Robustness Disparities in Phi-3-Mini and Mistral-7B on Adversarial GSM-Symbolic Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the robustness of Phi-3-mini and Mistral-7B-v0.1 vary when evaluated on adversarially perturbed GSM-Symbolic instances versus clean instances across multiple languages, using accuracy and. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. 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 the robustness of Phi-3-mini and Mistral-7B-v0.1 vary when evaluated on adversarially perturbed GSM-Symbolic instances versus clean instances across multiple languages, using accuracy and perplexity as metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

15 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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 achieves an accuracy of 69.69% on the MATH dataset.	×	0.05
The previous state-of-the-art result on the MATH dataset was 53.90%.	×	0.06
Adding explicit code-based self-verification to GPT4-Code improves accuracy on the MATH dataset to 73.54%.	×	0.14
Adding both explicit code-based self-verification and verification-guided weighted majority voting (with 16 sampled path	×	0.13
In the verification-guided weighted majority voting example, the score for candidate answer '2' is calculated as 3.5 usi	×	0.04
In the verification-guided weighted majority voting example, the score for candidate answer '5' is calculated as 2.3 usi	×	0.04
Using Prompt 2 (code allowed 1 time) results in an overall accuracy of 74.48% on the MATH dataset.	×	0.05
Using the Basic Prompt results in an overall accuracy of 76% on the MATH dataset.	×	0.04
The average accuracy across different levels for Prompt 1 (no code allowed) is 60.80%.	×	0.04
The configuration with weights 1/0.5/0.2 achieves an accuracy of 73.54%.	×	0.02
The configuration using Majority Voting (weights 1/1/1) achieves an accuracy of 79.11%.	×	0.03
The proposed method achieves an average accuracy of 95.88% in the specific metric reported in Table (p11).	×	0.02

References

- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2602.10661v2
- http://arxiv.org/abs/2308.07921v1