DeepSeek R1 and Codestral Generalization in Cross-Language Code Repair Benchmarks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How do multimodal models like DeepSeek R1 generalize to out-of-domain code repair tasks compared to Codestral when evaluated on cross-language benchmarks like VulDeePecker and Devign. 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. 7 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.5/10. 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 do multimodal models like DeepSeek R1 generalize to out-of-domain code repair tasks compared to Codestral when evaluated on cross-language benchmarks like VulDeePecker and Devign?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

7 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.5/10.

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
Large language models (LLMs) have demonstrated significant potential in various tasks, including those requiring human-l	✓	0.30
Recent efforts to use LLMs for vulnerability detection remain preliminary, as they lack a deep understanding of whether	✓	0.44
LLM4Vuln is a unified evaluation framework that separates and assesses LLMs' vulnerability reasoning capabilities and ex	✓	0.41
UniVul is the first benchmark that provides retrievable knowledge and context-supplementable code across three represent	✓	0.32
Six representative LLMs (GPT-4.1, Phi-3, Llama-3, o4-mini, DeepSeek-R1, and QwQ-32B) were tested for 147 ground-truth vu	✓	0.37
The findings reveal the varying impacts of knowledge enhancement, context supplementation, and prompt schemes.	✓	0.30
14 zero-day vulnerabilities were identified in four pilot bug bounty programs, resulting in $3,576 in bounties.	✓	0.24

References

- https://doi.org/10.3390/electronics14224449
- https://doi.org/10.48550/arxiv.2401.16185

- https://doi.org/10.48550/arxiv.2502.07049