Benchmark Performance of GPT-5.2-Thinking Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of GPT-5.2-Thinking on reasoning mathematics coding and language understanding tasks. 5 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Sparks of Artificial General Intelligence: Early experiments with GPT-4. Research question: What are the benchmark performance scores of GPT-5.2-Thinking on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

11 papers retrieved. 5 claims extracted; 5 independently verified. Quality review score: 8.2/10.

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
GPT-4 was trained using an unprecedented scale of compute and data.	✓	0.21
GPT-4 is part of a new cohort of LLMs that exhibit more general intelligence than previous AI models.	✓	0.30
GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, wit	✓	0.33
GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as Chat	✓	0.29
GPT-4 could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI)	✓	0.32

References

- https://doi.org/10.48550/arxiv.2303.12712
- https://doi.org/10.48550/arxiv.2403.05530
- https://doi.org/10.1007/s11704-026-60308-3