Scaling Laws and Performance-Efficiency Trade-offs in Language Model Reasoning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the effect of model size on language model performance on logical reasoning tasks v18. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Task-Specific Efficiency Analysis: When Small Language Models Outperform Large Language Models. Research question: What is the effect of model size on language model performance on logical reasoning tasks v18.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.4/10.

3 Results

10 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 3.4/10.

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 evaluates 16 representative open-source language models ranging from 0.5B to 72B parameters.	×	0.09
The IMDB Movie Reviews dataset benchmark consists of 1,000 movie reviews for binary sentiment classification.	×	0.01
The HellaSwag benchmark evaluates commonsense reasoning across 10,042 examples using log-likelihood scoring.	×	0.04
The ARC-Easy benchmark consists of 2,376 multiple-choice questions regarding elementary scientific knowledge.	×	0.09
The SQuAD 2.0 benchmark includes 11,873 examples requiring answer generation or unanswerable question recognition.	×	0.03
The GSM8K benchmark contains 1,319 grade-school problems requiring multi-step mathematical reasoning.	×	0.05
The Performance-Efficiency Ratio (PER) metric combines accuracy, throughput, memory usage, and latency.	✓	0.21
The PER metric is calculated using the geometric mean of four dimensions after min-max normalization to the [0, 1] range	×	0.07
On the GSM8K benchmark, the Llama-3.1-8B model achieved an accuracy of 0.8097.	×	0.04
On the GSM8K benchmark, the Qwen2.5-0.5B model achieved a throughput of 7927 tokens/s.	×	0.02
On the GSM8K benchmark, the Llama-3.1-8B model has a latency of 3.74 ms/token.	×	0.02
On the HellaSwag benchmark, the Qwen2.5-72B model achieved an accuracy of 0.64.	×	0.02
On the HellaSwag benchmark, the Llama-3.1-8B model has a latency of 132 ms/sample.	×	0.02
The Qwen2.5-72B model achieved a PER score of 0.1957 on the GSM8K benchmark.	×	0.02
The Vicuna-13B model achieved a PER score of 0 on the GSM8K benchmark.	×	0.02

References

- http://arxiv.org/abs/2603.21389v1
- http://arxiv.org/abs/2601.01982v1
- http://arxiv.org/abs/2407.04973v1