iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83405

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Federated Learning and Large Language Models: Recent Advances, Challenges, and Future Directions for Privacy-Preserving AI

Roopesh Kumar Sharma[1], Parag Jain[2], Kirti Sharma[3], Ujjval Mishra[4]
[1, 2, 3, 4]*Assistant professor, Department of Computer Science & Engineering, Chameli Devi Group of Institutions, Indore*

Abstract: ***The rapid advancement of Large Language Models (LLMs) has transformed artificial intelligence by enabling remarkable performance in natural language understanding, generation, and decision-making tasks. However, conventional centralized training approaches require massive amounts of data aggregation, raising significant concerns regarding privacy, security, and regulatory compliance. Federated Learning (FL) has emerged as a promising distributed learning paradigm that enables collaborative model training across decentralized data sources without sharing raw data. The integration of Federated Learning and Large Language Models, commonly referred to as Federated Large Language Models (FedLLMs), offers a novel approach to developing privacy-preserving artificial intelligence systems while maintaining model effectiveness. This review examines recent advances in Federated Large Language Models, focusing on federated pre-training, federated fine-tuning, parameter-efficient adaptation techniques, privacy-preserving mechanisms, and personalized learning strategies. The study analyzes key challenges including communication overhead, data heterogeneity, security vulnerabilities, model bias, scalability limitations, and explainability concerns. Furthermore, the review identifies critical research gaps and proposes an integrated framework that combines privacy protection, secure aggregation, parameter-efficient fine-tuning, personalized learning, and explainable artificial intelligence. Finally, future research directions are discussed to guide the development of scalable, trustworthy, and privacy-aware AI systems. The findings suggest that FedLLMs represent a significant step toward achieving secure and decentralized artificial intelligence across diverse application domains such as healthcare, finance, education, cybersecurity, and Industry 4.0.***

Keywords: Federated Learning (FL), Large Language Models (LLMs), Federated Large Language Models (FedLLMs), Privacy-Preserving AI, Distributed Learning, Federated Fine-Tuning, Parameter-Efficient Fine-Tuning (PEFT), Differential Privacy, Secure Aggregation, Explainable Artificial Intelligence (XAI), Personalized Learning, Trustworthy AI.

I. INTRODUCTION

Artificial Intelligence (AI) has experienced unprecedented growth in recent years, driven largely by the emergence of Large Language Models (LLMs) such as GPT, PaLM, LLaMA, and other foundation models. These models have demonstrated exceptional capabilities in natural language processing, content generation, question answering, reasoning, and decision support applications. Their success is primarily attributed to large-scale training on massive datasets collected from diverse sources. However, the centralized nature of traditional LLM training requires aggregating enormous amounts of data into a single location, creating significant concerns regarding data privacy, security, ownership, and regulatory compliance.

As organizations increasingly handle sensitive information in sectors such as healthcare, finance, education, government, and cybersecurity, privacy-preserving machine learning approaches have become essential. Federated Learning (FL), introduced as a decentralized machine learning paradigm, enables multiple participants to collaboratively train models without exchanging raw data. Instead, only model updates are shared and aggregated, thereby reducing privacy risks while maintaining collaborative learning benefits. This characteristic makes Federated Learning particularly attractive for applications involving confidential and distributed datasets. The convergence of Federated Learning and Large Language Models has given rise to Federated Large Language Models (FedLLMs), an emerging research area that seeks to combine the powerful capabilities of LLMs with the privacy-preserving properties of federated learning. FedLLMs enable organizations and devices to collaboratively train, adapt, and deploy language models while retaining data locally. Recent studies have explored federated pre-training, federated fine-tuning, parameter-efficient adaptation techniques, secure aggregation protocols, and personalized learning mechanisms to improve the practicality and effectiveness of FedLLMs.

©IJRASET: All Rights are Reserved

Despite significant progress, several challenges continue to hinder the widespread adoption of Federated Large Language Models. The enormous size of modern language models introduces substantial communication and computational overhead. Additionally, heterogeneous data distributions, adversarial attacks, model poisoning threats, privacy leakage risks, fairness concerns, and explainability limitations remain active areas of research. Addressing these challenges is crucial for enabling scalable and trustworthy deployment of FedLLMs in real-world environments.

Motivated by these developments, this review provides a comprehensive examination of recent advances, challenges, and future directions in Federated Large Language Models. The study systematically analyzes existing research on federated training architectures, parameter-efficient fine-tuning techniques, privacy-preserving mechanisms, security frameworks, and domain-specific applications. Furthermore, research gaps are identified and an integrated conceptual framework is proposed to support the development of next-generation privacy-preserving AI systems.

The remainder of this paper is organized as follows. Section 2 presents a comprehensive literature review on the evolution of Federated Large Language Models, federated fine-tuning approaches, and key challenges. Section 3 discusses the research gaps identified from existing studies. Section 4 outlines future research directions, while Section 5 provides a detailed discussion of current trends and deployment considerations. Section 6 proposes an integrated FedLLM framework, and Section 7 concludes the review by summarizing key findings and highlighting future opportunities for privacy-preserving artificial intelligence.

II. LITERATURE REVIEW

A. Evolution and Integration of Federated Learning with Large Language Models

The rapid advancement of Large Language Models (LLMs) has significantly improved natural language processing capabilities across diverse applications. However, traditional centralized training approaches require massive data aggregation, raising concerns regarding privacy, security, and regulatory compliance. Federated Learning (FL) has emerged as a promising paradigm that enables collaborative model training across decentralized devices while preserving data privacy. Zhuang et al. (2023) highlighted the motivations and challenges of integrating foundation models with federated learning, emphasizing privacy preservation and distributed intelligence. Similarly, Hu et al. (2024) presented comprehensive solutions and future directions for Federated Large Language Models (FedLLMs), discussing communication efficiency, model heterogeneity, and privacy protection.

Recent studies have further explored the convergence of FL and LLMs. Ye et al. (2024) introduced OpenFedLLM, demonstrating the feasibility of training large language models using decentralized private datasets. Sani et al. (2024) argued that the future of LLM pre-training lies in federated approaches due to increasing concerns over data ownership and governance. Thakur et al. (2025) analyzed the fusion of FL and LLMs, identifying key opportunities in privacy-aware AI development. Kenteris and Kotis (2026) extended this integration by combining Federated Learning, Knowledge Graphs, and LLMs for language learning applications, demonstrating the growing interdisciplinary adoption of federated AI frameworks.

Moreover, industry-oriented reviews such as Jing et al. (2026) highlighted the importance of Federated Large Language Models in Industry 4.0 environments, where decentralized industrial data can be utilized without compromising confidentiality. These developments indicate that FL has become a critical enabler for deploying large-scale language models in privacy-sensitive and distributed computing environments.

