On-Device vs. Cloud Deployment Trade-offs for SLMs and LLMs in CWE Detection for Private Python Codebases

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the trade-off between inference throughput and pass@1 accuracy for SLMs vs. LLMs in CWE detection tasks on private Python codebases when deployed on-device vs. in cloud environments. 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. 3 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

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 trade-off between inference throughput and pass@1 accuracy for SLMs vs. LLMs in CWE detection tasks on private Python codebases when deployed on-device vs. in cloud environments?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.8/10.

3 Results

15 papers retrieved. 3 claims extracted; 2 independently verified. Quality review score: 7.8/10.

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 in baseline evaluation on 100 examples.	×	0.11
Fine-tuned codegen-mono model achieved 99% accuracy, $\approx$98.08% precision, 100% recall, and $\approx$99.04% F1-Score.	✓	0.28
The fine-tuned codegen-mono model achieved near-perfect accuracy of 99% on the CWE detection task.	✓	0.16

References

- http://arxiv.org/abs/2504.16584v1
- http://arxiv.org/abs/2604.23361v1
- http://arxiv.org/abs/2503.09433v2