SOVEREIGN: What is the impact of model size scaling (7B vs 34B vs 70B) on Code Llama's performance in code infilling task

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

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

Analysis of: Code Llama: Open Foundation Models for Code. Research goal: What is the impact of model size scaling (7B vs 34B vs 70B) on Code Llama's performance in code infilling tasks on the MBPP benchmark, as measured by exact match accuracy and edit distance?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

9 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 7.2/10 $\rightarrow$ REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

References

- https://doi.org/10.48550/arxiv.2401.04088
- https://doi.org/10.48550/arxiv.2310.06825
- https://doi.org/10.48550/arxiv.2308.12950