Context Window Size Effects on BigCodeBench Pass@1 Accuracy in Llama 2 and Code Llama Python Variants

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the impact of context window size on BigCodeBench pass@1 accuracy degradation across Llama 2 and Code Llama Python variants during cross-library API usage scenarios. Transformer language models typically operate with a fixed-length context window, which has grown in step with large-scale pretraining datasets. In the BabyLM Challenge, however, many past submissions have defaulted to using much shorter sequence lengths. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: What is the Best Sequence Length for BABYLM?. Research question: What is the impact of context window size on BigCodeBench pass@1 accuracy degradation across Llama 2 and Code Llama Python variants during cross-library API usage scenarios.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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/2212.08399v2
- http://arxiv.org/abs/1811.01121v3
- http://arxiv.org/abs/2510.19493v1