Code-Based Self-Verification Enhances Robustness in Adversarial Math Word Problems

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: To what extent does code-based self-verification improve robustness against adversarial perturbations in math word problems compared to standard multimodal fusion approaches. 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. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. 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: To what extent does code-based self-verification improve robustness against adversarial perturbations in math word problems compared to standard multimodal fusion approaches?.

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 2.8/10.

3 Results

14 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
The MATH dataset is recognized as the most challenging math word problem dataset.	×	0.13
GPT4-Code reaches 69.69% on MATH, surpassing the previous state of the art result of 53.90%.	×	0.04
Adding explicit code-based self-verification improves the accuracy of GPT4-Code to 73.54%.	×	0.11
Adding both explicit code-based self-verification and verification-guided weighted majority voting improves the accuracy	×	0.13
The repetend in the decimal representation of 1/19 contains 18 digits.	×	0.02
The 39th digit in the decimal representation of 1/19 is the same as the 3rd digit.	×	0.00
The 3rd digit in the decimal representation of 1/19 is 2.	×	0.00
The pattern of 18 repeating digits in the decimal representation of 1/19 is '052631578947368421'.	×	0.00
The 21st digit in the decimal representation of 1/19 is '5'.	×	0.00
The overall accuracy of different prompts on the MATH dataset varies, with the highest accuracy being 76%.	×	0.06
The code usage frequency varies with different prompts, with the highest frequency being 100%.	×	0.08
The accuracy of the model varies with different levels of the MATH dataset, with the highest accuracy being 74.48%.	×	0.06
The precision, recall, and accuracy values vary with different numbers of sampled reasoning paths, with the highest accu	×	0.03
The average accuracy, precision, and recall values are 79.11%, 95.88%, and 82%, respectively.	×	0.01

References

- http://arxiv.org/abs/2308.07921v1
- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2103.15670v3