How does 4-bit quantization affect the reasoning capability of LLaMA 3.2 and Mistral on the HumanEval benchmar

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: How does 4-bit quantization affect the reasoning capability of LLaMA 3.2 and Mistral on the HumanEval benchmark compared to FP16 baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

9 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.8/10.

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.	×	0.10
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, demanding versatile capabilities	×	0.14
HumanEval-V uses code generation tasks for evaluation instead of the multiple-choice or short-answer questions commonly	×	0.08
The visual context must be essential for solving the task in HumanEval-V, with all relevant information contained in a s	×	0.04
Tasks in HumanEval-V should be designed around the visual context with minimal textual description.	×	0.05
HumanEval-V utilizes a two-stage evaluation pipeline that supports LMMs with limited coding abilities by first prompting	×	0.07
Extensive experiments with 22 LMMs were conducted on HumanEval-V.	✓	0.15

References

- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2310.06825v1
- http://arxiv.org/abs/2410.12381v3