Inference Efficiency of Llama-3.1-8B, Mistral-7B, and Qwen3-8B for Code Generation on MBPP

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do Llama-3.1-8B, Mistral-7B-v0.1, and Qwen3-8B compare in terms of inference efficiency (throughput and latency) when generating code on MBPP under constrained hardware conditions. Romanized Nepali, the Nepali language written in the Latin alphabet, is the dominant medium for informal digital communication in Nepal, yet it remains critically underresourced in the landscape of Large Language Models (LLMs). This study presents a systematic benchmarking of. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking Linguistic Adaptation in Comparable-Sized LLMs: A Study of Llama-3.1-8B, Mistral-7B-v0.1, and Qwen3-8B on Romanized Nepali. Research question: How do Llama-3.1-8B, Mistral-7B-v0.1, and Qwen3-8B compare in terms of inference efficiency (throughput and latency) when generating code on MBPP under constrained hardware conditions?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

16 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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/2508.15478v2
- http://arxiv.org/abs/2604.14171v1
- http://arxiv.org/abs/2306.08568v2