Instruction Fine-Tuning Effects on Language Model Mathematical Problem-Solving Accuracy

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: What is the effect of instruction fine-tuning on language model mathematical problem-solving accuracy v20. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: SmallToLarge (S2L): Scalable Data Selection for Fine-tuning Large Language Models by Summarizing Training Trajectories of Small Models. Research question: What is the effect of instruction fine-tuning on language model mathematical problem-solving accuracy v20.

2 Methodology

Systematic literature search across multiple databases yielded 6 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

6 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
S2L is compared against Random Sampling, Least Confidence, Middle Perplexity, High Learnability, Facility Locations, and	×	0.08
MathInstruct dataset contains 262,040 training examples.	×	0.07
Pythia, Phi-2, and Llama-2 are used as base models for validating S2L.	×	0.04
Evaluation datasets include GSM8K, MATH, NumGLUE, SVAMP, Mathematics, and SimulEq.	×	0.04
Exact match is used as the evaluation metric for open-formed questions.	×	0.02
S2L achieves 69.1 on GSM8K, 32.6 on MATH, and 65.7 on NumGLUE with Phi-2 (2.7B) model.	×	0.06
S2L achieves 78.4 on SVAMP, 58.4 on Mathematics, and 44.2 on SimulEq with Phi-2 (2.7B) model.	×	0.04
S2L achieves an average accuracy of 57.7 with Phi-2 (2.7B) model.	×	0.08

References

- http://arxiv.org/abs/2305.03937v1
- http://arxiv.org/abs/2605.14071v1
- http://arxiv.org/abs/2403.07384v2