Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079321

REVIEW

When Federated Learning Meets Large Language Models: Taxonomy, Challenges, and Opportunities

Shan Jiang[1], **Wenxin You**[2], **Haoran Zhang**[3], **Shichang Xuan**[3,*] **and Jiaxing Shen**[4]

[1]School of Software Engineering, Sun Yat-Sen University, Zhuhai, China
[2]School of Art and Design, Guangzhou Institute of Science and Technology, Guangzhou, China
[3]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[4]School of Data Science, Lingnan University, Hong Kong SAR, China

*Corresponding Author: Shichang Xuan. Email: xuanshichang@hrbeu.edu.cn

Received: 19 January 2026; Accepted: 20 April 2026; Published: 15 June 2026

ABSTRACT: Large Language Models (LLMs) have been playing a transformative role in natural language understanding and generation, yet adapting LLMs to domain-specific and privacy-sensitive data remains challenging under centralized training. Federated Learning (FL) provides a promising alternative by enabling training LLMs collaboratively without sharing raw data. However, integrating FL and LLMs introduces new challenges, including model size, device heterogeneity, non-IID data, and alignment requirements. This survey offers a structured overview of the federated LLM ecosystem. We present a comprehensive taxonomy encompassing system architectures, advanced data strategies for addressing heterogeneity, and retrieval-augmented generation in federated contexts. Additionally, we review efficient adaptation methods that enable LLM tuning on resource-constrained clients and analyze data security and privacy concerns. We conclude by summarizing emerging applications in healthcare, industry, software engineering, and finance, and by outlining open problems and research opportunities for scalable, secure, and responsible federated LLM deployment.

KEYWORDS: Large language models; federated learning; foundation models; federated large language models

1 Introduction

The advent of Large Language Models (LLMs) has revolutionized artificial intelligence and empowered machines with powerful capabilities in natural language understanding, reasoning, and generation [1]. LLMs, such as GPT [2] and LLaMA [3], leverage massive corpora and billions of parameters to achieve emergent behaviors that were previously unattainable. Despite the advancements, deploying LLMs in real-world applications encounters a critical challenge: the need for vast, diverse datasets to fine-tune LLMs conflicts directly with increasingly stringent data privacy regulations and the proprietary character of domain-specific information [4]. Centralized training, which requires assembling raw data into a central authority, incurs high risks of data leakage and violates data sovereignty laws.

Federated Learning (FL) has been considered as a transformative paradigm to resolve the tension of data privacy [5]. By enabling collaborative model training across decentralized devices without exchanging raw data, FL provides a mechanism to harness the collective intelligence of siloed datasets while preserving privacy [6]. The integration of FL and LLM, i.e., federated LLM, promises to unlock new frontiers in personalized healthcare, secure financial analysis, and industrial automation. Unlike traditional FL, which typically

focuses on training small-scale models from scratch, federated LLM primarily addresses the challenges of fine-tuning and aligning pre-trained LLMs in resource-constrained, heterogeneous environments.

The rapid evolution of federated LLM has catalyzed a surge in academic research, resulting in a plethora of architectures, optimization techniques, specialized toolkits, and benchmarks [7,8]. Frameworks such as FederatedScope-LLM [9] and OpenFedLLM [10] have been developed to standardize the deployment of LLMs on decentralized infrastructure, providing researchers with robust environments for benchmarking performance. Similarly, comprehensive toolkits bridging continuous pre-training and alignment [11] have streamlined the transition from general-purpose models to domain-specific experts.

While recent surveys [12–17] have explored the intersection of FL and LLMs, they often treat federated LLMs as a general extension of LLMs. This survey distinguishes itself by focusing on the system-level operations of LLMs in federated settings. We analyze the unresolved tension between the massive memory requirements of LLMs and the limited resources of edge devices, providing a critical comparison of solutions spanning system architecture, model fine-tuning, data security and privacy, and applications.

Specifically, Cheng et al. [12] and Chen et al. [13] provide foundational motivations but often overlook recent advancements in parameter-efficient fine-tuning and trustworthy alignment. Other studies [14,15,17], offer valuable insights into edge computing and fusion strategies, respectively, but may not fully address the complex interplay between security, data heterogeneity, and retrieval-augmented generation. Ren et al. [16] discuss LLMs broadly, yet a dedicated, granular analysis of the specific methodologies for federated LLM tuning remains necessary. Table 1 compares this survey against existing works highlighting our distinct contributions.

Table 1: Comparison of this survey with existing literature on Federated LLMs.

Reference	System Architecture	Model Fine-Tuning	Data Security & Privacy	Applications
Cheng et al. [12]	○	○	●	●
Chen et al. [13]	○	●	○	○
Thakur et al. [14]	●	●	○	○
Piccialli et al. [15]	●	○	○	●
Ren et al. [16]	○	●	●	○
Hu et al. [17]	○	●	●	○
This Work	●	●	●	●

Note: ●: Detailed coverage; ○: Partial or limited coverage.

Fig. 1 depicts the structure of this survey. Section 3 summarizes the advanced system architectures of federated LLMs compared with the naive centralized one. Section 4 investigates the advanced model fine-tuning methods in contrast to naive full fine-tuning. Section 5 identifies common threats to federated LLM systems and the mitigation approaches from the perspectives of privacy preservation, security and robustness, and alignment and fairness. Section 6 presents the prototypes and applications of federated LLMs in academia and industries. Finally, in Section 7, we identify the desired properties of federated LLMs, introduce the trilemma of balancing efficiency, privacy, and utility, and outline open challenges and future directions.

We target at providing a comprehensive taxonomy and a critical analysis of the current federated LLM ecosystem. This survey moves beyond simple enumeration of methods to investigate the how, where, and safeguards of federated LLMs. Table 2 illustrates a unified taxonomy of federated LLMs from the perspectives of system, model, and data and articulates the common strategies, goals, key bottlenecks, and representative

techniques in each perspective. Note that although Table 2 separates federated LLM research into system, model, data, and application dimensions for clarity, these dimensions are not independent in practice. In particular, privacy, security, and alignment requirements act as cross-cutting constraints that shape feasible choices in system architecture, fine-tuning strategy, and deployment settings.

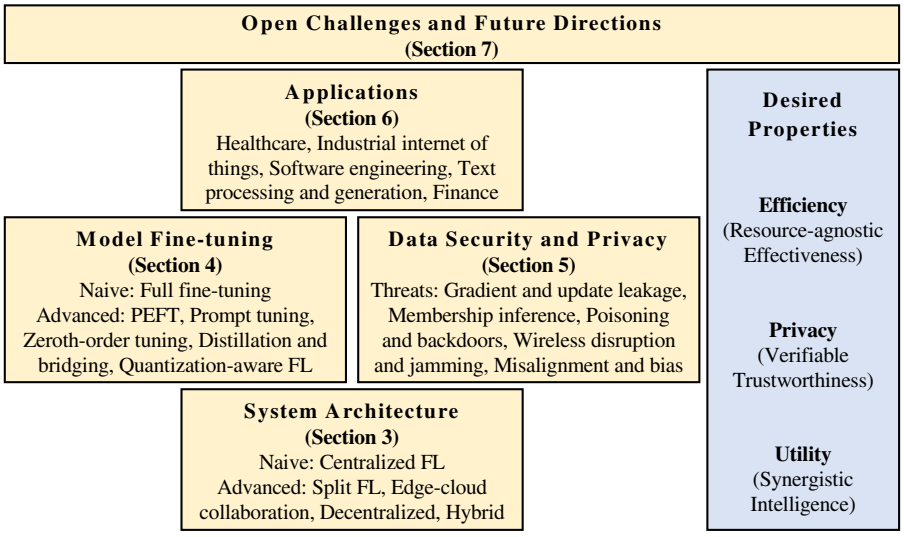

Figure 1: Survey structure.

Table 2: A unified taxonomy of federated LLMs, where privacy, security, and alignment act as cross-cutting constraints across system, model, and application design.

Dimension	Common Categories	Goal	Key Bottlenecks	Representative Techniques
System architecture	Centralized; split FL; edge-cloud collaboration; decentralized	Scale training under heterogeneous resources	Communication, stragglers, trust	FedAvg variants, asynchronous FL, split FL cut-layer design, P2P aggregation, distributed ledger, federated RAG
Model fine-tuning	Full fine-tuning; PEFT; prompt tuning	Reduce memory and communication while keeping quality	Bandwidth, GPU memory	Federated PEFT, federated prompt engineering, knowledge distillation, in-context learning, ZO optimizers
Data security and privacy	Data leakage and privacy; data security against attacks; alignment and fairness	Protect confidentiality, integrity, safety	Data leakage, poisoning, misalignment	Differential privacy, secure aggregation, selective encryption, robust aggregation, federated RLHF and DPO, fairness constraints

(Continued)

Table 2 (continued)

Dimension	Common Categories	Goal	Key Bottlenecks	Representative Techniques
Applications	Healthcare; industrial IoT; software engineering; text generation and legal; finance	Domain adaptation under regulation	Compliance, evaluation, deployment	Domain PEFT, private RAG, distillation to edge

To ensure a thorough and cutting-edge analysis of this rapidly evolving field, we conduct a systematic literature search using the Scopus database, as shown in Fig. 2. The search strategy uses the keywords federated learning and large language model, focusing on publications published in or before 2025 to capture the most recent developments. The initial query yields 613 results. Subsequently, we apply a manual screening process based on two primary exclusion criteria. On the one hand, we exclude purely conceptual studies that present simplistic ideas without concrete methodological frameworks or validation. On the other hand, we exclude papers that use LLMs solely as auxiliary tools (e.g., for validating results generated by other models) rather than as the subject of federated integration. The screening finally yields 104 research papers. The literature review method ensures that the survey focuses exclusively on substantive methodological contributions to the intersection of FL and LLMs.

Topics	#Papers	Percentage
Toolkits and benchmarks	5	4.81%
Reviews	6	5.77%
System architecture	21	20.19%
Model fine-tuning	26	25.00%
Data security & privacy	18	17.31%
Applications	28	26.92%
Total	104	100%

Figure 2: Survey method and distribution of research papers on different topics.

We adopt a quantitative analytical method and classify the reviewed corpus of the 104 primary studies into six distinct categories. As illustrated in Fig. 2, the distribution reveals that applications and model fine-tuning are the dominant trajectories, accounting for approximately 27% (28 papers) and 25% (26 papers) of the corpus, respectively. It indicates a dual focus within the community: expanding the practical utility of federated LLMs across domains while simultaneously addressing the computational constraints of training them. System architecture (21 papers) and data security and privacy (18 papers) form the field's structural backbone, collectively representing nearly 38% of the work. Conversely, the scarcity of contributions in toolkits and benchmarks (5 papers) highlights a critical gap, suggesting that while the field

is innovating rapidly in methods and use cases, it currently lacks standardized evaluation frameworks and unified development platforms.

The main contributions of this work are as follows:

- We explore the system architecture and deployment strategies (Section 3) required to support federated LLM, ranging from split FL and edge-cloud collaboration to decentralized topologies, including a discussion on advanced data strategies such as handling non-IID distributions and implementing federated retrieval-augmented generation.
- We provide a detailed examination of efficient fine-tuning methodologies (Section 4), categorizing cutting-edge techniques in federated parameter-efficient fine-tuning, prompt engineering, and memory-optimized instruction tuning. We analyze how these methods mitigate the communication and computational bottlenecks inherent to massive models.
- We conduct a rigorous analysis of data security and privacy (Section 5), synthesizing research on privacy preservation, security against poisoning attacks, and model alignment with human preferences via reinforcement learning from human feedback.
- We survey diverse applications of federated LLMs across healthcare, industry, and finance (Section 6), and outline the ongoing challenges and future research directions (Section 7).

2 Preliminaries

The convergence of LLMs and FL represents a synthesis of advanced natural language processing capabilities with decentralized, privacy-preserving computation. This section establishes the foundational concepts of LLMs and FL and defines federated LLMs.

2.1 Large Language Models

The evolution of natural language processing has been revolutionized by the introduction of the Transformer architecture, which utilizes self-attention mechanisms to capture long-range dependencies in textual data [18]. Modern LLMs, such as GPT [2] and LLaMA [3], scale the architecture to billions of parameters, training on massive corpora to develop emergent abilities in reasoning, coding, and general knowledge generation [19]. LLMs typically require two-phase training: pre-training to learn statistical language patterns, followed by fine-tuning (or instruction tuning) to align the model with specific tasks or human preferences [20]. While pre-training establishes the model's knowledge base, it requires immense computational resources, often limiting it to centralized data centers equipped with high-performance GPU clusters [21]. Consequently, adapting LLMs to private, domain-specific data remains a challenge, as transferring sensitive information to a central server for fine-tuning often violates data sovereignty regulations.

2.2 Federated Learning

FL addresses the constraints of data isolation by enabling collaborative model training without exchanging raw data. In the most classical FedAvg algorithm [22], a central server coordinates a global model and distribute it selected clients. Each participating client k performs local training using its local private dataset with the objective of minimizing a local loss function $F_k(w)$, producing a local update. The server then aggregates the clients' updates, typically via a weighted average, to update the global model parameters w. Mathematically, the objective is to minimize the global loss function $F(w)$:

$$\min_{w} F(w) = \sum_{k=1}^{K} p_k F_k(w) \tag{1}$$

where K is the total number of clients and p_k represents the relative weight of the k-th client, often proportional to the size of its local dataset. The FL paradigm ensures that raw data never leaves the local device, significantly reducing privacy risks. However, traditional FL faces hurdles such as statistical heterogeneity (Non-IID data), where the data distribution across clients varies significantly, leading to model drift and slow convergence [23].