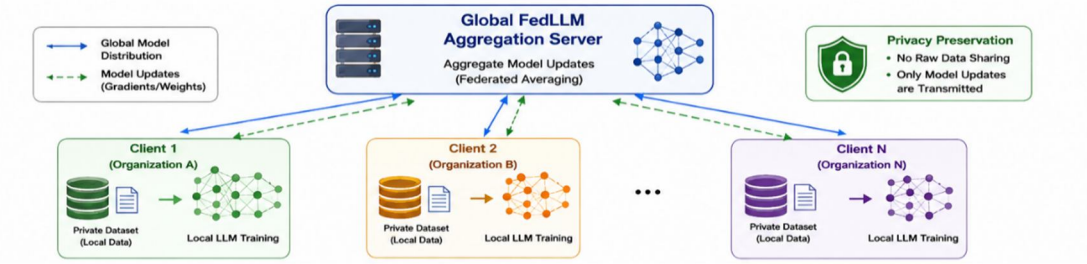


Figure 1: Architecture of Federated Large Language Model (FedLLM) Training Framework

This figure1 illustrates the overall architecture of a Federated Large Language Model system, where multiple client devices or organizations locally train an LLM using their private datasets. Instead of sharing raw data, only model parameters or gradients are transmitted to a central aggregation server. The server combines the updates using federated aggregation algorithms and distributes the improved global model back to participants. The figure highlights privacy preservation, decentralized learning, and collaborative model development enabled by the integration of Federated Learning and Large Language Models.

B. Federated Fine-Tuning and Parameter-Efficient Adaptation of Large Language Models

Fine-tuning large language models in federated environments has gained significant attention due to the high computational and communication costs associated with training billion-parameter models. Early work by Hilmkil et al. (2021) demonstrated the feasibility of scaling federated learning for LLM fine-tuning, providing foundational insights into distributed model adaptation. Building upon this foundation, Wu et al. (2025) conducted a comprehensive survey on federated fine-tuning techniques, highlighting challenges related to communication overhead, client heterogeneity, and convergence stability.

To address these limitations, researchers have investigated parameter-efficient fine-tuning (PEFT) methods. Wen et al. (2025) surveyed federated parameter-efficient fine-tuning approaches, emphasizing techniques such as Low-Rank Adaptation (LoRA) to reduce communication costs while maintaining model performance. Alzahrani and Yang (2026) proposed PrivLoRA, which enhances privacy protection during LoRA-based federated fine-tuning by minimizing information leakage from transmitted updates.

Several studies have also explored advanced optimization strategies for heterogeneous federated environments. Liao et al. (2025) introduced importance-aware model updating mechanisms to improve performance under non-identically distributed (non-IID) data conditions. Colombi et al. (2025) investigated edge-based federated fine-tuning of LLMs, demonstrating the practical deployment of language models on resource-constrained devices. Zhu et al. (2026) further extended personalized federated learning through efficient fine-tuning and uncertainty-aware mechanisms for medical vision-language models. These studies collectively demonstrate that parameter-efficient and personalized fine-tuning techniques are essential for making federated LLM deployment practical and scalable.

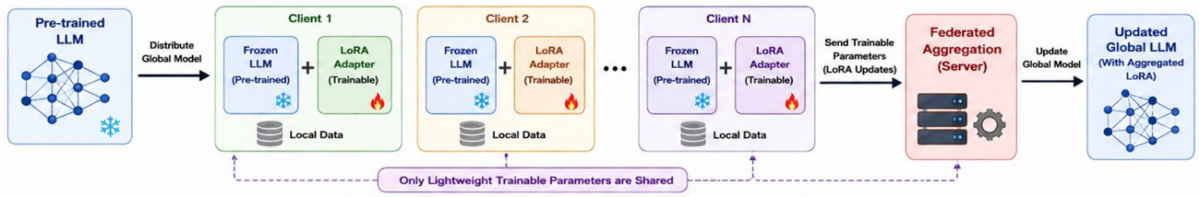


Figure 2: Federated Parameter-Efficient Fine-Tuning (PEFT) Process for Large Language Models

This figure 2 presents the workflow of parameter-efficient federated fine-tuning using techniques such as LoRA (Low-Rank Adaptation). The pre-trained global LLM is distributed to multiple clients, where only a small subset of trainable parameters is updated locally. These lightweight updates are aggregated centrally and used to refine the global model. The figure demonstrates how PEFT significantly reduces computational cost, communication overhead, and storage requirements while maintaining model performance in heterogeneous federated environments.

C. Challenges, Security, Privacy, and Future Directions of Federated Large Language Models

Despite their potential, Federated Large Language Models face numerous challenges related to security, robustness, privacy, fairness, and system scalability. Hu et al. (2024) identified communication bottlenecks, model synchronization issues, and privacy risks as major barriers to widespread adoption. Jiang et al. (2025) further examined the feasibility, robustness, and security aspects of Federated LLMs, emphasizing the vulnerability of distributed training systems to adversarial attacks and model poisoning.

Privacy remains a central concern in federated environments. Although raw data remain local, model updates may inadvertently reveal sensitive information. To address these concerns, researchers have proposed privacy-preserving techniques such as differential privacy, secure aggregation, and encrypted model updates. Alzahrani and Yang (2026) demonstrated privacy-enhancing mechanisms through PrivLoRA, while Kaur et al. (2026) developed FedLLM for privacy-preserving and explainable traffic flow prediction.

Bias propagation represents another emerging challenge. Zhao et al. (2026) investigated how social biases can spread during federated fine-tuning processes, highlighting the need for fairness-aware aggregation methods. Additionally, personalized federated learning frameworks have been proposed to accommodate client-specific requirements, as reviewed by Akhmetov et al. (2026) for sovereign personal AI agents. Emerging applications in healthcare (Tang & Deng, 2024), cybersecurity (AlHayan & Al-Muhtadi, 2026), quantum federated learning (Gurung & Pokhrel, 2025), and 6G semantic networks (Mittal et al., 2026) further demonstrate the expanding scope of Federated LLM research.

Overall, current literature suggests that future research should focus on improving communication efficiency, strengthening privacy guarantees, mitigating bias, enhancing personalization, and developing robust federated architectures capable of supporting increasingly complex large language models across diverse application domains.

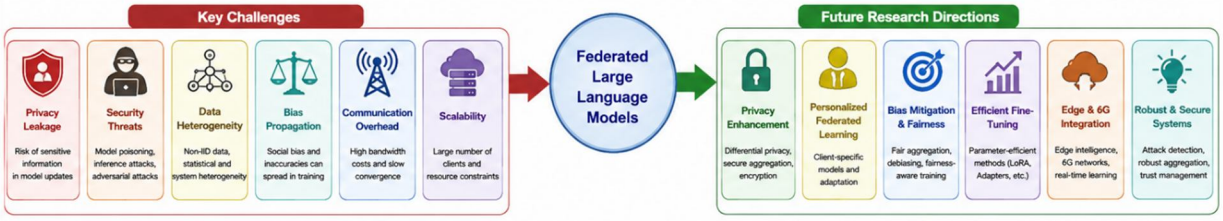


Figure 3: Key Challenges and Future Research Directions in Federated Large Language Models

This figure 3 depicts the major challenges associated with Federated Large Language Models, including privacy leakage, communication inefficiency, model poisoning attacks, data heterogeneity, bias propagation, and scalability issues. It also highlights future research directions such as differential privacy, secure aggregation, personalized federated learning, fairness-aware training, parameter-efficient adaptation, and edge intelligence integration. The figure provides a comprehensive overview of the research landscape and emerging opportunities for advancing secure and efficient FedLLM systems.

