Qwen3 and Frontier Models on GPQA Diamond and Advanced Reasoning Benchmarks

Assignee Research

June 5, 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 v8. 10 claims were extracted from source literature; 10 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 9.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Qwen3 Technical Report. Research question: Which frontier language models achieve highest scores on GPQA Diamond Humanity Last Exam and difficult reasoning benchmarks v8.

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 9.3/10.

3 Results

13 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 9.3/10.

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
Qwen3 is the latest version of the Qwen model family.	✓	0.20
Qwen3 comprises a series of large language models (LLMs) designed to advance performance, efficiency, and multilingual c	✓	0.30
The Qwen3 series includes models of both dense and Mixture-of-Expert (MoE) architectures, with parameter scales ranging	✓	0.32
Qwen3 integrates thinking mode (for complex, multi-step reasoning) and non-thinking mode (for rapid, context-driven resp	✓	0.32
Qwen3 eliminates the need to switch between different models–such as chat-optimized models (e.g., GPT-4o) and dedicated	✓	0.42
Qwen3 introduces a thinking budget mechanism, allowing users to allocate computational resources adaptively during infer	✓	0.38
Qwen3 leverages the knowledge from the flagship models to significantly reduce the computational resources required to b	✓	0.33
Empirical evaluations demonstrate that Qwen3 achieves state-of-the-art results across diverse benchmarks, including task	✓	0.34
Qwen3 is competitive against larger MoE models and proprietary models.	✓	0.21
Compared to its predecessor Qwen2.5, Qwen3 expands multilingual support.	✓	0.19

References

- https://doi.org/10.4230/oasics.icpec.2025.4
- https://doi.org/10.48550/arxiv.2412.19437
- https://doi.org/10.48550/arxiv.2505.09388