Frontier Model Performance on the Humanity Last Exam Benchmark: A Multi-Study Synthesis

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Humanity Last Exam benchmark frontier model evaluation comparison. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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: Humanity Last Exam benchmark frontier model evaluation comparison.

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. 14 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	×	0.01
All models were evaluated using the system prompt recommended by the HLE official guidelines and each model's default re	×	0.04
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.	×	0.09
On the Full Set, GPT-5.2-High achieved an accuracy of 33.35 on Raw HLE and 43.3 on Revised HLE-Verified.	×	0.08
On the Full Set, Claude-Opus4.6 achieved an accuracy of 38.95 on Raw HLE and 46.8 on Revised HLE-Verified.	×	0.08
On the Revised Subset, Gemini3-pro achieved an accuracy of 48.93 on Revised HLE-Verified compared to 18.99 on Raw HLE.	×	0.07
On the Revised Subset, GPT-5.2-High achieved an accuracy of 52.48 on Revised HLE-Verified compared to 14.44 on Raw HLE.	×	0.07
The Revised Subset includes only items for which at least one of the problem or answer fields was revised.	×	0.08
Under standard HLE benchmark evaluation, items with errors only in the rationale do not affect final scores because main	×	0.09
The comparison rules for evaluation determine mathematical equivalence, consider alternative forms, consider answer orde	×	0.02
The LLM Judge Prompt outputs are treated as diagnostic signals rather than definitive correctness labels.	×	0.02
Stage II Structured Extraction prompts require extracting the final answer (preferably from \boxed{}) and core reasoning	×	0.07

References

- http://arxiv.org/abs/2602.13964v3
- http://arxiv.org/abs/2603.04454v1
- http://arxiv.org/abs/2501.14249v10