Table 2.1 Systematic Literature Review on Federated Large Language Models

Ref.	Author(s) & Year	Research Focus	Methodology	Key Findings	Research Gap
1	Kenteris & Kotis (2026)	FL, Knowledge Graphs, and LLMs for language learning	Scoping Review	Demonstrated convergence of FL, KG, and LLMs for privacy-preserving education	Limited real-world implementation
2	Thakur et al. (2025)	Fusion of FL and LLMs	Analytical Study	Identified opportunities and challenges in integrating FL with LLMs	Lack of performance evaluation
3	Jing et al. (2026)	Federated LLMs in Industry 4.0	Review Study	Highlighted industrial applications and deployment requirements	Need for scalable industrial frameworks
4	Wu et al. (2025)	Federated Fine-Tuning of LLMs	Survey	Categorized fine-tuning strategies and challenges	Limited practical benchmarking
5	Ye et al. (2024)	OpenFedLLM Framework	Experimental Framework	Demonstrated decentralized training on private data	High communication overhead
6	Hu et al. (2024)	FedLLM Solutions and Challenges	Review	Discussed privacy, efficiency, and scalability challenges	Limited security mechanisms
7	Hilmkil et al. (2021)	Scaling FL for LLM Fine-Tuning	Experimental Study	Proved feasibility of distributed LLM fine-tuning	Resource constraints not fully addressed
8	AlHayan & Al-Muhtadi (2026)	Intrusion Detection using FL and LLMs	Deep Learning Framework	Improved behavioral intrusion detection accuracy	Generalization across environments
9	Sani et al. (2024)	Federated LLM Pre-training	Conceptual Analysis	Advocated federated pre-training as future direction	Lack of deployment validation
10	Tang & Deng (2024)	Medical Federated LLMs	Literature Review	Identified healthcare applications and privacy benefits	Limited clinical implementation
11	Mishra & Yadav (2026)	Decentralized LLM Training	Framework Design	Proposed global decentralized training architecture	Scalability testing required
12	Gurung & Pokhrel (2025)	Quantum Federated Learning with LLMs	Distillation-based Approach	Combined LLM knowledge distillation with QFL	Early-stage research
13	Jiang et al. (2025)	Feasibility, Robustness and Security	Comprehensive Survey	Identified robustness and security concerns	Need for practical solutions

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

14	Jiang et al. (2025)	Federated LLM Security	Survey	Discussed adversarial attacks and privacy risks	Limited mitigation evaluation
15	Mittal et al. (2026)	FedLLM for 6G Networks	Integrated Architecture	Improved semantic reasoning for self-organizing networks	Real-world deployment missing
16	Zhuang et al. (2023)	Foundation Models and FL	Review	Established motivations and future directions	Limited implementation studies
17	Colombi et al. (2025)	Edge Fine-Tuning in FL	Experimental Investigation	Demonstrated edge-based FedLLM deployment	Hardware constraints remain
18	Liao et al. (2025)	Heterogeneous FL with LLMs	Importance-Aware Updating	Improved performance under non-IID data	Communication efficiency challenges
19	Wen et al. (2025)	Parameter-Efficient Fine-Tuning	Survey	Reviewed LoRA and PEFT methods in FL	Lack of unified frameworks
20	Zhao et al. (2026)	Social Bias in Federated Fine-Tuning	Experimental Study	Revealed bias propagation risks	Need for fairness-aware aggregation
21	Zhu et al. (2026)	Personalized Medical Vision-LLMs	Personalized FL Framework	Improved personalization and uncertainty handling	Limited datasets
22	Alzahrani & Yang (2026)	PrivLoRA	Privacy-Preserving Fine-Tuning	Enhanced privacy in LoRA-based FL	Computational overhead
23	Akhmetov et al. (2026)	Sovereign Personal AI Agents	Review	Explored personalized FL for AI agents	Limited practical architectures
24	Kaur et al. (2026)	FedLLM for Traffic Prediction	Privacy-Preserving Framework	Improved explainability and privacy	Large-scale validation needed

From the reviewed literature, it is evident that recent research has increasingly focused on integrating Federated Learning with Large Language Models to address privacy, security, and data governance concerns. Most studies concentrate on federated fine-tuning, parameter-efficient adaptation, and privacy-preserving mechanisms such as LoRA and secure aggregation. However, significant research gaps remain in terms of communication efficiency, robustness against adversarial attacks, fairness-aware learning, and large-scale real-world deployment. These limitations provide opportunities for developing more efficient, secure, and scalable Federated Large Language Model frameworks in future research.

Based on the systematic review of existing studies on Federated Learning (FL) and Large Language Models (LLMs), several critical research gaps have been identified. Although substantial progress has been made in developing Federated Large Language Models (FedLLMs), many challenges remain unresolved, limiting their practical deployment and scalability.

III. RESEARCH GAP

A. Limited Scalability of Federated LLMs

Most existing studies focus on the conceptual design and feasibility of Federated LLMs; however, large-scale deployment across thousands or millions of distributed clients remains largely unexplored. The enormous size of modern LLMs introduces significant communication, storage, and computational overhead during federated training. Current frameworks such as OpenFedLLM and federated fine-tuning approaches have demonstrated promising results, but their effectiveness in large-scale real-world environments requires further investigation.

B. Communication and Resource Constraints

Federated learning relies on frequent exchange of model parameters between clients and aggregation servers. In the case of LLMs containing billions of parameters, communication costs become a major bottleneck. Although parameter-efficient fine-tuning techniques such as LoRA and PEFT have reduced communication overhead, there is still a lack of optimized mechanisms for resource-constrained devices, edge environments, and heterogeneous networks.

C. Privacy and Security Vulnerabilities

While Federated Learning improves data privacy by keeping raw data local, recent studies indicate that model updates may still leak sensitive information. Existing privacy-preserving techniques such as differential privacy, secure aggregation, and encrypted communication introduce additional computational complexity and often degrade model performance. Furthermore, protection against model poisoning, adversarial attacks, and inference attacks remains an open research challenge.

D. Handling Data Heterogeneity

Most federated environments contain highly non-identically distributed (non-IID) data across participating clients. Current aggregation methods struggle to maintain model accuracy under heterogeneous data distributions. Although importance-aware updating and personalized federated learning approaches have shown potential, more robust algorithms are needed to effectively handle client diversity while preserving global model performance.

