Phi-3 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 Phi-3 on reasoning mathematics coding and language understanding tasks. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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 Phi-3 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 4.2/10.

3 Results

16 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
HumanEval-V consists of 253 human-annotated coding tasks.	✓	0.17
Each task in HumanEval-V features a diagram, a function signature, and test cases.	×	0.13
HumanEval-V diagrams span six task types.	×	0.12
Claude 3.5 Sonnet achieves a 36.8% pass@1 score on HumanEval-V.	×	0.10
Pixtral 124B achieves a 21.3% pass@1 score on HumanEval-V.	×	0.03
Claude 3.5 Sonnet achieves a 74.3% pass rate with 100 samples.	×	0.04
Claude 3.5 Sonnet reaches a 55.3% pass@1 score with four self-refining iterations based on test case execution feedback.	×	0.03
Experiments were conducted with 22 Large Multimodal Models (LMMs).	×	0.15
GPT-4o achieved a 27.7% score in one metric and 40.0% in another metric as shown in the Proprietary LMMs table.	×	0.05
Pixtral 124B is an open-weight LMM with more than 70B parameters.	×	0.03
The evaluation pipeline includes a variant where the model generates a structured textual problem specification consisti	×	0.04

References

- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2406.10515v2