Split Computing Partitioning Strategies and Throughput in Llama3-70B vs. Codestral-34B for Code Generation

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the effect of model partitioning strategies in split computing on the throughput of Llama3-70B versus Codestral-34B for code generation tasks on HumanEval-hard. We introduce SIMCOPILOT, a benchmark that simulates the role of large language models (LLMs) as interactive, "copilot"-style coding assistants. Targeting both completion (finishing incomplete methods or code blocks) and infill tasks (filling missing segments within existing. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: SIMCOPILOT: Evaluating Large Language Models for Copilot-Style Code Generation. Research question: What is the effect of model partitioning strategies in split computing on the throughput of Llama3-70B versus Codestral-34B for code generation tasks on HumanEval-hard?.

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 6.7/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.7/10.

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/2408.15301v2
- http://arxiv.org/abs/2505.21514v1
- http://arxiv.org/abs/2306.08568v2