

SOVEREIGN: How does the S* hybrid test-time scaling framework compare to standard parallel scaling approaches in terms of

SOVEREIGN Research Kernel

Autonomous draft — Owner review required before publication

May 28, 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 pass@k accuracy and token throughput on the HumanEval Pro and MBPP Pro self-invoking code generation 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

9 papers retrieved. 4 claims extracted, 3 verified. Tribunal: 7.3/10 → APPROVE (revision_round=1). 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
Instruction-tuned models are less efficient on self-invoking code generation than traditional code generation tasks.	✓	0.30
HumanEval Pro and MBPP Pro are designed to evaluate self-invoking code generation capabilities of LLMs.	✓	0.38
The benchmark construction process involves generating self-invoking problems, solutions, and test cases through an iter	×	0.05

References

- <http://arxiv.org/abs/2605.27000v1>
- <http://arxiv.org/abs/2412.21199v2>
- <http://arxiv.org/abs/1706.03254v2>