2.3 Federated Large Language Models

Fig. 3 depicts the general pipeline of fine-tuning and personalization of LLMs in a federated setting. Federated LLMs extend the FL paradigm to the training and fine-tuning of transformer-based architectures. Unlike traditional FL, which often trains models from scratch, federated LLMs typically focus on federated fine-tuning of pre-trained LLMs. Specifically, the global model is initialized with pre-trained weights. Clients collaborate to adjust the weights (or a subset thereof) based on private instruction sets or domain-specific corpora.

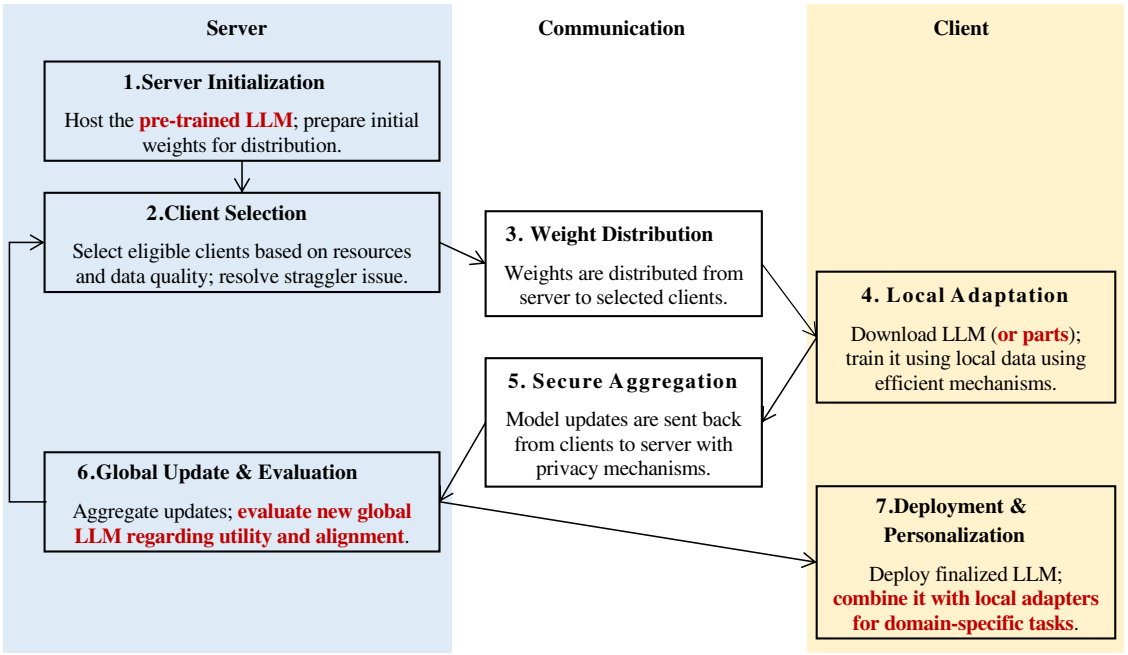

Figure 3: A general end-to-end federated LLM pipeline.

The integration of LLMs into federated settings introduces unique system challenges not present in standard FL. The sheer size of LLM parameters, ranging from billions to trillions, imposes severe communication bottlenecks when transmitting massive model updates between clients and the central server. In addition, the memory requirements for backpropagation often exceed the hardware capabilities of edge devices or consumer-grade GPUs found in decentralized nodes. Consequently, the standard FedAvg approach is often computationally infeasible for LLMs, necessitating the adoption of parameter-efficient techniques and communication-compression strategies to make distributed training viable.

In federated LLMs, heterogeneity arises along at least three distinct dimensions: data, system, and tasks. Data heterogeneity refers to the discrepancies in local data distributions, label spaces, and data quality across clients, which often induce client drift and unstable convergence. System heterogeneity refers to varieties in hardware capabilities, memory, bandwidth, and availability, leading to stragglers, stale updates,

and unequal participation. Task heterogeneity refers to differences in downstream tasks, personalization targets, or alignment preferences across clients, which may render a single global optimum ill-defined and lead to negative transfer under naive aggregation.

3 System Architecture

Table 3 summarizes the system-level design space of federated LLMs, focusing on where computation takes place and what information must be exchanged during training. Unlike conventional FL with small models, federated LLMs are dominated by the scale of transformer parameters and activations, which makes the choice of architecture a first-order determinant of feasibility, efficiency, and trust. We categorize existing systems into five representative architectures: (i) centralized FL, which retains the classic server-client aggregation pipeline; (ii) split FL, which partitions model layers to shift memory and compute to the server at the cost of transmitting intermediate activations; (iii) edge-cloud collaboration, which generalizes split execution via resource-aware offloading and scheduling; (iv) decentralized topologies, which remove the central coordinator to improve resilience and reduce trust assumptions; and (v) hybrid designs that combine the first four primitives. For each category, Table 3 highlights the principal communication object, practical advantages, and key limitations, providing a concise guide for selecting architectures under different resource, privacy, and deployment constraints.

Table 3: System architecture taxonomy for federated LLMs and major trade-offs.

Architecture	Where LLM Runs	Communication Object	Advantages	Limitations
Centralized FL	Client trains locally; server aggregates	Gradients/weights/adapter deltas	Simple; strong coordination; mature tooling	Bandwidth heavy for LLMs; stragglers; single point of failure
Split FL	Client runs early layers; server runs later layers	Activations and gradients at the cut layer	Enables training when clients cannot host the full model	Activation leakage risk; high uplink usage; cut-layer tuning complexity
Edge-cloud collaboration	Dynamic partitioning/offloading across edge and cloud	Mixed (deltas, activations, partial states)	Resource-aware scheduling; better latency/throughput	Orchestration complexity; privacy boundary management
Decentralized	Peers exchange/validate updates without a central server	Peer updates and ledger proofs/records	Removes central trust; resilient; incentives possible	Consensus overhead; aggregation quality control; incentive design
Hybrid	Combine above (e.g., split FL, asynchronous FL, and PEFT)	Task-dependent	Integration of advantages above	Hard to analyze; evaluation and reproducibility challenges

From the perspective of heterogeneity, the architectural choices primarily respond to system heterogeneity, i.e., variation in client resources, connectivity, and availability. Although architectural decisions also interact with privacy and data diversity, their primary role is to ensure end-to-end training feasibility under uneven computational and communication capabilities. In contrast, data heterogeneity is discussed in Section 3.4.1, while task heterogeneity becomes especially visible in personalized fine-tuning and multi-task adaptation settings discussed in Section 4.

Beyond efficiency and scalability, architectural choices in federated LLMs also determine the system's trust model and attack surface. For instance, centralized FL concentrates coordination but creates a single point of failure, split FL alleviates client memory limits at the cost of activation leakage risk, and decentralized designs reduce central trust assumptions while introducing new validation and consensus challenges.

3.1 Split Federated Learning

The prohibitive memory requirements of LLMs often render standard FL infeasible on edge devices with limited resources. Split FL addresses the bottleneck by partitioning the model architecture between the client and server. In split FL, the client executes the initial layers (the head) to generate intermediate activations, which are then transmitted to the server for forward propagation through the remaining layers (the tail) [24]. Gradients flow in reverse during backpropagation. The structural division allows clients to train massive models while holding only a fraction of the parameters locally.

Implementing split FL for LLMs requires robust frameworks that can handle the high communication overhead of transmitting activation maps. FedsLLM [25] introduces a parallelized split FL architecture specifically optimized for communication networks. By enabling parallel training across multiple clients and synchronizing the split-layer boundaries, the framework mitigates the straggler effect common in sequential split learning. FedsLLM significantly reduces the computational load on edge devices compared to full-model FL, although it introduces a heavy dependency on uplink bandwidth.

VFLAIR-LLM [26] is proposed to rigorously evaluate different split federated LLM architectures. It provides a comprehensive benchmark suite. VFLAIR-LLM elucidates the trade-offs between cut-layer selection, communication latency, and model performance, and serves as a critical tool for system designers, offering metrics that quantify how different splitting strategies affect the convergence rate and resource consumption of various LLM backbones.

While architectural splitting solves memory constraints, it introduces optimization challenges, particularly when data distributions across clients are non-IID. The separation of layers can decouple the learning of low-level features (client-side) from high-level semantic reasoning (server-side).

To counter potential degradation in generalization capability, recent research has integrated advanced minimization techniques into the split FL workflow. Tan et al. [27] propose incorporating sharpness-aware minimization into the local client updates. By seeking parameters that lie in neighborhoods of uniformly low loss rather than sharp minima, the method improves the model's robustness to heterogeneous data. The integration ensures that the split configuration does not compromise the global model's ability to generalize across diverse user prompts.

Split FL is frequently cited as a privacy-preserving solution because raw data remains local. However, the transmission of intermediate activations creates a new attack surface. A critical analysis [28] challenges the assumption of inherent security. In particular, inversion attacks can reconstruct original inputs from cut-layer activations, especially in the context of LLMs with high semantic density. The finding necessitates the integration of differential privacy or activation compression mechanisms to obfuscate the transmitted signals.

Furthermore, the reliance on continuous communication makes split FL vulnerable to physical-layer disruptions. R-SFLLM [29] addresses the issue by proposing a jamming-resilient framework. Recognizing that wireless channels are susceptible to interference, the authors design a robust transmission protocol that maintains training stability even under active jamming attacks. R-SFLLM ensures that the collaborative fine-tuning process remains viable in hostile or unstable network environments.

3.2 Edge-Cloud Collaboration

The deployment of LLM within federated settings necessitates a paradigm shift from simple parameter aggregation to complex edge-cloud orchestration. Unlike traditional FL, where models are small enough for trivial on-device processing, LLMs impose severe computational and memory demands. Consequently, recent literature focuses on intelligent scheduling, resource-aware offloading, and collaborative architectures that bridge the gap between high-capacity cloud servers and resource-constrained edge devices.

Standard FedAvg suffers from the straggler effect, where the global training speed is bottlenecked by the slowest device. The straggler issue is exacerbated in federated LLM due to the heterogeneity of edge hardware. To address the issue, Tri-AFLLM [30] introduces a resource-efficient adaptive framework. By abandoning synchronous lock-step updates in favor of an asynchronous protocol, the system allows faster clients to contribute more frequently while slower nodes update partially. Tri-AFLLM significantly improves convergence speed without idling powerful resources.

Beyond timing, the qualitative match between a client's data and the model's objective is crucial. FedCLLM [31] shows that random client selection is inefficient for LLM fine-tuning. Instead, it utilizes domain descriptions to match clients with specific downstream tasks. By filtering participants based on the semantic relevance of local data, the framework ensures that the global model aggregates high-value updates, reducing communication rounds and improving task-specific performance.

Once participants are selected, orchestrating the compute resources becomes important. LTQA [32] proposes a delay-optimization strategy. It continuously monitors network latency and computational throughput, dynamically assigning training loads to nodes that minimize the overall system delay.

Software optimization alone is often inadequate to bridge the resource gap for LLMs. Recent approaches have begun to exploit specialized hardware at the edge. Huang et al. [33] integrate Embedded Data Processing Units (DPUs) into the workflow. By offloading specific tensor operations and data preprocessing tasks to DPUs, the main CPU/GPU is freed for core model updates, effectively expanding the compute envelope of edge devices.

For environments where individual devices cannot hold even a quantized model, DisLLM [34] proposes a distributed inference and training architecture. The framework partitions the LLM across a mesh of resource-constrained devices, treating the edge network as a single cohesive computer. While DisLLM enables deploying larger models, it requires rigorous privacy assurances to prevent data leakage between collaborating nodes.

To unify the methods above, comprehensive frameworks are required. MPCTF [35] establishes a multi-party collaborative training protocol that standardizes the interaction between data owners, compute providers, and model architects. The abstraction layer simplifies the setup of decentralized LLM training.

Complementing the training phase, FoRA [36] focuses on efficiently propagating knowledge from the cloud to the edge. It optimizes the fine-tuning process for on-device LLMs by selectively transferring parameters that yield the highest accuracy gains, thereby minimizing bandwidth consumption during the downlink phase. Meanwhile, ensuring the federated LLM systems operate securely in production can

be addressed by federated data modeling [37], which outlines cloud-native architectures for deploying collaborative models while adhering to strict security compliance standards.

3.3 Decentralized Approaches

Centralized orchestration in federated LLMs often suffers from a single point of failure and limited scalability. Furthermore, the dependence on a central server raises concerns regarding censorship and trust. To mitigate the risks, the architecture of federated LLM is increasingly moving towards decentralized, peer-to-peer topologies, often underpinned by blockchain technology to ensure integrity and incentivize participation.

