LLaMA-1B Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of LLaMA-1B on reasoning mathematics coding and language understanding tasks. 14 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: EchoMind: An Interrelated Multi-level Benchmark for Evaluating Empathetic Speech Language Models. Research question: What are the benchmark performance scores of LLaMA-1B on reasoning mathematics coding and language understanding tasks.

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 4.2/10.

3 Results

13 papers retrieved. 14 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
EchoMind is an interrelated multi-level benchmark for evaluating empathetic speech language models.	✓	0.32
EchoMind evaluates both text and audio inputs and outputs.	×	0.02
EchoMind includes tasks for understanding, reasoning, conversation, content, and voice.	×	0.08
EchoMind supports multi-level evaluation (M).	×	0.08
EchoMind includes metrics for WER, SemSim, Acc, NISQA, DNMOS, EmoAlign, and VES.	×	0.02
GPT-4o-Audio has a WER of 2.93, SemSim of 99.18, and Acc of 64.29.	×	0.01
GPT-4o-Audio has a NISQA of 97.58 and DNMOS of 65.16.	×	0.02
GPT-4o-Audio has an EmoAlign of 41.20 and VES of 5.39.	×	0.01
GPT-4o-Audio has a Text-CCtxFit score of 3.93 and Text-CRespNat score of 4.21.	×	0.02
GPT-4o-Audio has a Text-CColloqDeg score of 4.28 and Text-CSpeechRel score of 3.06.	×	0.02
GPT-4o-Audio has an Audio-VES score of 3.81 and Audio-Quality score of 4.23.	×	0.02
GPT-4o-Audio has a NISQA score of 4.60 and Human NISQA score of 4.57.	×	0.02
GPT-4o-Audio has a CCtxFit score of 4.00 (+0.14) and CSpeechRel score of 3.68 (+0.76).	×	0.02
GPT-4o-Audio has a VES score of 3.75 (+0.51).	×	0.03

References

- `http://arxiv.org/abs/2510.22758v2`

- http://arxiv.org/abs/2601.01982v1
- http://arxiv.org/abs/2510.00071v2