Fine-Tuned Codestral-7B and Llama3-70B Cross-Domain Generalization in Security Vulnerability Classification

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the cross-domain generalization accuracy of fine-tuned Codestral-7B versus Llama3-70B on unseen programming languages beyond Python for security vulnerability classification. Many ML-based approaches have been proposed to automatically detect, localize, and repair software vulnerabilities. While ML-based methods are more effective than program analysis-based vulnerability analysis tools, few have been integrated into modern IDEs, hindering practical. 6 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: AIBugHunter: A Practical Tool for Predicting, Classifying and Repairing Software Vulnerabilities. Research question: What is the cross-domain generalization accuracy of fine-tuned Codestral-7B versus Llama3-70B on unseen programming languages beyond Python for security vulnerability classification.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

16 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
The AIBugHunter approach achieves 74% CWE-ID multiclass accuracy.	×	0.06
BERT-base model achieves 65% CWE-ID multiclass accuracy.	×	0.02
CodeBERT model achieves 59% CWE-Type multiclass accuracy.	×	0.02
CodeBERT model achieves 50% CWE-ID multiclass accuracy.	×	0.02
BoW+RF model achieves 27% CWE-ID multiclass accuracy.	×	0.02
BoW+NB model achieves 27% CWE-Type multiclass accuracy.	×	0.02

References

- http://arxiv.org/abs/2305.16615v1
- http://arxiv.org/abs/2011.08508v3
- http://arxiv.org/abs/2408.15301v2