Frontier Language Models on GPQA Diamond and High-Difficulty Reasoning Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 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 v18. 15 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.4/10. 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 v18.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.4/10.

3 Results

16 papers retrieved. 15 claims extracted; 2 independently verified. Quality review score: 5.4/10.

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
Current benchmarks suffer from several critical limitations and do not fully capture the advanced reasoning and code gen	×	0.11
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, highlighting the persisting challenges in truly	×	0.06
HLCE is a novel benchmark comprising 235 competitive programming problems from IOI and ICPC World Finals (2010-2024), fe	✓	0.17
Even the most advanced LLMs achieve only 15.1% and 11.4% pass@1 rates on HLCE.	✓	0.18
A novel self-recognition task was proposed to measure models' abilities to recognize the correctness of their own genera	×	0.11
Test-time scaling laws on HLCE demonstrate that current LLMs have not yet reached their performance upper bounds and hig	×	0.13
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, all	×	0.05
Reasoning-enhanced models have made substantial progress in the code generation domain, with claude-3.7 achieving 62.3%	×	0.07
Claude-3.7-thinking model underperforms compared to non-reasoning models and shows notably weaker results on IOI problem	×	0.05
Claude-3.7-thinking achieves 0% pass rates on IOI problems, representing a significant decline despite its enhanced reas	×	0.05
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/2501.14249v10
- http://arxiv.org/abs/2506.12713v2
- http://arxiv.org/abs/2604.23730v1