Mistral-7B-Instruct-v0.2 vs. Llama-2-7B and Gemma-7B on MathOdyssey Calculus Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the exact match accuracy of Mistral-7B-Instruct-v0.2 compare to Llama-2-7B and Gemma-7B on university-level calculus problems in the MathOdyssey dataset. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MathOdyssey: Benchmarking Mathematical Problem-Solving Skills in Large Language Models Using Odyssey Math Data. Research question: How does the exact match accuracy of Mistral-7B-Instruct-v0.2 compare to Llama-2-7B and Gemma-7B on university-level calculus problems in the MathOdyssey dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.9/10.

3 Results

12 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 7.9/10.

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
Large language models (LLMs) have significantly advanced natural language understanding and demonstrated strong problem-	✓	0.31
Most LLMs still struggle with solving mathematical problems due to the intricate reasoning required.	✓	0.26
The MathOdyssey dataset includes diverse mathematical problems at high school and university levels, created by experts	✓	0.34
The MathOdyssey dataset is designed to rigorously test LLMs in advanced problem-solving scenarios and cover a wider rang	✓	0.33
Benchmarking was conducted on open-source models, such as Llama-3 and DBRX-Instruct, and closed-source models from the G	✓	0.29
LLMs perform well on routine and moderately difficult tasks but face significant challenges with Olympiad-level problems	✓	0.34
There is a narrowing performance gap between open-source and closed-source models, yet substantial challenges remain, pa	✓	0.34
The dataset, results, and code are publicly available.	✓	0.18

References

- https://doi.org/10.4230/oasics.icpec.2025.4
- https://doi.org/10.48550/arxiv.2406.18321
- https://doi.org/10.1038/s41597-025-05283-3