GPT-4T Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of GPT-4T on reasoning mathematics coding and language understanding tasks. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: HumanEval-V: Benchmarking High-Level Visual Reasoning with Complex Diagrams in Coding Tasks. Research question: What are the benchmark performance scores of GPT-4T on reasoning mathematics coding and language understanding tasks.

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 4.5/10.

3 Results

16 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
HumanEval-V consists of 253 human-annotated coding tasks.	✓	0.17
Each task in HumanEval-V features a diagram encoding the problem context, a function signature defining the task's input	×	0.08
The top-performing model, Claude 3.5 Sonnet, achieves 36.8% pass@1 on HumanEval-V.	×	0.11
The best open-weight model, Pixtral 124B, reaches 21.3% pass@1 on HumanEval-V.	×	0.03
Claude 3.5 Sonnet achieves a 74.3% pass rate with 100 samples on HumanEval-V.	×	0.06
Claude 3.5 Sonnet can reach 55.3% pass@1 with four self-refining iterations based on test case execution feedback on Hum	×	0.05
HumanEval-V offers a more diverse and complex set of diagrams spanning six task types.	×	0.10
The visual context must be essential for solving the task, with all relevant information contained in a single image.	×	0.04
Tasks should be designed around the visual context with minimal textual description.	×	0.04
Test cases could rigorously verify whether the model captures all critical visual information.	×	0.05
The evaluation pipeline supports LMMs with limited coding abilities by first prompting them to generate a structured dia	×	0.07
The evaluation pipeline prioritizes visual understanding over coding proficiency.	×	0.06
The evaluation pipeline uses a two-stage evaluation pipeline.	×	0.03
The evaluation pipeline involves 22 LMMs.	×	0.09
The evaluation pipeline includes three variants: PV 2C(D, σ), PV 2C(D, σ, ICoT), and PV 2T(D, σ).	×	0.02

References

- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2410.12381v3

- http://arxiv.org/abs/2507.16746v2