4-mini, DeepSeek-R1, and QwQ-32B.	✓	0.20
The evaluation involved 147 ground-truth vulnerabilities and 147 non-vulnerable cases.	✓	0.20
The evaluation was conducted across 3,528 controlled scenarios.	×	0.10
The study identified 14 zero-day vulnerabilities in four pilot bug bounty programs.	✓	0.18
The identified zero-day vulnerabilities resulted in $3,576 in bounties.	×	0.10

References

- https://doi.org/10.3390/app152212150
- https://doi.org/10.48550/arxiv.2401.16185
- https://openalex.org/W7162149865