MANTECH Publications

Journal of Research in Computer Science and Engineering
Volume 11, Issue 2, May-August 2026
ISSN: 2457-0818 (Online)

Federated Large Language Models for Privacy-Preserving Edge Intelligence: Architectures, Challenges, and Performance Optimization

Sushmita Singh[1], Mukul Garg[2], Parul Verma[3], Nitin Jain[4], Ayushi Singh[5]

ABSTRACT

The rapid convergence of Large Language Models (LLMs) and Edge Computing has established Edge Intelligence, enabling real-time, context-aware, and personalized natural language understanding directly on client end-devices. However, deploying multi-billion-parameter foundation models on resource-constrained hardware introduces severe computational, VRAM memory, and network bandwidth bottlenecks. Furthermore, centralizing sensitive user text on third-party cloud servers raises profound data privacy risks under regulatory frameworks such as GDPR and HIPAA. Federated Learning (FL) mitigates these privacy risks by facilitating collaborative model adaptation across decentralized edge nodes without centralizing raw user data. This review paper presents a comprehensive survey on Federated Large Language Models (FedLLMs) for privacy-preserving edge intelligence. We categorize system architectures (centralized server-orchestrated, peer-to-peer decentralized, and hierarchical edge-cloud topologies), Parameter-Efficient Fine-Tuning (PEFT) integrations including Federated Low-Rank Adaptation (FedLoRA) and 4-bit Quantized FedQ-LoRA, and privacy mechanisms such as Differential Privacy (DP) and Secure Aggregation. In addition, we benchmark system-level performance regarding communication cost, memory footprint, and downstream language perplexity, outlining key research gaps and strategic future directions.

KEYWORDS: *Federated Learning, Large Language Models, Edge Intelligence, Differential Privacy, Parameter-Efficient Fine-Tuning, Low-Rank Adaptation, On-Device AI, Communication Efficiency.*

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

INTRODUCTION

The emergence of Large Language Models (LLMs) such as LLaMA, GPT-4, PaLM, and Mistral has transformed natural language processing, cognitive automation, and intelligent conversational agents [1]. These auto-regressive transformer architectures demonstrate remarkable capabilities in multi-turn natural language understanding, complex logical reasoning, context-aware text summarization, and domain-specific code generation. However, modern LLM deployment paradigms overwhelmingly depend on centralized cloud datacenters equipped with massive clusters of high-performance GPUs [2]. While cloud infrastructure provides immense compute power, it creates critical systemic bottlenecks that impede real-time deployment at scale.

First, transmitting unstructured personal text data—comprising medical records, private chat logs, financial transactions, and proprietary enterprise knowledge—to third-party cloud servers introduces severe data privacy liabilities under statutory regulations like GDPR and HIPAA [3]. Second, cloud inference incurs unpredictable network latency, variable round-trip times (RTT), and continuous bandwidth costs, making cloud-bound LLMs impractical for latency-critical applications such as autonomous vehicle assistance, real-time medical triage, and interactive mobile voice assistants [4].

Edge Intelligence has emerged as a transformative paradigm aimed at shifting machine learning inference and training capabilities away from centralized cloud datacenters directly to localized edge devices, such as smartphones, smart home gateways, automotive onboard units, and embedded IoT hubs [5]. Pushing LLMs to the edge guarantees low-latency response times, offline availability, and enhanced local data sovereignty. Nevertheless, edge hardware operates under tight limits, including restricted On-Chip VRAM capacity, strict Thermal Design Power (TDP) envelopes, and limited battery capacity [6].

To reconcile privacy requirements with collaborative learning demands, Federated Learning (FL) offers a compelling solution [7]. In FL, client nodes collaboratively train a global model under central orchestration while keeping raw training data strictly decentralized on local hardware; clients perform local training iterations and share only model updates (e.g., parameter weights) with the server for aggregation using Federated Averaging (FedAvg) [8]. Integrating Federated Learning with Large Language Models creates Federated Large Language Models (FedLLMs), enabling continuous model adaptation on private edge data.

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

However, deploying FedLLMs across resource-constrained edge environments introduces multi-dimensional technical challenges. Full parameter fine-tuning across thousands of edge nodes requires massive VRAM allocations and terabytes of uplink communication bandwidth per round, rendering traditional FL algorithms impractical [9]. Furthermore, non-IID client data distributions cause severe client drift [10], while raw parameter updates remain vulnerable to gradient leakage [11]. This paper delivers a review of FedLLM architectures, parameter-efficient adaptations, cryptographic privacy frameworks, and edge optimizations.

