Claude-3-Opus Benchmark Performance Across Reasoning Mathematics and Language Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of Claude-3-Opus on reasoning mathematics coding and language understanding tasks. 13 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Qwen2 Technical Report. Research question: What are the benchmark performance scores of Claude-3-Opus on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

7 papers retrieved. 13 claims extracted; 3 independently verified. Quality review score: 6.5/10.

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 Qwen2 series includes models with parameter counts ranging from 0.5 billion to 72 billion.	×	0.09
The Qwen2 series includes both dense models and a Mixture-of-Experts model.	✓	0.16
The Qwen2-72B base model achieves a score of 84.2 on the MMLU benchmark.	×	0.10
The Qwen2-72B base model achieves a score of 37.9 on the GPQA benchmark.	×	0.10
The Qwen2-72B base model achieves a score of 64.6 on the HumanEval benchmark.	×	0.10
The Qwen2-72B base model achieves a score of 89.5 on the GSM8K benchmark.	×	0.10
The Qwen2-72B base model achieves a score of 82.4 on the BBH benchmark.	×	0.10
The Qwen2-72B-Instruct model achieves a score of 9.1 on the MT-Bench benchmark.	×	0.10
The Qwen2-72B-Instruct model achieves a score of 48.1 on the Arena-Hard benchmark.	×	0.13
The Qwen2-72B-Instruct model achieves a score of 35.7 on the LiveCodeBench benchmark.	×	0.11
Qwen2 models support approximately 30 languages.	×	0.11
Qwen2 model weights are available on Hugging Face and ModelScope.	✓	0.18
Supplementary materials including example code for Qwen2 are available on GitHub.	✓	0.16

References

- https://doi.org/10.48550/arxiv.2406.11931
- https://doi.org/10.48550/arxiv.2407.10671
- https://doi.org/10.48550/arxiv.2404.14219