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

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does extended thinking time affect language model accuracy on competition-level mathematics v6. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. Research question: How does extended thinking time affect language model accuracy on competition-level mathematics v6.

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 4.0/10.

3 Results

4 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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 VSI-Bench benchmark includes tasks such as counting objects, measuring distances, determining room sizes, and route	×	0.03
The VSI-Bench benchmark uses datasets like ScanNet, ScanNet++, and ARKitScenes for video scans and 3D annotations.	×	0.05
The benchmark curation pipeline involves dataset unification, QA pair generation through templates and human annotation,	×	0.01
Human-level performance on the VSI-Bench is compared against models like LLaVA-Video-72B, Gemini-1.5 Pro, LLaVA-OneVisio	×	0.05
The error breakdown by task shows varying levels of accuracy across different models and tasks.	×	0.03
The cognitive maps generated by MLLMs are visualized and compared to ground truth in Figure 9.	×	0.06
The accuracy of MLLM's predicted cognitive maps decreases with increasing distance, as shown in Figure 10.	×	0.04
The benchmark includes tasks like counting objects, measuring distances, determining room sizes, and route planning.	×	0.01
The benchmark uses datasets like ScanNet, ScanNet++, and ARKitScenes for video scans and 3D annotations.	×	0.03
The benchmark curation pipeline involves dataset unification, QA pair generation through templates and human annotation,	×	0.01
Human-level performance on the VSI-Bench is compared against models like LLaVA-Video-72B, Gemini-1.5 Pro, LLaVA-OneVisio	×	0.05
The error breakdown by task shows varying levels of accuracy across different models and tasks.	×	0.03
The cognitive maps generated by MLLMs are visualized and compared to ground truth in Figure 9.	×	0.06
The accuracy of MLLM's predicted cognitive maps decreases with increasing distance, as shown in Figure 10.	×	0.04

References

- http://arxiv.org/abs/2307.02422v2
- http://arxiv.org/abs/2503.21676v2
- http://arxiv.org/abs/2412.14171v2