Test-Time Compute Scaling vs. Inference Efficiency Techniques in Medical Reasoning Benchmarks

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does test-time compute scaling compare to other inference efficiency techniques (e.g., distillation, quantization) in improving reasoning performance on medical question-answering benchmarks like. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Sleep-time Compute: Beyond Inference Scaling at Test-time. Research question: How does test-time compute scaling compare to other inference efficiency techniques (e.g., distillation, quantization) in improving reasoning performance on medical question-answering benchmarks like USMLE or MedQA, measured by accuracy and latency trade-offs?.

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 4.5/10.

3 Results

4 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
Adding sleep-time compute shifts the Pareto frontier of test-time compute versus accuracy beyond the baseline curve for	✓	0.22
At lower test-time budgets, sleep-time compute achieves performance comparable to the baseline while using 5$\times$ fewer test	×	0.12
At high test-time compute budgets, the test-time compute only baseline slightly outperforms sleep-time compute.	✓	0.17
The study uses GPT-4o and GPT-4o-mini as non-reasoning models for experiments on Stateful GSM-Symbolic.	×	0.11
Test-time compute was varied for non-reasoning models by constructing prompts that instruct different amounts of verbosi	×	0.05
The experiments for non-reasoning models used a temperature of 0 for generation.	×	0.03
For reasoning models o1, o3-mini, and Claude Sonnet 3.7 on Stateful AIME, test-time compute scaling was based on API ava	×	0.13
Deepseek-R1 API does not provide a method to control test-time compute.	×	0.08
Budget forcing and extension prompts from Muennighoff et al. (2025) were applied to Deepseek-R1 to control test-time com	×	0.05
Results for o1, o3-mini, and R1 on Stateful AIME were averaged over 3 runs.	×	0.06
In the provided example context, a juggler can juggle 800 balls, 1/4 are tennis balls, 1/2 of the tennis balls are indig	×	0.01
The correct answer to the question 'How many marked indigo tennis balls are there?' based on the provided context is 10.	×	0.04

References

- http://arxiv.org/abs/2504.13171v1
- http://arxiv.org/abs/2501.05464v2
- http://arxiv.org/abs/2209.10326v2