Mistral-Large-2 and GPT-4 Inference Latency and Throughput on MBPP Coding Tasks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the relative inference latency and throughput trade-off between Mistral-Large-2 and GPT-4 when executing complex coding tasks on the MBPP dataset. The advent of Large Language Models (LLMs) has raised concerns about their enormous carbon footprint, starting with energy-intensive training and continuing through repeated inference. This study investigates the potential of using fine-tuned Small Language Models (SLMs) as a. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Emissions and Performance Trade-off Between Small and Large Language Models. Research question: What is the relative inference latency and throughput trade-off between Mistral-Large-2 and GPT-4 when executing complex coding tasks on the MBPP dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.5/10.

3 Results

10 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.5/10.

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/2501.06658v1
- http://arxiv.org/abs/2601.08844v1
- http://arxiv.org/abs/2305.14982v2