DeepSeek-V4-Pro and GPT-4 Performance on HumanEval Code Generation Benchmarks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the performance difference between DeepSeek-V4-Pro and GPT-4 on HumanEval code generation benchmark scores. 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. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. 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 performance difference between DeepSeek-V4-Pro and GPT-4 on HumanEval code generation benchmark scores?.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 10 claims extracted; 1 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.	×	0.15
HumanEval-V offers a more diverse and complex set of diagrams spanning six task types.	×	0.10
Current LMMs exhibit stronger vision-to-language alignment than vision-to-code.	×	0.06
Claude 3.5 Sonnet achieves a 74.3% pass rate with 100 samples.	×	0.05
LMMs still have difficulty understanding diagrams that are trivial for humans, particularly understanding spatial transf	✓	0.18
HumanEval-V includes 253 human-annotated coding tasks.	×	0.15
Each HumanEval-V task features a diagram encoding the problem context, a function signature, and test cases to verify so	×	0.09
The visual context in HumanEval-V tasks must be essential for solving the task, with all relevant information contained	×	0.06
HumanEval-V uses code generation tasks for evaluation instead of multiple-choice or short-answer questions.	×	0.07
Proprietary LMMs' best performance occurs when they serve as diagram describers, with GPT-4o acting as the coder model.	×	0.05

References

- http://arxiv.org/abs/2308.07921v1
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2303.13375v2