Adversarial Training Robustness Transfer from CodeT5 to InCoder and CodeGen

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the robustness improvement from adversarial training (FGSM vs. PGD) on CodeT5 generalize to other code generation models like InCoder or CodeGen on MBXP and HumanEval benchmarks. 0 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: Robust Image Classification: Defensive Strategies against FGSM and PGD Adversarial Attacks. Research question: How does the robustness improvement from adversarial training (FGSM vs. PGD) on CodeT5 generalize to other code generation models like InCoder or CodeGen on MBXP and HumanEval benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 10 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

10 papers retrieved. 0 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.

References

- http://arxiv.org/abs/2408.13274v1
- http://arxiv.org/abs/2011.05157v2
- http://arxiv.org/abs/2502.13141v1