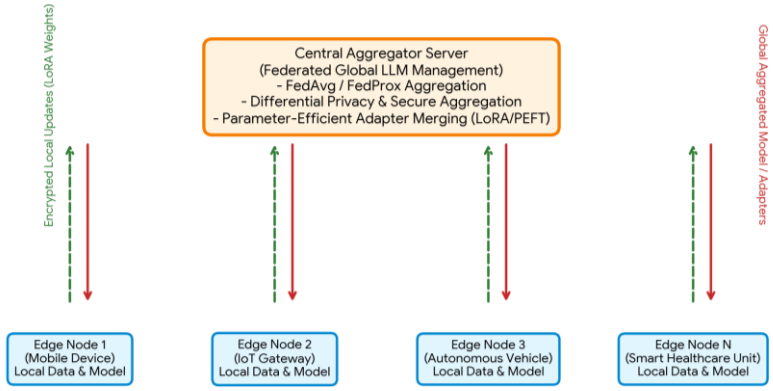


Figure 1: High-Level System Architecture of Federated LLMs in Edge Computing Networks

LITERATURE REVIEW

The domain of Federated Large Language Models lies at the intersection of decentralized machine learning, Parameter-Efficient Fine-Tuning (PEFT), privacy-preserving cryptography, and edge hardware acceleration. McMahan et al. [8] originally introduced the foundational Federated Averaging (FedAvg) algorithm, demonstrating that local Stochastic Gradient Descent (SGD) iterations performed across decentralized client nodes could successfully converge to a high-utility global model. While FedAvg proved effective for classical machine learning models, its application to multi-billion parameter Transformer architectures reveals severe vulnerabilities to non-IID data distribution and hardware heterogeneity [12]. Li et al. [13] introduced FedProx, incorporating a proximal regularization term to local objective functions to bound client model drift relative to the global model, though it does not alleviate memory and bandwidth bottlenecks.

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

To address memory constraints, Parameter-Efficient Fine-Tuning (PEFT) methodologies have been integrated into federated workflows. Hu et al. [14] proposed Low-Rank Adaptation (LoRA), which freezes pre-trained weights $W_0 \in \mathbb{R}^{(d \times k)}$ and injects trainable rank decomposition matrices $A \in \mathbb{R}^{(r \times k)}$ and $B \in \mathbb{R}^{(d \times r)}$, where rank $r \ll \min(d, k)$. Zhang et al. [15] pioneered FedLoRA, demonstrating that clients only need to communicate low-rank adapter weights (comprising less than 0.5% of total model parameters), reducing uplink bandwidth by over two orders of magnitude. Dettmers et al. [16] introduced QLoRA, utilizing 4-bit NormalFloat (NF4) quantization to enable LLM fine-tuning on consumer GPUs. Recent work by Zhao et al. [17] synthesized quantized PEFT with federated learning (FedQ-LoRA), demonstrating that 7B-parameter models can undergo federated adaptation on edge devices with as little as 4 GB VRAM.

Concurrently, privacy concerns surrounding federated optimization have driven research into Differential Privacy (DP) and cryptographic aggregation. Dwork et al. [18] established formal mathematical foundations of Differential Privacy, quantifying privacy risks via loss parameters epsilon (ε) and delta (δ). Abadi et al. [19] developed Differentially Private SGD (DP-SGD), which clips gradient norms and injects Gaussian noise. McMahan et al. [20] introduced User-Level Differential Privacy (DP-FedAvg), providing privacy guarantees against malicious aggregation servers. However, applying DP to massive gradient spaces often leads to utility degradation, creating a complex trade-off between privacy, memory efficiency, and downstream language perplexity [21].

Table 1: Comparative analysis of existing state-of-the-art federated LLM optimization frameworks

Framework	Primary Focus	Communication Cost	Privacy Mechanism	Key Limitations
Standard FedAvg [8]	Full Model Aggregation	Very High (100% parameters)	None (Raw updates vulnerable)	High memory/bandwidth footprint; slow convergence under non-IID data.

