Scaling Instruction-Tuned Datasets and Generalization in Large Multimodal Models for Chart Understanding

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the scaling of instruction-tuned datasets (e.g., MMC-Instruction) beyond 1M instances influence the generalization of LMMs across different chart types, as measured by accuracy on benchmarks. 14 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. 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 does the scaling of instruction-tuned datasets (e.g., MMC-Instruction) beyond 1M instances influence the generalization of LMMs across different chart types, as measured by accuracy on benchmarks like ChartQA or FigureQA?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

16 papers retrieved. 14 claims extracted; 2 independently verified. Quality review score: 5.8/10.

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
The MMC-Instruction dataset contains 600k samples.	×	0.10
The MMC-Instruction dataset supports free-form answers with an open vocabulary or multiple-choice question answering (MQ	×	0.06
The MMC-Instruction dataset has an average answer length of 23.7 words.	×	0.06
Existing LMMs, including GPT-4V, show limitations in correctly interpreting charts on the MMC-Benchmark.	✓	0.21
GPT-4V faces significant challenges specifically in 'Chart to Datatable' and 'Chart to Json' tasks within the MMC-Benchm	×	0.10
MMCA achieves state-of-the-art performance on current chart question-answer benchmarks compared with existing open-sourc	×	0.12
On the ChartQA benchmark, MMCA achieved a score of 57.4.	×	0.06
On the DocVQA benchmark, MMCA achieved a score of 72.5.	×	0.04
On the TextVQA benchmark, MMCA achieved a score of 59.6.	×	0.04
Removing fine-tuning of the vision encoder in MMCA reduces the ChartQA score from 57.4 to 54.2.	×	0.05
In a specific test case regarding land area, GPT-4V and LLaVA-v1.5 incorrectly identified China as the third largest cou	×	0.03
MMC-Benchmark includes tasks such as chart information extraction, chart reasoning, contextual chart understanding, char	×	0.11
MMC-Benchmark offers two quantitative evaluation methods: free-format Generation Ability Evaluation using GPT-4 and mult	×	0.09

References

- http://arxiv.org/abs/2401.02384v3
- http://arxiv.org/abs/2311.10774v2
- http://arxiv.org/abs/2407.14506v3