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

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What are the scaling laws for chain-of-thought reasoning in large language models. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Mobile-MMLU: A Mobile Intelligence Language Understanding Benchmark. Research question: What are the scaling laws for chain-of-thought reasoning in large language models.

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 4.3/10.

3 Results

16 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.3/10.

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
The study evaluates language models ranging from 1B to 9B parameters on Mobile-MMLU and Mobile-MMLU-Pro.	×	0.12
The evaluation framework uses lm-eval-harness to assess model performance.	×	0.03
Mobile-MMLU and Mobile-MMLU-Pro consist entirely of multiple-choice questions.	✓	0.17
A model with 65.4% performance on MMLU achieved 68.1% on Mobile-MMLU.	×	0.08
The performance spread on the MMLU benchmark ranges from 45.9% to 71.8%.	×	0.06
The performance spread on the MMLU-Pro benchmark ranges from 7.5% to 36.5%.	×	0.11
The performance spread on the Mobile-MMLU benchmark ranges from 34.5% to 75.0%.	×	0.08
Qwen2.5-3B-Instruct achieves 68.1% accuracy on Mobile-MMLU.	×	0.06
Llama-3.2-3B-Instruct scores 50.2% accuracy on Mobile-MMLU.	×	0.06
Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct have comparable parameter counts in the 3B range.	×	0.01
Phi-3.5-mini-instruct achieved a score of 63.7 on a specific evaluation metric.	×	0.02
The mean score reported in Table (p14) is 46.84.	×	0.03

References

- http://arxiv.org/abs/2409.17143v1
- http://arxiv.org/abs/2503.20786v1
- http://arxiv.org/abs/2503.09567v5