Frontier Language Models on GPQA Diamond and Reasoning Benchmarks v13

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v13. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ARB: Advanced Reasoning Benchmark for Large Language Models. Research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v13.

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 3.6/10.

3 Results

15 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.6/10.

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
Most tasks in the ARB benchmark are designed to be out of reach of current models, with many models scoring over 95% bei	×	0.05
The tested skills in the ARB benchmark should correlate with generally useful human skills.	×	0.03
The ARB benchmark is designed to be straightforward for model creators to compare the performances of different models,	×	0.04
The ARB benchmark aims to minimize data contamination, as recent LLMs may contain some tasks in their training data, lea	×	0.07
The ARB benchmark includes problems that are not pathological or overly adversarial to avoid the dangers of underclaimin	×	0.04
The ARB benchmark consists of three types of questions: multiple choice, short answer, and open response, in descending	×	0.07
Multiple choice questions in the ARB benchmark consist of a question and four to five possible answers, sourced from sta	×	0.08
Short answer questions in the ARB benchmark ask for final answers in the format of a short phrase or mathematical expres	×	0.03
Open response questions in the ARB benchmark are more challenging and require manual grading, sourced from problem books	×	0.04
The mathematics part of the ARB dataset is the most diverse, including contest mathematics problems and university mathe	×	0.06
The ARB benchmark categorizes errors into logical errors, hallucinating facts or theorems, and arithmetic/calculation er	×	0.03
GPT-4's performance on the ARB benchmark is graded based on guidelines that include logical errors, hallucinating facts	×	0.04
The ARB benchmark notes that errors might not be independent, with arithmetic mistakes being more or less frequent in wr	×	0.02
The ARB benchmark observes that GPT-4 is likely to make incorrect simplifications to get to some final answer in approac	×	0.03

References

- http://arxiv.org/abs/2503.15113v1
- http://arxiv.org/abs/2406.13803v3
- http://arxiv.org/abs/2307.13692v2