E. Bias and Fairness Issues

Recent research has revealed that social, cultural, and demographic biases can propagate during federated fine-tuning of large language models. Existing studies primarily focus on identifying bias rather than mitigating it. There is a significant need for fairness-aware aggregation strategies and bias mitigation mechanisms that ensure equitable model performance across diverse user populations.

F. Lack of Personalized Federated LLM Frameworks

Many current FedLLM architectures aim to develop a single global model for all clients. However, user requirements often vary significantly across domains and applications. Personalized Federated Learning has emerged as a promising solution, but existing frameworks are still in the early stages of development. More research is needed to balance global knowledge sharing with local personalization.

G. Insufficient Domain-Specific Implementations

Although applications of Federated LLMs have been explored in healthcare, cybersecurity, Industry 4.0, transportation, and 6G networks, most studies remain experimental or conceptual. There is limited evidence regarding real-world deployment, long-term performance evaluation, and integration with operational systems. Domain-specific frameworks tailored to practical industry requirements remain underdeveloped.

H. Need for Integrated and Explainable FedLLM Systems

Current research generally addresses privacy, efficiency, security, and explainability independently. Few studies propose a unified framework that simultaneously provides:

- Privacy preservation,
- Communication efficiency,
- Personalization,
- Explainability,
- Security robustness, and
- Scalability.

Developing an integrated Federated Large Language Model framework that balances these requirements represents a significant research opportunity.

The literature indicates that despite the growing interest in Federated Large Language Models, existing studies largely focus on individual challenges such as privacy, communication efficiency, or fine-tuning optimization. There remains a lack of comprehensive, scalable, secure, explainable, and personalized FedLLM frameworks capable of operating effectively in heterogeneous real-world environments. Therefore, future research should focus on designing integrated Federated LLM architectures that address communication constraints, privacy preservation, bias mitigation, personalization, and scalability simultaneously, thereby enabling practical deployment across diverse application domains.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

IV. FUTURE RESEARCH DIRECTIONS

The integration of Federated Learning (FL) and Large Language Models (LLMs) has emerged as a promising paradigm for developing privacy-preserving artificial intelligence systems. Despite significant advancements, several challenges remain unresolved, creating opportunities for future research. The following subsections discuss key research directions that can enhance the efficiency, security, scalability, and trustworthiness of Federated Large Language Models (FedLLMs).

A. Communication-Efficient Federated LLM Training

One of the most significant challenges in Federated LLM training is the large communication overhead caused by the transmission of model parameters between clients and the central server. Modern LLMs often contain billions of parameters, making frequent synchronization expensive in terms of bandwidth and latency. Future research should focus on developing communication-efficient techniques such as model compression, gradient sparsification, quantization, adaptive client participation, and asynchronous aggregation mechanisms. Additionally, hierarchical and clustered federated learning architectures can reduce communication costs by aggregating updates locally before transmitting them to the global server. Such approaches will enable the deployment of FedLLMs in large-scale and resource-constrained environments.

B. Privacy-Preserving Fine-Tuning Techniques

Although federated learning prevents direct sharing of raw data, model updates can still reveal sensitive information through inference attacks. Future research should explore advanced privacy-preserving fine-tuning methods that provide strong privacy guarantees while maintaining model performance. Techniques such as Differential Privacy (DP), Secure Multi-Party Computation (SMPC), Homomorphic Encryption (HE), and privacy-aware Low-Rank Adaptation (LoRA) mechanisms should be further optimized for large-scale language models. Developing lightweight privacy-preserving frameworks that minimize computational overhead while ensuring data confidentiality will be crucial for applications involving healthcare, finance, and government data.

C. Personalized Federated Large Language Models

Current federated learning frameworks typically aim to build a single global model that serves all participants. However, users often possess unique data distributions, preferences, and requirements that cannot be fully captured by a generalized model. Future studies should focus on personalized Federated LLM architectures that balance global knowledge sharing with local adaptation. Meta-learning, client-specific fine-tuning, and adaptive aggregation algorithms can help create personalized models that improve user experience while preserving privacy. Such personalization is particularly important for intelligent assistants, healthcare systems, educational platforms, and recommendation systems.

D. Bias Mitigation and Fairness-Aware Learning

Bias and fairness have become critical concerns in large language models. Federated environments may unintentionally amplify social, cultural, demographic, or linguistic biases due to heterogeneous client data distributions. Future research should investigate fairness-aware aggregation methods and bias mitigation techniques that ensure equitable model performance across diverse populations. Explainable bias detection frameworks, fairness-constrained optimization methods, and demographic-aware evaluation metrics can contribute to reducing bias propagation during federated training. Addressing fairness concerns is essential for developing responsible and inclusive AI systems.

E. Secure Aggregation and Adversarial Robustness

Security threats such as model poisoning, backdoor attacks, adversarial manipulation, and inference attacks pose serious risks to Federated LLM systems. Future research should focus on developing robust secure aggregation protocols capable of detecting and mitigating malicious client behavior without compromising privacy. Advanced anomaly detection techniques, trust-based client selection, blockchain-enabled verification mechanisms, and robust aggregation algorithms can strengthen the security of federated systems. Ensuring adversarial robustness will be essential for deploying FedLLMs in mission-critical domains such as cybersecurity, defense, healthcare, and financial services.

F. Edge Computing and Resource-Efficient Deployment

The deployment of large language models on edge devices remains challenging due to limited computational power, memory constraints, and energy consumption.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Future studies should investigate resource-efficient architectures that combine edge computing with federated learning. Model pruning, knowledge distillation, parameter-efficient fine-tuning, and lightweight transformer architectures can help reduce computational requirements. Integrating FedLLMs with edge intelligence frameworks will enable real-time decision-making, lower latency, reduced cloud dependency, and enhanced privacy in Internet of Things (IoT), smart city, and mobile computing applications.

G. Federated LLMs for Domain-Specific Applications

Although Federated LLMs have shown promise across multiple sectors, most existing implementations remain experimental. Future research should focus on developing domain-specific FedLLM frameworks tailored to healthcare, finance, education, cybersecurity, Industry 4.0, transportation, and 6G communication networks. These specialized models should incorporate domain knowledge, regulatory requirements, and application-specific privacy constraints. Furthermore, extensive real-world validation studies are required to assess their effectiveness, scalability, and reliability in practical operational environments.

H. Explainable and Trustworthy Federated AI Systems

The increasing complexity of Federated LLMs raises concerns regarding transparency, interpretability, and trust. Users and organizations often require explanations for model decisions, especially in high-stakes applications. Future research should focus on integrating Explainable Artificial Intelligence (XAI) techniques within federated learning frameworks to improve model interpretability without compromising privacy. Trustworthy AI systems should also address issues related to accountability, fairness, robustness, privacy compliance, and ethical governance. Developing explainable and trustworthy FedLLM architectures will enhance user confidence and support the broader adoption of privacy-preserving AI technologies.

