Current Language Model Benchmarks Fail to Measure Reasoning Capabilities

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What are the limitations of current language model evaluation benchmarks for measuring reasoning v11. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts. Research question: What are the limitations of current language model evaluation benchmarks for measuring reasoning v11.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.5/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.5/10.

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/1610.00031v1
- http://arxiv.org/abs/2307.12114v3
- http://arxiv.org/abs/2407.04973v1