GPT-4 Babilong Score Discrepancy: Evaluation Protocol Variations Across Studies

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Benchmark archaeology: investigate Babilong score discrepancy for GPT-4 — reported 10.0%–85.0% (spread 75.0pp) across 2 papers. Sources: 'BABILong: Testing the Limits of LLMs wit' (10.0%); 'BABILong:. 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 1.5/10. 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: Benchmark archaeology: investigate Babilong score discrepancy for GPT-4 — reported 10.0%–85.0% (spread 75.0pp) across 2 papers. Sources: 'BABILong: Testing the Limits of LLMs wit' (10.0%); 'BABILong: Testing the Limits of LLMs wit' (85.0%). Identify evaluation protocol differences (few-shot, prompting, preprocessing)..

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 1.5/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 1.5/10.

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.

References

- http://arxiv.org/abs/2406.10149v2
- http://arxiv.org/abs/2304.03277v1
- http://arxiv.org/abs/2303.13375v2