Qwen3-235B Performance Degradation Under PPTC-R Adversarial Instructions

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the performance of Qwen3-235B degrade under PPTC-R adversarial user instructions compared to standard instructions. The growing dependence on Large Language Models (LLMs) for finishing user instructions necessitates a comprehensive understanding of their robustness to complex task completion in real-world situations. To address this critical need, we propose the PowerPoint Task Completion. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.7/10. 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 performance of Qwen3-235B degrade under PPTC-R adversarial user instructions compared to standard instructions?.

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 5.7/10.

3 Results

15 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 5.7/10.

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 PPTC-R benchmark evaluates LLMs' robustness to user instruction and software version in PowerPoint task completion s	✓	0.32
PPTC-R includes 5 settings combining different perturbations: 3 for user instructions and 2 for APIs.	×	0.12
Language-level perturbation involves translating original English instructions into 14 non-English languages.	×	0.06
Semantic-level perturbation uses GPT-4 to rephrase original instructions with the same semantic meaning in 4 different w	×	0.03
Sentence-level perturbation adds 1-3 irrelevant chitchat non-instruction sentences into the original instructions.	×	0.03
The software version adjustments in APIs can simulate version update situations affecting LLM's API selection.	×	0.13
GPT-4 and ChatGPT are evaluated on 3 closed-source and 4 open-source LLMs in the PPTC-R benchmark.	✓	0.18
Davinci-003 shows a drop in performance of -7.8 for sentence-level perturbation and -5.2 for semantic-level perturbation	×	0.03
ChatGPT shows a performance drop of -9.3 for sentence-level perturbation and -5.6 for semantic-level perturbation in tas	×	0.10
GPT-4 shows a performance drop of -2.8 for sentence-level perturbation and -3.1 for semantic-level perturbation in sessi	×	0.03
LLaMa-2 shows a performance drop of -8.6 for software version updates in creating new slides.	×	0.06
WizardLM performs at 94.2 in the sentence-level perturbation setting for creating new slides.	×	0.03

References

- http://arxiv.org/abs/2403.03788v1
- http://arxiv.org/abs/2502.13141v1

- http://arxiv.org/abs/2306.14027v2