In environments lacking a trusted central authority, blockchain ledgers provide an immutable record of model updates. PureLLM [38] introduces a blockchain-driven decentralized framework specifically for personalized FL. By removing the central aggregator, PureLLM allows devices to exchange updates directly via a peer-to-peer mesh. The blockchain consensus mechanism validates these updates, filtering out malicious contributions before they are assimilated into the local models. The structure not only enhances robustness against poisoning attacks but also improves resource efficiency by allowing nodes to selectively assimilate knowledge relevant to specific tasks.

A critical challenge in decentralized federated LLM is the free-rider problem, where participants obtains the global model with limited or no contributions. DISM [39] addresses the free-rider problem by embedding a reward system within the blockchain protocols. Smart contracts automatically dispense tokens or reputation marks to clients based on the quality and volume of data contributions. Such an economic layer is essential for sustaining long-term collaborative fine-tuning, because the computational cost of training LLMs is too high for altruistic participation alone.

3.4 Advanced Data Strategies

The utility of federated LLMs relies heavily on the quality of distributed data and how the data is used. Unlike centralized training, where data is curated and shuffled, federated environments suffer from extreme statistical heterogeneity (Non-IID data) and varying data quality. Furthermore, the static nature of parametric knowledge in LLMs conflicts with the dynamic information available at the edge. Hence, we explore strategies to mitigate data heterogeneity through advanced filtering and distillation, as well as methods to augment generation using distributed knowledge bases.

3.4.1 Handling Non-IID Data

Data heterogeneity in federated LLMs refers to differences in local data distributions across clients, including concept drift, domain shift, task imbalance, and quality variance. The main technical effect is to increase divergence between local and global optimization trajectories, which can slow convergence, destabilize aggregation, and reduce final generalization performance. In contrast to system heterogeneity, which affects training speed and participation, data heterogeneity directly influences the statistical consistency of the learned global model. Data heterogeneity in federated LLM manifests in two primary forms: distribution shifts (concept drift) and quality variance. The standard FedAvg algorithm often fails when client datasets diverge significantly in task composition or noise levels.

Not all local data contributes positively to the global model. Low-quality or irrelevant samples can degrade performance and slow convergence. To address this, FedDDF [40] introduces a mechanism to assess data utility during training. By dynamically filtering out samples with high loss variance or low alignment with the global objective, the system ensures that the model learns only from high-value local interactions.

A related challenge is the scarcity of labeled instruction data on edge devices. Clients typically possess abundant unstructured text but lack the instruction-response pairs required for fine-tuning. FedIT-U2S [41] proposes an automated pipeline where an auxiliary teacher model synthesizes instructions from raw local text. It allows the federated network to utilize vast amounts of previously unusable data for instruction tuning without manual annotation.

When clients perform fundamentally different tasks, forcing a single global model to master all of them simultaneously can lead to negative transfer. MIRA [42] frames the problem as federated multi-task learning. Rather than aggregating all weights equally, MIRA employs a routing mechanism that allows the global model to specialize parameters for distinct tasks across client clusters. It preserves the unique capabilities required by specific edge cases while maintaining a shared knowledge base.

Deploying full-scale LLMs on resource-constrained clients is often infeasible. Knowledge distillation has been considered a potent solution to close the gap between large server-side models and smaller client-side networks. FedBiOT [43] introduces a bilevel optimization framework in which clients perform local fine-tuning without requiring access to the full model. Through a distillation process, the client optimizes a lightweight adapter or compressed model, which is then synchronized with the server. FedBiOT reduces communication costs significantly while retaining the performance benefits of larger architectures.

Similarly, FedBridgeICL [44] explores the synergy between small and large models via in-context learning. Instead of traditional gradient updates, FedBridgeICL uses the large server model to generate high-quality context vectors or demonstrations. The vectors or demonstrations are transmitted to smaller client models to guide inference. Such a bridging approach allows small edge models to emulate the reasoning capabilities of larger counterparts without incurring the computational cost of full parameter synchronization.

3.4.2 Federated Retrieval-Augmented Generation

Parametric knowledge in LLMs is prone to hallucinations and quickly becomes outdated. Retrieval-Augmented Generation (RAG) mitigates the issue by fetching relevant context from an external database. Federated RAG refers to the process of responding to user prompts by fusing local private knowledge with global shared knowledge (as illustrated in Fig. 4).

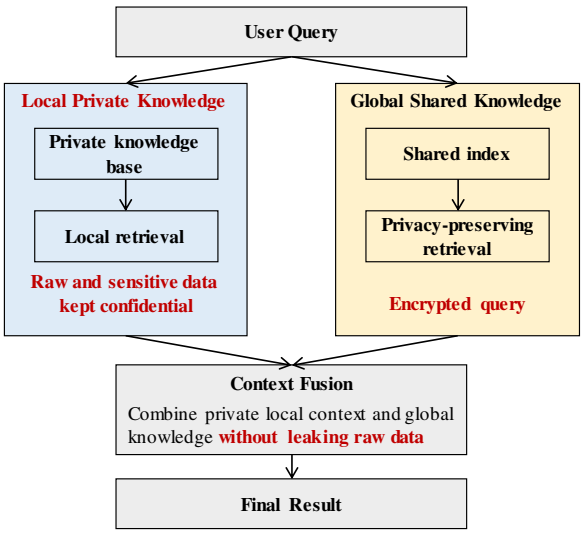

Figure 4: Workflow of federated retrieval-augmented generation.

DF-RAG [45] proposes a framework specifically designed for collaborative environments like healthcare. First, a local retrieval module accesses private patient records on the client device to answer specific queries. Second, a privacy-preserving global retrieval module accesses a shared index of medical knowledge contributed by other institutions. The separation ensures that sensitive local data (e.g., patient history) never leaves the device, while general medical insights are shared. By aggregating retrieval results rather than raw data, DF-RAG enables the LLM to generate informed, personalized, and globally consistent responses.

Federated RAG introduces additional privacy risks beyond those of standard federated fine-tuning because leakage may occur throughout the retrieval pipeline. In particular, sensitive information may be exposed through index leakage (e.g., embeddings, metadata, or index structure), query leakage (user intent or private prompts), retrieval-result leakage (returned passages, scores, or provenance), and update leakage from retriever or index optimization. The leakage channels are tightly coupled with retrieval quality: stronger privacy protection through obfuscation, restricted sharing, or encrypted retrieval may reduce recall, ranking precision, or latency performance. Hence, federated RAG should be evaluated not only by generation utility but also by the joint privacy-retrieval-effectiveness trade-off.

4 Model Fine-Tuning

Table 4 organizes the efficiency-centric adaptation strategies that make federated training of LLMs practical under tight client constraints. Because end devices and cross-silo participants often lack the memory, compute, and uplink bandwidth required for full-parameter backpropagation, federated LLM research increasingly focuses on shrinking the parameters to be trained, reducing model updates and communication overhead, and accommodating heterogeneous hardware. We group existing approaches into six method families: full fine-tuning, parameter-efficient fine-tuning (PEFT), prompt tuning, zeroth-order (gradient-free) tuning, distillation and bridging, and quantization-aware FL. Furthermore, we summarize the training target, client-side requirements, and the resulting communication cost for each method. The taxonomy serves as a conceptual decision aid for selecting federated fine-tuning approaches under different resource, privacy, and deployment constraints. We stress that the methods are extremely sensitive to hyperparameter settings and system conditions, including privacy considerations, quantization level, client participation, and data heterogeneity. For instance, PEFT and prompt tuning are often preferred when bandwidth and memory are limited, while distillation and quantization are attractive for deployment-oriented scenarios where the goal is to produce smaller or lower-precision models without moving raw data off-client. Meanwhile, note that fine-tuning methods are not chosen solely for resource efficiency. The methods also determine which parts of the model and intermediate states are exposed during FL, and what security assumptions and defense mechanisms are feasible.

4.1 Federated Parameter-Efficient Fine-Tuning

Federated LLMs faces a critical bottleneck: the prohibitive cost of full-parameter fine-tuning. With model sizes ranging from billions to trillions of parameters, transmitting full gradient updates saturates communication bandwidth and overwhelms the memory capacities of edge devices. Federated PEFT avoids transmitting full gradient updates by freezing the pre-trained backbone and updating only a small subset of parameters or additional adapter modules [46]. Among different strategies, Low-Rank Adaptation (LoRA) [47] is regarded as the de facto standard, enabling clients to train low-rank matrices that approximate weight updates. Recent literature has expanded beyond basic LoRA implementations to address specific federated challenges, including system heterogeneity, communication constraints, and data personalization.

Table 4: Taxonomy of efficient fine-tuning methods for federated LLMs.

Method Family	What Is Trained	Client Requirements	Communication Cost	Notes (Typical Use Cases)
Full fine-tuning	All parameters	Large GPU memory; stable connectivity	Very high	Rare for edge; mostly for datacenters or small LLMs
PEFT	Low-rank matrices or adapters; backbone frozen	Moderate memory; gradient access	Low to medium	Default choice in federated LLM; supports personalization and heterogeneity
Prompt tuning	Prompt vectors or tokens; model frozen	Low memory; sometimes gradient-free possible	Very low	Works well when the model is black-box or the devices are weak
Zeroth-order tuning	No backprop graph; gradient estimated from probes	Very low memory; more forward passes	Low to medium	Good for tight-memory clients; can be noisy or slow to converge
Distillation and bridging	Student (small) model or adapters learn from teacher outputs	Mostly inference on the teacher; training on the student	Low (often logits or demos)	Edge deployment; compress knowledge; security needs auditing
Quantization -aware FL	Low-bit weights or adapters (heterogeneous precision)	Device-specific precision support	Low	Addresses heterogeneity; aggregation across precisions is non-trivial

4.1.1 Communication and Resource Optimization

Standard implementations of PEFT often fail to consider the stochastic characteristics of wireless edge networks. Wireless channels introduce latency and packet loss, necessitating joint optimization of learning and transmission. AirFL-LoRA [48] formulates an optimization problem that balances the computation rank with wireless resource allocation. By adjusting the rank of LoRA adapters based on channel quality, the system maximizes convergence speed while adhering to strict energy budgets.

Sustainability has also become a primary objective. The carbon footprint of training LLMs is substantial, and federated settings exacerbate the sustainability concern because redundant local computations are required. Iftikhar et al. [49] illustrate that federated PEFT greatly saves energy consumption in comparison with centralized training by minimizing data movement and leveraging low-power edge processors. Furthermore, Jiang et al. [50] establish baseline protocols for reducing the trainable parameter space, proving that massive reductions in communication overhead can be achieved with negligible degradation in downstream task performance.

4.1.2 Dynamic and Adaptive Mechanisms

Static adapter configurations often yield sub-optimal results because different layers of an LLM contribute unequally to task adaptation [51]. DynamicFedPEFT [52] introduces a mechanism to dynamically adjust the trainable parameters during the training process. Rather than fixing the rank or the specific modules beforehand, the framework monitors gradient norms to allocate trainable parameters to the most sensitive layers, thereby accelerating convergence.

Similarly, not all parameters within a LoRA adapter are equally important. The adaptive importance-aware LoRA approach [53] integrates an importance scoring mechanism. Local clients prune less significant ranks during the update phase, transmitting only the most critical weight changes to the server. The method serves a dual purpose: it acts as a compression scheme to reduce uplink traffic and prevents overfitting by regularizing the adaptation process on local datasets.

4.1.3 System Heterogeneity and Aggregation Strategies

A pervasive challenge in FL is system heterogeneity, where clients possess varying computational capabilities. Enforcing a uniform LoRA rank across all clients forces the system to operate at the speed of the slowest device [54]. To mitigate the issue, Ning et al. [55] permit clients to train adapters with ranks proportional to local resources. However, aggregating matrices of different dimensions introduces structural errors. To this end, an error-compensated aggregation protocol is proposed that aligns diverse local updates into a coherent global model without discarding information from weaker clients.

Furthermore, Zhu et al. [56] focus on the algorithmic stability of aggregation. By refining the update rules, the framework dampens the noise introduced by non-IID data distributions, ensuring that the global aggregation of low-rank matrices remains stable even when local updates diverge significantly.

From a theoretical perspective, the stability of federated LoRA aggregation depends not only on update magnitude but also on the geometric compatibility of client-specific low-rank subspaces. When clients adopt heterogeneous ranks or use different layers under non-IID data, their low-rank updates may span mismatched directions, leading naive averaging to introduce projection error and amplify client drift. Existing studies suggest that aggregation becomes more stable when the principal adaptation subspaces are approximately aligned and when update norms are appropriately normalized or error-compensated [56]. However, a general theory covering heterogeneous LoRA ranks, partial participation, and strongly non-IID settings remains incomplete. We therefore view current results as an important first step rather than a complete characterization of the stability of federated PEFT.

Finally, the efficacy of federated PEFT relies heavily on which clients participate in training. Solat and Lee [57] argue that random selection is inefficient for LLM adaptation. The proposed strategy evaluates potential participants based on both computational readiness and informational value of local data, ensuring that communication rounds are utilized by the most impactful contributors.

4.1.4 Personalization

While the global model aims for generalization, local clients often require personalization. FedALoRA [58] proposes an adaptive local aggregation scheme. Instead of simply overwriting the local adapter with the global average, the method computes a weighted combination, allowing the local model to retain knowledge specific to the user's data distribution while benefiting from global knowledge.

