Instruction Fine-Tuning Improves Language Model Mathematical Problem-Solving Accuracy

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the effect of instruction fine-tuning on language model mathematical problem-solving accuracy v14. 19 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MathScale: Scaling Instruction Tuning for Mathematical Reasoning. Research question: What is the effect of instruction fine-tuning on language model mathematical problem-solving accuracy v14.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

12 papers retrieved. 19 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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 MWPBENCH training set comprises around 20K questions.	×	0.03
GPT-3.5-Turbo-0613 is used for concept extraction.	×	0.03
The concept extraction process obtains 2,018 topics and 8,892 knowledge points.	×	0.06
The edge weight in the concept graph is smoothed using Equation (1) with ε = 1e -5.	×	0.04
The concept composition process is repeated for approximately 1K epochs, resulting in 2 million unique concept compositi	×	0.03
GPT-3.5-Turbo-0613 is used to create 2 million question-answer pairs with the concept compositions.	×	0.11
The generated datasets are decontaminated by excluding all math questions in the test set of MWPBENCH.	×	0.07
The MathScaleQA dataset is created by combining the generated data with the training set of MWPBENCH.	×	0.06
The validation step (Section 3.4) is excluded from the final pipeline because it does not improve results.	×	0.03
MathScale-7B achieves a 35.0% (micro) and 37.5% (macro) accuracy across MWPBENCH.	×	0.11
MathScale-7B surpasses its best counterparts of equivalent size by 42.9% and 43.7% in micro and macro averages, respecti	×	0.13
MathScale-Mistral demonstrates performance parity in both micro and macro averages relative to GPT-3.5-Turbo.	×	0.06
When scaling the size of the MathScaleQA dataset, a nearly logarithmic growth in the performance of the MathScale-7b mod	×	0.11
GSM8K and MATH, the most popular math datasets, each contain around 7.5K training examples.	×	0.06
WizardMath introduces an array of operations for GPT-3.5 to generate math questions with increased complexity.	×	0.06
MetaMath bootstraps questions in GSM8K and MATH through answer augmentation, question rephrasing, self-verification, and	×	0.06
The newly generated examples by WizardMath and MetaMath exhibit substantial similarity to the original examples containe	×	0.02
MathScale proposes a conceptually simple and scalable method to generate training examples by extracting high-level conc	×	0.14

References

- https://arxiv.org/abs/2501.14002
- https://arxiv.org/abs/2403.02884
- https://arxiv.org/abs/2405.14804