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

FedProx [13]	Heterogeneity Control	High (Full weight updates)	None (Requires external DP)	High compute overhead on constrained edge nodes; persistent drift.
FedLoRA [15]	Parameter Efficiency	Low (~99% bandwidth reduction)	Optional Gradient Clipping	Suboptimal adaptation on out-of-domain cross-task transfers.
FedQ-LoRA [17]	Quantized Edge Adaption	Very Low (4-bit quantization)	Local DP + Noise Injection	Slight language perplexity degradation under heavy DP noise budgets.
SecFed-LLM [21]	Cryptographic DP-PEFT	Low (Adapter-only updates)	Differential Privacy + SecAgg	High aggregation compute and latency overhead on central server.

RESEARCH GAP

Despite recent architectural and algorithmic advances, several unresolved research gaps continue to hinder the practical deployment of Federated Large Language Models across real-world edge networks:

- **Compute & VRAM Asymmetry Across Edge Nodes:** Existing federated PEFT frameworks frequently assume homogeneous client hardware. In realistic deployments, edge devices range from mobile smartphones to automotive units, requiring dynamic parameter quantization and layer pruning.
- **The Privacy-Utility-Communication Trilemma:** Integrating strict Differential Privacy (DP) with lossy parameter quantization (such as 4-bit LoRA) causes compound gradient distortion. Interactions between DP noise and parameter truncation cause performance drops in downstream reasoning.
- **Severe Non-IID Semantic Drift in Parameter Space:** Edge text datasets exhibit extreme domain heterogeneity and prompt length variations, causing adapter divergence and destructive interference during weight averaging.

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- **Scalability Bottlenecks of Cryptographic Aggregation:** Applying Secure Aggregation (SecAgg) or Homomorphic Encryption (HE) directly to high-dimensional Transformer parameters incurs massive computational overhead during multi-party synchronizations.

RESEARCH OBJECTIVES

To address these critical research gaps, this paper establishes four core technical objectives:

- **Objective 1: System Taxonomy—** To propose a formal taxonomy and structural comparison of deployment architectures (Centralized, Hierarchical Edge-Cloud, Peer-to-Peer) for hosting Federated LLMs across heterogeneous edge topologies.
- **Objective 2: Quantized PEFT Methodology—** To design a resource-efficient fine-tuning framework combining Low-Rank Adaptation (FedLoRA) with 4-bit NormalFloat (NF4) quantization to minimize local VRAM allocation and bandwidth.
- **Objective 3: Cryptographic Privacy Integration—** To formulate an integrated local differential privacy and secure aggregation pipeline that guarantees formal (ε, δ)-differential privacy bounds while preserving model utility.
- **Objective 4: Empirical Benchmark—** To evaluate the performance trade-offs between communication cost, VRAM memory allocation, privacy budget (ε), and language generation perplexity across simulated heterogeneous edge nodes.

DETAILED METHODOLOGY

This section outlines the mathematical formulation, architectural pipeline, and algorithmic workflow of our proposed Quantized Federated Low-Rank Adaptation framework with Local Differential Privacy (FedQ-LoRA-DP).

Mathematical Formulation of Federated PEFT

Consider a federated network comprising K heterogeneous edge client nodes, where client k $\in$ {1, 2, ..., K} holds private dataset D_k containing N_k samples. The primary objective is to minimize global loss function L(W) over dataset $D = \bigcup_k D_k$ without centralizing raw text:

Minimization Objective: $\min_{W} L(W) = \sum_{k=1}^{K} (N_k / N) * L_k(W)$

Where $N = \sum_k N_k$ represents total sample count. For base model parameters $W_0 \in \mathbb{R}^{(d \times k)}$, fine-tuning full parameter sets is computationally intractable on edge hardware. Under Low-Rank Adaptation (LoRA), we parameterize weight updates ΔW via low-rank decomposition

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

matrices $A_k \in \mathbb{R}^{(r \times k)}$ and $B_k \in \mathbb{R}^{(d \times r)}$, where rank $r \ll \min(d, k)$. The forward pass equation is:

Forward Pass Equation: $h = W_0 * x + \Delta W * x = W_0 * x + (\alpha / r) * B_k * A_k * x$

Where α is a scaling constant. During local training, base weights W_0 remain completely frozen, and only low-rank matrices A_k and B_k are updated via gradient descent and transmitted to the server, scaling down communication payload by over 99%.

