State-of-the-Art Large Language Model Performance on Reasoning Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What are the state-of-the-art large language model results on reasoning benchmarks published recently v9. 12 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. Research question: What are the state-of-the-art large language model results on reasoning benchmarks published recently v9.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

14 papers retrieved. 12 claims extracted; 6 independently verified. Quality review score: 6.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
In the BIG-Bench paper (Srivastava et al., 2022), none of the evaluated models, including PaLM 540B, outperformed human-	✓	0.24
Few-shot evaluation of PaLM 540B with answer-only prompting outperforms the average human-rater on 6 out of 23 BBH tasks	✓	0.19
The few-shot evaluation of PaLM 540B with answer-only prompting is overall 1.4% better than the BIG-Bench reported resul	×	0.13
Chain-of-Thought (CoT) prompting provides double-digit performance improvements for PaLM, InstructGPT, and Codex models	×	0.12
Codex with CoT prompting outperforms the average human-rater score on 17 out of 23 BBH tasks.	✓	0.23
Codex with answer-only prompting outperforms the average human-rater score on 5 out of 23 BBH tasks.	✓	0.19
Codex with CoT prompting outperforms the average human-rater by more than 6%.	✓	0.19
Codex with CoT prompting lags behind the best human-rater performance by over 20%.	✓	0.16
For OpenAI models ranging from text-ada-001 to text-curie-002, CoT prompting results in negative or zero performance gai	×	0.05
For OpenAI models, the performance delta between CoT and no-CoT prompting increases with model scale up to the largest m	×	0.08
For PaLM models, CoT prompting yields negative performance gain for the smallest model size (8B).	×	0.06
For PaLM models, CoT prompting performance improves as the model size increases beyond 8B.	×	0.07

References

- http://arxiv.org/abs/2407.04973v1
- http://arxiv.org/abs/2305.17306v1
- http://arxiv.org/abs/2210.09261v1