Instruction Ambiguity and CodeT5 Failure Rates in PPTC-R PowerPoint Tasks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the correlation between sentence-level instruction ambiguity and CodeT5's failure rate in completing complex PowerPoint tasks within the PPTC-R evaluation framework. 16 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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: What is the correlation between sentence-level instruction ambiguity and CodeT5's failure rate in completing complex PowerPoint tasks within the PPTC-R evaluation framework?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

9 papers retrieved. 16 claims extracted; 3 independently verified. Quality review score: 4.5/10.

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
PPTC-R is a benchmark designed to measure and analyze Large Language Models' robustness to user instructions and softwar	✓	0.27
Previous robustness evaluations for LLMs are primarily based on traditional natural language tasks where the model only	×	0.06
The PPTC-R benchmark evaluates how adversarial perturbations influence LLMs' API calls for complex PowerPoint task compl	✓	0.18
The PPTC-R benchmark includes instruction perturbations involving translating original English instructions into 14 non-	×	0.04
The PPTC-R benchmark includes sentence-level perturbations where GPT-4 generated chitchat sentences are added to origina	×	0.09
The PPTC-R benchmark includes semantic-level perturbations where GPT-4 is prompted to express original instructions with	×	0.04
The PPTC-R benchmark tests LLM robustness to software version shifts by introducing many new APIs to simulate version up	×	0.14
The PPTC-R benchmark tests LLM robustness to software version shifts by removing many APIs to simulate situations where	✓	0.18
The PPTC-R benchmark consists of a total of 5 settings derived from 3 user instruction perturbations and 2 API perturbat	×	0.07
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.	×	0.06
According to Table (p6), Davinci-003 achieved a score of 72.6 on the 'Creating new slides' Turn-based Original task.	×	0.03
According to Table (p6), GPT-4's performance on 'Creating new slides' Turn-based tasks dropped by 10.4 points under Sent	×	0.05
According to Table (p7), GPT-4's score on 'Creating new slides' Turn-based tasks increased by 0.6 points in the Update A	×	0.04
According to Table (p7), Davinci-003's score on 'Creating new slides' Turn-based tasks decreased by 28.1 points in the U 4	×	0.03
According to the table on page 7, WizardLM achieved a score of 94.2 in the Session-based 'Creating new slides' task.	×	0.03

References

- http://arxiv.org/abs/2402.17717v4
- http://arxiv.org/abs/2403.03788v1
- http://arxiv.org/abs/2311.01767v2