Impact of Diagram Complexity on LMM and Human Reasoning Accuracy in HumanEval-V

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of increasing the number of diagram components on LMM reasoning accuracy in HumanEval-V, and how does this compare to human performance on the same tasks. 10 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 is the impact of increasing the number of diagram components on LMM reasoning accuracy in HumanEval-V, and how does this compare to human performance on the same 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 4.5/10.

3 Results

11 papers retrieved. 10 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.	✓	0.18
Each task in HumanEval-V features a diagram, a function signature, and test cases.	×	0.13
HumanEval-V spans six task types.	×	0.14
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.	×	0.04
Experiments were conducted with 22 Large Multimodal Models (LMMs).	✓	0.15
GPT-4o achieved a 27.7% score in one evaluation setting and 40.0% in another according to Table (p5).	×	0.02
The evaluation pipeline includes a variant where the model generates a structured textual problem specification consisti	×	0.05

References

- http://arxiv.org/abs/2505.04921v2
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2009.07935v3