Frontier Language Models on GPQA Diamond and HLCE Benchmark Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 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 v14. 14 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Humanity's Last Code Exam: Can Advanced LLMs Conquer Human's Hardest Code Competition?. Research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v14.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.6/10.

3 Results

13 papers retrieved. 14 claims extracted; 1 independently verified. Quality review score: 4.6/10.

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
Models such as o3-mini, Gemini-2.5-pro, and o4-mini are capable of achieving medal-level performance in ICPC competition	×	0.09
These models still underperform compared to human medalists in the IOI.	×	0.03
HLCE is a novel benchmark comprising 235 competitive programming problems from IOI and ICPC World Finals (2010-2024).	✓	0.16
HLCE features both standard and interactive programming challenges that significantly exceed the difficulty of existing	×	0.10
Comprehensive evaluations on 12 leading LLMs show that even the most advanced LLMs achieve only 15.1% and 11.4% pass@1 r	×	0.14
A novel self-recognition task is proposed to measure models' abilities to recognize the correctness of their own generat	×	0.11
Test-time scaling laws on HLCE demonstrate that current LLMs have not yet reached their performance upper bounds.	×	0.12
Comparative analyses with top human competitors reveal the gap between advanced LLMs and competition medalists.	×	0.08
Models such as Codex, StarCoder, and CodeLlama have demonstrated remarkable proficiency in understanding and generating	×	0.05
Instruction-tuned models like ChatGPT and Claude have further pushed the boundaries of code generation capabilities.	×	0.07
Reasoning-enhanced models have made substantial progress in the code generation domain.	×	0.10
Claude-3.7 Sonnet achieves 62.3% accuracy on SWE-bench Verified, outperforming o3-mini(high) (49.3%).	×	0.03
There is a significant performance gap of models between IOI and ICPC World Finals competitions.	×	0.08
O4-mini(high) achieves a pass@1 rate of 25.21%.	×	0.07

References

- http://arxiv.org/abs/2606.05405v1
- http://arxiv.org/abs/2407.04973v1
- http://arxiv.org/abs/2506.12713v2