Synthetic Training Data Enhancements in Language Model Mathematical Reasoning

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does synthetic training data improve language model performance on mathematical reasoning benchmarks v7. 20 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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: How does synthetic training data improve language model performance on mathematical reasoning benchmarks v7.

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 3.2/10.

3 Results

15 papers retrieved. 20 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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% accuracy on the MATH dataset.	×	0.07
The previous state-of-the-art result on the MATH dataset was 53.90%.	×	0.06
Adding explicit code-based self-verification improves GPT4-Code's accuracy to 73.54%.	×	0.14
Adding both explicit code-based self-verification and verification-guided weighted majority voting improves GPT4-Code's	×	0.14
The number of sampled paths used in verification-guided weighted majority voting is 16.	×	0.04
The 3rd digit in the decimal representation of 1/19 is 2.	×	0.01
The 39th digit in the decimal representation of 1/19 is 5.	×	0.01
The pattern of 18 repeating digits in the decimal representation of 1/19 is '052631578947368421'.	×	0.01
The 21st digit in the repeating pattern of 1/19 is '5'.	×	0.00
The overall accuracy of the 4 prompts on the MATH dataset is 74.48%.	×	0.05
The accuracy of Prompt 1 on the MATH dataset is 74%.	×	0.07
The accuracy of Prompt 2 on the MATH dataset is 72%.	×	0.07
The accuracy of the basic prompt on the MATH dataset is 70%.	×	0.05
The accuracy of the method with 16 sampled reasoning paths is 84%.	×	0.03
The average accuracy of the method with different reasoning paths is 79.11%.	×	0.04
The accuracy of the method with weights 1/0/0 is 73.54%.	×	0.02
The accuracy of the method with weights 0.5/0.5/1 is 74%.	×	0.02
The accuracy of the method with weights 1/0.5/0.2 is 74%.	×	0.02
The accuracy of the method with weights 1/1/1 (Majority Voting) is 72%.	×	0.06

References

- http://arxiv.org/abs/2604.25926v1
- http://arxiv.org/abs/2406.15444v5
- http://arxiv.org/abs/2308.07921v1