Multi-Source Human Feedback Integration Enhances CodeT5+ Robustness Against Adversarial Perturbations

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: To what extent does the integration of multi-source human feedback via RLHF-Blender improve the robustness of CodeT5+ against adversarial code perturbations in the HumanEval-plus dataset. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Feature Denoising for Improving Adversarial Robustness. Research question: To what extent does the integration of multi-source human feedback via RLHF-Blender improve the robustness of CodeT5+ against adversarial code perturbations in the HumanEval-plus dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

11 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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 ImageNet classification dataset contains approximately 1.28 million images across 1000 classes.	×	0.05
The evaluation uses 50,000 ImageNet validation images.	×	0.04
Adversarial perturbations are considered under the L∞ norm with a maximum value of ϵ=16 relative to a pixel intensity sc	×	0.04
The default configuration adds 4 denoising blocks to a ResNet, placed after the last residual block of res2, res3, res4,	×	0.02
ALP achieved 27.9% accuracy on ImageNet validation images under a 10-iteration PGD attack on an Inception-v3 backbone.	×	0.12
The ResNet-101 baseline model achieved 49.7% accuracy under a 10-iteration PGD attack.	×	0.09
The ResNet-152 baseline model achieved 52.5% accuracy under a 10-iteration PGD attack.	×	0.09
Adding four non-local Gaussian denoising blocks to ResNet-152 improved accuracy from 52.5% to 55.7% under a 10-iteration	×	0.12
PGD is identified as the strongest white-box attacker compared to FGSM, iterative FGSM, and its momentum variant.	×	0.07
The dot-product version of the non-local means denoising operation requires no extra parameters.	×	0.06
The bilateral filter is defined by restricting the neighborhood Ω(i) in the non-local means equation to a local region,	×	0.05
The mean filter is implemented as average pooling with a stride of 1.	×	0.02

References

- http://arxiv.org/abs/1812.03411v2
- http://arxiv.org/abs/2008.07651v1
- http://arxiv.org/abs/2103.15670v3