Multimodal vs. Text-Only Models on MATH Benchmark Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the performance of multimodal models trained on both textual and symbolic mathematical representations compare to text-only models on the MATH dataset. 10 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: CMM-Math: A Chinese Multimodal Math Dataset To Evaluate and Enhance the Mathematics Reasoning of Large Multimodal Models. Research question: How does the performance of multimodal models trained on both textual and symbolic mathematical representations compare to text-only models on the MATH dataset?.

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 4.5/10.

3 Results

13 papers retrieved. 10 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
CMM-Math is a high-quality Chinese multimodal mathematical dataset consisting of evaluation and training datasets.	✓	0.20
Math-LMM is a math-specific LMM that trains with three stages: foundational pre-training, foundational fine-tuning, and	✓	0.25
Math-LMM outperforms strong open-source LMMs over CMM-Math, MATHVISTA, and Math-V datasets in most cases.	×	0.14
GSM8K and MATH are two popular textual datasets to comprehensively evaluate the LLMs.	×	0.09
MMMU primarily evaluates the model's visual recognition abilities, with only a few questions involving simple mathematic	×	0.06
MATHVISTA consolidates and transforms existing FQA, GPS, MWP, TQA, and VQA datasets, enabling tests of basic mathematica	×	0.02
The mean accuracy of LMMs on CMM-Math is 35.66 for CogVLM2, 44.61 for InternLM-VL, 47.41 for Qwen2-VL-Instruct, 43.32 fo	×	0.04
The mean accuracy of LMMs on CMM-Math is 49.91 for Qwen-VL-Max, 41.88 for Gemini, and 29.02 for GPT-4o.	×	0.05
The mean accuracy of LMMs on CMM-Math is 64.91 for Qwen-VL-Max (3-Shot), 41.65 for Gemini (3-Shot), and 65.98 for GPT-4o	×	0.04
The mean accuracy of LMMs on CMM-Math is 32.10 for Math-LMM (Ours 7B) and 26.13 for Math-LMM (Ours 72B).	×	0.10

References

- http://arxiv.org/abs/2409.02834v3
- http://arxiv.org/abs/2505.23851v1
- http://arxiv.org/abs/2411.00387v3