Code Llama and StarCoder Robustness to Adversarial Identifiers on MBPP via PPTC-R

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the robustness of Code Llama to adversarial identifier perturbations on MBPP compare to StarCoder when measured by execution success rate under varying prompt complexity. 14 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PPTC-R benchmark: Towards Evaluating the Robustness of Large Language Models for PowerPoint Task Completion. Research question: How does the robustness of Code Llama to adversarial identifier perturbations on MBPP compare to StarCoder when measured by execution success rate under varying prompt complexity?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

15 papers retrieved. 14 claims extracted; 2 independently verified. Quality review score: 4.4/10.

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
PPTC-R is a benchmark designed to measure and analyze LLMs' robustness to user instructions and software versions in Pow	✓	0.32
Previous robustness evaluations are based on traditional natural language tasks where models only generate options or te	×	0.09
The PPTC-R benchmark includes instruction perturbations involving translating original English instructions into 14 non-	×	0.09
The PPTC-R benchmark includes sentence-level perturbations where 1 to 3 GPT-4 generated chitchat sentences are added to	×	0.05
The PPTC-R benchmark includes semantic-level perturbations where GPT-4 is prompted to express original instructions with	×	0.05
The PPTC-R benchmark tests robustness to software versions by introducing many new APIs to simulate version updates.	×	0.14
The PPTC-R benchmark tests robustness to software versions by removing many APIs to simulate situations where current so	✓	0.17
The study tested 3 closed-source LLMs, including GPT-4 and ChatGPT.	×	0.08
The study tested 4 representative open-source LLMs, including LLaMa-2 and WizardLM.	×	0.07
Under sentence-level perturbation, Davinci-003's performance on creating new slides (Turn-based) dropped from 72.6 to 64	×	0.03
Under semantic-level perturbation, GPT-4's performance on creating new slides (Session-based) dropped from 22.7 to 14.2.	×	0.03
When APIs were removed (Lack setting), Davinci-003's performance on creating new slides (Turn-based) dropped from 72.6 t	×	0.03
When APIs were updated (Update setting), GPT-4's performance on creating new slides (Turn-based) increased slightly from	×	0.04
In the session-based task of editing PPT templates, WizardLM achieved a score of 94.2.	×	0.03

References

- http://arxiv.org/abs/2401.10065v3
- http://arxiv.org/abs/2509.21843v1
- http://arxiv.org/abs/2403.03788v1