4.2 Federated Prompt Engineering

Federated PEFT reduces the computational burden by updating a subset of model parameters; however, it still requires access to the model's gradients and internal weights. Such a requirement poses a barrier for clients with severe resource constraints or when the LLM is deployed as a black-box service. Federated prompt engineering addresses these limitations by optimizing the input space rather than the model space [59]. By learning optimal discrete tokens or continuous embeddings, federated prompt engineering enables clients to steer the global model's behavior with minimal communication overhead.

4.2.1 Continuous Soft Prompt Optimization

Soft prompting involves prepending learnable continuous vectors to the input sequence. The vectors are optimized via backpropagation while keeping the LLM frozen. In the federated context, soft prompting allows for extreme parameter efficiency, as only the small prompt vectors need to be aggregated.

Recent applications have extended the soft prompting technique beyond traditional natural language processing tasks. For instance, the FPTuning-LLM framework [60] adapts LLMs for time-series forecasting in the hotel industry. By treating historical booking data as textual sequences and applying soft prompt tuning, the system leverages LLMs' semantic reasoning to predict future demand without exposing sensitive commercial data. FPTuning-LLM shows that learnable prompts can effectively bridge the modality gap between numerical time-series data and pre-trained linguistic representations.

4.2.2 Discrete Prompting and Synthetic Augmentation

Unlike soft prompts, which are continuous embeddings, discrete prompts consist of human-readable tokens. Optimizing these tokens is challenging due to the non-differentiable nature of discrete text. However, discrete prompts offer better interpretability and transferability.

To address data scarcity among local clients, Tanimura et al. [61] introduce a mechanism that uses synthetic examples. The framework augments local datasets with synthetically generated samples, allowing the prompt tuner to converge more robustly. By mixing real and synthetic data, the system mitigates the risk of overfitting to sparse local distributions while maintaining the privacy guarantees inherent to FL.

4.2.3 Black-Box and Edge-Centric Adaptation

In many real-world deployments, clients interact with LLMs via APIs (known as Model-as-a-Service) and do not have access to gradients. The model-as-a-service paradigm necessitates the use of derivative-free optimization strategies. FebBPT [62] targets edge environments where computational power is severely limited. Instead of backpropagating errors, the system employs evolution strategies or reinforcement learning signals to iteratively refine prompts based on the model's output. FebBPT shifts the computational load from gradient calculation to inference, enabling the deployment of sophisticated prompt tuning on lightweight edge devices.

4.2.4 Reasoning and Scheduling Efficiency

Beyond simple task adaptation, prompt engineering in federated settings is increasingly focused on enhancing LLMs' reasoning capabilities and optimizing system throughput.

Liu et al. [63] explore aggregating reasoning paths. Rather than merely averaging prompt vectors, the approach encourages clients to share successful Chain-of-Thought (CoT) templates. The global model

thereby learns to structure its reasoning more effectively, leading to improved accuracy on complex query-answering tasks.

Simultaneously, the efficiency of processing these prompts is critical. FedLLM-PPS [64] addresses the latency bottlenecks associated with handling multiple prompt requests. A parallel scheduling algorithm is proposed to optimize the order and batching of prompt evaluations across the federated network. It ensures that the computational resources of participating clients are used to their fullest, reducing the overall time-to-convergence of the global prompt model.

4.3 Quantization and Zeroth-Order Optimization

Training LLMs within a federated ecosystem imposes severe memory and communication constraints [65]. Standard backpropagation is often infeasible on consumer-grade edge devices because it requires storing activation maps and optimizer states. Consequently, the research community has pivoted toward advanced optimization paradigms that circumvent full-precision gradient computation. The techniques include Zeroth-Order optimization, heterogeneous quantization, and hybrid gradient strategies.

4.3.1 Zeroth-Order Optimization

Zeroth-Order optimization has emerged as a compelling alternative to First-Order methods. By approximating gradients through forward passes and random perturbations, Zeroth-Order methods eliminate the need to store the computation graph, thereby significantly reducing memory footprints.

Recent works have sought to stabilize Zeroth-Order convergence in federated settings. FedAdamZO [66] integrates adaptive momentum into the derivative-free process. FedAdamZO shows that combining Adam-style momentum with zeroth-order estimators helps overcome the high variance typically associated with random gradient approximations, making it suitable for memory-constrained fine-tuning. Besides algorithmic innovation, theoretical foundations have been strengthened by studies such as FedMeZO [67]. In particular, existing analyses derive convergence bounds for federated zeroth-order tuning under non-IID data by assuming smooth objectives, bounded estimator variance, and controlled client heterogeneity. The results are important because they show that derivative-free tuning can remain convergent even when local data distributions are biased. At the same time, the guarantees are still conditional on assumptions whose validity may weaken in practical federated LLM settings with heavy-tailed updates, partial participation, and highly heterogeneous tasks. Therefore, current theory provides a useful baseline justification for zeroth-order federated tuning, but its extension to more realistic large-scale LLM regimes remains an open problem.

Beyond efficiency, derivative-free methods offer inherent privacy advantages. FedDPZO [68] highlights that transmitting perturbed loss values or weights, rather than explicit gradients, reduces the attack surface for gradient leakage exploits. FedDPZO aligns well with the privacy-preservation mandate of FL while maintaining competitive performance on downstream tasks.

4.3.2 Heterogeneous Quantization

In practical deployments, client devices possess diverse hardware capabilities, ranging from high-end workstations to mobile phones. Uniform optimization strategies often fail to accommodate the disparity.

To address system heterogeneity, FAH-QLoRA [69] proposes a flexible framework. Clients with limited resources participate by training heavily quantized models (e.g., 4-bit), while capable clients utilize higher precision (e.g., 8- or 16-bit). The server aggregates the heterogeneous updates into a unified global model.

FAH-QLoRA combines the quantization technique with LoRA to ensure that no participant is excluded due to hardware limitations, effectively maximizing the available training data across the network.

4.3.3 Hybrid Strategies

While Zeroth-Order and quantization reduce resource demands, First-Order methods generally yield faster convergence. Researchers have thus developed hybrid, accelerated schemes to balance efficiency and training speed.

FedHO [70] introduces a memory-efficient protocol via hybrid gradient computation. By strategically alternating between precise gradient calculations and approximated updates, or by offloading specific computational chunks, the framework reduces the peak memory usage on local devices without sacrificing the accuracy benefits of gradient-based learning.

Furthermore, the choice of optimizer plays a critical role in instruction tuning. FEDNPAIT [71] investigates the application of Nesterov-accelerated Adaptive Moment Estimation and its partially adaptive variant in federated environments. The study reveals that advanced momentum-based optimizers can significantly accelerate convergence for instruction-following tasks compared to standard SGD or FedAvg, particularly in the complex loss landscapes of LLMs.

5 Data Security and Privacy

Table 5 summarizes the security- and privacy-critical threats for federated LLMs and the corresponding defense mechanisms commonly adopted in the literature. Although FL avoids the direct centralization of raw training data, federated LLM pipelines still expose multiple attack surfaces, including gradients and adapter updates, split-learning activations, model outputs accessible to queriers, and even the underlying communication channel, that can leak sensitive information or compromise model integrity. Moreover, the scale and memorization capacity of LLMs amplify risks, while heterogeneous and partially trusted participants create opportunities for poisoning and backdoor insertion. In this section, we organize threats by the attack surface and downstream impact, then map them to representative defenses. The final column highlights unresolved challenges, particularly the utility-privacy trade-off at LLM scale, the cost of robust validation, and the difficulty of auditing fairness and alignment when sensitive attributes remain local, motivating the open problems and future directions in Section 7. Importantly, the data security and privacy issues concerned are not independent of the technical choices of system architecture and fine-tuning methods. Instead, the issues emerge from and constrain the system architecture and fine-tuning mechanisms adopted by federated LLM systems.

5.1 Privacy Preservation

While FL fundamentally mitigates privacy risks by retaining raw data on local devices, integrating LLMs introduces novel vulnerabilities. The vast parameter space of LLMs allows for the unintended memorization of training data [72], and the exchange of high-dimensional gradients or parameter updates can be exploited to reconstruct original inputs [73]. Consequently, privacy mechanisms in federated LLM must evolve beyond standard aggregation to address specific threats ranging from sensitive data identification to cross-cloud leakage.

Effective privacy preservation begins before the training process initiates. Traditional FL assumes that keeping data local is sufficient, yet it overlooks the risk that models may inadvertently learn and regurgitate personally identifiable information. FedAPILLM [74] proposes utilizing the federated LLM framework itself to detect vulnerabilities. By training a federated model to recognize sensitive fields within API structures, the

system can automatically flag or redact such information in real time. FedAPILLM proactively ensures that the data fed into the fine-tuning process is sanitized, thereby reducing the surface area for potential privacy breaches during subsequent model interactions.

Table 5: Threat-defense taxonomy for data security and privacy in federated LLMs.

Threat	Attack Surface	Impact	Common Defenses	Open Issues
Gradient and update leakage	Gradients, adapter deltas, activations (split FL)	Reconstruction of private text or personally identifiable information	DP, secure aggregation, selective encryption, activation protection	Utility loss; tight privacy accounting for PEFT or prompting
Membership inference	Model outputs and representations	Whether a record/client participated	DP, query throttling, auditing	Hard for LLMs with memorization; eval standards lacking
Poisoning and backdoors	Malicious client updates; distillation channels	Targeted misbehavior; integrity loss	Robust aggregation, anomaly detection, update validation, attestation	Adaptive attackers; expensive validation at LLM scale
Wireless disruption and jamming	Physical layer or packet corruption	Training instability; degraded convergence	Channel-aware aggregation; redundancy; scheduling	Joint design of communications and learning is still immature
Misalignment and bias	Preference data; prompts; aggregation policies	Harmful or unfair outputs	Federated RLHF and DPO, fairness constraints, prompt screening	Pluralistic values conflict; fairness auditing with local-only attributes

The gradients exchanged during LLM fine-tuning contain significant information about the local batch data. Adversaries can employ gradient inversion techniques to reconstruct the original text. A recent study [75] highlights that the risk is particularly acute in LLMs due to the semantic richness of the text data. The authors advocate advanced noise injection mechanisms that go beyond standard Differential Privacy (DP) to address gradient inversion attacks. By analyzing the correlation between gradient sparsity and information leakage, the proposed methods selectively perturb updates to maximize privacy while preserving linguistic quality.

Applying heavy cryptographic protocols or uniform noise to billions of parameters is computationally prohibitive and detrimental to model convergence. Pan and Wu [76] address the efficiency bottleneck. The core premise is that not all model parameters contribute equally to privacy leakage. By identifying and encrypting only the most sensitive subsets of parameters (often those associated with rare tokens or specific attention heads), the system significantly reduces computational overhead while maintaining robust protection levels.

Besides selective encryption, sampling strategies offer a statistical shield. FCLM [77] introduces a method designed for non-IID environments. Instead of aggregating updates from all clients or random

subsets, the system clusters clients based on similarities in their data distributions. By sampling from the clusters, the aggregation process masks the contribution of any single device within the group variance. FCLM not only enhances privacy by breaking the direct lineage between a specific client and the global update but also stabilizes training on heterogeneous data.

5.2 Security and Robustness

While privacy mechanisms protect data confidentiality, they do not inherently secure the model against integrity attacks. The distributed nature of FL introduces significant attack surfaces, particularly Model Poisoning and Backdoor Attacks, where malicious clients inject deceptive updates to manipulate the global model's behavior [78]. In the context of LLMs, the threats are amplified by the model's vast parameter space and the opacity of deep neural networks. Furthermore, the deployment of federated LLM in wireless and edge environments necessitates robustness against not only malicious actors but also adversarial environmental conditions.

Backdoor attacks in FL typically involve a sophisticated adversary embedding a hidden trigger into the model, causing it to misclassify inputs only when the trigger is present [79]. In the era of LLMs, backdoor attacks have become more complex. The work LBKD [80] highlights a critical vulnerability in specialized domains. Specifically, standard aggregation is insufficient to filter out subtle triggers embedded in domain-specific data. The proposed bidirectional knowledge distillation framework reveals a dual nature: while distillation is often used for model compression, it can also serve as a sophisticated vector for implanting persistent backdoors that survive aggregation. It suggests that the very mechanisms used to make federated LLM efficient require rigorous security auditing to prevent the propagation of malicious traits.

Beyond malicious data injection, the physical transmission medium presents a security challenge. Federated LLM deployments often rely on wireless channels susceptible to noise and intentional jamming. ROFED-LLM [81] addresses the fragility of LLM training in adversarial wireless environments. The study shows that standard robust aggregation algorithms focus primarily on outlier updates caused by data poisoning but fail to account for channel-induced corruption. By modeling the adversarial interference in the wireless spectrum, ROFED-LLM introduces a channel-aware aggregation scheme. It ensures that the global model maintains high fidelity even when the communication links are actively compromised or heavily degraded, a prerequisite for deploying federated LLM in critical infrastructure.

Interestingly, recent research shifts the paradigm from protecting the model to using the model as a protection mechanism, as LLMs possess strong reasoning capabilities that can be harnessed to secure underlying systems.

FedITD [82] exemplifies the idea by applying PEFT to the domain of Insider Threat Detection. Traditional centralized detection systems risk exposing sensitive user logs. FedITD utilizes pre-trained LLMs to analyze behavioral logs locally. By fine-tuning the model via FL, the system learns to identify complex, non-linear patterns of insider threats across an organization without centralizing the raw audit trails. It demonstrates that federated LLM can serve as a potent cybersecurity tool, provided the training process itself remains secure.

