Frontier Large Language Models in Mathematical Reasoning, Code Generation, and Scientific Knowledge

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Comprehensive comparison of frontier large language models on mathematical reasoning code generation and scientific knowledge v13. 18 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.4/10. 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: Comprehensive comparison of frontier large language models on mathematical reasoning code generation and scientific knowledge v13.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.4/10.

3 Results

12 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 2.4/10.

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), accuracy of the first-error-s	×	0.01
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	×	0.06
The inference temperature was set to zero across all models to ensure reproducibility and minimize variance.	×	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
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 MR-Score for Qwen-1.8B under zero-shot setting is 0.1.	×	0.01
The MR-Score for Phi3-3.8B under few-shot setting is 21.9.	×	0.01
The MR-Score for Deepseek-Math-7B-RL under zero-shot setting is 0.1.	×	0.04
The MR-Score for Llama3-8B under few-shot setting is 17.4.	×	0.01
The MR-Score for MAmmoTH-70B under zero-shot setting is 5.0.	×	0.01
The MR-Score for Llama3-70B under few-shot setting is 34.2.	×	0.01
The MR-Score for Qwen1.5-72B under few-shot setting is 23.3.	×	0.01
The MR-Score for Deepseek-v2-236B under few-shot setting is 34.1.	×	0.06
The MR-Score for Claude3-Haiku under few-shot setting is 4.9.	×	0.03
The MR-Score for GPT-3.5-Turbo under few-shot setting is 17.9.	×	0.03
The MR-Score for Claude3-Sonnet under few-shot setting is 20.8.	×	0.09
The MR-Score for GPT-4-Turbo under few-shot setting is 53.0.	×	0.02

References

- http://arxiv.org/abs/2405.07551v1
- http://arxiv.org/abs/2312.17080v4
- http://arxiv.org/abs/2310.03731v1