LongLLaVA-9B Benchmark Performance Across Reasoning Mathematics Coding and Language 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 LongLLaVA-9B on reasoning mathematics coding and language understanding tasks. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. Research question: What are the benchmark performance scores of LongLLaVA-9B 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 3.3/10.

3 Results

16 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
GPT-4o achieves an overall accuracy of 72.6% on the MMLU-Pro benchmark.	×	0.06
Llama-3-70B-Instruct achieves an accuracy of 56.2% on the MMLU-Pro benchmark.	×	0.06
Top-tier closed-source models outperform open-source models on the MMLU-Pro benchmark.	×	0.09
GPT-4o achieves over 70% accuracy in Math and Physics subjects on the MMLU-Pro benchmark.	×	0.06
Mistral-7B-v0.1 achieves just over 20% accuracy in Math and Physics subjects on the MMLU-Pro benchmark.	×	0.06
Models generally show a higher performance floor in knowledge-intensive subjects (History, Psychology) compared to reaso	×	0.08
DeepSeek-V2-Chat underperforms relative to its peers in History and Psychology subjects on the MMLU-Pro benchmark.	×	0.04
Engineering and Law subjects consistently scored lower than other subjects among the 14 evaluated in MMLU-Pro.	×	0.03
Lower scores in Engineering on MMLU-Pro are largely due to new questions sourced from the STEM Website requiring complex	×	0.06
Law scores on MMLU-Pro suffer because questions include more intricate details and additional options necessitating deep	×	0.07
An error analysis was conducted on 120 randomly selected erroneous predictions made by GPT-4o on the MMLU-Pro benchmark.	×	0.05
Phi-3-medium-4k-instruct has 14B parameters.	×	0.02
Phi-3-mini-4k-instruct has 3.8B parameters.	×	0.02

References

- http://arxiv.org/abs/2406.01574v6
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2503.20786v1