GPT-OSS-120B Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of GPT-OSS-120B on reasoning mathematics coding and language understanding tasks. 13 claims were extracted from source literature; 2 were 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-OSS-120B on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 13 claims extracted; 2 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.	✓	0.15
Each task in HumanEval-V features a diagram encoding the problem context, a function signature defining the task's input	×	0.07
The top-performing model, Claude 3.5 Sonnet, achieves 36.8% pass@1 on HumanEval-V.	×	0.10
The best open-weight model, Pixtral 124B, reaches 21.3% pass@1 on HumanEval-V.	×	0.05
Claude 3.5 Sonnet achieves a 74.3% pass rate with 100 samples on HumanEval-V.	×	0.05
Claude 3.5 Sonnet reaches 55.3% pass@1 with four self-refining iterations based on test case execution feedback on Human	×	0.04
HumanEval-V offers a more diverse and complex set of diagrams spanning six task types.	×	0.10
HumanEval-V demands versatile capabilities for diagram understanding and reasoning.	×	0.10
The visual context must be essential for solving the task in HumanEval-V, with all relevant information contained in a s	×	0.04
Tasks in HumanEval-V should be designed around the visual context with minimal textual description.	×	0.05
Test cases in HumanEval-V could rigorously verify whether the model captures all critical visual information.	×	0.06
HumanEval-V utilizes a two-stage evaluation pipeline that supports LMMs with limited coding abilities.	×	0.09
Extensive experiments with 22 LMMs were conducted on HumanEval-V.	✓	0.15

References

- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2312.17080v4