Meta-Reasoning Performance and Few-Shot Arithmetic Benchmark Correlations

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does meta-reasoning performance on MR-GSM8K correlate with few-shot reasoning accuracy on other arithmetic benchmarks like MATH or SVAMP. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Deep reflective reasoning in interdependence constrained structured data extraction from clinical notes for digital health. Research question: Does meta-reasoning performance on MR-GSM8K correlate with few-shot reasoning accuracy on other arithmetic benchmarks like MATH or SVAMP?.

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 3.7/10.

3 Results

4 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
Deep reflective reasoning is a multi-round iterative process that facilitates adjusting variable-value assignments for c	×	0.09
The comprehensive score for all variables, average F1 score, is 0.9112.	×	0.05
Margin Status for Invasive Carcinoma (with 5 classes) achieved best performance with accuracy of 0.9697 and F1 score of	×	0.02
Tumor Extent (with 9 classes) achieved an accuracy of 0.8198 and F1 score of 0.7933, which is the lowest performance sco	×	0.04
Correct rate is defined as the number of correctly predicted cases divided by the total number of cases evaluated for th	×	0.02
Correct rate is equivalent to accuracy in the context of numeric variables.	×	0.07
Tumor Size - Greatest dimension (cm) has a Correct Rate of 0.9680, MAE of 0.1079, RMSE of 0.7094, and Coverage of 0.8756	×	0.02
Distance of Tumor from Radial Margin (cm) has a Correct Rate of 0.8670, MAE of 0.5663, RMSE of 1.7153, and Coverage of 0	×	0.02
Distance of Tumor from Distal Margin (cm) has a Correct Rate of 0.9760, MAE of 0.1207, RMSE of 0.8367, and Coverage of 0	×	0.02
Distance of Tumor from Closest Margin (cm) has a Correct Rate of 0.7700, MAE of 0.8469, RMSE of 2.0639, and Coverage of	×	0.02
The average Correct Rate for numeric variables is 0.8953, MAE is 0.4104, RMSE is 1.3313.	×	0.07

References

- http://arxiv.org/abs/2603.20435v2
- http://arxiv.org/abs/2502.17848v4
- http://arxiv.org/abs/2510.08540v2