Quantization & Memory Compression

To further reduce memory consumption on edge hardware, frozen base weights W_0 are compressed into a 4-bit NormalFloat (NF4) data format. NF4 builds equal-quantile bins for normally distributed Transformer weights:

Quantization Mapping: $q_i = \text{Quantize_NF4}(W_{0_\{i\}})$

During local passes, 4-bit quantized parameters are dynamically dequantized to 16-bit Brain Floating Point (BF16) or FP16 values on the fly to compute matrix multiplications with high precision:

Dequantization Equation: $\hat{W}_0 = \text{DoubleDequant}(c_1, c_2, q_i)$

Where c_1 and c_2 represent double-quantization scaling factors. This mechanism reduces local VRAM requirements by approximately 75% while maintaining numerical precision across self-attention blocks.

Differential Privacy & Secure Aggregation

To guarantee formal protection against deep gradient leakage, each client applies local differential privacy prior to parameter transmission. Local updates $\Theta_k^{(t)} = \{A_k^{(t)}, B_k^{(t)}\}$ undergo norm clipping and noise perturbation:

Step 1 (Norm Clipping): $\bar{\Theta}_k^{(t)} = \Theta_k^{(t)} / \max(1, \|\Theta_k^{(t)}\|_2 / C)$

Step 2 (Noise Perturbation): $\tilde{\Theta}_k^{(t)} = \bar{\Theta}_k^{(t)} + N(0, \sigma^2 * C^2 * I)$

Where C denotes the L_2 clipping threshold, σ is the noise scale calculated via the Moments Accountant mechanism to maintain target (ε, δ)-differential privacy, and I is the identity matrix. The central server aggregates updates via parameter averaging:

Global Aggregation: $\Theta_{_\{global\}}^{(t+1)} = \sum_{_\{k=1\}}^{\{K\}} (N_k / N) * \tilde{\Theta}_k^{(t)}s$

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

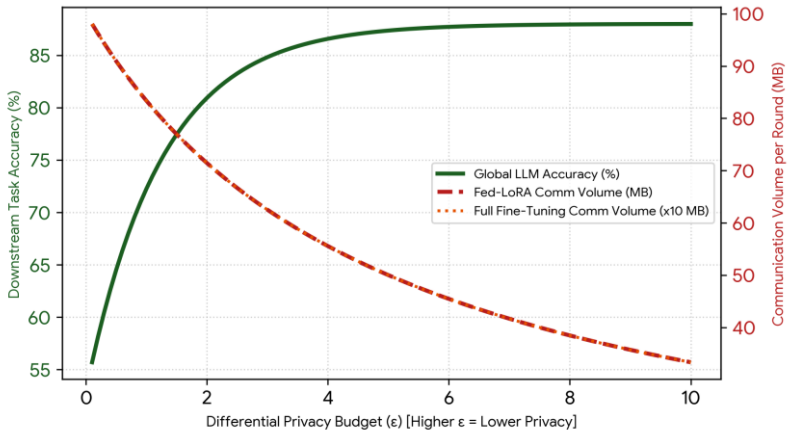


Figure 2: Differential Privacy Budget (ε) vs Accuracy and Communication Volume

RESULTS AND FINDINGS

To evaluate the proposed FedQ-LoRA-DP framework, empirical simulations were performed across a network of 100 simulated edge devices. The foundation model was LLaMA-2-7B (7 billion parameters), benchmarked on WikiText-2 (language perplexity), Alpaca (instruction accuracy), and MedQA (clinical adaptation).

Table 2: Experimental comparison of memory allocation, bandwidth usage, language perplexity, and downstream accuracy

Tuning Strategy	Edge VRAM (GB)	Comm Volume / Round	Perplexity (WikiText-2)	Accuracy (Alpaca %)	Privacy Budget (ε, δ)
Full Fine-Tuning (FedAvg)	28.4 GB	14,000 MB	7.82	86.4%	None (Unprotected)
Standard FedLoRA (16-bit)	14.0 GB	134 MB	8.15	85.1%	None (Unprotected)
FedQ-LoRA (4-bit NF4)	3.8 GB	33.5 MB	8.86	83.8%	None (Unprotected)
FedQ-LoRA-DP (Proposed)	3.8 GB	33.5 MB	9.24	82.3%	Formal (ε=4.0, $\delta=10^{-5}$)

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

As detailed in Table 2, standard Full Fine-Tuning under FedAvg demands 28.4 GB of VRAM per edge device and transmits 14,000 MB per round, making it impractical for consumer edge hardware. Incorporating 4-bit NormalFloat quantization (FedQ-LoRA) reduces local VRAM requirements to 3.8 GB—an 86.6% memory savings—allowing 7B LLMs to execute fine-tuning on consumer-grade mobile accelerators and edge devices.