Future advancements in Federated Large Language Models will depend on addressing key challenges related to communication efficiency, privacy protection, personalization, fairness, security, resource optimization, domain adaptation, and explainability. By overcoming these limitations, researchers can develop scalable and trustworthy privacy-preserving AI systems capable of supporting next-generation intelligent applications across diverse domains.

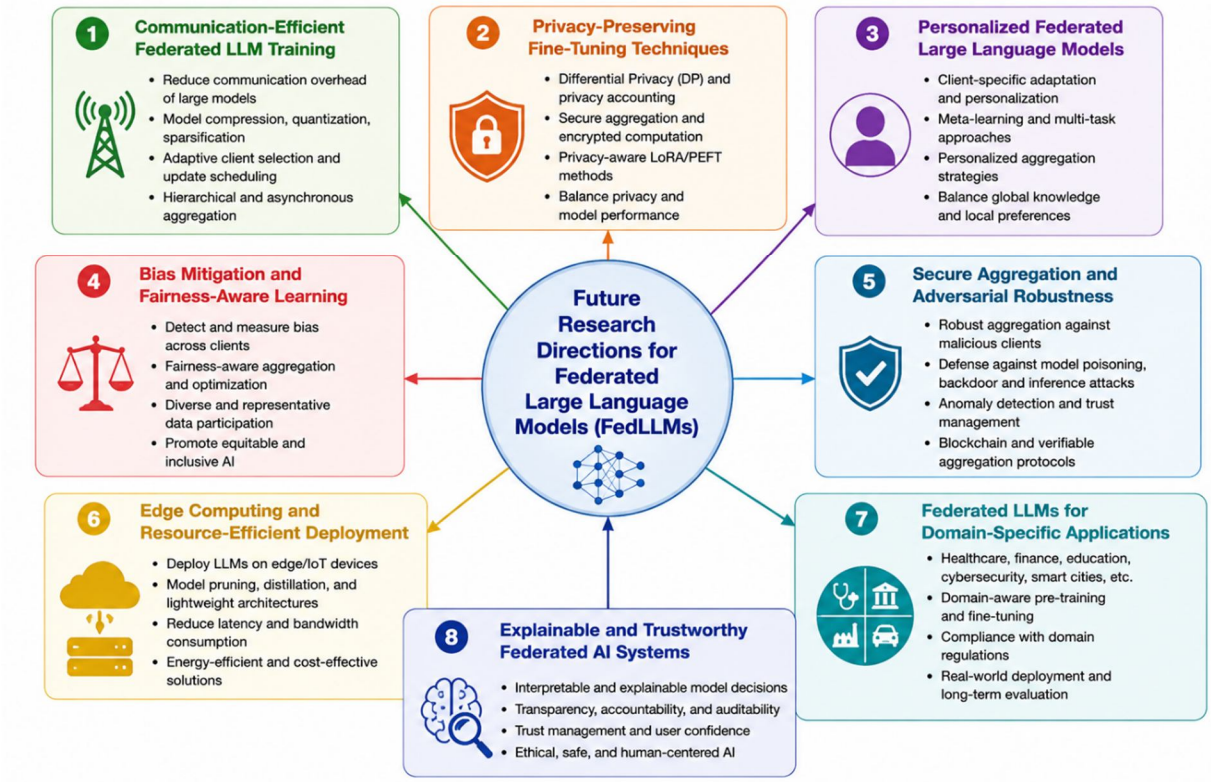


Figure 4: Future Research Directions for Federated Large Language Models (FedLLMs)

Figure 4 presents a comprehensive roadmap of the major future research directions in Federated Large Language Models (FedLLMs). At the center of the figure is the concept of Future Research Directions for Federated Large Language Models, which acts as the foundation for advancing privacy-preserving and decentralized artificial intelligence systems. Surrounding the central node are eight interconnected research domains that collectively address the limitations of current Federated LLM frameworks. The first research direction, Communication-Efficient Federated LLM Training, focuses on reducing communication overhead through techniques such as model compression, quantization, gradient sparsification, and hierarchical aggregation. Privacy-Preserving Fine-Tuning Techniques emphasize the integration of differential privacy, secure aggregation, encryption, and privacy-aware parameter-efficient tuning methods to protect sensitive information during distributed training. The figure also highlights Personalized Federated Large Language Models, which aim to adapt global models to client-specific requirements while maintaining collaborative learning benefits. Bias Mitigation and Fairness-Aware Learning addresses the need for equitable model performance by reducing demographic, social, and cultural biases that may emerge during federated training. Furthermore, Secure Aggregation and Adversarial Robustness focus on defending Federated LLM systems against model poisoning, backdoor attacks, malicious clients, and inference attacks through robust aggregation and trust management mechanisms. Edge Computing and Resource-Efficient Deployment explore lightweight model architectures, pruning, knowledge distillation, and energy-efficient solutions to support deployment on edge and Internet of Things (IoT) devices. The figure also identifies Federated LLMs for Domain-Specific Applications as an important future direction, emphasizing customized implementations for healthcare, finance, education, cybersecurity, Industry 4.0, and smart city applications. Finally, Explainable and Trustworthy Federated AI Systems aim to improve transparency, interpretability, accountability, and user trust by integrating explainable artificial intelligence techniques into federated learning environments. The figure illustrates how these interconnected research areas can collectively contribute to the development of scalable, secure, privacy-preserving, and trustworthy Federated Large Language Model ecosystems capable of supporting next-generation intelligent applications.

V. DISCUSSION

The integration of Federated Learning (FL) and Large Language Models (LLMs) has emerged as a promising research area for developing privacy-preserving and decentralized artificial intelligence systems. The reviewed literature demonstrates significant advancements in Federated Large Language Models (FedLLMs), particularly in privacy protection, parameter-efficient fine-tuning, personalized learning, and distributed model training. However, several technical and operational challenges continue to hinder their large-scale adoption. This section discusses the comparative performance of existing approaches, the trade-offs involved in FedLLM development, practical deployment challenges, and emerging trends shaping the future of federated foundation models.

A. Comparative Analysis of Existing FedLLM Approaches

Existing FedLLM approaches can be broadly categorized into federated pre-training, federated fine-tuning, parameter-efficient federated learning, personalized federated learning, and domain-specific federated architectures. Studies such as OpenFedLLM and federated pre-training frameworks focus on training large language models directly on decentralized data sources, enabling collaborative learning without centralizing sensitive information. While these approaches provide strong privacy guarantees, they often suffer from substantial communication and computational overhead.

