To what extent does instruction length in BigCodeBench correlate with syntax error rates in generated code for

Assignee Research

May 29, 2026

Abstract

Chain-of-thought (CoT) has emerged as a groundbreaking tool in NLP, notably for its efficacy in complex reasoning tasks, such as mathematical proofs. However, its application in code generation faces a distinct challenge, i.e., although the code generated with CoT reasoning is logically correct, it faces the problem of syntax error (e.g., invalid syntax error report) during code execution, which causes the CoT result's pass@1 in HumanEval even lower than the zero-shot result. In this paper, we present Code Chain-of-Thought (CodeCoT) that integrates CoT with a self-examination process for cod

1 Introduction

This paper examines: CodeCoT: Tackling Code Syntax Errors in CoT Reasoning for Code Generation. Research question: To what extent does instruction length in BigCodeBench correlate with syntax error rates in generated code for pandas versus scikit-learn specific tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

13 papers retrieved. 11 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
CodeCoT increases pass@1 from 75.6% to 79.3% for the HumanEval dataset.	✓	0.16
CodeCoT increases pass@1 from 75.6% and 69.8% to 79.3% and 89.5% for HumanEval and MBPP datasets.	×	0.11
The trajectory of language model development has witnessed a consistent emphasis on scaling, both in terms of the model	×	0.04
Brants et al. (2007) demonstrated the advantages of models trained on a colossal 2 trillion tokens, resulting in the gen	×	0.04
The transformative potential of scaling was further emphasized with the evolution of transformer architectures, which ca	×	0.03
GPT-3 by Brown et al. (2020a) is a behemoth with 175 billion parameters.	×	0.00
Studies like that by Hestness et al. (2017) and Rosenfeld et al. (2019) evaluated the relationship between model and dat	×	0.03
The concept of chain-of-thought prompting was introduced to harness the reasoning capabilities of large language models,	×	0.06
Chain-of-thought prompting was initially proposed by Wei et al. (2022).	×	0.05
CodeCoT process is divided into four pivotal components: the CoT Prompt, Test Cases Generation, Code Generation, and Sel	✓	0.21
The iterative procedure of self-examination will be conducted through a series of multi-step iterations, allowing user-s	×	0.04

References

- http://arxiv.org/abs/2303.12869v1
- http://arxiv.org/abs/2308.08784v2
- http://arxiv.org/abs/2406.15877v4