Mistral-Large-2 Inference Efficiency on MATH vs. Specialized Math Models

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the inference efficiency (tokens/sec or latency) of Mistral-Large-2 when solving MATH problems compared to smaller specialized math-focused models. Large language models (LLMs) have been explored in a variety of reasoning tasks including solving of mathematical problems. Each math dataset typically includes its own specially designed evaluation script, which, while suitable for its intended use, lacks generalizability. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MARIO Eval: Evaluate Your Math LLM with your Math LLM–A mathematical dataset evaluation toolkit. Research question: What is the inference efficiency (tokens/sec or latency) of Mistral-Large-2 when solving MATH problems compared to smaller specialized math-focused models?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

14 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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
gpt-3.5-turbo appears to surpass the three open-source mathematical evaluation toolkits in terms of making correctness j	×	0.05
Our toolkit in the configuration of the basic design attains superior accuracy compared to gpt-3.5-turbo.	×	0.04
The incorporation of LLM technology provides an additional improvement of about 1% when integrated with prior toolkits.	×	0.04
On the two testsets, our basic design can achieve about 97% accuracy.	×	0.02
Our basic design still achieves better than ToRA toolkit, which was specifically tailored for its output.	×	0.03
On GaoKao2023-ToRA, most of the inferred results are incorrect.	×	0.01
All toolkits exhibited satisfactory performance, indicating that integrating the LLM may not be necessary.	×	0.02

References

- http://arxiv.org/abs/2604.25926v1
- http://arxiv.org/abs/2404.13925v1
- http://arxiv.org/abs/2503.11495v1