CodeT5 Robustness in Low-Resource Language Vulnerability Detection Under Adversarial Perturbations

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the robustness of CodeT5's vulnerability detection vary across different low-resource languages when evaluated against adversarial code perturbations. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Detecting the Machine: A Comprehensive Benchmark of AI-Generated Text Detectors Across Architectures, Domains, and Adversarial Conditions. Research question: How does the robustness of CodeT5's vulnerability detection vary across different low-resource languages when evaluated against adversarial code perturbations?.

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 3.8/10.

3 Results

13 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
Random Forest achieves an AUROC of 0.9767 on the HC3 to HC3 evaluation condition.	×	0.01
Random Forest achieves an AUROC of 0.6337 on the eli5 to hc3 evaluation condition.	×	0.02
Logistic Regression achieves an AUROC of 0.8882 on the HC3 to HC3 evaluation condition.	×	0.01
SVM with RBF kernel achieves an AUROC of 0.7993 on the HC3 to HC3 evaluation condition.	×	0.01
Random Forest suffers the largest cross-domain degradation among classical detectors, dropping from an AUROC of 0.977 on	×	0.08
The study utilizes Mistral-7B-Instruct-v0.2 with FP16 precision and no quantization.	×	0.03
The Mistral-7B-Instruct-v0.2 model was configured with Flash Attention 2 and torch.compile (reduce-overhead).	×	0.04
Text generation experiments used a batch size of 48, a maximum of 150 new tokens, a temperature of 0.7, and a top-p of 0	×	0.03
AI-generated text detection in this study is formulated as a binary classification problem where each individual answer	×	0.07
In the dataset preparation, human answers are labeled as 0 and LLM answers are labeled as 1.	×	0.06
Generated tokens were decoded with the prompt stripped to yield clean answer strings.	×	0.03
Each question was formatted using Mistral's [INST] instruction template.	×	0.05

References

- http://arxiv.org/abs/2504.07887v2

- http://arxiv.org/abs/2204.08143v2
- http://arxiv.org/abs/2603.17522v1