Fine-Tuning on Single-Language Code Corpora and Robustness Degradation in Sub-10B Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent does domain adaptation via fine-tuning on single-language code corpora degrade the robustness of sub-10B models against obfuscated vulnerability patterns in mixed-language evaluation. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Fine-Tuning Code Language Models to Detect Cross-Language Bugs. Research question: To what extent does domain adaptation via fine-tuning on single-language code corpora degrade the robustness of sub-10B models against obfuscated vulnerability patterns in mixed-language evaluation sets?.

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 3.7/10.

3 Results

13 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 study uses a binary classification method to detect CLBs, determining whether a sample contains a bug (label = 1 or	×	0.04
The performance of CodeLMs is evaluated using several classification metrics, including accuracy, precision, recall, F1	×	0.07
Accuracy is calculated as (TP + TN) / (TP + TN + FP + FN).	×	0.02
Precision is calculated as TP / (TP + FP).	×	0.01
Recall is calculated as TP / (TP + FN).	×	0.01
F1 score is calculated as 2 $\times$ (Precision $\times$ Recall) / (Precision + Recall).	×	0.06
AUC is calculated as the integral of TPR with respect to FPR from 0 to 1.	×	0.02
TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives, respectively.	×	0.01
TPR (True Positive Rate), also known as Recall, is calculated as TP / (TP + FN).	×	0.02
FPR (False Positive Rate) is calculated as FP / (FP + TN).	×	0.01
For small models, full fine-tuning is performed.	×	0.15
For large models with more parameters, the LoRA (Low-Rank Adaptation) technique is employed due to limited GPU resources	×	0.05
The input context length of the CodeLMs is uniformly limited to no more than 512 tokens, except for the token sequence r	×	0.07
CLBs tend to manifest as non-fatal yet hard-to-detect runtime anomalies.	×	0.03
Behavioral deviation is the most frequently observed symptom of CLBs.	×	0.02

References

- http://arxiv.org/abs/2507.21954v2
- http://arxiv.org/abs/2412.09565v2
- http://arxiv.org/abs/2504.16584v1