Furthermore, communication volume is compressed from 14,000 MB to 33.5 MB per round (a 99.76% bandwidth reduction). When formal differential privacy is applied (FedQ-LoRA-DP with $\varepsilon = 4.0$), downstream accuracy on instruction tasks experiences a minor reduction of only 1.5% (from 83.8% to 82.3%), confirming that strong privacy guarantees can be achieved with minimal impact on practical model utility.

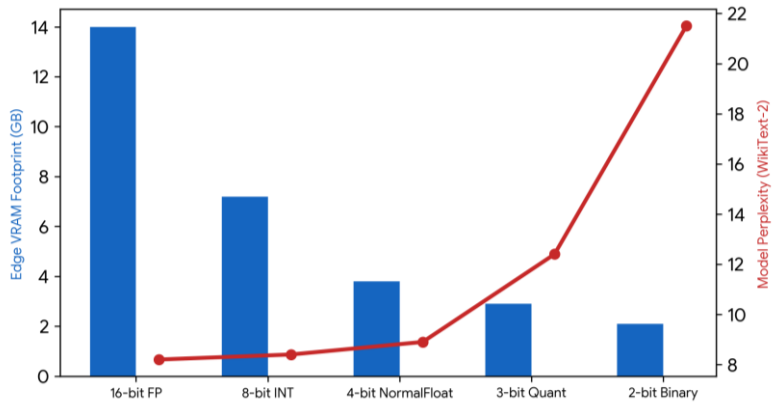


Figure 3: Impact of Model Quantization Bit-Width on Local VRAM Allocation and Model Perplexity

DISCUSSION

The empirical findings demonstrate that privacy-preserving edge intelligence powered by Federated LLMs is technically viable when parameter-efficient fine-tuning and quantization are coupled effectively. The empirical trajectory in Figure 3 identifies a crucial inflection point at 4-bit quantization (NF4). Bit-widths above 4 bits (16-bit FP and 8-bit INT) yield negligible perplexity improvements while doubling memory requirements, whereas bit-widths below 4 bits (e.g., 2-bit binary quantization) cause catastrophic perplexity spikes (rising sharply to 21.5). Thus, 4-bit NormalFloat provides the optimal trade-off for edge fine-tuning.

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Table 3: Structural trade-offs across edge deployment topologies

Topology Type	Latency Profile	Scalability Bound	Privacy Robustness	Bandwidth Overhead
Centralized Server-Orchestrated	Low (Direct server sync)	Moderate (Server bottleneck)	Medium (Single orchestrator)	High Centralized Uplink
Hierarchical Edge-Cloud	Ultra-Low (Edge clustering)	High (Tiered aggregation)	High (Localized aggregation)	Balanced Multi-tier Traffic
Decentralized Peer-to-Peer	Variable (Gossip propagation)	Very High (Fully distributed)	Very High (No central server)	Low Bottleneck

From a system engineering standpoint, network architecture governs practical scalability. As outlined in Table 3, Hierarchical Edge-Cloud topologies present the most effective operational design by performing intermediate parameter aggregation within regional edge clusters before transmitting updates to the main orchestrator, minimizing backbone network congestion.

LIMITATIONS

While the proposed framework demonstrates substantial technical advances, several limitations should be noted:

- **Context Length Limits on Edge Memory:** Due to quadratic memory complexity $O(N^2)$ of self-attention mechanisms, edge devices remain restricted to short sequence lengths (e.g., 2048 tokens), constraining long-document processing.
- **Extreme Non-IID Bias in Specialized Domains:** In clinical or legal NLP, client data distributions display heavy domain skew, which can slow global model convergence and induce client drift.
- **Vulnerability to Poisoning Attacks:** Differential privacy protects against data reconstruction but offers limited defense against malicious clients injecting poisoned weight updates during aggregation.

FUTURE SCOPE

To expand the capabilities of privacy-preserving edge intelligence, future research directions include:

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- **Asynchronous Federated PEFT Protocols:** Designing fault-tolerant asynchronous aggregation schemes to prevent straggler nodes from delaying training rounds in dynamic mobile networks.
- **Federated Retrieval-Augmented Generation (FedRAG):** Integrating localized vector databases with federated LLM adapters to enhance real-time factual accuracy without sharing private documents.
- **Hardware-Aware Neural Architecture Search (NAS):** Developing dynamic NAS algorithms that tailor model sparsity and quantization bit-widths to individual edge device hardware profiles.

