How does the defect localization performance of LLaMA 3.2 compare to Codex or CodeGen when using sliding windo

Assignee Research

June 10, 2026

Abstract

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 Robustness benchmark (PPTC-R) to measure LLMs' robustness to the user PPT task instruction and software version. Specifically, we construct adversarial user instructions by attacking user instructions at sentence, semantic, and multi-language levels. To assess the robustness of Language Models to software versions, we vary

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 defect localization performance of LLaMA 3.2 compare to Codex or CodeGen when using sliding window truncation across different programming languages in BugsInPy-like benchmarks?.

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 3.7/10.

3 Results

15 papers retrieved. 0 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.

References

- http://arxiv.org/abs/2403.03788v1
- http://arxiv.org/abs/2508.21256v1
- http://arxiv.org/abs/2604.18404v1