Language Model Performance on BIG-Bench Hard Reasoning Tasks: A Multi-Study Synthesis

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: BIG-Bench Hard reasoning task language model evaluation comparison. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: BIG-Bench Extra Hard. Research question: BIG-Bench Hard reasoning task language model evaluation comparison.

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 5.2/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.2/10.

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/2601.01982v1
- http://arxiv.org/abs/2210.09261v1
- http://arxiv.org/abs/2502.19187v2