MathGPT-8B Benchmark Performance Across Mathematical Reasoning and Language Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of MathGPT-8B on reasoning mathematics coding and language understanding tasks. 10 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Survey of Mathematical Reasoning in the Era of Multimodal Large Language Model: Benchmark, Method & Challenges. Research question: What are the benchmark performance scores of MathGPT-8B on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

4 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 7.2/10.

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
Mathematical reasoning is a core aspect of human cognition and is vital across many domains, from educational problem-so	✓	0.35
As artificial general intelligence (AGI) progresses, integrating large language models (LLMs) with mathematical reasonin	✓	0.42
This survey provides the first comprehensive analysis of mathematical reasoning in the era of multimodal large language	✓	0.42
The survey reviews over 200 studies published since 2021.	×	0.15
The survey examines the state-of-the-art developments in Math-LLMs, with a focus on multimodal settings.	✓	0.25
The field is categorized into three dimensions: benchmarks, methodologies, and challenges.	✓	0.16
The survey explores the multimodal mathematical reasoning pipeline, as well as the role of (M)LLMs and the associated me	✓	0.27
The survey identifies seven major challenges hindering the realization of AGI in the domain of mathematical reasoning.	✓	0.26
The survey offers insights into the future direction for enhancing multimodal reasoning capabilities.	✓	0.24
This survey serves as a critical resource for the research community in advancing the capabilities of LLMs to tackle com	✓	0.39

References

- https://doi.org/10.48550/arxiv.2412.11936

- https://doi.org/10.48550/arxiv.2501.14002
- https://doi.org/10.18653/v1/2025.findings-acl.614