Frontier Large Language Models in Mathematical Reasoning and Scientific Knowledge Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 5 peer-reviewed papers addressing the following research question: Comprehensive comparison of frontier large language models on mathematical reasoning code generation and scientific knowledge v10. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Sample-Efficient Human Evaluation of Large Language Models via Maximum Discrepancy Competition. Research question: Comprehensive comparison of frontier large language models on mathematical reasoning code generation and scientific knowledge v10.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

5 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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.

References

- http://arxiv.org/abs/2512.08959v1
- http://arxiv.org/abs/2404.08008v2
- http://arxiv.org/abs/2506.13817v1