Multimodal Fine-Tuned Small Language Models vs. Large Multimodal LLMs in CWE Detection Accuracy

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How do small language models (SLMs) fine-tuned with multimodal context compare to larger multimodal LLMs in terms of CWE detection accuracy and alignment metrics on the extended Big-Vul dataset. In this paper, we present a challenging code reasoning task: vulnerability detection. Large Language Models (LLMs) have shown promising results in natural-language and math reasoning, but state-of-the-art (SOTA) models reported only 54.5% Balanced Accuracy in our vulnerability. 6 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: To Err is Machine: Vulnerability Detection Challenges LLM Reasoning. Research question: How do small language models (SLMs) fine-tuned with multimodal context compare to larger multimodal LLMs in terms of CWE detection accuracy and alignment metrics on the extended Big-Vul dataset?.

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 8.3/10.

3 Results

7 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 8.3/10.

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
State-of-the-art (SOTA) models reported only 54.5% Balanced Accuracy in vulnerability detection evaluation.	✓	0.27
Models struggle to reason about the code semantics relevant to identifying vulnerabilities, especially subtle semantic d	✓	0.33
Explored models and training settings (better prompts, larger models, more pre-training data, and fine-tuning) did not l	✓	0.31
Adding domain knowledge to prompts helped certain models understand some code semantics but still failed in multi-step r	✓	0.33
New models, new training methods, or more execution-specific pretraining data may be needed to conquer vulnerability det	✓	0.31
Auto-regressive models may not be sufficient for complex code reasoning tasks like vulnerability detection.	✓	0.19

References

- https://doi.org/10.48550/arxiv.2502.07049
- https://doi.org/10.48550/arxiv.2401.16185
- https://doi.org/10.48550/arxiv.2403.17218