

The Benchmark Performance Scores Of Llama-4 On Reasoning Mathematics Coding And Language Understanding Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of Llama-4 on reasoning mathematics coding and language understanding tasks. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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 benchmark performance scores of Llama-4 on reasoning mathematics coding and language understanding tasks.

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 4.2/10.

3 Results

16 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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	×	0.05
The evaluation framework used is lm-eval-harness.	×	0.02
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 despite its modest performance on MMLU (65.4%).	×	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%.	×	0.06
The performance spread on Mobile-MMLU ranges from 34.5% to 75.0%.	×	0.06
Qwen2.5-3B-Instruct achieves 68.1% accuracy on Mobile-MMLU.	×	0.05
Llama-3.2-3B-Instruct scores 50.2% on Mobile-MMLU.	×	0.05
The mean accuracy of the models on Mobile-MMLU is 46.84%.	×	0.06
Phi-3.5-mini-instruct achieves 63.7% accuracy on Mobile-MMLU.	×	0.05

References

- <http://arxiv.org/abs/2509.25160v1>
- <http://arxiv.org/abs/2210.09261v1>
- <http://arxiv.org/abs/2503.20786v1>