To what extent does fine-tuning Deepseek R1 on JavaScript-specific vulnerability patterns improve its detectio

Assignee Research

May 29, 2026

Abstract

Large language models (LLMs) have demonstrated significant potential in various tasks, including those requiring human-level intelligence, such as vulnerability detection. However, recent efforts to use LLMs for vulnerability detection remain preliminary, as they lack a deep understanding of whether a subject LLM's vulnerability reasoning capability stems from the model itself or from external aids such as knowledge retrieval and tooling support. In this paper, we aim to decouple LLMs' vulnerability reasoning from other capabilities, such as vulnerability knowledge adoption, context informatio

1 Introduction

This paper examines: LLM4Vuln: A Unified Evaluation Framework for Decoupling and Enhancing LLMs' Vulnerability Reasoning. Research question: To what extent does fine-tuning Deepseek R1 on JavaScript-specific vulnerability patterns improve its detection accuracy over pre-trained weights when evaluated on the Big-Vul benchmark?.

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 7.3/10.

3 Results

11 papers retrieved. 7 claims extracted; 5 independently verified. Quality review score: 7.3/10.

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
LLM4Vuln is a unified evaluation framework designed to decouple LLMs' vulnerability reasoning from capabilities such as	✓	0.45
UniVul is the first benchmark that provides retrievable knowledge and context-supplementable code across Solidity, Java,	✓	0.25
The study tested six representative LLMs: GPT-4.1, Phi-3, Llama-3, o4-mini, DeepSeek-R1, and QwQ-32B.	✓	0.24
The evaluation involved 147 ground-truth vulnerabilities and 147 non-vulnerable cases.	✓	0.21
The evaluation was conducted across 3,528 controlled scenarios.	×	0.10
The study identified 14 zero-day vulnerabilities in four pilot bug bounty programs.	✓	0.19
The identified zero-day vulnerabilities resulted in $3,576 in bounties.	×	0.11

References

- https://doi.org/10.48550/arxiv.2305.06161
- https://doi.org/10.48550/arxiv.2401.16185
- https://doi.org/10.48550/arxiv.2403.05530