Test-Time Compute Scaling and Accuracy Trade-offs in Reasoning Benchmarks

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Test-time compute scaling reasoning benchmark performance accuracy tradeoff. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Do LLMs Overthink Basic Math Reasoning? Benchmarking the Accuracy-Efficiency Tradeoff in Language Models. Research question: Test-time compute scaling reasoning benchmark performance accuracy tradeoff.

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 2.4/10.

3 Results

15 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 2.4/10.

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 study evaluated 53 models on the LLMTHINKBENCH benchmark across basic mathematical tasks, quantization settings, and	×	0.10
Models achieving over 95% accuracy on GSM8K scored below 75% on the authors' basic math tasks within the Qwen2.5 family.	×	0.10
Phi-4 achieves 78.92% accuracy using an average of approximately 378.6 tokens.	×	0.02
Phi-4-reasoning scores 72.23% accuracy while using approximately 6066.2 tokens.	×	0.05
Constraining Phi-4-reasoning to a 1,024-token budget reduces its accuracy to 53.48%.	×	0.04
The accuracy drop for Phi-4-reasoning when constrained to 1,024 tokens represents an 18.75 point absolute loss.	×	0.03
For the Qwen3 family, accuracy collapses when generation is constrained below approximately 512 tokens.	×	0.03
The LLMTHINKBENCH task space comprises 14 deterministic arithmetic operations.	×	0.03
Each task in the benchmark is defined as a mapping from an input domain to an output domain with a computationally deter	×	0.04
A public leaderboard for the benchmark results is hosted at ctrl-gaurav.github.io/LLMThinkBench.	×	0.05
Quantization often preserves basic arithmetic ability for large models.	×	0.04
Non-reasoning models remain relatively stable in performance when output length is restricted, whereas reasoning models	×	0.06

References

- http://arxiv.org/abs/2507.04023v3

- http://arxiv.org/abs/2504.13171v1
- http://arxiv.org/abs/2506.23424v1