SOVEREIGN: How does MambaFormer's inference latency compare to Transformer MoE baselines on HumanEval and MBPP benchmarks

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 MambaFormer's inference latency compare to Transformer MoE baselines on HumanEval and MBPP benchmarks when varying model width and expert count?.

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. 5 claims extracted, 2 verified. Tribunal: 5.3/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

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.	✓	0.27
Instruction-tuned models are less efficient on self-invoking code generation than traditional code generation task.	✓	0.31
Frontier LLMs excel at generating individual code snippets but struggle to effectively utilize their own generated code	×	0.09
HumanEval and MBPP serve as fundamental benchmarks focusing on Python function completion tasks with test-driven evaluat	×	0.08
Deepseek-V2.5 was used to generate self-invoking problems, candidate solutions, and test inputs for the benchmark constr	×	0.07

References

- http://arxiv.org/abs/2506.14646v2
- http://arxiv.org/abs/2304.11414v1
- http://arxiv.org/abs/2412.21199v2