Continued Pretraining on Model-Translated Mathematical Code and MATH Benchmark Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does continued pretraining on model-translated mathematical code affect small decoder-only models' accuracy on the MATH benchmark compared to standard mathematical text pretraining. 18 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. 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: How does continued pretraining on model-translated mathematical code affect small decoder-only models' accuracy on the MATH benchmark compared to standard mathematical text pretraining?.

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 3.5/10.

3 Results

14 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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.	×	0.11
GPT4-Code reaches 69.69% accuracy on the MATH dataset.	×	0.07
The previous state-of-the-art result on the MATH dataset is 53.90%.	×	0.07
Adding explicit code-based self-verification improves the accuracy of GPT4-Code to 73.54% on the MATH dataset.	×	0.13
Adding both explicit code-based self-verification and verification-guided weighted majority voting improves the accuracy	×	0.14
The repetend in the decimal representation of 1/19 contains 18 digits.	×	0.02
The 3rd digit in the decimal representation of 1/19 is 2.	×	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'.	×	0.03
The overall accuracy of the model using Basic Prompt is 74.48%.	×	0.02
The overall accuracy of the model using Prompt 1 is 74%.	×	0.03
The overall accuracy of the model using Prompt 2 is 72%.	×	0.02
The average accuracy of the model across different levels using Basic Prompt is 95.88%.	×	0.02
The average accuracy of the model across different levels using Prompt 1 is 79.11%.	×	0.03
The average accuracy of the model across different levels using Prompt 2 is 73.54%.	×	0.03
The average accuracy of the model using 16 sampled reasoning paths is 84%.	×	0.03
The average precision of the model using 16 sampled reasoning paths is 90%.	×	0.04
The average recall of the model using 16 sampled reasoning paths is 82%.	×	0.03

References

- http://arxiv.org/abs/2601.21725v2
- http://arxiv.org/abs/2410.08196v1
- http://arxiv.org/abs/2308.07921v1