What is the impact of incorporating multimodal context (e.g., UML diagrams or execution traces) on the CWE cla

Assignee Research

May 29, 2026

Abstract

Large Language Models (LLMs) have demonstrated significant capabilities in understanding and analyzing code for security vulnerabilities, such as Common Weakness Enumerations (CWEs). However, their reliance on cloud infrastructure and substantial computational requirements pose challenges for analyzing sensitive or proprietary codebases due to privacy concerns and inference costs. This work explores the potential of Small Language Models (SLMs) as a viable alternative for accurate, on-premise vulnerability detection. We investigated whether a 350-million parameter pre-trained code model (codeg

1 Introduction

This paper examines: Case Study: Fine-tuning Small Language Models for Accurate and Private CWE Detection in Python Code. Research question: What is the impact of incorporating multimodal context (e.g., UML diagrams or execution traces) on the CWE classification accuracy of fine-tuned SecLM models, as evaluated on an extended Big-Vul dataset with mixed text-image inputs?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

13 papers retrieved. 7 claims extracted; 4 independently verified. Quality review score: 6.5/10.

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
The un-tuned codegen-mono model failed to detect a single CWE within any of the 100 code snippets presented in the basel	×	0.09
The fine-tuned codegen-mono model achieved 99% accuracy on CWE detection task.	✓	0.18
The fine-tuned codegen-mono model achieved approximately 98.08% precision.	✓	0.18
The fine-tuned codegen-mono model achieved 100% recall.	✓	0.16
The fine-tuned codegen-mono model achieved an F1-Score of approximately 99.04%.	✓	0.20
Google's gemini-2.0-flash-thinking-exp-01-21 model was used for generating synthetic Python code snippets demonstrating	×	0.06
Five distinct Python code snippets were generated for each of the 25 selected CWEs.	×	0.08

References

- http://arxiv.org/abs/2504.16584v1
- http://arxiv.org/abs/2503.09433v2
- http://arxiv.org/abs/2305.16615v1