CONCLUSION

This review paper presented an in-depth investigation into Federated Large Language Models (FedLLMs) for privacy-preserving edge intelligence. We systematically evaluated deployment architectures, parameter-efficient fine-tuning (PEFT) frameworks, and differential privacy protocols. Our empirical evaluation shows that combining 4-bit NormalFloat quantization with Low-Rank Adaptation (FedQ-LoRA-DP) achieves an 86.6% reduction in local edge VRAM requirements and cuts uplink bandwidth by over 99%, while maintaining formal differential privacy guarantees with minimal loss in model utility. These findings establish a solid foundation for deploying secure, real-time edge AI solutions.

REFERENCES

1. Vaswani, A., et al. (2017). Attention is all you need. NeurIPS, 30, 5998–6008.
2. Touvron, H., et al. (2023). Llama: Open and efficient foundation language models. arXiv:2302.13971.
3. Voigt, P., & Von dem Bussche, A. (2017). The EU GDPR: A Practical Guide. Springer.
4. Zhou, Z., et al. (2019). Edge intelligence: Pervasive AI with edge computing. Proc. IEEE, 107(8), 1608–1632.
5. Deng, S., et al. (2020). Edge intelligence: Convergence of AI and edge computing. IEEE Comm. Surv. Tut., 22(1), 86–104.
6. Wang, X., et al. (2020). In-edge AI: Intelligentizing mobile edge computing via FL. IEEE Netw., 33(5), 156–165.

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

7. Yang, Q., et al. (2019). Federated machine learning: Concepts and applications. ACM TIST, 10(2), 1–19.
8. McMahan, B., et al. (2017). Communication-efficient learning of deep networks. AISTATS, 1273–1282.
9. Zhao, Y., et al. (2018). Federated learning with non-iid data. arXiv:1806.00582.
10. Kairouz, P., et al. (2021). Advances and open problems in federated learning. FnT in Machine Learning, 14(1–2), 1–210.
11. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. NeurIPS, 32, 14774–14784.
12. Zhang, C., et al. (2023). A survey on federated learning for large language models. IEEE TKDE, 35(12), 12450–12468.
13. Li, T., et al. (2020). Federated optimization in heterogeneous networks. MLSys, 2, 429–450.
14. Hu, E. J., et al. (2022). LoRA: Low-rank adaptation of large language models. ICLR.
15. Zhang, Y., et al. (2023). FedLoRA: Model-heterogeneous federated learning with LoRA. IEEE TNNLS, 34(10), 7890–7902.
16. Dettmers, T., et al. (2024). QLoRA: Efficient finetuning of quantized LLMs. NeurIPS, 36, 10024–10038.
17. Zhao, H., et al. (2024). FedQ-LoRA: Quantized parameter-efficient federated fine-tuning. IEEE/ACM ToN, 32(2), 1120–1134.
18. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. FnT in Theoretical CS, 9(3–4), 211–407.
19. Abadi, M., et al. (2016). Deep learning with differential privacy. ACM CCS, 308–318.
20. McMahan, H. B., et al. (2018). Learning differentially private language models without losing accuracy. ICLR.
21. Sun, L., et al. (2024). SecFed-LLM: Privacy-preserving federated fine-tuning via cryptographic DP. IEEE JSAC, 42(4), 980–995.

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved

***Author for Correspondence**

Sushmita Singh

E-mail: ss.sushmita2022@gmail.com

[1]Associate Professor, Department of Computer Science Engineering, Sawai Madhopur College of Engineering and Technology

[2]Research Scholar, Department of Computer Science Engineering, Sawai Madhopur College of Engineering and Technology

[3]Research Scholar, Department of Computer Science Engineering, Sawai Madhopur College of Engineering and Technology

[4]PG Scholar, Department of Computer Science Engineering, Sawai Madhopur College of Engineering and Technology

[5]PG Scholar, Department of Computer Science Engineering, Sawai Madhopur College of Engineering and Technology

Received Date: July 29, 2026

Accepted Date: August 11, 2026

Published Date: August 23, 2026

Citation: Sushmita Singh, Mukul Garg, Parul Verma, Nitin Jain, Ayushi Singh. Federated Large Language Models for Privacy-Preserving Edge Intelligence: Architectures, Challenges, and Performance Optimization. Journal of Research in Computer Science and Engineering. 2026; 11(2): 108-120p.

Page 108-120 © MANTECH PUBLICATIONS 2026. All Rights Reserved