Similarly, Luo and Ji [83] propose a framework to enhance the security of Edge-Cloud AI systems. LLMs are employed to monitor data collaboration flows between edge devices and the cloud. The LLM acts as a semantic guardian, identifying anomalous data exchanges that deviate from established security protocols. It creates a symbiotic relationship where the FL framework updates the security model, and the security model, in turn, protects the FL infrastructure.

5.3 Alignment and Fairness

Ensuring that LLMs align with human intent and ethical standards is essential. In centralized settings, alignment and fairness are typically achieved through Reinforcement Learning from Human Feedback (RLHF) [20]. However, the federated paradigm introduces unique challenges: human preferences are heterogeneous across clients, and sensitive demographic data required for fairness auditing remains local. Hence, decentralized alignment strategies have been popular to balance global convergence with pluralistic values and rigorous fairness guarantees.

The direct translation of RLHF to federated environments is non-trivial due to the communication overhead of maintaining multiple models (actor, critic, reward, and reference models). FedRLHF [84] addresses the structural impediments by proposing a framework that ensures convergence with privacy preservation. Theoretical guarantees are provided that federated policy optimization can match centralized performance, assuming a coherent global preference exists.

However, the assumption of a single global preference is often flawed. Different cultures and user groups possess distinct values. PluralLLM [85] challenges the one-size-fits-all alignment paradigm. Instead of aggregating conflicting feedback into a diluted global average, the framework facilitates pluralistic alignment. By leveraging the natural data partitioning of FL, PluralLLM allows the model to maintain multiple alignment heads or adapt to diverse value systems, thereby respecting the heterogeneity of the user base rather than suppressing it.

In the literature, Proximal Policy Optimization (PPO) is often considered the standard for RLHF; however, it is computationally intensive on edge devices. Direct Preference Optimization (DPO) [86] has emerged as a resource-efficient alternative. Recent work explores the intersection of DPO and behavioral economics in federated LLMs. KTO [87] investigates how human cognitive biases, specifically loss aversion defined in Prospect Theory, influence federated fine-tuning. The authors argue that standard DPO assumes rational preference labeling. By modeling the non-linear value perception of human annotators (where losses loom larger than gains), the proposed method refines the loss function to better capture true user intent, leading to more robust alignment in distributed settings.

Besides alignment, fairness is also important in LLMs. Fairness in federated LLM is two-fold: ensuring the model does not discriminate based on protected attributes, and ensuring fair representation of client contributions.

In the context of prompt engineering, bias often propagates from the input phrasing. FedPSF-LLM [88] introduces a mechanism to vet prompts locally before the prompts influence the global model. By evaluating the response disparities across demographic groups at the client level, the system filters out prompts that trigger discriminatory outputs, preventing bias from polluting the global aggregation.

Specific domains require even stricter adherence to fairness and privacy. In the Internet of Medical Things, an incorrect or biased recommendation can have life-altering consequences. PFFPO [89] integrates differential privacy with fairness constraints directly into the optimization objective. The method employs a multi-objective approach that maximizes helpfulness while simultaneously minimizing the statistical distance between outputs across different patient demographics. It ensures that the alignment process does not inadvertently favor specific medical profiles over others.

The evaluation of security, privacy, and fairness mechanisms in federated LLMs depends critically on the assumed adversary model. Relevant dimensions include whether the server is honest-but-curious or malicious, whether attackers control one or multiple colluding clients, and whether the attack surface lies in model updates, communication messages, outputs, or retrieval components. Accordingly, reported defense performance should be interpreted together with the attacker's capabilities and prior knowledge.

In practice, the defense mechanisms are commonly evaluated along several complementary dimensions: attack success or backdoor persistence for integrity attacks; exposure risk under membership, reconstruction, or attribute inference for privacy attacks; utility degradation in downstream task performance; and system overhead in communication, latency, or computation. We therefore emphasize that rigorous comparison requires reporting not only defense effectiveness but also the associated privacy-utility-efficiency trade-offs under clearly stated threat assumptions.

6 Applications

Federated LLMs are increasingly adopted in application domains where data is valuable but difficult to centralize due to privacy, regulation, intellectual property, or operational constraints. In these settings, organizations or devices can collaboratively adapt an LLM to domain-specific language and tasks while keeping raw data local. Compared with conventional deployment of a fixed LLM, federated LLMs offer a path to continuous, distributed improvement from heterogeneous participants, enabling personalization and faster adaptation to shifting terminology, policies, and context. At the same time, real-world applications impose non-trivial requirements beyond model quality, including communication efficiency, on-device resource limits, robustness to non-IID data, compliance with data-governance rules, and protections against leakage and poisoning. The following applications illustrate how the constraints shape practical system designs and highlight recurring evaluation criteria such as utility, latency, privacy risk, and operational reliability.

However, it is important to note that the application literature is uneven in maturity. While some studies report empirical gains on real or institutionally sourced datasets, many others remain proof-of-concept, simulation-based, or prototype-oriented. Therefore, the following discussion distinguishes between demonstrated capabilities under constrained validation settings and aspirational deployment scenarios, particularly in safety-critical domains such as healthcare and finance, where privacy preservation alone is insufficient for real-world adoption.

6.1 Healthcare

The biomedical domain is among the most critical yet challenging environments for deploying LLMs. While general-purpose LLMs demonstrate remarkable capabilities in natural language understanding, the direct application in healthcare is impeded by strict privacy regulations, such as GDPR [90], and the siloed nature of medical records. FL has consequently emerged as a vital paradigm, enabling the training of robust biomedical models across distributed institutions without the need for centralized data aggregation. Recent surveys highlight that integrating LLMs into federated healthcare networks offers unique opportunities to address data heterogeneity and scarcity, though significant challenges remain regarding interpretability and communication overhead [91].

One of the most actively explored applications of federated LLMs in healthcare is the extraction of insights from unstructured clinical text, such as physician notes and discharge summaries, although current evidence is still concentrated in prototype studies and controlled evaluations rather than routine clinical deployment. Pharmacovigilance, specifically the identification of Adverse Drug Reactions (ADR), benefits significantly from federated LLMs. By leveraging federated architectures, researchers can aggregate knowledge from diverse patient populations to identify rare side effects without compromising patient anonymity [92]. However, the computational cost of fine-tuning massive language models on local hospital servers, which often lack high-end GPU clusters, acts as a barrier to adoption. To mitigate the resource constraints, recent methodologies propose selective layer fine-tuning. Instead of updating all model parameters or utilizing standard adapters, there are techniques to identify and train only the most relevant layers of

the transformer architecture, thereby achieving performance comparable to full fine-tuning while drastically reducing memory usage and communication costs [93].

Beyond static text analysis, modern healthcare requires predictive modeling that accounts for the longitudinal nature of patient history. Electronic Health Records (EHR) are inherently temporal, containing sequences of visits, diagnoses, and treatments. Standard LLMs often treat input data as static context, failing to capture the dynamic progression of chronic conditions. Novel frameworks now incorporate temporal-aware mechanisms within the federated setting. These approaches utilize time-series capable prompts and specialized attention mechanisms to model disease progression, allowing the global model to learn temporal patterns of patient deterioration across multiple hospitals while keeping the raw temporal sequences local [94].

The scope of federated LLM in medicine extends beyond text to multimodal applications, particularly in medical imaging and report generation. In colonoscopy analysis, precise polyp segmentation is critical for early cancer detection. Integrating vision-language foundation models with federated strategies allows for the development of clinically applicable tools. By employing LoRA within a federated framework, institutions can collaboratively refine large vision models for specific segmentation tasks, ensuring high accuracy and privacy preservation simultaneously [95].

More broadly, multimodal federated LLMs deserve explicit attention because many realistic deployments combine text with images, video, waveforms, or sensor streams rather than relying on text alone. A representative case is medical image-report collaboration, where one institution may hold radiology images while another contributes report corpora or downstream annotation expertise. In such settings, federated multimodal adaptation must address not only standard non-IID effects but also cross-modal representation alignment, missing modalities across clients, and stronger privacy risks from perceptual data. Parameter-efficient adaptation can be particularly attractive because it can localize updates to modality-specific encoders or fusion layers, but aggregating such heterogeneous multimodal adapters remains less mature than in text-only federated LLMs. It suggests that multimodal federated LLM applications are promising but remain constrained by limited benchmarks, evaluation protocols, and theory.

Furthermore, the generative capabilities of LLMs have been explored for administrative and diagnostic documentation tasks such as medical report generation and summarization. Existing studies indicate encouraging performance in controlled settings, but broader validation is still needed before these systems can be considered reliable for routine clinical workflows. Generating medical reports from heterogeneous data sources poses a challenge due to the varying data formats and equipment standards across different hospitals. Communication-efficient heterogeneous FL frameworks have been developed to address the issue. The developed systems allow hospitals with different local model architectures or computational capabilities to collaborate on a shared objective, such as generating coherent radiology reports, by exchanging knowledge through prototype alignment or distilled representations rather than raw gradients [96]. It ensures that even smaller clinics with limited infrastructure can benefit from and contribute to state-of-the-art medical report generation systems.

6.2 Industrial Internet of Things

The integration of LLMs into the Industrial Internet of Things (IIoT) and next-generation networks marks a transition from static sensing to intelligent, semantic reasoning at the edge. A comprehensive review of the synergy suggests that FL is essential for deploying these models in networked systems, primarily to navigate the trade-offs between the computational demands of LLMs and the strict privacy requirements of industrial data [97].

In the realm of autonomous mobility, vehicles act as mobile computing nodes that require real-time decision-making capabilities. Traditional cloud-centric training struggles with the high latency and bandwidth limitations inherent in vehicular networks. To address the issue, recent frameworks like iFLOW have introduced scalable multi-model FL architectures specifically designed for computing on the wheels, enabling cars to collaboratively learn from diverse road conditions without sharing raw sensor streams [98]. Furthermore, the heterogeneity of driving scenarios necessitates robust adaptation techniques. Federated instruction tuning strategies have been proposed to enhance feature diversity, allowing autonomous systems to generalize better across varying traffic environments and rare edge cases [99]. Besides control systems, in-cabin intelligent assistants are also evolving through federated strategies. By applying LoRA to automotive systems, manufacturers can fine-tune LLMs for personalized driver interactions while minimizing the communication overhead associated with transmitting full model updates [100].

Unmanned Aerial Vehicle (UAV) swarms pose distinct challenges, characterized by high mobility and intermittent connectivity. Standard federated protocols often fail when nodes frequently drop out of the network. Specialized algorithms for low-altitude UAV networks now incorporate dropout resilient mechanisms, ensuring that the collaborative fine-tuning of LLMs remains stable even when swarm members disconnect unexpectedly due to interference or battery constraints [101].

Beyond mobility, the manufacturing sector is leveraging federated LLM to enhance the intelligence of digital twins and production lines. Edge-centric architectures are being developed to bring LLM inference closer to the factory floor, addressing the latency requirements of real-time process monitoring [102]. A critical application in manufacturing is the maintenance of Digital Twins, which requires massive amounts of labeled data. Novel asynchronous FL frameworks empower the digital replicas by utilizing LLMs for intelligent data labeling. Such an approach facilitates secure data sharing and model training across different manufacturing sites, effectively bridging the gap between physical assets and the virtual counterparts [103].

The underlying network infrastructure supporting the applications above also benefits from LLMs. As 5G networks become increasingly complex, identifying security threats is critical. FedLLMGuard utilizes federated LLMs to detect anomalies in network traffic, leveraging LLMs' semantic understanding to identify subtle attack patterns that traditional rule-based systems might miss [104]. Similarly, in video surveillance, FedVAD enhances anomaly detection by employing GPT-driven semantic distillation. The technique transfers the rich semantic knowledge of large models into lightweight edge detectors, improving the identification of unusual events in video streams while preserving privacy [105].

Finally, the application of federated LLM extends to the design of the hardware itself and scientific discovery. In the complex field of chip design, the FedChip framework demonstrates how federated LLMs can assist in generating and optimizing Verilog code for AI accelerators, allowing disparate design houses to collaborate on better chip architectures without revealing proprietary intellectual property [106]. In scientific modeling, adaptive FL with local LLMs has proven effective for simulating complex photonic and chemical systems, accelerating discovery by aggregating insights from distributed experimental data [107].

6.3 Software Engineering

The application of LLMs to software engineering has revolutionized coding workflows, yet the proprietary nature of commercial source code presents a barrier to centralized training. FL offers a viable path to leverage private code repositories for model improvement without exposing intellectual property. To address the practical constraints of distributed development environments, the F-CodeLLM framework introduces a methodology for adapting LLMs to software tasks across decentralized clients, ensuring that local coding patterns contribute to global model intelligence without raw data leakage [108]. Similarly, adaptive fine-tuning frameworks have been proposed to address the heterogeneity of developer environments, enabling

models to dynamically adapt to specific programming languages and project requirements across different organizations [109].

