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

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What are the state-of-the-art large language model results on reasoning benchmarks published recently v13. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Gemini in Reasoning: Unveiling Commonsense in Multimodal Large Language Models. Research question: What are the state-of-the-art large language model results on reasoning benchmarks published recently v13.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

12 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
Gemini Pro lags behind GPT-4 Turbo in accuracy on language-only commonsense reasoning tasks.	✓	0.17
Gemini Pro encounters challenges in temporal reasoning.	×	0.05
Gemini Pro encounters challenges in social reasoning.	×	0.05
Gemini Pro encounters challenges in emotion recognition in images.	×	0.03
On the CommonsenseQA benchmark, GPT-4 Turbo (k-shot, CoT) achieved an accuracy of 80.0.	×	0.05
On the HellaSWAG benchmark, GPT-4 Turbo (k-shot, CoT) achieved an accuracy of 95.0.	×	0.05
On the RiddleSense benchmark, GPT-4 Turbo (k-shot, CoT) achieved an accuracy of 95.0.	×	0.05
On the ETHICS benchmark, GPT-4 Turbo (k-shot, CoT) achieved an accuracy of 98.0.	×	0.05
GPT-4V outperforms Gemini Pro Vision across all subtasks (Q$\rightarrow$A, QA$\rightarrow$R, Q$\rightarrow$AR) on the VCR dataset.	×	0.04
On the VCR dataset Q$\rightarrow$A subtask, GPT-4V achieved an accuracy of 80.0 while Gemini Pro Vision achieved 74.0.	×	0.04
On the VCR dataset QA$\rightarrow$R subtask, GPT-4V achieved an accuracy of 72.0 while Gemini Pro Vision achieved 70.0.	×	0.04
On the VCR dataset Q$\rightarrow$AR subtask, GPT-4V achieved an accuracy of 56.0 while Gemini Pro Vision achieved 48.0.	×	0.04
On the VCR dataset, Gemini Pro Vision surpasses GPT-4V in temporal-type questions.	×	0.04
General Commonsense involves understanding basic everyday knowledge, such as recognizing that birds typically fly.	×	0.04
Physical Commonsense includes knowing that a glass will break if dropped on a hard floor.	×	0.03
Science Commonsense involves understanding that water boils at a higher temperature at sea level than in the mountains.	×	0.03

References

- http://arxiv.org/abs/2006.01205v2
- http://arxiv.org/abs/2312.17661v1
- http://arxiv.org/abs/2109.13006v1