Language Model Performance on the AIME Mathematical Competition Benchmark

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: AIME mathematical competition language model benchmark evaluation. 20 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Is Mathematical Problem-Solving Expertise in Large Language Models Associated with Assessment Performance?. Research question: AIME mathematical competition language model benchmark evaluation.

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 3.5/10.

3 Results

16 papers retrieved. 20 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
GPT-4 achieves 94.9% accuracy on the GSM8K problem-solving task.	×	0.08
GPT-5 achieves 97.4% accuracy on the GSM8K problem-solving task.	×	0.08
GPT-4 achieves 29.8% accuracy on the MATH problem-solving task.	×	0.13
GPT-5 achieves 30.5% accuracy on the MATH problem-solving task.	×	0.13
GPT-4 achieves 46.4% accuracy on the GSM8K step-level assessment task.	×	0.10
GPT-5 achieves 49.3% accuracy on the GSM8K step-level assessment task.	×	0.10
GPT-4 achieves 34.5% accuracy on the MATH step-level assessment task.	×	0.11
GPT-5 achieves 38.6% accuracy on the MATH step-level assessment task.	×	0.11
GPT-5 consistently outperforms GPT-4 across both problem-solving and step-level assessment tasks.	×	0.13
GPT-4 achieves 48.6% assessment accuracy on solved-correct GSM8K items.	×	0.06
GPT-4 achieves 6.6% assessment accuracy on solved-incorrect GSM8K items.	×	0.07
GPT-5 achieves 50.2% assessment accuracy on solved-correct GSM8K items.	×	0.06
GPT-5 achieves 16.1% assessment accuracy on solved-incorrect GSM8K items.	×	0.06
GPT-4 achieves 61.5% assessment accuracy on solved-correct MATH items.	×	0.07
GPT-4 achieves 23.0% assessment accuracy on solved-incorrect MATH items.	×	0.08
GPT-5 achieves 70.5% assessment accuracy on solved-correct MATH items.	×	0.07
GPT-5 achieves 24.6% assessment accuracy on solved-incorrect MATH items.	×	0.07
The dataset consists of 400 GSM8K items and a randomly sampled 400-item subset of MATH.	×	0.03
The models are evaluated on the same set of items in two independent tasks: problem solving and assessment.	×	0.12
The same model deployment is used across both tasks for each configuration.	×	0.05

References

- http://arxiv.org/abs/2505.12925v2
- http://arxiv.org/abs/2408.15971v1
- http://arxiv.org/abs/2603.25633v1