MFOUR Vibe Framework Effects on Llama3 Inference Latency and Throughput in Code Generation

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the MFOUR Vibe Framework impact the inference latency and throughput of Llama3 compared to baseline stochastic decoding in code generation benchmarks. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Accelerating Diffusion Large Language Models with SlowFast Sampling: The Three Golden Principles. Research question: How does the MFOUR Vibe Framework impact the inference latency and throughput of Llama3 compared to baseline stochastic decoding in code generation benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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://arxiv.org/abs/2506.10848
- https://arxiv.org/abs/2503.17793
- https://arxiv.org/abs/2401.03221