Scaling Laws of Chain-of-Thought Reasoning in Large Language Models

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What are the scaling laws for chain-of-thought reasoning in large language models v11. 20 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Reasoning Effort and Problem Complexity: A Scaling Analysis in LLMs. Research question: What are the scaling laws for chain-of-thought reasoning in large language models v11.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

13 papers retrieved. 20 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 success rate is assessed as a binary measure: whether the LLM successfully outputs a valid solution to the Tents puz	×	0.05
Problem complexity is quantified by the problem size, defined as the product of the grid dimensions (rows $\times$ columns).	×	0.12
Reasoning effort is measured by the total number of tokens generated by the LLMs to produce the final answer, including	×	0.05
The hypothesis is a linear scaling relationship between problem size and reasoning effort.	×	0.14
The goodness of fit is quantified using the R2 metric, where scores closer to 1 indicate that a larger proportion of the	×	0.07
No model solved problems larger than size 100.	×	0.02
o3-mini achieves the highest success rate, followed by DeepSeek R1 and Gemini 2.0 Flash Thinking.	×	0.03
Qwen/QwQ-32B-Preview struggles with problem instances larger than size 20.	×	0.03
For DeepSeek R1 and o3-mini, there is a roughly linear increase in reasoning effort with problem size.	×	0.14
The slopes of the linear fits for DeepSeek R1 and o3-mini are very similar, suggesting comparable scaling behavior in re	×	0.11
DeepSeek R1 consistently uses more tokens than o3-mini.	×	0.03
Gemini 2.0 Flash Thinking is excluded due to unknown number of thinking tokens.	×	0.03
The R2 value for DeepSeek R1 is 0.667.	×	0.03
The R2 value for o3-mini is 0.833.	×	0.00
The R2 value for Qwen/QwQ-32B-Preview is 0.087.	×	0.00
The R2 value for low reasoning effort is 0.489.	×	0.07
The R2 value for medium reasoning effort is 0.833.	×	0.07
The R2 value for high reasoning effort is 0.813.	×	0.07
The R2 value for easy difficulty is 0.468.	×	0.02
The R2 value for tricky difficulty is 0.502.	×	0.02

References

- http://arxiv.org/abs/2503.09567v5
- http://arxiv.org/abs/2503.15113v1
- http://arxiv.org/abs/2410.03595v1