GPT-4 vs. Specialized Models in Mathematical Reasoning Accuracy and Depth

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does GPT-4's performance on mathematical reasoning tasks compare to specialized models like Codex or AlphaCode in terms of accuracy and problem-solving depth. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification. Research question: How does GPT-4's performance on mathematical reasoning tasks compare to specialized models like Codex or AlphaCode in terms of accuracy and problem-solving depth?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

8 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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/2311.15732v2
- http://arxiv.org/abs/2303.13375v2
- http://arxiv.org/abs/2308.07921v1