A critical task in software maintenance is understanding legacy code. Recent research explores code summarization techniques that function without direct access to the source text. By leveraging federated strategies, systems can generate natural language descriptions of code functionality while keeping the underlying logic on local servers, thereby maintaining strict confidentiality [110]. Security within the deployment pipeline also benefits from this distributed approach. Novel prompt engineering strategies have been integrated into FL workflows to enhance the detection of malicious code during deployment, ensuring that security checks evolve collaboratively across different nodes to identify emerging threats [111].

6.4 Text Processing and Generation

Beyond software code, the utility of federated LLMs extends to broader text processing and rule-generation tasks. In the domain of complex event processing (CEP), defining precise rules for event detection is often labor-intensive. Federated LLMs facilitate the automated generation and refinement of CEP rules by aggregating diverse event patterns from multiple sources, thereby improving the system's ability to interpret complex scenarios [112]. Archival science has also seen the introduction of federated intelligence, in which LLMs assist in generating automated descriptions of archival materials. The application allows institutions to modernize their catalogs collaboratively while respecting the privacy or sensitivity of specific historical records [113].

Specialized text and audio processing domains further demonstrate the versatility of federated LLMs. The FL-former architecture adapts the Transformer model for Chinese automatic speech recognition in a federated setting, addressing data scarcity issues often encountered in dialect- or domain-specific datasets [114]. In the entertainment industry, the FedNPC framework utilizes FL to power Non-Player Characters (NPCs) in games. By training on distributed player interactions, game developers can create more dynamic and responsive character dialogues without centralizing massive logs of player behavior [115]. Furthermore, robust aggregation techniques are being applied to multimodal knowledge discovery in computational social systems, allowing researchers to analyze social trends and text data across disparate platforms while mitigating the noise and reliability issues inherent in user-generated content [116]. Besides, FedJudge reports recent advances in federated LLMs for the legal profession [117].

6.5 Finance

The financial sector operates under some of the most stringent regulatory frameworks regarding data privacy and security, making the centralization of transaction records or customer profiles nearly impossible. Consequently, federated LLMs are increasingly being explored as a way to unlock the value of financial data while remaining more compatible with privacy and compliance constraints [118]. However, much of the current evidence remains at the prototype system and task-specific experimental validation stages. A primary challenge in finance is the computational cost of deploying massive models on the secure, often resource-constrained infrastructure of smaller banks or local branches.

To bridge the gap between model capability and deployment constraints, researchers have introduced Federated Financial Reasoning Distillation [119]. The approach involves a student-teacher paradigm where a compact financial expert model is trained by distilling reasoning capabilities from multiple larger teacher models distributed across different institutions. It allows the smaller model to inherit complex financial reasoning skills, including risk assessment and market trend analysis, without ever directly accessing the teachers' private training data. Such strategies democratize access to high-level financial AI, enabling smaller entities to leverage the collective intelligence of the market while maintaining absolute data sovereignty.

Overall, current federated LLM applications in finance are promising for tasks such as risk analysis, reasoning distillation, and panic-index or market-sentiment related analysis, but evidence for deployment in high-stakes decision pipelines remains limited and demands stress testing and regulatory evaluation.

7 Open Challenges and Future Directions

While the preceding sections show that federated LLMs are technically feasible, a significant gap remains to transition from current experimental prototypes to production-ready deployment. The transition from being workable to being usable requires academia and industry to move beyond simple parameter efficiency and address the systemic tensions inherent in federated LLMs.

This section defines the desired properties of an ideal federated LLM system and then introduces the open challenges and future directions. Specifically, a fully mature federated LLM framework should satisfy the following properties:

- Efficiency (resource-agnostic effectiveness): The system should enable fine-tuning and inference on heterogeneous edge devices without imposing prohibitive memory or bandwidth costs, and without relying on significant cloud-side computational offloading.
- Privacy (verifiable trustworthiness): The system must provide rigorous guarantees that the LLM has not been poisoned, that private data cannot be reconstructed from updates, and that the resulting model aligns with safety guidelines across diverse cultural contexts.
- Utility (synergistic intelligence): The federated LLMs should outperform local isolated training not just in generalization, but in reasoning capabilities, effectively synthesizing fragmented knowledge into a coherent global intelligence without suffering from catastrophic forgetting of the pre-trained base.

Current methodologies often satisfy one property at the expense of others. As illustrated in Fig. 5, we term this phenomenon the federated LLM trilemma (balancing efficiency, privacy, and utility). For instance, aggressive quantization (efficiency) may degrade reasoning capabilities (utility), while complex cryptographic defenses (privacy) often introduce latency that makes real-time training on edge devices impossible (efficiency).

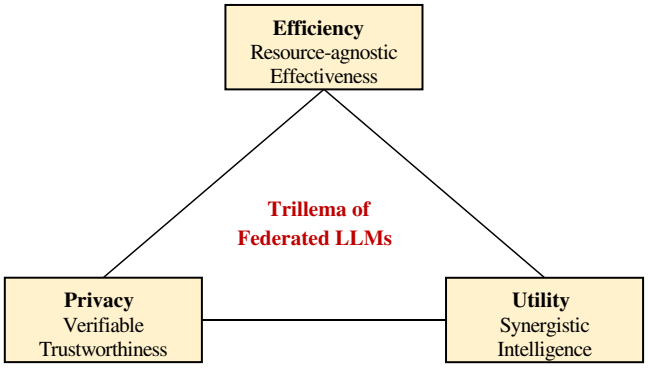

Figure 5: Trilemma of federated LLMs: balancing efficiency, privacy, and utility.

In the following, we analyze the critical bottlenecks preventing the realization of the properties above and outline some high-impact research directions to bridge the gaps.

Communication- and computation-efficient federated fine-tuning. As discussed in Section 4, edge devices are simultaneously the most privacy-sensitive and the most resource-limited participants in federated LLM systems. They often host intimate user data, yet they must operate under stringent constraints on

compute, memory, battery, and uplink bandwidth. A critical open problem is therefore to design federated fine-tuning pipelines whose end-to-end cost profiles match the realities of heterogeneous edge fleets, where participation must be robust to intermittent connectivity, stragglers, and highly variable device capabilities. Progress requires moving beyond parameter-efficient as a proxy for being system-level efficient. Even though PEFT reduces the number of trainable parameters, local training can remain bottlenecked by activation memory, optimizer state, and repeated forward and backward passes. Future research should co-design adaptation algorithms and FL protocols to minimize both local computation and communication while preserving convergence under non-IID data. It includes principled approaches to controlling update size and frequency, per-client adaptive training budgets that explicitly incorporate device telemetry, and compression mechanisms that preserve utility when the transmitted signal is already low-dimensional but semantically rich (as in LoRA or prompt vectors). The overarching goal is to make federated fine-tuning a routine, sustainable workload on commodity edge hardware rather than an occasional, high-cost operation reserved for powerful clients. The key research question can be summarized as: how can federated LLM systems jointly minimize communication, memory, and latency without sacrificing adaptation quality?

Federated RAG at scale. Section 3.4.2 positions federated RAG as a promising strategy for knowledge-intensive tasks when the knowledge base is distributed and cannot be centrally pooled. However, scaling federated RAG introduces a distinctive set of privacy and systems challenges that are not resolved by the assumption that raw documents remain local. In practice, data can be compromised through the retrieval pipeline itself: embeddings, vector indices, retrieval traces, and retrieved contexts can leak sensitive semantics even when underlying documents are never transmitted. As a result, federated RAG requires an explicit threat model and defenses that treat retrieval artifacts as first-class leakage channels. A second open question concerns the secure representation of knowledge. The community lacks a mature understanding of when an embedding space, index structure, or retrieval interface is privacy-preserving against modern inference and reconstruction attacks, particularly when downstream generators can amplify subtle semantic hints. Establishing rigorous notions of leakage for retrieval representations, along with practical mechanisms that mitigate it while retaining retrieval quality, is essential for deployable systems. Finally, federated RAG at real scale must address the fact that global knowledge typically cannot be hosted on any single device. It motivates federated indexing and routing mechanisms that support sharded or hierarchical search across device and infrastructure tiers, enable efficient incremental updates, and enforce access control and auditability across administrative boundaries. The central research challenge is to reconcile retrieval quality, scalability, and privacy guarantees within a single end-to-end design. Key research questions include: how to protect retrieval privacy without significantly harming recall and ranking quality, and what evaluation protocol to use to jointly measure privacy leakage and retrieval effectiveness in federated RAG pipelines.

Privacy mechanisms tailored to LLM adaptation. Section 5 surveys privacy risks and defenses in FL, but federated LLMs increasingly rely on adaptation regimes that diverge from the classical full-gradient assumptions underlying much of the privacy literature. A key future direction is to develop privacy accounting and protection mechanisms that reflect what is actually trained and communicated in federated LLMs, including adapter updates and prompt parameters. Because the objects are smaller but often highly informative, it remains an open empirical and theoretical question whether parameter efficiency improves privacy in practice or merely concentrates sensitive information into a lower-dimensional channel that is easier to analyze and exploit. Split FL further sharpens the problem: intermediate activations can encode substantial information about private inputs, and activation-based leakage may persist even when model updates are protected by secure aggregation. Defenses must therefore be tailored to the representational structure and semantic density of LLM activations, while remaining feasible under the bandwidth and latency constraints of edge deployments. Equally important is the evaluation methodology. Privacy should

be treated as a measurable system property rather than an implicit assumption, and leakage testing should be integrated into the standard evaluation loop alongside utility, robustness, and efficiency. In particular, privacy evaluation should explicitly cover inversion and reconstruction risks, membership inference, memorization and regurgitation behaviors, and leakage through retrieval contexts and activations, using standardized protocols that enable meaningful comparisons across methods. Key research questions include: how to evaluate privacy guarantees under realistic federated LLM threat models and whether defenses can provide meaningful protection without unacceptable degradation in utility or system cost.

Standardized federated LLM benchmarks. A major obstacle to rigorous progress in federated LLM research is the lack of standardized benchmarks and evaluation protocols. Existing studies differ widely in model backbones, parameter scales, hardware constraints, network bandwidth, client participation patterns, privacy mechanisms, and non-IID data distributions, making cross-paper numerical comparisons difficult and sometimes misleading. Future research should therefore develop standardized federated LLM benchmarks with common tasks, shared data partitioning schemes, clearly specified system settings, and unified metrics that cover not only model quality but also communication cost, memory footprint, training time, privacy leakage risk, robustness to attacks, and alignment-related behavior. Such benchmarks would improve reproducibility, enable more meaningful side-by-side comparisons of methods, and help the community distinguish gains owing to algorithmic advances from those caused by differences in experimental setup. In addition, current studies rarely report systematic sensitivity analyses over key hyperparameters that govern the trade-off among efficiency, privacy, and utility, such as LoRA rank, privacy considerations, quantization precision, and client participation ratio. Such an omission makes it difficult to determine the degree of the reported advantages. Future benchmark suites should therefore include controlled ablations and sensitivity sweeps to validate decision frameworks under comparable conditions.

Federated multi-modal foundation models for embodied AI. Embodied AI is rapidly adopting vision-language and vision-language-action foundation models, and FL is a natural fit because robots and agents collect private, high-dimensional sensor data in homes, workplaces, and other sensitive environments. At the same time, multi-modal adaptation amplifies the central bottlenecks of federated LLMs. Model footprints increase due to vision encoders and fusion modules, activation memory becomes more demanding, and bandwidth constraints become more binding when any intermediate representations must be exchanged. Moreover, heterogeneity is often more severe than in text-only settings: differences in sensors, viewpoints, environments, and tasks induce pronounced distribution shift and continual-learning dynamics. An open problem is thus to develop FL methods that can efficiently and safely adapt multi-modal foundation models under strict memory and bandwidth budgets while keeping raw sensory streams local. It includes principled choices of where and how to apply parameter-efficient adaptation across modality-specific and cross-modal components, and system designs that minimize the exposure of sensitive perceptual information during training and inference. Finally, embodied deployments elevate the importance of robustness and alignment because errors can have physical consequences. Future federated pipelines for embodied AI must therefore integrate safety-critical evaluation and aggregation considerations into the training loop, ensuring that improved average performance does not come at the cost of rare but catastrophic behaviors.

8 Conclusion

Federated LLMs represent a promising convergence of foundation-model capability and privacy-preserving collaborative learning. By keeping sensitive data on-device, federated LLMs enable domain adaptation in settings where centralized training is infeasible due to regulatory constraints, confidentiality requirements, or data ownership. At the same time, the scale and complexity of LLMs fundamentally change the design space of federated systems: naive full-parameter federated fine-tuning is often blocked by

communication and memory limits, while heterogeneous devices and non-IID data amplify optimization instability. This survey summarizes the rapid progress along three fronts. First, system architectures such as split FL, edge-cloud orchestration, and decentralized protocols extend feasibility across constrained and trust-limited environments. Second, efficient adaptation methods, including federated PEFT, prompt tuning, quantization, zeroth-order optimization, and distillation, substantially reduce resource costs while maintaining strong downstream performance. Third, data security and privacy are becoming first-class concerns, with growing attention to defenses against leakage, robustness against poisoning and backdoors, and federated alignment across heterogeneous human preferences. Despite the advances, federated fine-tuning with higher efficiency, federated RAG at scale, LLM adaptation-specific privacy mechanisms, and federated multi-modal foundation models remain open challenges. Addressing them will be essential for deploying federated LLMs as reliable, compliant, and socially responsible infrastructure across critical domains.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the HK RGC Theme-Based Research Scheme (No. T43-513/23-N) and the Pearl River Talent Plan (No. 2024QN11X183).

