Contrastive Loss Margins and CodeT5 Transferability Robustness Under Black-Box Adversarial Attacks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the correlation between contrastive loss margins and CodeT5's transferability robustness when subjected to black-box adversarial perturbations generated by substitute models on code. 11 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Transferable Backdoor Attacks for Code Models via Sharpness-Aware Adversarial Perturbation. Research question: What is the correlation between contrastive loss margins and CodeT5's transferability robustness when subjected to black-box adversarial perturbations generated by substitute models on code generation tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

12 papers retrieved. 11 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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
STAB is evaluated on three widely-used Python code datasets: Py150, CodeSearchNet (CSN), and PyTorch (PyT).	×	0.05
Py150 contains 150K Python files extracted from GitHub before 2016 for machine learning research.	×	0.04
CodeSearchNet (CSN) provides over 400K Python functions sourced from GitHub.	×	0.01
PyTorch (PyT) includes 218K Python package libraries crawled from the PyPI and Anaconda.	×	0.00
Backdoor attacks are conducted on generation tasks, which are more challenging.	×	0.07
STAB is compared with existing backdoor attacks in terms of transferability across different datasets.	✓	0.17
STAB is evaluated for its ability to evade state-of-the-art backdoor defense mechanisms.	×	0.09
The impact of key components and hyperparameters on STAB's effectiveness is analyzed.	×	0.06
STAB uses Sharpness-Aware Minimization for optimizing the surrogate model.	✓	0.16
STAB involves a forward pass and backward pass during the optimization process.	×	0.02
STAB updates the weight matrix during the optimization process.	×	0.02

References

- http://arxiv.org/abs/2602.11213v1
- http://arxiv.org/abs/2407.13111v1
- http://arxiv.org/abs/2412.15924v1