Quantized Large Language Model Inference Latency Scaling in Federated Edge Deployments

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the inference latency of quantized large language models scale with the number of concurrent edge devices in a federated learning setup for real-time threat detection. Successful integration of deep neural networks (DNNs) or deep learning (DL) has resulted in breakthroughs in many areas. However, deploying these highly accurate models for data-driven, learned, automatic, and practical machine learning (ML) solutions to end-user applications. 9 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review. Research question: How does the inference latency of quantized large language models scale with the number of concurrent edge devices in a federated learning setup for real-time threat detection?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

14 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 8.2/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Successful integration of deep neural networks (DNNs) or deep learning (DL) has resulted in breakthroughs in many areas.	✓	0.29
DL algorithms are often computationally expensive, power-hungry, and require large memory to process complex and iterati	✓	0.33
Training and inference of DL models are typically performed on high-performance computing (HPC) clusters in the cloud.	✓	0.30
Data transmission to the cloud results in high latency, round-trip delay, security and privacy concerns, and the inabili	✓	0.33
Edge devices are end devices closest to the user, such as mobile phones, cyber–physical systems (CPSs), wearables, the I	✓	0.38
These devices have limited memory, computing resources, and power-handling capability.	✓	0.26
Optimization techniques at both the hardware and software levels have been developed to handle the DL deployment efficie	✓	0.28
Understanding the existing research, challenges, and opportunities is fundamental to leveraging the next generation of e	✓	0.30
Four research directions have been pursued for efficient DL.	✓	0.19

References

- https://doi.org/10.1109/jproc.2022.3226481
- https://doi.org/10.1016/j.iotcps.2023.02.004
- https://doi.org/10.1109/jproc.2019.2918951