Chain-of-Thought Prompting Enhances Mistral-Large-2 Robustness Over GPT-4 on MBPP Edge Cases

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: To what extent does chain-of-thought prompting improve the robustness of Mistral-Large-2 versus GPT-4 on edge-case scenarios within the MBPP benchmark. Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation across various domains, including medicine. We present a comprehensive evaluation of GPT-4, a state-of-the-art LLM, on medical competency examinations and. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Capabilities of GPT-4 on Medical Challenge Problems. Research question: To what extent does chain-of-thought prompting improve the robustness of Mistral-Large-2 versus GPT-4 on edge-case scenarios within the MBPP benchmark?.

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 4.3/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.3/10.

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/2303.13375v2
- http://arxiv.org/abs/2407.04973v1
- http://arxiv.org/abs/2601.01982v1