Frontier Large Language Models in Mathematical Reasoning and Scientific Knowledge Synthesis

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Comprehensive comparison of frontier large language models on mathematical reasoning code generation and scientific knowledge v11. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Critical Review of Causal Reasoning Benchmarks for Large Language Models. Research question: Comprehensive comparison of frontier large language models on mathematical reasoning code generation and scientific knowledge 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 3.5/10.

3 Results

4 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
The e-CARE dataset requires a model to choose a correct hypothesis for a given premise from two candidates to form a val	×	0.03
The e-CARE dataset contains examples where the options are not well-crafted, making the correct decision difficult for h	×	0.02
The Forecasting Subquestions task of BigBench evaluates the log-probability assigned by human-generated questions relate	×	0.01
The Forecasting Subquestions task of BigBench does not contain subquestions that are not causes of the question where a	×	0.02
The BIGbench entailed polarity task evaluates an LLM's ability to detect entailed polarities from implicative verbs.	×	0.01
LogiQA, Dream, and RACE are referred to as causal reasoning datasets in Yang et al. 2022.	×	0.05
The BigBench speech detection dataset assesses whether a given figure of speech is a simile, metaphor, or pun.	×	0.01
The BigBench 'Indic cause and effect' dataset mainly assesses language translation.	×	0.02
The authors propose a Causal Language Understanding Evaluation (CLUE) framework consisting of a minimal but exhaustive s	×	0.12
Some existing causal reasoning benchmarks allow answers to be looked up due to the use of informative or non-fictional c	×	0.11

References

- http://arxiv.org/abs/2407.08029v1

- http://arxiv.org/abs/2504.19565v3
- http://arxiv.org/abs/2310.03731v1