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

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 19 peer-reviewed papers addressing the following research question: How does extended thinking time affect language model accuracy on competition-level mathematics v9. 14 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PARAMANU-GANITA: Can Small Math Language Models Rival with Large Language Models on Mathematical Reasoning?. Research question: How does extended thinking time affect language model accuracy on competition-level mathematics v9.

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 6.9/10.

3 Results

19 papers retrieved. 14 claims extracted; 9 independently verified. Quality review score: 6.9/10.

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
LLEMMA 7B was trained on 256 A100 40GB GPUs for roughly 23,000 A100 training hours.	×	0.05
The study defines a Small Language Model (SLM) as containing less than 300M parameters.	×	0.06
The decontamination process removed around 170,346,325 words from the pretraining corpus.	×	0.04
The final pretraining corpus contains 5,578,762,486 words.	×	0.04
PARAMANU-GANITA is a 208 million-parameter decoder-only Auto Regressive SLM.	✓	0.23
PARAMANU-GANITA was pretrained from scratch on 31.5 billion tokens.	✓	0.17
PARAMANU-GANITA training required 170 A100 hours.	×	0.12
PARAMANU-GANITA was trained using a context size of 4096.	✓	0.17
PARAMANU-GANITA is 34 times smaller than 7B LLMs.	✓	0.19
On the GSM8K test accuracy metric, PARAMANU-GANITA outperforms generalist LLMs by approximately 30 percentage points.	✓	0.22
On the GSM8K test accuracy metric, PARAMANU-GANITA outperforms math-specialised LLMs by 3-23 percentage points.	✓	0.22
On the MATH benchmark, PARAMANU-GANITA outperformed various models by 6-8 percentage points.	✓	0.22
On benchmarks including LogiQA, MMLU, and AGIEVAL, PARAMANU-GANITA outperformed other models by 1-4 percentage points.	✓	0.16
The PARAMANU-GANITA model is available at https://huggingface.co/gyanai/paramanu-ganita-208M-hf.	✓	0.24

References

- https://www.semanticscholar.org/paper/0c0f38b95741010c386e95270bfa2d1ac0726b1a
- http://arxiv.org/abs/2502.07154v4

- https://arxiv.org/abs/2404.14395