Frontier Language Models on GPQA Diamond and Advanced Reasoning Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 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 v19. 7 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Qwen3 and Frontier Models on GPQA Diamond and Advanced Reasoning Benchmarks. Research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v19.

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 6.3/10.

3 Results

14 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 6.3/10.

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
Qwen3 and other frontier models were evaluated on GPQA Diamond and advanced reasoning benchmarks v8.	✓	0.35
10 claims were extracted from source literature and independently verified against retrieved documents.	✓	0.29
An automated multi-reviewer quality assessment produced a score of 9.3/10.	✓	0.32
The report is a machine-generated literature synthesis and does not constitute original research.	✓	0.35
The research goal is to determine which frontier language models achieve the highest scores on GPQA Diamond Humanity Las	✓	0.58
The report has an automated review score of 9.3/10.	✓	0.24
Full text and citation are available at Assignee Research.	✓	0.23

References

- https://openalex.org/W7161452897
- https://doi.org/10.5281/zenodo.20563498
- https://doi.org/10.5281/zenodo.20563497