Open-Source vs Proprietary Multimodal Models Scaling Laws on HumanEval-V

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the scaling law of open-source multimodal models on HumanEval-V compare to proprietary models in terms of pass@1 accuracy across different parameter ranges. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. 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: How does the scaling law of open-source multimodal models on HumanEval-V compare to proprietary models in terms of pass@1 accuracy across different parameter ranges?.

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 4.3/10.

3 Results

8 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.3/10.

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
HumanEval-V uses code generation tasks for evaluation instead of multiple-choice or short-answer questions.	×	0.08
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 rate with four self-refining iterations based on test case execution feedback.	×	0.04
Experiments were conducted with 22 LMMs.	×	0.11
The evaluation pipeline includes a variant denoted as PV2C(D, σ) which involves direct visual-to-code generation.	×	0.06
The Chain of Thought (CoT) variant incorporates a zero-shot CoT instruction to outline reasoning before code generation.	×	0.05
The Intermediate Textual Representation variant produces a structured textual problem specification consisting of Proble	×	0.03
GPT-4o achieves a 27.7% pass@1 score in the baseline setting according to Table (p5).	×	0.01
Gemini 1.5 Pro achieves a 22.9% pass@1 score in the baseline setting according to Table (p5).	×	0.02
Pixtral 124B has more than 70B parameters.	×	0.02

References

- http://arxiv.org/abs/2204.07288v1
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2407.17856v4