Scaling Laws of LLM Performance in Self-Invoking Code Generation Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the performance of LLMs on self-invoking code generation tasks scale with model size when evaluated using HumanEval+ and MBPP+ benchmarks across 1-10B parameter models. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: COrAL: Order-Agnostic Language Modeling for Efficient Iterative Refinement. Research question: How does the performance of LLMs on self-invoking code generation tasks scale with model size when evaluated using HumanEval+ and MBPP+ benchmarks across 1-10B parameter models?.

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 2.2/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 2.2/10.

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.

References

- http://arxiv.org/abs/2410.09675v1
- http://arxiv.org/abs/2505.21514v1
- http://arxiv.org/abs/2306.09896v5