Inference Latency Scaling of Mistral-Large-2 on MBPP Code Completion Tasks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the inference latency of Mistral-Large-2 scale with input sequence length on MBPP code completion tasks. 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. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Code Llama: Open Foundation Models for Code. Research question: How does the inference latency of Mistral-Large-2 scale with input sequence length on MBPP code completion tasks?.

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 3.3/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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

- https://doi.org/10.48550/arxiv.2308.12950
- https://doi.org/10.1007/s11704-026-60308-3
- https://doi.org/10.48550/arxiv.2403.05530