Multimodal vs. Text-Only Models in Math Word Problem Performance and Efficiency

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Do multimodal models exhibit higher PER than text-only models on math word problems (e.g., SVAMP, AQuA) when evaluated with equal compute budgets, and how does modality fusion impact efficiency. Recent progress in large language models (LLMs) like GPT-4 and PaLM-2 has brought significant advancements in addressing math reasoning problems. In particular, OpenAI's latest version of GPT-4, known as GPT-4 Code Interpreter, shows remarkable performance on challenging math. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification. Research question: Do multimodal models exhibit higher PER than text-only models on math word problems (e.g., SVAMP, AQuA) when evaluated with equal compute budgets, and how does modality fusion impact efficiency metrics?.

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 2.5/10.

3 Results

15 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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 dataset is recognized as the most challenging math word problem dataset.	×	0.13
GPT4-Code achieves an accuracy of 69.69% on the MATH benchmark.	×	0.04
The previous state-of-the-art result on the MATH benchmark was 53.90%.	×	0.04
Adding explicit code-based self-verification to GPT4-Code improves accuracy on the MATH benchmark to 73.54%.	×	0.13
Adding both explicit code-based self-verification and verification-guided weighted majority voting (with 16 sampled path	×	0.12
In the verification-guided weighted majority voting example, the score for candidate answer '2' is calculated as 3.5.	×	0.04
In the verification-guided weighted majority voting example, the score for candidate answer '5' is calculated as 2.3.	×	0.04
Using Prompt 2 (allowing code usage 1 time) results in an overall accuracy of 74.48% on the MATH dataset.	×	0.08
Using Prompt 1 (no code allowed) results in an overall accuracy of 60.80% on the MATH dataset.	×	0.06
The average precision of the proposed method is 95.88%.	×	0.02
The average recall of the proposed method is 79.11%.	×	0.04
The configuration with weights 1/0.5/0.2 achieves an accuracy of 73.54%.	×	0.02

References

- http://arxiv.org/abs/2603.07394v1
- http://arxiv.org/abs/2308.07921v1
- http://arxiv.org/abs/2508.19294v2