# Robustness of Meta-Reasoning Benchmarks Across Language and Code Domains

Assignee Research

June 7, 2026

**Abstract**

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the robustness of meta-reasoning benchmarks like MR-GSM8K when applied to cross-domain tasks such as language understanding or code generation. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Meta Reasoning for Large Language Models. Research question: What is the robustness of meta-reasoning benchmarks like MR-GSM8K when applied to cross-domain tasks such as language understanding or code generation?.

## 2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

## 3 Results

11 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.5/10.

## 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 |
|---|---|---|
| MRP achieves the second-best performance in 4 of 7 tasks, including Gameof24, TriviaQA, BigToM, and Code. | × | 0.06 |
| MRP attains the highest overall performance across the 7 tasks, with an average of 0.772. | × | 0.04 |
| TOT excels in certain tasks such as GSM8K and Gameof24 but performs less impressively in others. | × | 0.04 |
| There are noticeable performance gaps compared with MRP in tasks such as BigToM (0.43 VS 0.57) and Code (0.765 VS 0.867) | × | 0.03 |
| The performance with GPT-4 is satisfactory, but the experimental results with GPT-3.5 indicate that the effectiveness of | × | 0.05 |
| Error analysis revealed the main issues: Scoring Error, Self-opinion, Factual Error, and Reasoning Error. | × | 0.04 |
| When the model's capabilities are limited, it cannot have sufficient awareness of its own reasoning abilities and the me | × | 0.07 |
| This performance drop also appears in other reasoning methods. | × | 0.08 |

## References


- http://arxiv.org/abs/2409.08687v4
- http://arxiv.org/abs/2306.06371v1
- http://arxiv.org/abs/2406.11698v1