Dynamic Problem Difficulty Adaptation and Model Robustness in Self-Invoking Code Generation Benchmarks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the incorporation of dynamic problem difficulty adaptation in self-invoking code generation benchmarks (e.g., HumanEval Pro) affect model robustness when compared to static difficulty. 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. 5 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. 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 incorporation of dynamic problem difficulty adaptation in self-invoking code generation benchmarks (e.g., HumanEval Pro) affect model robustness when compared to static difficulty benchmarks, measured by pass@k accuracy across varying problem complexities?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

14 papers retrieved. 5 claims extracted; 2 independently verified. Quality review score: 4.8/10.

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.	✓	0.28
Instruction-tuned models are less efficient on self-invoking code generation than traditional code generation tasks.	✓	0.29
HumanEval and MBPP serve as fundamental benchmarks for code generation.	×	0.13
The evaluation landscape for Code LLMs has evolved significantly.	×	0.08
Self-invoking problem generation uses Deepseek-V2.5 to generate problems, solutions, and test inputs.	×	0.10

References

- http://arxiv.org/abs/2412.21199v2
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2503.03656v2