

Multimodal Alignment in Code Generation Models: Accuracy and Throughput Trade-offs on HumanEval

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does multimodal alignment (e.g., text + code images) in models like GPT-4 affect code generation throughput on HumanEval while maintaining <5% drop in accuracy compared to text-only alignment. 11 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: HumanEval Pro and MBPP Pro: Evaluating Large Language Models on Self-invoking Code Generation. Research question: How does multimodal alignment (e.g., text + code images) in models like GPT-4 affect code generation throughput on HumanEval while maintaining <5% drop in accuracy compared to text-only alignment?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

15 papers retrieved. 11 claims extracted; 2 independently verified. Quality review score: 4.7/10.

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 |
|---|----------|------------|
| o1-mini achieves 96.2% pass@1 on HumanEval but only 76.2% on HumanEval Pro. | ✓ | 0.25 |
| Instruction-tuned models are less efficient on self-invoking code generation than traditional code generation tasks. | ✓ | 0.30 |
| HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) serve as fundamental benchmarks, focusing on Python function | × | 0.05 |
| Several benchmarks have expanded code evaluation benchmarks to encompass multiple programming languages (Zheng et al., 2 | × | 0.05 |
| The benchmark construction process involves three steps: (1) Self-invoking problem Generation, (2) Solutions Generation, | × | 0.11 |
| Deepseek-V2.5 is used to generate the self-invoking problems, as well as their candidate solutions and test inputs. | × | 0.07 |
| The final execution results are used to construct complete test cases with assert command. | × | 0.03 |
| Qwen2.5-Coder-7B-base achieves 59.6% on HumanEval Pro and 38.6% on MBPP Pro. | × | 0.12 |
| Qwen2.5-Coder-7B-instruct achieves 64.9% on HumanEval Pro and 35.1% on MBPP Pro. | × | 0.09 |
| DeepseekCoder-33B-base achieves 71.9% on HumanEval Pro and 38.6% on MBPP Pro. | × | 0.13 |
| DeepseekCoder-33B-instruct achieves 80.7% on HumanEval Pro and 43.9% on MBPP Pro. | × | 0.10 |

References

- <http://arxiv.org/abs/2410.12381v3>

- <http://arxiv.org/abs/2303.13375v2>
- <http://arxiv.org/abs/2412.21199v2>