Federated fine-tuning methods have emerged as a more practical alternative, allowing pre-trained models to be adapted locally using client-specific data. Parameter-efficient techniques such as Low-Rank Adaptation (LoRA) and other PEFT methods significantly reduce the number of trainable parameters exchanged during training, thereby improving communication efficiency. Personalized federated learning approaches further enhance model performance by adapting global models to individual client requirements, making them suitable for applications with highly heterogeneous data distributions. Domain-specific implementations in healthcare, cybersecurity, transportation, and Industry 4.0 demonstrate the versatility of FedLLMs; however, their effectiveness often depends on the availability of specialized datasets and infrastructure. Overall, no single approach fully addresses privacy, efficiency, personalization, scalability, and explainability simultaneously, indicating the need for integrated FedLLM frameworks.

B. Trade-offs Between Privacy, Performance, and Scalability

A major challenge in Federated Large Language Models is balancing privacy preservation, model performance, and scalability. Strong privacy-preserving techniques such as differential privacy, secure aggregation, and homomorphic encryption effectively protect sensitive information but often introduce computational overhead and reduce model accuracy. Similarly, parameter-efficient fine-tuning methods improve communication efficiency but may limit the model's ability to capture complex patterns compared to full-model training.

Scalability presents another important trade-off. Increasing the number of participating clients enhances data diversity and model generalization but also increases communication costs, synchronization complexity, and convergence time. Personalized federated learning improves local model performance but may reduce the consistency and generalizability of the global model. Consequently, future FedLLM systems must carefully balance these competing objectives to achieve practical deployment while maintaining privacy, accuracy, and efficiency.

C. Practical Challenges in Real-World Deployment

Although numerous studies have demonstrated the feasibility of Federated Large Language Models in controlled environments, real-world deployment remains challenging. One major issue is the heterogeneity of client devices, network conditions, and data distributions. Clients often possess varying computational capabilities, storage capacities, and communication bandwidth, making uniform training strategies ineffective. Furthermore, non-identically distributed (non-IID) data can significantly affect model convergence and performance.

Security concerns also remain a critical barrier to deployment. Federated systems are vulnerable to model poisoning attacks, backdoor attacks, malicious client participation, and inference-based privacy leakage. Regulatory compliance and data governance requirements further complicate implementation, particularly in sensitive sectors such as healthcare and finance. In addition, deploying LLMs on edge devices introduces challenges related to memory consumption, energy efficiency, and real-time inference. Addressing these practical issues requires the development of robust, adaptive, and resource-efficient federated learning architectures capable of operating under diverse real-world conditions.

D. Emerging Trends in Federated Foundation Models

Recent research indicates a growing shift from traditional federated learning toward federated foundation models that integrate large-scale pre-trained architectures with decentralized learning paradigms. One notable trend is the adoption of parameter-efficient fine-tuning techniques, including LoRA and adapter-based learning, which significantly reduce communication and computation requirements. Personalized federated learning is also gaining attention as researchers seek to accommodate individual user preferences while preserving collaborative knowledge sharing.

Another emerging trend is the integration of Federated Learning with edge computing, Internet of Things (IoT) infrastructures, and next-generation 6G communication networks. These developments aim to support low-latency, privacy-preserving AI services at the network edge. Additionally, explainable and trustworthy AI frameworks are becoming increasingly important, particularly for applications involving healthcare, finance, and public services where transparency and accountability are essential. Researchers are also exploring the combination of Federated Learning with Knowledge Graphs, multimodal foundation models, blockchain technologies, and quantum computing to enhance intelligence, security, and scalability. These trends suggest that federated foundation models will play a central role in the future development of secure, distributed, and privacy-aware artificial intelligence systems.

The discussion highlights that Federated Large Language Models offer substantial benefits in terms of privacy preservation, decentralized intelligence, and collaborative learning. However, significant challenges remain regarding communication efficiency, scalability, security, personalization, and real-world deployment. Future research should focus on developing integrated federated foundation model frameworks that balance privacy, performance, scalability, explainability, and trustworthiness while supporting diverse application domains.

VI. PROPOSED RESEARCH FRAMEWORK

To address the limitations identified in existing Federated Large Language Model (FedLLM) research, this study proposes an integrated framework that combines privacy preservation, security enhancement, parameter-efficient learning, personalization, and explainability within a unified federated environment. The proposed framework aims to facilitate secure and efficient collaborative training of Large Language Models across distributed clients without requiring the sharing of sensitive raw data. By integrating advanced privacy-preserving mechanisms, lightweight fine-tuning techniques, personalized learning strategies, and explainable artificial intelligence components, the framework seeks to improve scalability, trustworthiness, and practical applicability across diverse domains such as healthcare, finance, education, cybersecurity, and Industry 4.0.

A. Integrated Federated LLM Architecture

The proposed architecture consists of multiple distributed clients, a federated aggregation server, and several functional layers that support collaborative model training.

Each client maintains local datasets and performs model training or fine-tuning independently. Instead of sharing raw data, clients transmit encrypted model updates to the central aggregation server. The server combines local updates using federated aggregation algorithms and generates an improved global model, which is subsequently redistributed to participating clients. This architecture enables decentralized learning while maintaining data privacy and regulatory compliance. Furthermore, the framework incorporates adaptive aggregation strategies to handle heterogeneous client environments and varying computational capabilities.

B. Privacy and Security Layer

The Privacy and Security Layer forms the foundation of the proposed framework by ensuring secure communication and protecting sensitive information throughout the federated learning process. This layer integrates Differential Privacy (DP), Secure Aggregation, and encryption-based mechanisms to minimize information leakage from model updates. Differential privacy introduces carefully calibrated noise into local model parameters, preventing the reconstruction of private data. Secure aggregation protocols ensure that the server can only access aggregated updates rather than individual client contributions. Additionally, anomaly detection and trust management mechanisms are incorporated to identify malicious clients and mitigate model poisoning, backdoor attacks, and adversarial threats. These security measures collectively enhance the confidentiality, integrity, and robustness of the Federated LLM ecosystem.

C. Parameter-Efficient Fine-Tuning Module

Given the enormous size of modern Large Language Models, full-scale federated training can be computationally expensive and communication-intensive. Therefore, the proposed framework incorporates a Parameter-Efficient Fine-Tuning (PEFT) module that utilizes techniques such as Low-Rank Adaptation (LoRA), adapters, and prompt tuning. Rather than updating all model parameters, only a small subset of trainable parameters is optimized and transmitted during federated learning. This significantly reduces communication overhead, computational cost, memory requirements, and energy consumption. The PEFT module enables resource-constrained devices and edge environments to participate effectively in federated training while maintaining competitive model performance.

D. Personalized Learning Component

Client data distributions often vary significantly across organizations, users, and application domains. To address this challenge, the proposed framework incorporates a Personalized Learning Component that balances global knowledge sharing with local adaptation. After receiving the global model from the aggregation server, each client applies personalized fine-tuning using local datasets and user-specific preferences. Adaptive aggregation techniques and client-specific parameter layers allow the framework to capture local characteristics without sacrificing collaborative learning benefits. This personalized approach improves model relevance, user satisfaction, and predictive accuracy, particularly in domains such as healthcare, education, recommendation systems, and intelligent personal assistants.

