# Frontier Language Models on GPQA Diamond and Reasoning Benchmarks V11

Assignee Research

June 6, 2026

**Abstract**

This report synthesises findings from 4 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 v11. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Agents' Last Exam. Research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v11.

## 2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

## 3 Results

4 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.8/10.

## 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 benchmark requires workflows to match real professional practice and use software that domain experts would actually | × | 0.05 |
| A task should be an end-to-end deliverable that would take an expert substantial time, rather than only a few UI operati | × | 0.02 |
| The output should admit deterministic checking or an unambiguous rubric tied to observable artifacts. | × | 0.00 |
| The benchmark taxonomy is grounded in SOC 2018 and O*NET, clustering occupations with similar software-mediated workflow | × | 0.09 |
| The union of 16 major prior benchmarks leaves 13 of 55 subdomains entirely uncovered. | × | 0.03 |
| The task construction pipeline involves expert outreach, task submission, editing, auto-review, implementation, executio | × | 0.02 |
| The typical GCUA harness architecture includes a main agent loop, context building, LLM inference, action decision, tool | × | 0.01 |
| Early agents revolved around a thin reasoning loop in the style of ReAct, while contemporary harnesses share a richer ma | × | 0.03 |

## References


- http://arxiv.org/abs/2606.05405v1
- http://arxiv.org/abs/2508.15478v2
- http://arxiv.org/abs/2510.22758v2