Robustness of JaCoText and CodeT5 Against Adversarial Natural Language Perturbations in MBPP

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the difference in robustness between JaCoText and CodeT5 against adversarial natural language perturbations in the MBPP benchmark as measured by drop in execution accuracy. 10 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: Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge. Research question: What is the difference in robustness between JaCoText and CodeT5 against adversarial natural language perturbations in the MBPP benchmark as measured by drop in execution accuracy?.

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. 10 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.

5 Extracted Claims

Claim	Verified	Confidence
The safety threshold τ is defined as 0.5, and a model is considered safe if its safety score exceeds this threshold.	×	0.04
All SLMs, excluding GPT-4o mini, were tested locally on an NVIDIA A30 GPU using the Ollama service, requiring a total of	×	0.01
For the remaining models, accessed via API, the total cost was approximately 35 USD, with querying the judge LLM account	×	0.05
The control set was constructed by randomly sampling a small subset of prompts from the base prompts in CLEAR-Bias and m	×	0.05
Five candidate large models—GPT-4o, Claude 3.5 Sonnet, Llama 3.1 405B, Gem—were assessed for the role of the judge.	×	0.03
The computational resources used for evaluations included 10 GPU hours for SLMs and approximately 35 USD for API access	×	0.02
The safety score threshold τ = 0.5 is used to determine if a model is considered safe.	×	0.04
The models assessed include Gemma2 2B, Gemma2 27B, Phi-4 14B, Llama 3.1 8B, GPT-4o mini for SLMs, and Gemini 2.0 Flash,	×	0.03
The control set provided comprehensive coverage of both isolated and intersectional bias categories, as well as SC and C	×	0.04
The approximate computational resources used for evaluations were 10 GPU hours for SLMs and approximately 35 USD for API	×	0.02

References

- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2306.06371v1
- http://arxiv.org/abs/2306.11066v2