Alignment Fine-Tuning Effects on LLM Robustness in Adversarial Code Generation

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does alignment fine-tuning on self-invoking code generation tasks affect the robustness of LLMs against adversarial perturbations, measured by pass@1 degradation on perturbed versions of. 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: Improving the Robustness of Large Language Models for Code Tasks via Fine-tuning with Perturbed Data. Research question: How does alignment fine-tuning on self-invoking code generation tasks affect the robustness of LLMs against adversarial perturbations, measured by pass@1 degradation on perturbed versions of HumanEval+ and MBPP+ benchmarks?.

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/2306.11066v2
- http://arxiv.org/abs/2602.11411v1
- http://arxiv.org/abs/2412.21199v2