DeepSeek R1 and Claude Efficiency-Accuracy Trade-offs in Secure Code Review Pipelines

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: What is the efficiency-accuracy trade-off when deploying Deepseek R1 and Claude in secure code review pipelines, measured by inference latency and vulnerability detection F1-scores on the Big-Vul. Large language models (LLMs) have demonstrated strong capability for code understanding and vulnerability detection. However, most existing approaches rely on static prompting and treat the model as a passive predictor, limiting adaptability under uncertainty, particularly in. 10 claims were extracted from source literature; 10 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 9.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adaptive Self-Prompting in Agentic LLM Frameworks for Code Fault Detection. Research question: What is the efficiency-accuracy trade-off when deploying Deepseek R1 and Claude in secure code review pipelines, measured by inference latency and vulnerability detection F1-scores on the Big-Vul dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.2/10.

3 Results

1 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 9.2/10.

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
Large language models (LLMs) have demonstrated strong capability for code understanding and vulnerability detection.	✓	0.28
Most existing approaches rely on static prompting and treat the model as a passive predictor, limiting adaptability unde	✓	0.35
This paper introduces adaptive self-prompting as a core mechanism for agentic LLM-based fault detection in C-language em	✓	0.42
We propose two complementary frameworks: Agentic Retrieval-Augmented Generation (A-RAG) and Agentic Supervised Fine-Tuni	✓	0.34
A-RAG performs confidence-triggered, reasoning-conditioned retrieval from CWE and SEI CERT knowledge bases at inference	✓	0.32
A-SFT internalizes improvements through a self-evaluation sweep that refines instructions and training exemplars during	✓	0.29
Experiments are conducted on a unified dataset constructed from the Toyota ITC benchmark and a curated subset of Big-Vul	✓	0.35
Results show that adaptive self-prompting substantially improves predictive performance and error calibration compared t	✓	0.44
Adaptive self-prompting achieves up to 86.3% F1 score while significantly reducing high-confidence misclassifications.	✓	0.26
Confidence-aware reflection and adaptive reasoning enhance both robustness and safety in LLM-based fault detection for e	✓	0.35

References

- https://doi.org/10.3390/software5020016