SOVEREIGN: HumanEval Pro and MBPP Pro: Evaluating Large Language Models on Self-invoking Co

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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

Analysis of: HumanEval Pro and MBPP Pro: Evaluating Large Language Models on Self-invoking Code Generation. Research goal: How does the S* hybrid test-time scaling framework compare to standard parallel scaling approaches in terms of code generation accuracy and throughput on the HumanEval and MBPP benchmarks for LLMs?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

8 papers retrieved. 4 claims extracted, 4 verified. Tribunal: 7.8/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
o1-mini achieves 96.2% pass@1 on HumanEval but only 76.2% on HumanEval Pro.	✓	0.26
Most LLMs excel in traditional code generation benchmarks like HumanEval and MBPP, but their performance declines on sel	✓	0.40
On self-invoking code generation task, the instruction-tuned models demonstrate only marginal improvements compared to t	✓	0.39
Three new benchmarks were created: HumanEval Pro, MBPP Pro, and BigCodeBench-Lite Pro.	✓	0.27

References

- http://arxiv.org/abs/2412.21199v2
- http://arxiv.org/abs/2601.16175v2
- http://arxiv.org/abs/2504.13171v1