Tokenization Efficiency and Inference Latency in Romanized Nepali Across Llama-3.1, Mistral, and Qwen

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 3 peer-reviewed papers addressing the following research question: What is the correlation between tokenization efficiency and inference latency for Romanized Nepali tasks across Llama-3.1, Mistral, and Qwen architectures. 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. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.7/10. 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: What is the correlation between tokenization efficiency and inference latency for Romanized Nepali tasks across Llama-3.1, Mistral, and Qwen architectures?.

2 Methodology

Systematic literature search across multiple databases yielded 3 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

3 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 7.7/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Romanized Nepali is the dominant medium for informal digital communication in Nepal.	✓	0.21
Romanized Nepali remains critically underresourced in the landscape of Large Language Models (LLMs).	✓	0.24
The study evaluates Llama-3.1-8B, Mistral-7B-v0.1, and Qwen3-8B under zero-shot and fine-tuned settings using a curated	✓	0.41
Performance is quantified across five metrics spanning seven measurement dimensions: Perplexity (PPL), BERTScore, chrF++	✓	0.30
Models were fine-tuned using Quantized Low-Rank Adaptation (QLoRA) with Rank-Stabilized LoRA (rsLoRA) at rank r=32 on du	✓	0.40
At zero-shot, all three models fail to generate Romanized Nepali, each exhibiting a distinct architecture-specific failu	✓	0.29
Following fine-tuning, all three models resolve these failures and converge to BERTScore approximately 0.75 and chrF++ g	✓	0.25
Overall dimension-wise assessment across ten criteria identifies Qwen3-8B as the overall recommended architecture, being	✓	0.38

References

- https://openalex.org/W4392267988
- https://openalex.org/W7154865894
- https://openalex.org/W7115207296