Fine-Tuning on Self-Invoking Code Generation Benchmarks Enhances Multi-Step Reasoning Performance

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: To what extent does fine-tuning on self-invoking code generation benchmarks (vs. standard benchmarks) improve performance on multi-step reasoning tasks like GSM8K or MATH, as measured by accuracy at. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MateInfoUB: A Real-World Benchmark for Testing LLMs in Competitive, Multilingual, and Multimodal Educational Tasks. Research question: To what extent does fine-tuning on self-invoking code generation benchmarks (vs. standard benchmarks) improve performance on multi-step reasoning tasks like GSM8K or MATH, as measured by accuracy at different model scales?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

4 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.7/10.

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
Large Language Models (LLMs) have transformed various domains, particularly computer science (CS) education.	✓	0.31
LLMs exhibit remarkable capabilities in code-related tasks and problem-solving.	✓	0.16
A novel bilingual (English-Romanian) multimodal (text and image) dataset of multiple-choice questions derived from a hig	✓	0.36
The dataset includes problems that are easier solved using reasoning on paper, while for others writing code is more eff	✓	0.20
State of The Art LLMs have been systematically evaluated on this dataset, analyzing their performance on theoretical pro	✓	0.20
The findings reveal the strengths and limitations of current LLMs, including the influence of language choice (English v	✓	0.31
The dataset will be made publicly available in both English and Romanian.	✓	0.20
An educational application tailored for Romanian students has been released, enabling them to self-assess using the data	✓	0.27

References

- https://doi.org/10.48550/arxiv.2511.00527
- https://doi.org/10.18653/v1/2025.bea-1.3
- https://openalex.org/W7161916717