Fine-Tuning Mistral-Large-2 On Domain-Specific Math Datasets (E.G., Math-Pt) Performance On Its Math Benchmark Scores

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does fine-tuning Mistral-Large-2 on domain-specific math datasets (e.g., Math-PT) improve its MATH benchmark scores compared to zero-shot or few-shot evaluation. The use of large language models (LLMs) for complex mathematical reasoning is an emergent area of research, with fast progress in methods, models, and benchmark datasets. However, most mathematical reasoning evaluations exhibit a significant linguistic bias, with the vast. 8 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MATH-PT: A Math Reasoning Benchmark for European and Brazilian Portuguese. Research question: How does fine-tuning Mistral-Large-2 on domain-specific math datasets (e.g., Math-PT) improve its MATH benchmark scores compared to zero-shot or few-shot evaluation?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

15 papers retrieved. 8 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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 MATH-PT benchmark uses a standardized prompting strategy adapted to each linguistic variant (European and Brazilian	✓	0.19
For multiple-choice questions in MATH-PT, a direct instruction prompt in Portuguese is used	✓	0.17
The prompt explicitly instructs the model to output only the letter of the correct option inside a \boxed{} command	×	0.02
European Portuguese questions may contain figures while Brazilian Portuguese questions do not	×	0.13
For questions containing figures, an additional block enumerating the referenced visual content is included	×	0.02
The Brazilian Portuguese subset (pt-BR) consists exclusively of questions without figures	×	0.13
For open-ended questions, models are asked to place the final answer inside a \boxed{} command	×	0.08
All models are evaluated in a zero-shot setting, without chain-of-thought supervision or few-shot examples	×	0.10

References

- http://arxiv.org/abs/2604.25926v1
- http://arxiv.org/abs/2508.11281v3

- http://arxiv.org/abs/2508.04848v1