Multimodal Language Models on Visual Mathematical and Scientific Reasoning Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do multimodal language models perform on visual mathematical and scientific reasoning v13. 10 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. Research question: How do multimodal language models perform on visual mathematical and scientific reasoning v13.

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 5.3/10.

3 Results

15 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 5.3/10.

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
MMC-Benchmark is a human-annotated benchmark containing nine distinct tasks for evaluating reasoning capabilities over c	✓	0.25
Existing LMMs, including GPT-4V, show limitations in correctly interpreting charts when evaluated on MMC-Benchmark.	✓	0.20
MMC-Instruction contains 600k samples, making it larger than FigureQA (180k), DVQA (300k), PlotQA (224k), ChartQA (21.9k	×	0.04
MMC-Instruction supports free-form answers and open-ended/MQA formats, whereas FigureQA, DVQA, and PlotQA use fixed voca	×	0.03
MMCA achieves state-of-the-art performance on current chart question-answer benchmarks compared with existing open-sourc	×	0.13
MMC-Benchmark includes tasks such as chart-to-datatable and chart-to-json.	×	0.13
GPT-4V faces significant challenges on MMC-Benchmark, specifically in Chart to Datatable and Chart to Json tasks.	×	0.11
On a specific chart regarding land area, GPT-4V and LLaVA-v1.5 incorrectly identified China as the third largest country	×	0.04
MMCA achieved a score of 57.4 on ChartQA, 72.5 on DocVQA, and 59.6 on TextVQA.	×	0.01
Removing fine-tuning of the vision encoder in MMCA results in lower performance scores on ChartQA (54.2), DocVQA (67.8),	×	0.06

References

- `http://arxiv.org/abs/2306.09265v1`

- http://arxiv.org/abs/2311.10774v2
- http://arxiv.org/abs/2407.04973v1