Frontier Model Performance on the GPQA Diamond Benchmark: A Literature Synthesis

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: GPQA Diamond benchmark frontier model performance evaluation recent literature. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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: GPQA Diamond benchmark frontier model performance evaluation recent literature.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

15 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
Improvements in LLM performance on certain tasks cannot be disentangled from model improvements (e.g., more parameters,	×	0.04
The e-CARE dataset contains examples where selecting the correct hypothesis is difficult for humans because the options	×	0.03
The Forecasting Subquestions task of BigBench evaluates log-probability assigned to human-generated questions related as	×	0.03
The BIGbench entailed polarity task evaluates an LLM's ability to detect entailed polarities from implicative verbs.	×	0.02
The BIGbench entailed polarity task is not a causal reasoning task.	×	0.08
LogiQA, Dream, and RACE datasets are referred to as causal reasoning datasets in Yang et al. 2022.	×	0.06
LogiQA, Dream, and RACE datasets assess reading comprehension rather than the deduction of causal relationships.	×	0.02
The BigBench speech detection dataset assesses whether a figure of speech is a simile, metaphor, or pun.	×	0.04
The BigBench 'Indic cause and effect' task mainly assesses language translation.	×	0.03
Good performance of LLMs on existing causal tasks may be attributed to data-processing and retrieval capabilities rather	×	0.12
Some existing causal reasoning tasks offer multiple-choice answers to the LLM.	×	0.08

References

- http://arxiv.org/abs/2403.03788v1
- http://arxiv.org/abs/2402.11651v2

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