E. Explainability and Trust Module

The Explainability and Trust Module enhances transparency and user confidence by providing interpretable insights into model predictions and learning processes. Explainable Artificial Intelligence (XAI) techniques are integrated to identify influential features, generate decision explanations, and support model auditing. The module also incorporates fairness monitoring mechanisms to detect and mitigate potential biases arising from heterogeneous client data. Additionally, accountability and compliance features enable organizations to verify that federated training processes adhere to ethical and regulatory requirements. By improving transparency, fairness, and trustworthiness, this module facilitates the responsible deployment of Federated Large Language Models in high-stakes applications where interpretability is essential.

F. Summary of the Proposed Framework

The proposed framework integrates federated learning, privacy-preserving technologies, parameter-efficient fine-tuning, personalized learning, and explainable AI into a unified architecture for Federated Large Language Models. The framework addresses major challenges identified in the literature, including communication overhead, privacy risks, security vulnerabilities, data heterogeneity, and lack of explainability. Through its modular design, the framework provides a scalable, secure, efficient, and trustworthy solution for deploying privacy-preserving AI systems across distributed environments. This integrated approach offers a promising foundation for future research and real-world implementation of Federated Large Language Models.

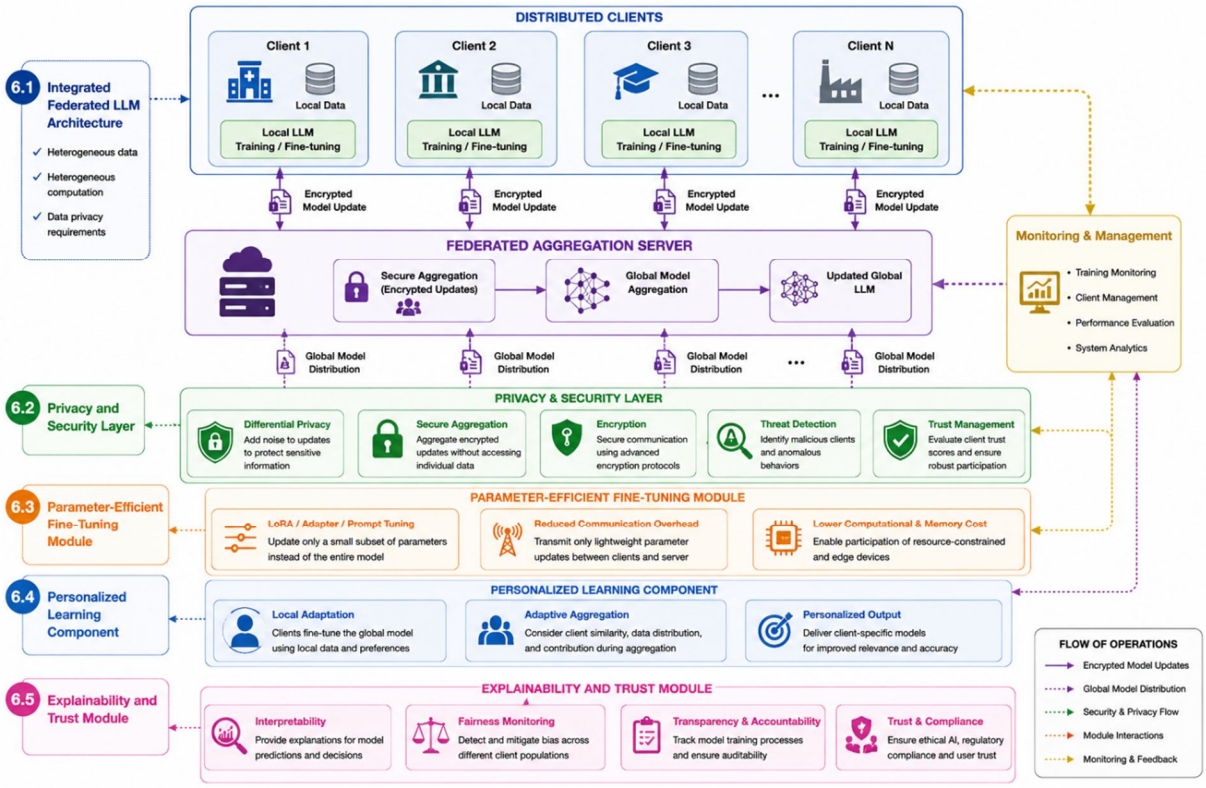


Figure 5: Proposed Research Framework for Federated Large Language Models (FedLLMs)

Figure 5 illustrates the proposed integrated research framework for Federated Large Language Models (FedLLMs), designed to address the key challenges identified in existing federated AI systems, including privacy preservation, communication efficiency, personalization, security, and explainability. The framework consists of five interconnected modules that collaboratively enable secure and efficient decentralized training of large language models across multiple distributed clients.

At the top of the framework, the Integrated Federated LLM Architecture represents the distributed learning environment where multiple clients from diverse domains such as healthcare, education, finance, and industry maintain local datasets and perform local model training or fine-tuning. Instead of sharing raw data, clients transmit encrypted model updates to a centralized Federated Aggregation Server, which performs secure model aggregation and distributes the updated global model back to participating clients. This architecture supports heterogeneous data sources, varying computational capabilities, and privacy-sensitive applications.

The Privacy and Security Layer serves as the foundation of the framework by ensuring secure communication and protecting sensitive information throughout the federated learning process. This layer integrates Differential Privacy, Secure Aggregation, Encryption, Threat Detection, and Trust Management mechanisms. These technologies collectively prevent data leakage, defend against malicious clients, and ensure confidentiality, integrity, and robustness during model training and update exchange.

To address the computational challenges associated with training large language models, the framework incorporates a Parameter-Efficient Fine-Tuning Module. This module utilizes techniques such as LoRA (Low-Rank Adaptation), Adapter Tuning, and Prompt Tuning to update only a small subset of model parameters instead of the entire model. As a result, communication overhead, memory consumption, and computational costs are significantly reduced, enabling efficient participation of edge devices and resource-constrained clients.

The framework further includes a Personalized Learning Component that enhances model adaptability by combining global knowledge with local customization. Through local adaptation and adaptive aggregation mechanisms, clients can personalize the global model according to their specific data distributions and user requirements. This approach improves prediction accuracy, relevance, and user satisfaction while maintaining the collaborative benefits of federated learning.

At the bottom of the architecture, the Explainability and Trust Module improves transparency and accountability by integrating explainable artificial intelligence techniques.

This module provides model interpretability, fairness monitoring, transparency tracking, and regulatory compliance support. It enables stakeholders to understand model decisions, detect potential biases, and ensure ethical AI deployment in sensitive application domains.

