FA(LLaMA) Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of FA(LLaMA) on reasoning mathematics coding and language understanding tasks. 14 claims were extracted from source literature; 1 was 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: GSM8K-V: Can Vision Language Models Solve Grade School Math Word Problems in Visual Contexts. Research question: What are the benchmark performance scores of FA(LLaMA) on reasoning mathematics coding and language understanding tasks.

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 4.5/10.

3 Results

14 papers retrieved. 14 claims extracted; 1 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
GSM8K-V is a vision-based benchmark constructed by systematically mapping each case in the GSM8K benchmark into a visual	×	0.11
GSM8K-V preserves the exact set of problems as GSM8K, with problem statements expressed in visual rather than textual fo	×	0.05
The construction pipeline for GSM8K-V consists of three stages: extracting mathematical information and allocating it ac	×	0.08
Human annotation was conducted during the construction of GSM8K-V to ensure accuracy and reliability.	×	0.05
Gemini-2.5-Pro achieved an accuracy of 46.93% on the GSM8K-V benchmark.	×	0.12
Existing models consistently reach 80%–90% accuracy on the text-based GSM8K benchmarks.	×	0.09
Gemini-2.5-Pro achieved an accuracy of 95.22% on the text-based GSM8K benchmark.	✓	0.16
GPT-5 achieved an accuracy of 41.54% on GSM8K-V.	×	0.03
GPT-4o achieved an accuracy of 41.91% on GSM8K-V.	×	0.09
InternVL3.5-8B achieved an accuracy of 16.91% on GSM8K-V.	×	0.02
Llama-4-17B-128E-Instruct achieved an accuracy of 42.28% on GSM8K-V.	×	0.03
Qwen2.5-VL-72B-Instruct achieved an accuracy of 28.31% on GSM8K-V.	×	0.01
Ovis2.5-9B achieved an accuracy of 37.13% on GSM8K-V.	×	0.01
GSM8K-V enables a within-item comparison by evaluating models on both the original text question and the visualized vers	×	0.03

References

- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2406.10515v2