Test-Time Compute Scaling and Language Model Performance on Reasoning Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does test-time compute scaling improve language model performance on reasoning benchmarks v9. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: The Cost of Dynamic Reasoning: Demystifying AI Agents and Test-Time Scaling from an AI Infrastructure Perspective. Research question: How does test-time compute scaling improve language model performance on reasoning benchmarks v9.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

16 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
HotpotQA is a question-answering benchmark that assesses the agent's ability to accurately retrieve relevant evidence to	×	0.05
WebShop is a web-shopping benchmark where agents find the best-fit item.	×	0.03
MATH is a benchmark for math problem solving.	×	0.03
HumanEval is a benchmark for programming tasks.	×	0.04
Reflection allows agents to evaluate past decisions and revise strategies accordingly.	×	0.05
LATS leverages Monte Carlo Tree Search to simulate multiple branches of reasoning and action.	×	0.02
LLMCompiler incorporates structured multi-step planning and streaming for asynchronous task execution to minimize latenc	×	0.04
The original version of LATS executes LLM inference and tool invocation sequentially, aggravating end-to-end latency.	×	0.02
The 95%-ile latency for ShareGPT (Chatbot) is 9.7s.	×	0.01
The 95%-ile latency for HotpotQA (ReAct) is 20.7s.	×	0.01
The 95%-ile latency for WebShop (ReAct) is 50.8s.	×	0.02

References

- https://www.semanticscholar.org/paper/4b1a7908f5de651a36d3e1892114da9d47cd9fab
- https://arxiv.org/abs/2502.05171
- https://arxiv.org/abs/2506.04301