Synthetic Training Data Enhances Language Model Performance on Mathematical Reasoning Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does synthetic training data improve language model performance on mathematical reasoning benchmarks v18. 14 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLM-ProS: Analyzing Large Language Models' Performance in Competitive Problem Solving. Research question: How does synthetic training data improve language model performance on mathematical reasoning benchmarks v18.

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 4.2/10.

3 Results

14 papers retrieved. 14 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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 study uses a curated dataset of 166 ICPC World Finals problems from 2011 to 2024.	✓	0.20
Five models were evaluated: GPT-4o, Mistral Large, Llama-3.1-405B, o1-mini, and o1-preview.	✓	0.23
In the benchmark of 166 problems, o1-mini achieved 16 Accepted (AC) verdicts.	×	0.07
In the benchmark of 166 problems, o1-preview achieved 15 Accepted (AC) verdicts.	×	0.07
In the benchmark of 166 problems, GPT-4o achieved 0 Accepted (AC) verdicts.	×	0.07
In the benchmark of 166 problems, Mistral-Large achieved 0 Accepted (AC) verdicts.	×	0.08
In the benchmark of 166 problems, Llama-3.1 achieved 0 Accepted (AC) verdicts.	×	0.05
o1-mini received 124 Wrong Answer (WA) verdicts in the evaluation.	×	0.04
o1-preview received 120 Wrong Answer (WA) verdicts in the evaluation.	×	0.04
GPT-4o received 41 Compile Error (CE) verdicts and 39 Runtime Error (RE) verdicts.	×	0.02
o1-mini and o1-preview consistently outperform other evaluated LLMs in accuracy, verdict distribution, and resource effi	×	0.10
Models with specialized training for chain-of-thought reasoning exhibit greater robustness and adaptability to unseen pr	×	0.09
General-purpose models showed a significant performance drop on unseen data compared to specialized models.	×	0.04
The problem 'Allied Chute Manufacturers' involves calculating properties based on an integer n representing points in a	×	0.03

References

- http://arxiv.org/abs/2410.02152v1
- http://arxiv.org/abs/2405.07551v1
- http://arxiv.org/abs/2502.04355v1