

# Adversarial Perturbations and Robustness of JaCoText Explanations Across Programming Benchmarks

Assignee Research

June 7, 2026

## Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of adversarial perturbations on the robustness of JaCoText’s explanations across different programming language benchmarks, quantified using explanation consistency metrics. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. 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 impact of adversarial perturbations on the robustness of JaCoText’s explanations across different programming language benchmarks, quantified using explanation consistency metrics?.

## 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 3.5/10.

## 3 Results

14 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.5/10.

## 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 $\tau$ 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 about 30% of this cost attribu	×	0.07
The control set for judge evaluation was constructed by randomly sampling a small subset of prompts from the base prompt	×	0.06
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 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 computational resources used for evaluations include 10 GPU hours for SLMs and approximately 35 USD for API access t	×	0.01
The safety scores for models are presented in the benchmark tables, with values indicating the level of safety for each	×	0.07
The refusal and debiasing rates for models are presented in Table (p18), with values indicating the proportion of respon	×	0.01
The stereotype and counter-stereotype rates for models are presented in Table (p18), with values indicating the proporti	×	0.01

## References

- <http://arxiv.org/abs/2504.07887v2>
- <http://arxiv.org/abs/2008.07651v1>
- <http://arxiv.org/abs/2306.11066v2>