Llama3 vs. Optimized LSTM Inference Latency for Edge Time-Series Forecasting

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the inference latency of Llama3 compare to optimized LSTM architectures when performing minute-level time-series forecasting on edge devices. The deployment of transformer-based models on resource-constrained edge devices represents a critical challenge in enabling real-time artificial intelligence applications. This comprehensive survey examines lightweight transformer architectures specifically designed for edge. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Lightweight Transformer Architectures for Edge Devices in Real-Time Applications. Research question: How does the inference latency of Llama3 compare to optimized LSTM architectures when performing minute-level time-series forecasting on edge devices?.

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 4.2/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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

- http://arxiv.org/abs/2601.03290v1
- http://arxiv.org/abs/2511.18613v1
- http://arxiv.org/abs/2006.10996v3