Frontier Language Models on GPQA Diamond and Reasoning Benchmarks V6

Assignee Research

June 5, 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 v6. 13 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: HLE-Verified: A Systematic Verification and Structured Revision of Humanity's Last Exam. Research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v6.

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 3.7/10.

3 Results

14 papers retrieved. 13 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
The study evaluates GPT-5.2-Thinking, Gemini3-Pro-Preview, Claude-Opus4.5, Claude-Opus4.6, Grok-4.1 (fast-reasoning), De	×	0.03
All models were evaluated using the system prompt recommended by the HLE official guidelines and each model's default re	×	0.03
To reduce variance from stochastic decoding, five independent rollouts per item were run, and avg5 accuracy (average cor	×	0.03
Calibration Error (Cali Err) is computed from the model's self-reported confidence and the binary correctness label usin	×	0.02
On the Full Set, Gemini3-pro achieved an accuracy of 40.42 on Raw HLE and 48.2 on Revised HLE-Verified.	×	0.10
On the Full Set, GPT-5.2-High achieved an accuracy of 33.35 on Raw HLE and 43.3 on Revised HLE-Verified.	×	0.09
On the Full Set, Claude-Opus4.6 achieved an accuracy of 38.95 on Raw HLE and 46.8 on Revised HLE-Verified.	×	0.09
On the Revised Subset, Gemini3-pro achieved an accuracy of 74 on Raw HLE Subset and 48.93 on Revised HLE-Verified Subset	×	0.08
The Revised Subset comparison is limited to items that were edited or flagged during the verification process where at l	×	0.08
Under standard HLE benchmark evaluation, items with errors only in the rationale do not affect final scores because main	×	0.09
The methodology for comparison rules includes determining mathematical equivalence, considering alternative forms, consi	×	0.02
The LLM Judge Prompt outputs are treated as diagnostic signals rather than definitive correctness labels.	×	0.03
Stage II of the pipeline involves structured extraction, a repair prompt for an expert correction model, and a final adj	×	0.11

References

- http://arxiv.org/abs/2602.13964v3
- http://arxiv.org/abs/2505.03786v1
- http://arxiv.org/abs/2606.05405v1