Qwen2.5 Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 19 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of Qwen2.5 on reasoning mathematics coding and language understanding tasks. 13 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: The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. Research question: What are the benchmark performance scores of Qwen2.5 on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 19 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

19 papers retrieved. 13 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 experiments were conducted on reasoning models and their non-reasoning counterparts, specifically Claude 3.7 Sonnet	×	0.06
Claude 3.7 Sonnet and DeepSeek-R1/V3 were selected because they allow access to thinking traces, unlike OpenAI's o-serie	×	0.04
Results on the o3-mini model were reported for experiments focused solely on final accuracy.	×	0.04
A maximum token budget of 64k was allowed for Claude 3.7 Sonnet models.	×	0.05
A maximum length of up to 64k tokens was allowed for DeepSeek-R1/V3 models on local servers.	×	0.03
For each puzzle instance and complexity level, 25 samples per model were analyzed.	×	0.03
A filtering process was applied to ensure analyzed samples followed the requested response format, including sequence of	×	0.03
Problem complexity was varied by manipulating problem size N, representing disk count, checker count, block count, or cr	×	0.04
Figure 4 shows the upper bound performance capabilities (pass@k) of model pairs under equivalent inference token compute	×	0.06
In the first regime of low problem complexity, non-thinking models achieve performance comparable to or better than thin	×	0.07
In the second regime of medium complexity, the performance gap between thinking and non-thinking models increases as the	×	0.04
In the third regime of high problem complexity, the performance of both thinking and non-thinking models collapses to ze	×	0.07
Thinking models delay performance collapse compared to non-thinking models in high complexity regimes.	×	0.07

References

- https://arxiv.org/abs/2509.06266

- https://arxiv.org/abs/2505.19914
- https://arxiv.org/abs/2506.06941