Scaling Laws of InternLM Performance in Mathematics and Language Understanding Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of model size scaling (e.g., 7B vs. 20B) on InternLM's performance in mathematics and language understanding tasks, as measured by benchmarks like MMLU or HELM, and how does this. 11 claims were extracted from source literature; 1 was 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: Mobile-MMLU: A Mobile Intelligence Language Understanding Benchmark. Research question: What is the impact of model size scaling (e.g., 7B vs. 20B) on InternLM's performance in mathematics and language understanding tasks, as measured by benchmarks like MMLU or HELM, and how does this compare to scaling trends in other LLMs?.

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 4.5/10.

3 Results

13 papers retrieved. 11 claims extracted; 1 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
Gemma-2-9B-it, Qwen2.5-7B-instruct, Llama-3.1-8B-instruct, Qwen2.5-3B-instruct, Phi-3.5-mini-instruct, Llama-3.2-3B-inst	×	0.07
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.	✓	0.16
Phi-3.5-mini-instruct achieves 68.1% accuracy on Mobile-MMLU.	×	0.05
The performance spread on MMLU ranges from 45.9% to 71.8%.	×	0.02
The performance spread on MMLU-Pro ranges from 7.5% to 36.5%.	×	0.05
The performance spread on Mobile-MMLU ranges from 34.5% to 75.0%.	×	0.05
Qwen2.5-3B-Instruct achieves 68.1% accuracy on Mobile-MMLU.	×	0.05
Llama-3.2-3B-Instruct scores 50.2% accuracy on Mobile-MMLU.	×	0.05
The mean accuracy of the models on Mobile-MMLU is 46.84%.	×	0.05
Phi-3.5-mini-instruct achieves 63.7% accuracy on Mobile-MMLU.	×	0.05

References

- http://arxiv.org/abs/2403.09832v1
- http://arxiv.org/abs/2401.16420v1
- http://arxiv.org/abs/2503.20786v1