How do specialized code models like Code Llama perform relative to general foundation models on BigCodeBench t

Assignee Research

May 29, 2026

Abstract

We release Code Llama, a family of large language models for code based on Llama 2 providing state-of-the-art performance among open models, infilling capabilities, support for large input contexts, and zero-shot instruction following ability for programming tasks. We provide multiple flavors to cover a wide range of applications: foundation models (Code Llama), Python specializations (Code Llama - Python), and instruction-following models (Code Llama - Instruct) with 7B, 13B, 34B and 70B parameters each. All models are trained on sequences of 16k tokens and show improvements on inputs with up

1 Introduction

This paper examines: Code Llama: Open Foundation Models for Code. Research question: How do specialized code models like Code Llama perform relative to general foundation models on BigCodeBench tasks measuring cross-library function composition versus isolated library invocation?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

8 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 7.5/10.

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.

References

- http://arxiv.org/abs/2504.06006v4
- http://arxiv.org/abs/2306.09896v5
- http://arxiv.org/abs/2308.12950v3