Scaling Performance of Self-Invoking Code Generation Models on HumanEval Pro and MBPP Pro

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the performance of self-invoking code generation models scale with dataset size when evaluated on HumanEval Pro and MBPP Pro benchmarks. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.6/10. 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 the performance of self-invoking code generation models scale with dataset size when evaluated on HumanEval Pro and MBPP Pro benchmarks?.

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 4.6/10.

3 Results

9 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.6/10.

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.	✓	0.23
Instruction-tuned models are less efficient on self-invoking code generation than traditional code generation tasks.	✓	0.32
HumanEval and MBPP serve as fundamental benchmarks, focusing on Python function completion tasks with test-driven evalua	×	0.07
Several benchmarks have expanded code evaluation benchmarks to encompass multiple programming languages, complex tasks l	×	0.12
The benchmark construction process involves three steps: Self-invoking problem Generation, Solutions Generation, and Tes	×	0.11
Deepseek-V2.5 is used to generate self-invoking problems, candidate solutions, and test inputs.	×	0.07
An iterative method involving Python execution check and manual review is employed to ensure that all test cases pass su	×	0.02
The final execution results are used to construct complete test cases with assert command.	×	0.02
Qwen2.5-Coder-7B-base achieves 59.6% on HumanEval Pro and 38.6% on MBPP Pro.	×	0.10
Qwen2.5-Coder-7B-instruct achieves 64.9% on HumanEval Pro and 35.1% on MBPP Pro.	×	0.08
DeepseekCoder-33B-base achieves 71.9% on HumanEval Pro and 38.6% on MBPP Pro.	×	0.11
DeepseekCoder-33B-instruct achieves 80.7% on HumanEval Pro and 43.9% on MBPP Pro.	×	0.08

References

- http://arxiv.org/abs/2604.26923v1
- http://arxiv.org/abs/2410.02073v2
- http://arxiv.org/abs/2412.21199v2