Author Contributions: The authors confirm contributions to the paper as follows: Conceptualization, Shan Jiang and Shichang Xuan; investigation, Wenxin You and Haoran Zhang; writing—original draft preparation, Shan Jiang, Wenxin You and Haoran Zhang; writing—review and editing, Shan Jiang, Shichang Xuan and Jiaxing Shen; visualization, Wenxin You and Haoran Zhang; supervision, Shan Jiang and Shichang Xuan; funding acquisition, Shan Jiang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. ACM Trans Intell Syst Technol. 2025;16(5):1–72. doi:10.1145/3744746.
2. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv:2303.08774. 2023.
3. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: open and efficient foundation language models. arXiv:2302.13971. 2023.
4. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. J Mach Learn Res. 2024;25(70):1–53.
5. Aggarwal M, Khullar V, Rani S, Prola T, Bhattacharjee SB, Shawon SM, et al. Federated learning on internet of things: extensive and systematic review. Comput Mater Contin. 2024;79(2):1795–834. doi:10.32604/cmc.2024.049846.
6. Zhang H, Jiang S, Xuan S. Decentralized federated learning based on blockchain: concepts, framework, and challenges. Comput Commun. 2024;216(1):140–50. doi:10.1016/j.comcom.2023.12.042.
7. Hilmkil A, Callh S, Barbieri M, Sütfeld LR, Zec EL, Mogren O. Scaling federated learning for fine-tuning of large language models. In: International Conference on Applications of Natural Language to Information Systems (NLDB). Cham, Switzerland: Springer; 2021. p. 15–23.
8. Ye R, Ge R, Zhu X, Chai J, Yaxin D, Liu Y, et al. FedLLM-bench: realistic benchmarks for federated learning of large language models. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2024. p. 111106–30.

9. Kuang W, Qian B, Li Z, Chen D, Gao D, Pan X, et al. Federatedscope-LLM: a comprehensive package for fine-tuning large language models in federated learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2024. p. 5260–71.
10. Ye R, Wang W, Chai J, Li D, Li Z, Xu Y, et al. OpenfedLLM: training large language models on decentralized private data via federated learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2024. p. 6137–47.
11. Zhang Z, Zhang Y, Chen G, Qu L, Zhou X, Wang H, et al. From continuous pre-training to alignment: a comprehensive toolkit for large language models in federated learning. Neurocomputing. 2025;647:130572.
12. Cheng Y, Zhang W, Zhang Z, Zhang C, Wang S, Mao S. Towards federated large language models: motivations, methods, and future directions. IEEE Commun Surv Tutor. 2025;27(4):2733–64.
13. Chen C, Feng X, Li Y, Lyu L, Zhou J, Zheng X, et al. Integration of large language models and federated learning. Patterns. 2024;5(12):101098. doi:10.1016/j.patter.2024.101098.
14. Thakur D, Guzzo A, Fortino G. Analyzing the fusion of federated learning and large language model. In: 2025 IEEE 5th International Conference on Human-Machine Systems (ICHMS). Piscataway, NJ, USA: IEEE; 2025. p. 282–8.
15. Piccialli F, Chiaro D, Qi P, Bellandi V, Damiani E. Federated and edge learning for large language models. Inf Fusion. 2025;117(1):102840. doi:10.1016/j.inffus.2024.102840.
16. Ren C, Yu H, Peng H, Tang X, Zhao B, Yi L, et al. Advances and open challenges in federated foundation models. IEEE Commun Surv Tutor. 2025;28(1):2087–126. doi:10.1109/comst.2025.3552524.
17. Hu J, Wang D, Wang Z, Pang X, Xu H, Ren J, et al. Federated large language model: solutions, challenges and future directions. IEEE Wirel Commun. 2025;32(4):82–9.
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 6000–10.
19. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2020. p. 1877–901.
20. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2022. p. 27730–44.
21. Jiang S, Zhou X, Zhang M, Xu C, Liao G, Chen J, et al. Edge large language models: a comprehensive survey. CCF Trans Pervasive Comput Interact. 2026;2(2):129. doi:10.1007/s42486-025-00227-7.
22. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. London, UK: PMLR; 2017. p. 1273–82.
23. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37(3):50–60.
24. Zhang M, Shen X, Cao J, Cui Z, Jiang S. Edgeshard: efficient LLM inference via collaborative edge computing. IEEE Internet Things J. 2025;12(10):13119–31. doi:10.1109/jiot.2024.3524255.
25. Zhao K, Yang Z, Huang C, Chen X, Zhang Z. FedsLLM: federated split learning for large language models over communication networks. In: 2024 International Conference on Ubiquitous Communication (Ucom). Piscataway, NJ, USA: IEEE; 2024. p. 438–43.
26. Gu Z, Fan Q, Sun L, Liu Y, Ye X. VFLAIR-LLM: a comprehensive framework and benchmark for split learning of LLMs. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2025. p. 5470–81.
27. Tan B, Ren J, Li Y, Ding S, Chaddad A. Enhancing large language model fine-tuning with sharpness-aware minimization under split federated learning. In: International Conference on Intelligent Computing (ICIC). Berlin/Heidelberg, Germany: Springer; 2025. p. 260–71.
28. Yao D, Li B. Is split learning privacy-preserving for fine-tuning large language models? IEEE Trans Big Data. 2024. doi:10.1109/tbdata.2024.3524101.

29. Djuhera A, Andrei VC, Li X, Mönich UJ, Boche H, Saad W. R-SFLLM: jamming resilient framework for split federated learning with large language models. IEEE Trans Inf Forensics Secur. 2025;20:8296–311.
30. Qiao D, Ao X, Liu Y, Chen X, Song F, Qin Z, et al. Tri-AFLLM: resource-efficient adaptive asynchronous accelerated federated LLMs. IEEE Trans Circuits Syst Video Technol. 2025;35(5):4198–211.
31. Iwan I, Tanjung SY, Yahya BN, Lee SL. FedCLLM: federated client selection assisted large language model utilizing domain description. Internet Things. 2025;30:101506.
32. Dun J, Li Z. Dynamic node scheduling for delay optimization in federated large language model training. In: 2025 34th International Conference on Computer Communications and Networks (ICCCN). Piscataway, NJ, USA: IEEE; 2025. p. 1–6.
33. Huang W, Xu C, Li J, Wang S, Zhao S. Edge-assisted collaborative training method for large language model with embedded data processing unit. In: International Conference on Information Processing and Network Provisioning. Berlin/Heidelberg, Germany: Springer; 2024. p. 9–19.
34. Sadeepa S, Kavinda K, Hashika E, Sandeepa C, Gamage T, Liyanage M. DisLLM: distributed LLMs for privacy assurance in resource-constrained environments. In: 2024 IEEE Conference on Communications and Network Security (CNS). Piscataway, NJ, USA: IEEE; 2024. p. 1–9.
35. Liu N, Liu D. MPCTF: a multi-party collaborative training framework for large language models. Electronics. 2025;14(16):3253.
36. Li C, Gu B, Zhao Z, Qu Y, Xin G, Huo J, et al. Federated transfer learning for on-device LLMs efficient fine tuning optimization. Big Data Min Anal. 2025;8(2):430–46.
37. Jonnalagadda AK, Madupati B, Vegesna RV, Vududala SK. Federated data modeling for LLM deployment in secure cloud-native architectures. In: 2025 International Conference on Computing Technologies & Data Communication (ICCTDC). Piscataway, NJ, USA: IEEE; 2025. p. 1–8.
38. Adnan MT, Oroceo PA, Lee JM, Kim DS. PureLLM: a blockchain-driven decentralized PFL for robust and resource-efficient next-gen LLMs. In: 2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN). Piscataway, NJ, USA: IEEE; 2025. p. 526–31.
39. Zhang J, Pan Y, Wu Z, Zhou R, Yang Y, Wang P, et al. Incentive mechanisms for collaborative intelligence sharing in blockchain-based federated LLM fine-tuning. In: Blockchain and Web3 Technology Innovation and Application Exchange Conference (BWTAC). Berlin/Heidelberg, Germany: Springer; 2025. p. 323–33.
40. Nguyen NLB, Tran TQ, Wong KS. FedDDF: dynamic dataset filtering in federated large language model training. In: ASIA CCS'25: Proceedings of the International Workshop on Secure and Efficient Federated Learning. New York, NY, USA: ACM; 2025. p. 1–6.
41. Ye R, Ge R, Yuchi F, Chai J, Wang Y, Chen S. Leveraging unstructured text data for federated instruction tuning of large language models. In: International Workshop on Trustworthy Federated Learning. Berlin/Heidelberg, Germany: Springer; 2024. p. 119–31.
42. Elbakary A, Issaid CB, ElBatt T, Seddik K, Bennis M. Mira: a method of federated multi-task learning for large language models. IEEE Netw Lett. 2025;7(3):171–5.
43. Wu F, Li Z, Li Y, Ding B, Gao J. Fedbiot: LLM local fine-tuning in federated learning without full model. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2024. p. 3345–55.
44. Zhang H, Pang J, Huang Y, Xie Z, Liu Z. FedBridgeICL: federated bridging of small and large models for in-context learning. In: International Conference on Wireless Artificial Intelligent Computing Systems and Applications (WASA). Berlin/Heidelberg, Germany: Springer; 2025. p. 1–10.
45. Garcia J, Gong J, Zajac M, Hahn A. DF-RAG: a dual federated retrieval-augmented generation framework for collaborative medical AI. In: Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies. Piscataway, NJ, USA: IEEE; 2025. p. 418–23.
46. Xu M, Cai D, Wu Y, Li X, Wang S. FwdLLM: efficient federated finetuning of large language models with perturbed inferences. In: 2024 USENIX Annual Technical Conference. Berkeley, CA, USA: USENIX Association; 2024. p. 579–96.

47. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv:2106.09685. 2021.
48. Sun H, Tian H, Ni W, Zheng J, Niyato D, Zhang P. Federated low-rank adaptation for large models fine-tuning over wireless networks. IEEE Trans Wirel Commun. 2025;24(1):659–75. doi:10.1109/twc.2024.3497998.
49. Iftikhar S, Alsamhi SH, Davy S. Enhancing sustainability in LLM training: leveraging federated learning and parameter-efficient fine-tuning. IEEE Trans Sustain Comput. 2025;10(6):1158–72.
50. Jiang J, Jiang H, Ma Y, Liu X, Fan C. Low-parameter federated learning with large language models. In: International Conference on Web Information Systems and Applications (WISA). Berlin/Heidelberg, Germany: Springer; 2024. p. 319–30.
51. Bai J, Chen D, Qian B, Yao L, Li Y. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2024. p. 14457–83.
52. Luo X, Jiang C, Wang S, Zhang Y. DynamicFedPEFT: efficient fine-tuning of dynamic federated parameters for large language models. In: International Conference on Knowledge Science, Engineering and Management (KSEM). Berlin/Heidelberg, Germany: Springer; 2025. p. 89–100.
53. Su Y, Yan N, Deng Y. Federated LLMS fine-tuned with adaptive importance-aware lora. In: IEEE International Conference on Communications (ICC). Piscataway, NJ, USA: IEEE; 2025. p. 6112–7.
54. Wang Z, Shen Z, He Y, Sun G, Wang H, Lyu L, et al. Flora: federated fine-tuning large language models with heterogeneous low-rank adaptations. In: NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.;2024. p. 22513–33.
55. Ning W, Wang J, Qi Q, Sun H, Cheng D, Liu C, et al. Federated fine-tuning on heterogeneous LoRAs with error-compensated aggregation. IEEE Trans Neural Netw Learn Syst. 2025;36(10):17826–40. doi:10.1109/tnnls.2025.3586545.
56. Zhu J, Lu Y, Yang W, Zhang J. An enhanced low-rank fine-tuning framework for federated large language models. Neurocomputing. 2026;669:132475. doi:10.1016/j.neucom.2025.132475.
57. Solat F, Lee J. Optimizing client participation in communication-constrained federated LLM adaptation with LoRA. Sensors. 2025;25(21):6538. doi:10.3390/s25216538.
58. Yi X, Hu C, Cai B, Huang H, Chen Y, Wang K. FedALoRA: adaptive local LoRA aggregation for personalized federated learning in LLM. IEEE Internet Things J. 2025;12(24):51854–65.
59. Che T, Liu J, Zhou Y, Ren J, Zhou J, Sheng V, et al. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2023. p. 7871–88.
60. Gao C. Federated p-tuning based Time-LLM for hotel booking prediction. In: Proceedings of the 2025 International Conference on Generative Artificial Intelligence for Business. New York, NY, USA: ACM; 2025. p. 7–11.
61. Tanimura T, Nakano W, Kitagawa Y, Takase M. Federated discrete prompt tuning for language models using synthetic examples. In: 2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC). Piscataway, NJ, USA: IEEE; 2025. p. 1–4.
62. Li Y, Sun J, Liu Y, Zhang Y, Li A, Chen B, et al. Federated black-box prompt tuning system for large language models on the edge. In: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom). New York, NY, USA: ACM; 2024. p. 1775–7.
63. Liu X, Pang T, Fan C. Federated prompting and chain-of-thought reasoning for improving LLMS answering. In: International Conference on Knowledge Science, Engineering and Management (KSEM). Berlin/Heidelberg, Germany: Springer; 2023. p. 3–11.
64. Lv G, Gu B, Jia X, Gao L, Qu Y, Cui L. Federated learning and parallel prompt scheduling strategies for large language models. In: International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP). Berlin/Heidelberg, Germany: Springer; 2024. p. 317–26.

