Adversarial Training with PGD Attacks and CodeT5 Alignment to Human Preferences

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of adversarial training with PGD attacks on the alignment of CodeT5's generated code with human preferences, as measured by the CodeT5-HumanEval alignment score. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: RRHF: Rank Responses to Align Language Models with Human Feedback without tears. Research question: What is the impact of adversarial training with PGD attacks on the alignment of CodeT5's generated code with human preferences, as measured by the CodeT5-HumanEval alignment score?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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.

References

- http://arxiv.org/abs/2408.13274v1
- http://arxiv.org/abs/2309.02144v1
- http://arxiv.org/abs/2304.05302v3