CodeT5 Robustness to Adversarial Code Perturbations on QuixBugs Benchmark

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the robustness of CodeT5 against adversarial code perturbations on the QuixBugs benchmark compare to its performance on clean code inputs when measured by exact match accuracy. 7 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On the Adversarial Robustness of Vision Transformers. Research question: How does the robustness of CodeT5 against adversarial code perturbations on the QuixBugs benchmark compare to its performance on clean code inputs when measured by exact match accuracy?.

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 4.7/10.

3 Results

15 papers retrieved. 7 claims extracted; 1 independently verified. Quality review score: 4.7/10.

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
ViTs are more robust to high-frequency perturbations than CNNs using crafted frequency-filtered attacks and feature visu	✓	0.21
ViTs pay more attention to the low-level structures and their feature maps become noisier when ResNet features are intro	×	0.06
Clean Accuracy (CA) stands for the accuracy evaluated on the entire ImageNet-1k test set.	×	0.07
Robust Accuracy (RA) stands for the accuracy on the adversarial examples generated with 1,000 test samples.	×	0.06
ViT-S/16 has a robust accuracy of 74.0% against low-pass frequency-filtered PGD attack with perturbation strength 0.001.	×	0.04
DeiT-S/16 has a robust accuracy of 63.52 against adversarial examples with perturbation strength 0.0.	×	0.06
Deit-T/16 has a clean accuracy of 72.3% and a robust accuracy of 36.8% against adversarial examples with perturbation st	×	0.12

References

- http://arxiv.org/abs/2207.04129v3
- http://arxiv.org/abs/2103.15670v3
- http://arxiv.org/abs/2408.14728v2