What are the benchmark performance scores of Kimi-Audio-7B on reasoning mathematics coding and language unders

Assignee Research

June 10, 2026

Abstract

Understanding and reasoning over diagrams is a fundamental aspect of human intelligence. While Large Multimodal Models (LMMs) have demonstrated impressive capabilities across various tasks, existing benchmarks lack comprehensive evaluation of their diagram interpretation and reasoning abilities, particularly in coding contexts. We present HumanEval-V, a rigorous benchmark of human-annotated coding tasks that spans six task types and evaluates diverse visual reasoning capabilities. Each task features carefully crafted diagrams paired with function signatures and test cases, employing novel code

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 Kimi-Audio-7B 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 4.5/10.

3 Results

16 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.	✓	0.16
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.	×	0.11
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.	×	0.04
Claude 3.5 Sonnet can reach 55.3% pass@1 with four self-refining iterations based on test case execution feedback.	×	0.03
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.	×	0.09
The visual context must be essential for solving the task, with all relevant information contained in a single image.	×	0.05
Tasks should be designed around the visual context with minimal textual description.	×	0.04
Test cases could rigorously verify whether the model captures all critical visual information.	×	0.05
The evaluation pipeline supports LMMs with limited coding abilities by first prompting them to generate a structured dia	×	0.07
Extensive experiments with 22 LMMs were conducted.	✓	0.15

References

- http://arxiv.org/abs/2407.04973v1
- http://arxiv.org/abs/2412.21199v2
- http://arxiv.org/abs/2410.12381v3