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

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does extended thinking time affect language model accuracy on competition-level mathematics v17. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Omni-MATH: A Universal Olympiad Level Mathematic Benchmark For Large Language Models. Research question: How does extended thinking time affect language model accuracy on competition-level mathematics v17.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

12 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
Omni-MATH benchmark consists of 4428 problems.	×	0.09
Omni-MATH uses GPT-4o and OmniJudge for evaluation.	×	0.05
Omni-MATH includes problems from various competitions like IMO and IMC.	×	0.05
MathPix is used to convert PDF documents into LaTeX format.	×	0.01
AoPS Wiki and AoPS forum are used as data sources for Omni-MATH.	×	0.09
Omni-MATH includes a consistency check for solutions from AoPS forum.	×	0.08
Omni-MATH classifies competitions into five levels based on difficulty, scale, and prestige.	×	0.05
Omni-MATH uses a rule-based filtering and difficulty tagging process.	×	0.03
Omni-MATH includes a human check process for problem and solution pairs.	×	0.06

References

- http://arxiv.org/abs/2502.07154v4
- http://arxiv.org/abs/2410.07985v3
- http://arxiv.org/abs/2501.14275v2