InternLM-20B-Reward Performance on MR-GSM8K and MATH Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the performance of InternLM-20B-Reward on calculus reasoning tasks compare to other state-of-the-art LLMs when evaluated on a standardized benchmark like GSM8K or MATH, and what metrics. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MR-GSM8K: A Meta-Reasoning Benchmark for Large Language Model Evaluation. Research question: How does the performance of InternLM-20B-Reward on calculus reasoning tasks compare to other state-of-the-art LLMs when evaluated on a standardized benchmark like GSM8K or MATH, and what metrics (e.g., exact match accuracy) differentiate their capabilities?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

13 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.0/10.

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 MR-Score metric consists of three sub-metrics: Matthews Correlation Coefficient (MCC) for binary classification of s	×	0.02
The MCC score ranges from -1 to +1, where -1 indicates total disagreement between prediction and observation, 0 suggests	×	0.02
The models evaluated include Qwen-v1.5-1.8B, Llama3-70B, Deepseek-v2-236B, WizardMath-v1.1-7B, MAmmoTH-70B, DeepseekMath	×	0.06
Each model was evaluated under a zero-shot setting to assess their ability to follow instructions and their mathematical	×	0.07
The inference temperature was set to zero across all models to ensure reproducibility and minimize variance.	×	0.02
In the context of this paper, negative values are interpreted as no better than random guesses, and 0 is set as the cut-	×	0.03
The second metric is the accuracy of the first-error-step prediction, calculated as ACCstep = Ncorrect_first_error_step	×	0.01
The third metric calculates the accuracy of identifying both the first-error-step and explaining the error-reason.	×	0.01
The evaluation results for various models on the MR-GSM8K benchmark are provided in Table 2.	✓	0.17
The models were also tested under a few-shot setting to leverage their in-context learning abilities for understanding m	×	0.05

References

- http://arxiv.org/abs/2604.25926v1
- http://arxiv.org/abs/2312.17080v4
- http://arxiv.org/abs/2308.07921v1