65. Qiu W, Zhou Y, Wang J, Sheng QZ, Cui L. FLM-TopK: expediting federated large language model tuning by sparsifying intervalized gradients. In: IEEE INFOCOM 2025—IEEE Conference on Computer Communications. Piscataway, NJ, USA: IEEE; 2025. p. 1–10.
66. Ma B, Gao Y, Liu Y. FedAdamZO: a zeroth-order adaptive momentum method for memory-efficient fine-tuning of federated large language models. In: IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ, USA: IEEE; 2025. p. 1–6.
67. Ling Z, Chen D, Yao L, Li Y, Shen Y. On the convergence of zeroth-order federated tuning for large language models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2024. p. 1827–38.
68. Zhang X, Lin Y, Miao M, Lou J, Li J, Chen X. Zeroth-order federated private tuning for pretrained large language models. In: Australasian Conference on Information Security and Privacy (ACISP). Berlin/Heidelberg, Germany: Springer; 2025. p. 285–306.
69. Gao Z, Zhang Z, Guo Y, Gong Y. Federated adaptive fine-tuning of large language models with heterogeneous quantization and LoRA. In: IEEE Conference on Computer Communications (INFOCOM). Piscataway, NJ, USA: IEEE; 2025. p. 1–10.
70. Zhang Y, Cao J, Zhang M, Yang R. FedHO: memory-efficient federated fine-tuning for large models via hybrid gradient computation. In: FLEdge-AI '25: Proceedings of the Federated Learning and Edge AI for Privacy and Mobility. New York, NY, USA: ACM; 2025. p. 85–92.
71. Gao Z, Li Y, Yu X. FEDNPAIT: federated learning with NADAM and PADAM for instruction tuning. In: International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP). Berlin/Heidelberg, Germany: Springer; 2024. p. 185–203.
72. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21). Berkeley, CA, USA: USENIX Association; 2021. p. 2633–50.
73. Zheng JY, Zhang H, Wang L, Qiu W, Zheng HW, Zheng ZM. Safely learning with private data: a federated learning framework for large language model. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2024. p. 5293–306.
74. Wu J, Chen L, Fang S, Wu C. An application programming interface (API) sensitive data identification method based on the federated large language model. Appl Sci. 2024;14(22):10162. doi:10.3390/app142210162.
75. Wang F, Li B. Data reconstruction and protection in federated learning for fine-tuning large language models. IEEE Trans Big Data. 2024:1–13. doi:10.1109/TBDATA.2024.3524105.
76. Pan Q, Wu J. Selective privacy-preserving federated learning for large language model fine-tuning. In: 2025 International Wireless Communications and Mobile Computing (IWCMC). Piscataway, NJ, USA: IEEE; 2025. p. 1626–31.
77. Yun S, Bhuiyan ZA, Sadi MTAH, Su S. Privacy-preserving federated learning through clustered sampling on fine-tuning distributed non-iid large language models. In: 2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking. Piscataway, NJ, USA: IEEE; 2023. p. 531–8.
78. Zhang T, Yu H, Yang Z, Chen Y, Yu S. ELaVFL: efficient verifiable federated learning for large language models. IEEE Trans Dependable Secur Comput. 2025;22(6):6214–29. doi:10.1109/tdsc.2025.3581728.
79. Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V. How to backdoor federated learning. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. London, UK: PMLR.;2020. p. 2938–48.
80. Wang H, Li B, Chen P, Wu L, Li Z, Quek TQS. LBKD: rethinking federated backdoors for low-altitude economy via LLMs and bidirectional knowledge distillation. IEEE Trans Netw Sci Eng. 2025;13(8):4422–39. doi:10.1109/tnse.2025.3626056.
81. Wang H, Yin Z, Chen B, Zeng Y, Yan X, Zhou C, et al. Rofed-LLM: robust federated learning for large language models in adversarial wireless environments. IEEE Trans Netw Sci Eng. 2025;13:1084–96.

82. Wang ZQ, Wang H, El Saddik A. FedITD: a federated parameter-efficient tuning with pre-trained large language models and transfer learning framework for insider threat detection. IEEE Access. 2024;12:160396–417.
83. Luo H, Ji C. Federated learning-based data collaboration method for enhancing edge cloud AI system security using large language models. In: 2025 5th International Symposium on Computer Technology and Information Science (ISCTIS). Piscataway, NJ, USA: IEEE; 2025. p. 163–6.
84. Fan FX, Tan C, Ong YS, Wattenhofer R, Ooi WT. FedRLHF: a convergence-guaranteed federated framework for privacy-preserving and personalized RLHF. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems. Richland, SC, USA: International Foundation for Autonomous Agents and Multiagent Systems; 2025. p. 713–21.
85. Srewa M, Zhao T, Elmalaki S. PluralLLM: pluralistic alignment in LLMS via federated learning. In: Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems. New York, NY, USA: ACM; 2025. p. 64–9.
86. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2023. p. 53728–41.
87. Spadea F, Seneviratne O. Federated fine-tuning of large language models: Kahneman-Tversky vs. direct preference optimization. In: Companion Proceedings of the ACM on Web Conference 2025. New York, NY, USA: ACM; 2025. p. 1757–60.
88. Jiang Y, Li Z, Song B. Fine-tuning large language models in federated learning with fairness-aware prompt selection. Neural Netw. 2025;194(8):108160. doi:10.1016/j.neunet.2025.108160.
89. Ma T, Luo X, Tan R, Gao H. Privacy and fairness-guaranteed federated preference optimization for large language models in internet of medical things. IEEE Trans Consum Electron. 2025. doi:10.1109/tce.2025.3595092.
90. Li H, Yu L, He W. The impact of GDPR on global technology development. J Glob Inf Technol Manag. 2019;22(1):1–6. doi:10.1080/1097198x.2019.1569186.
91. Li X, Peng L, Wang YP, Zhang W. Open challenges and opportunities in federated foundation models towards biomedical healthcare. BioData Min. 2025;18(1):2. doi:10.1186/s13040-024-00414-9.
92. Guo D, Choo KKR. Applications of federated large language model for adverse drug reactions prediction: scoping review. J Med Internet Res. 2025;27(12):e68291. doi:10.2196/68291.
93. Zhang L, Li Y. Selective layer fine-tuning for federated healthcare NLP: a cost-efficient approach. In: 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA). Piscataway, NJ, USA: IEEE; 2025. p. 1–6.
94. Yue F, Qiu R, Li J, Li G. Privacy-preserving federated learning framework for disease progression prediction via temporal-aware large language modeling. In: 2025 International Conference on Sensor-Cloud and Edge Computing System (SCECS). Piscataway, NJ, USA: IEEE; 2025. p. 414–7.
95. Chen X, Zhu F, Li D, Li Q, Anwar MS, Shan G, et al. Towards clinically applicable large-model-based privacy-preserving polyp segmentation: a federated LoRA approach to colonoscopy. IEEE J Biomed Health Inform. 2025:1–13. doi:10.1109/JBHI.2025.3639279.
96. Che H, Jin H, Gu Z, Lin Y, Jin C, Chen H. LLM-driven medical report generation via communication-efficient heterogeneous federated learning. IEEE Trans Med Imaging. 2026;45(1):28–39. doi:10.1109/tmi.2025.3591185.
97. Yang H, Liu H, Yuan X, Wu K, Ni W, Zhang JA, et al. Synergizing intelligence and privacy: a review of integrating internet of things, large language models, and federated learning in advanced networked systems. Appl Sci. 2025;15(12):6587.
98. Wang Q, Yao Y, Ammar N, Shi W. iFLOW: an intelligent and scalable multi-model federated learning framework on the wheels. IEEE Tran Intell Trans Syst. 2025;26(10):15903–14. doi:10.1109/tits.2025.3578586.
99. Chen J, He J, Chen F, Lv Z, Tang J, Jia Y. Empowering IoT-based autonomous driving via federated instruction tuning with feature diversity. IEEE Internet Things J. 2025;12(6):6095–108. doi:10.1109/jiot.2024.3518615.
100. Chen J, Messou FJA, Zhang S, Liu T, Yu K, Niyato D. Federated fine-tuning of large language models for intelligent automotive systems with low-rank adaptation. In: 2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring). Piscataway, NJ, USA: IEEE; 2025. p. 1–6.

101. Bao Y, Cheng X, Nie L, Tao J. Enabling privacy-preserving and drop-out resilient federated LLM fine-tuning for the Low-altitude UAV swarm networks. IEEE Trans Cogn Commun Netw. 2025;12:2919–36.
102. Doğruluk E, Açıkgöz H. Edge-centric federated learning for LLMs in smart manufacturing: architectures, challenges, and opportunities. In: 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). Piscataway, NJ, USA: IEEE; 2025. p. 1250–6.
103. Sheng X, Yu C, Cui X, Zhou Y. Large language model and digital twins empowered asynchronous federated learning for secure data sharing in intelligent labeling. Mathematics. 2024;12(22):3550. doi:10.3390/math12223550.
104. Rezaei H, Taheri R, Shojafar M. FedLLMGuard: a federated large language model for anomaly detection in 5G networks. Comput Netw. 2025;269:111473.
105. Qi F, Pan R, Zhang H, Xu C. Enhancing federated video anomaly detection with GPT-driven semantic distillation. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2024. p. 234–51.
106. Nazzal M, Nguyen K, Vungarala D, Zand R, Angizi S, Phan H, et al. FedChip: federated LLM for artificial intelligence accelerator chip design. In: 2025 IEEE International Conference on LLM-Aided Design (ICLAD). Piscataway, NJ, USA: IEEE; 2025. p. 93–9.
107. Khan K. Adaptive federated learning with local large language models for modeling photonic and chemical systems. IEEE Access. 2025;13(3):160559–75. doi:10.1109/access.2025.3606855.
108. Cai Z, Chen J, Chen W, Wang W, Zhu X, Ouyang A. F-codeLLM: a federated learning framework for adapting large language models to practical software development. In: Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings. Piscataway, NJ, USA: IEEE; 2024. p. 416–7.
109. Chen J, Cai Z, Chen W, Wang W, Zheng Z, Yu PS. A federated adaptive large language model fine-tuning framework for software development. IEEE Trans Serv Comput. 2026;19(1):32–43. doi:10.1109/tsc.2025.3623626.
110. Kumar J, Chimalakonda S. Code summarization without direct access to code-towards exploring federated LLMS for software engineering. In: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering. New York, NY, USA: ACM; 2024. p. 100–9.
111. Seo J, Zhang N, Rong C. Flexible and secure code deployment in federated learning using large language models: prompt engineering to enhance malicious code detection. In: 2023 IEEE International Conference on Cloud Computing Technology and Science (CloudCom). Piscataway, NJ, USA: IEEE; 2023. p. 341–9.
112. Delouee ML, Pernes DG, Degeler V, Koldehofe B. Towards federated LLM-powered CEP rule generation and refinement. In: Proceedings of the 18th ACM International Conference on Distributed and Event-Based Systems. New York, NY, USA: ACM; 2024. p. 185–6.
113. Groppe J, Marquet A, Walz A, Groppe S. Automated archival descriptions with federated intelligence of LLMs. In: International Conference on Database and Expert Systems Applications (DEXA). Berlin/Heidelberg, Germany: Springer; 2025. p. 53–67.
114. Chen J, Li Z, Shen S, Yang J, Gao Z, Huang L, et al. FL-former: Chinese automatic speech recognition architecture under the federated large model. In: Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education. New York, NY, USA: ACM; 2024. p. 7–10.
115. Hong M, Zhang K, Zhang S, He Z. FedNPC: a federated learning framework for large language models in game NPCs. In: 2023 IEEE 21st Student Conference on Research and Development (SCOReD). Piscataway, NJ, USA: IEEE; 2023. p. 363–8.
116. Chen J, Chakraborty C, Polavarapu A, Qiu Y, Zhao Q, Alfarraj O, et al. A robust aggregation of federated large language models for multimodal knowledge discovery in computational social systems. IEEE Trans Comput Soc Syst. 2025;12(6):5433–48.
117. Yue L, Liu Q, Du Y, Gao W, Liu Y, Yao F. Fedjudge: federated legal large language model. In: International Conference on Database Systems for Advanced Applications (DASFAA). Berlin/Heidelberg, Germany: Springer; 2024. p. 268–85.
118. Jiang S, You W, Xuan S, Shen J. Decentralized finance security: a survey of attacks, defenses, and open challenges. High Confid Comput. 2026;6(2):100383.
119. Liu S, Yan J, Wang X, Jiang Y, Chen L, Fan T, et al. Federated financial reasoning distillation: training a small financial expert by learning from multiple teachers. In: Proceedings of the 6th ACM International Conference on AI in Finance. New York, NY, USA: ACM; 2025. p. 623–31.