Open-Llama-3B Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 18 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of Open-Llama-3B on reasoning mathematics coding and language understanding tasks. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Does your data spark joy? Performance gains from domain upsampling at the end of training. Research question: What are the benchmark performance scores of Open-Llama-3B on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 18 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

18 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
The Gauntlet v0.3 aggregates scores on benchmarks across 6 categories.	×	0.02
The study uses an inverse square root learning schedule similar to Zhai et al., 2022.	×	0.08
On the MMLU benchmark, the 7B models trained with the specified data mix have errors at or below the error vs. FLOP scal	×	0.07
On benchmarks other than MMLU (GSM8K, HumanEval, Gauntlet v0.3 Core Average), the 7B models trained with the specified d	×	0.08
The baseline data mix groups publicly-available datasets into four categories: Large-Scale Common Crawl, Small-Scale Com	×	0.11
The baseline training duration is set to 1 trillion tokens.	×	0.10
In the baseline data mix, Small-Scale Common Crawl and Domain Specific data are sampled for 0.5 epochs.	×	0.11
In the baseline data mix, Code data is sampled for 1 epoch.	×	0.04
Initial experiments indicated that a code data percentage around 20% boosted programming and reasoning ability without n	×	0.06
GSM8K scores improve monotonically as the duration of domain upsampling is increased.	×	0.12
An experiment was conducted applying domain upsampling for the last 10% of the training duration while removing math-rel	×	0.12

References

- http://arxiv.org/abs/2509.25160v1

- https://arxiv.org/abs/2406.03476
- http://arxiv.org/abs/2210.09261v1