Performance Comparison of Quantized and Full-Precision SLMs in Heterogeneous Federated NLP

Assignee Research

June 12, 2026

Abstract

Federated Learning (FL) has emerged as a promising paradigm for enabling collaborative machine learning while preserving data privacy, making it particularly suitable for Internet of Things (IoT) environments. However, resource-constrained IoT devices face significant challenges due to limited energy,unreliable communication channels, and the impracticality of assuming infinite blocklength transmission. This paper proposes a federated learning framework for IoT networks that integrates finite blocklength transmission, model quantization, and an error-aware aggregation mechanism to enhance ener

1 Introduction

This paper examines: Energy-Efficient Quantized Federated Learning for Resource-constrained IoT devices. Research question: How does the performance of 1B-10B parameter quantized SLMs compare to their full-precision counterparts across NLP tasks in federated learning setups under heterogeneous edge device conditions?.

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 8.7/10.

3 Results

4 papers retrieved. 5 claims extracted; 5 independently verified. Quality review score: 8.7/10.

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
Federated learning (FL) enables numerous clients, such as Internet of Things (IoT) devices, to collaboratively train a g	✓	0.26
FL mitigates privacy concerns by keeping data on devices.	✓	0.16
IoT devices, including sensors, drones, and low-power computing units, are often limited in both computational and commu	✓	0.17
Frequent transmission of large model updates from each device to a central server can overwhelm available bandwidth and	✓	0.31
Many FL frameworks assume ideal communication conditions, where updates from clients are transmitted without considering	✓	0.17

References

- http://arxiv.org/abs/2509.12814v1
- http://arxiv.org/abs/2602.14301v1
- http://arxiv.org/abs/2104.03042v1