Lugha-Llama-8B-wura-Pre-TLI Performance on HumanEval-V with Complex Diagram Reasoning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the performance of Lugha-Llama-8B-wura-Pre-TLI compare to other multimodal models on HumanEval-V, particularly in tasks requiring complex diagram interpretation and reasoning?,What is the. 18 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: How does the performance of Lugha-Llama-8B-wura-Pre-TLI compare to other multimodal models on HumanEval-V, particularly in tasks requiring complex diagram interpretation and reasoning?,What is the impact of pre-training Lugha-Llama-8B-wura with visual-language data on its reasoning and coding capabilities, as measured by HumanEval-V and other code generation benchmarks?,How does the scaling behavior of Lugha-Llama-8B-wura-Pre-TLI compare to larger multimodal models on HumanEval-V, particularly in terms of accuracy and efficiency trade-offs?.

2 Methodology

Systematic literature search across multiple databases yielded 13 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

13 papers retrieved. 18 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.	×	0.14
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.	×	0.03
Claude 3.5 Sonnet achieves a 74.3% pass rate with 100 samples on HumanEval-V.	×	0.05
Claude 3.5 Sonnet can reach 55.3% pass@1 with four self-refining iterations based on test case execution feedback on Hum	×	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
HumanEval-V uses code generation tasks for evaluation instead of multiple-choice or short-answer questions.	×	0.12
HumanEval-V addresses the lack of benchmarks that assess visual reasoning in programming contexts.	×	0.13
HumanEval-V features diagrams that are trivial for humans but challenging for current LMMs, particularly in understandin	✓	0.22
HumanEval-V includes a two-stage evaluation pipeline that supports LMMs with limited coding abilities by first prompting	×	0.06
HumanEval-V utilizes a rigorous annotation pipeline to ensure high-quality coding tasks.	×	0.08
HumanEval-V's tasks are designed around the visual context with minimal textual description.	×	0.04
HumanEval-V's test cases rigorously verify whether the model captures all critical visual information.	×	0.06
HumanEval-V's evaluation prioritizes visual understanding over coding proficiency.	×	0.07
HumanEval-V was evaluated with 22 LMMs.	×	0.10
HumanEval-V's evaluation includes variants such as direct code generation, Chain of Thought (CoT), and Intermediate Text	×	0.04

References

- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2506.15415v1
- http://arxiv.org/abs/2407.04973v1