Claude-Sonnet-3.5 Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of Claude-Sonet-3.5 on reasoning mathematics coding and language understanding tasks. 9 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: Mobile-MMLU: A Mobile Intelligence Language Understanding Benchmark. Research question: What are the benchmark performance scores of Claude-Sonet-3.5 on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 9 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 model suite includes Gemma-2-9B-it, Qwen2.5-7B-instruct, Llama-3.1-8B-instruct, Qwen2.5-3B-instruct, Phi-3.5-mini-in	×	0.01
The evaluation framework uses lm-eval-harness to assess the performance of the models.	×	0.04
Phi-3.5-mini-instruct achieves 63.7% accuracy on Mobile-MMLU.	×	0.04
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.	×	0.04
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%.	×	0.08
The performance spread on Mobile-MMLU ranges from 34.5% to 75.0%.	×	0.06
The mean accuracy on Mobile-MMLU is 46.84%.	×	0.05

References

- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2312.17080v4
- http://arxiv.org/abs/2503.20786v1