MedGemma Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of MedGemma on reasoning mathematics coding and language understanding tasks. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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: What are the benchmark performance scores of MedGemma on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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	×	0.03
The MCC score ranges from -1 to +1, where -1 indicates total disagreement between prediction and observation, 0 suggests	×	0.02
The evaluated models were tested under both zero-shot and few-shot settings to assess their ability to follow instructio	×	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-	×	0.04
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.	×	0.02
The models evaluated include Qwen-v1.5-1.8B, Llama3-70B, Deepseek-v2-236B, WizardMath-v1.1-7B, MAmmoTH-70B, DeepseekMath	×	0.06
The models vary greatly in size, ranging from a few billion parameters to 236 billion parameters.	×	0.08
The evaluation results are presented in Table 2, showing performance metrics for each model under zero-shot (k=0) and fe	×	0.04

References

- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2406.10515v2
- http://arxiv.org/abs/2312.17080v4