Extended Thinking Time Improves Language Model Accuracy in Competition-Level Mathematics

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does extended thinking time affect language model accuracy on competition-level mathematics v10. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking Large Language Models for Calculus Problem-Solving: A Comparative Analysis. Research question: How does extended thinking time affect language model accuracy on competition-level mathematics v10.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

15 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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 first derivative of () = -2 + 3 - 4 by the limit process is () = -4 + 3.	×	0.00
Chat GPT 4o, Copilot Pro, Gemini Advanced, Claude Pro, and Meta AI all provided correct answers and showed all detailed	✓	0.20
All five language models achieved perfect scores (100/100) on problems involving differentiation using the limit definit	×	0.07
The slope, m, at = -2 in () = -3 - 5 + 6 is part of the tangent line equation = 7 + 18.	×	0.00
Chat GPT 4o, Copilot Pro, Gemini Advanced, Claude Pro, and Meta AI all provided correct answers for finding the slope an	✓	0.21
Chat GPT 4o successfully solved most problems created by the five LLMs, with a success rate of 98% (98 out of 100).	×	0.13
The total score for Chat GPT 4o across all problems is 100/100 in the first table and 98/100 in the second table.	×	0.08
The total score for Copilot Pro across all problems is 100/100 in the first table and 61/100 in the second table.	×	0.06
The total score for Gemini Advanced across all problems is 100/100 in the first table and 47/100 in the second table.	×	0.05
The total score for Claude Pro across all problems is 100/100 in the first table and 68/100 in the second table.	×	0.07
The total score for Meta AI across all problems is 100/100 in the first table and 65/100 in the second table.	×	0.05
The overall total score across all models and problems is 500/500 in the first table and 293/500 in the second table.	×	0.02

References

- http://arxiv.org/abs/2504.13187v1
- http://arxiv.org/abs/2604.25926v1

- http://arxiv.org/abs/2503.15113v1