Robustness Metrics of Skywork-Reward-Gemma-2-27B and State-of-the-Art LMMs on HumanEval-V

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What are the robustness metrics (e.g., accuracy under adversarial diagrams) of Skywork-Reward-Gemma-2-27B compared to state-of-the-art LMMs on HumanEval-V. 20 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. 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 are the robustness metrics (e.g., accuracy under adversarial diagrams) of Skywork-Reward-Gemma-2-27B compared to state-of-the-art LMMs on HumanEval-V?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

12 papers retrieved. 20 claims extracted; 1 independently verified. Quality review score: 4.0/10.

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	×	0.08
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.	×	0.12
HumanEval-V uses code generation tasks for evaluation instead of multiple-choice or short-answer questions.	×	0.08
HumanEval-V focuses on assessing visual capabilities through a rigorous annotation pipeline.	×	0.06
HumanEval-V features diagrams that are self-explanatory with minimal textual clues.	×	0.04
HumanEval-V includes tasks that require understanding fine-grained visual elements such as matrices, arrow directions, a	×	0.03
HumanEval-V aligns with the ARC-AGI benchmark in terms of requiring comprehension from limited visual examples.	×	0.05
HumanEval-V offers a more diverse and complex set of diagrams compared to ARC's matrix-formatted diagrams.	×	0.06
HumanEval-V uses a two-stage evaluation pipeline that supports LMMs with limited coding abilities.	×	0.07
HumanEval-V prioritizes visual understanding over coding proficiency in its evaluation.	×	0.07
HumanEval-V was tested with 22 LMMs.	×	0.09
HumanEval-V includes a variant that enhances the V2C pipeline by incorporating a zero-shot CoT instruction.	×	0.08
HumanEval-V includes a variant where the model first produces a structured textual problem specification based on the di	×	0.03
The structured textual problem specification in HumanEval-V consists of three key sections: Problem Restatement, Visual	×	0.04

References

- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2008.07651v1
- http://arxiv.org/abs/2407.17856v4