Performance Comparison of CodeLlama and StarCoder on HumanEval Pro with Multi-Step Java Generation Tasks

Assignee Research

June 15, 2026

Abstract

We introduce self-invoking code generation, a new task designed to evaluate the progressive reasoning and problem-solving capabilities of LLMs. In this task, models are presented with a base problem and a related, more complex problem. They must solve the base problem and then utilize its solution to address the more complex one. This work features three key contributions. First, we propose a general recipe for generating more challenging versions of existing benchmarks, resulting in three new benchmarks: HumanEval Pro, MBPP Pro, and BigCodeBench-Lite Pro, specifically designed to assess LLMs

1 Introduction

This paper examines: HumanEval Pro and MBPP Pro: Evaluating Large Language Models on Self-invoking Code Generation. Research question: How does the pass@1 performance of CodeLlama and StarCoder compare on HumanEval Pro when evaluated with multi-step self-invoking Java code generation tasks, and what is the correlation with their performance on the original HumanEval benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.6/10.

3 Results

8 papers retrieved. 10 claims extracted; 8 independently verified. Quality review score: 7.6/10.

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.	✓	0.22
Instruction-tuned models are less efficient on self-invoking code generation than traditional code generation tasks.	✓	0.23
HumanEval and MBPP serve as fundamental benchmarks, focusing on Python function completion tasks with test-driven evalua	×	0.14
Several benchmarks have expanded code evaluation benchmarks to encompass multiple programming languages, complex tasks l	✓	0.17
Deepseek-V2.5 is used to generate self-invoking problems, candidate solutions, and test inputs.	✓	0.19
An iterative method involving Python execution check and manual review is employed to ensure that all test cases pass su	✓	0.22
Qwen2.5-Coder-7B-base achieves 59.6% on HumanEval Pro and 38.6% on MBPP Pro.	✓	0.21
Qwen2.5-Coder-7B-instruct achieves 64.9% on HumanEval Pro and 35.1% on MBPP Pro.	✓	0.23
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.	✓	0.16

References

- http://arxiv.org/abs/2412.21199v2
- http://arxiv.org/abs/2405.04520v1
- http://arxiv.org/abs/2410.12381v3