Additionally, the framework includes a Monitoring and Management Layer that continuously supervises training activities, client participation, performance evaluation, and system analytics. This layer provides feedback to all modules, ensuring efficient operation, security monitoring, and adaptive system optimization.

Overall, Figure 6.1 presents a comprehensive and scalable Federated Large Language Model framework that integrates privacy preservation, security enhancement, parameter-efficient learning, personalization, and explainability within a unified architecture. The proposed framework offers a practical foundation for developing trustworthy, efficient, and privacy-preserving AI systems capable of supporting next-generation distributed intelligent applications.

VII. CONCLUSION

The integration of Federated Learning (FL) and Large Language Models (LLMs) represents a promising approach for developing privacy-preserving, secure, and decentralized artificial intelligence systems. This review examined recent advances, challenges, and future research opportunities in Federated Large Language Models (FedLLMs), highlighting their potential to enable collaborative learning without compromising sensitive data. The findings indicate that FedLLMs can significantly enhance privacy protection while supporting diverse applications across healthcare, finance, education, cybersecurity, Industry 4.0, and intelligent communication networks.

A. Summary of Findings

The review revealed that Federated Learning provides an effective framework for training and fine-tuning Large Language Models on distributed data sources while preserving user privacy. Recent advancements in federated fine-tuning, parameter-efficient learning techniques, personalized learning, and secure aggregation have improved the practicality of FedLLMs. However, challenges related to communication efficiency, data heterogeneity, security vulnerabilities, scalability, and model explainability continue to limit widespread deployment.

B. Key Contributions of the Review

This review provides a comprehensive analysis of existing research on Federated Large Language Models by examining their architectures, training approaches, privacy-preserving mechanisms, security considerations, and domain-specific applications. It identifies major research trends, highlights critical research gaps, and proposes an integrated FedLLM framework that combines privacy protection, parameter-efficient fine-tuning, personalization, and explainability. The study also outlines future research directions that can guide the development of next-generation privacy-preserving AI systems.

C. Limitations of Current Research

Despite significant progress, current FedLLM research remains largely experimental and faces several limitations. Most studies focus on specific aspects such as privacy, fine-tuning, or security rather than providing comprehensive solutions. Large-scale real-world deployments are still limited, and challenges related to communication overhead, adversarial attacks, fairness, and resource constraints remain unresolved. Additionally, there is a lack of standardized evaluation frameworks for assessing the performance, security, and scalability of Federated LLM systems.

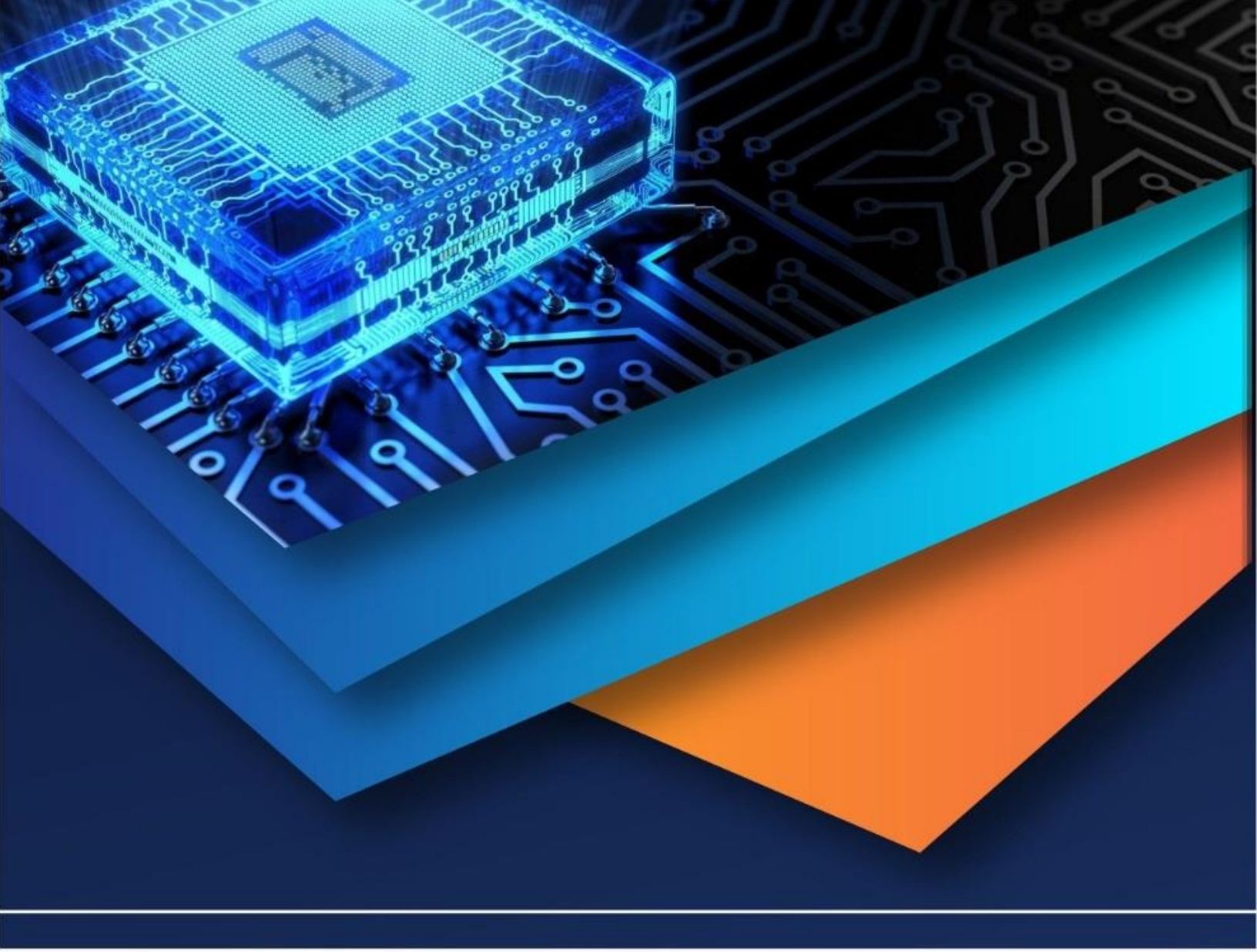
D. Future Outlook for Privacy-Preserving AI

The future of privacy-preserving AI is expected to be driven by the continued convergence of Federated Learning, Large Language Models, edge computing, explainable AI, and secure distributed systems. Emerging technologies such as parameter-efficient fine-tuning, personalized federated learning, differential privacy, blockchain-based security, and trustworthy AI frameworks will play a crucial role in enabling scalable and secure AI deployment. As research progresses, Federated Large Language Models are expected to become a foundational technology for building intelligent, transparent, and privacy-aware AI systems capable of operating across diverse real-world environments.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Kenteris, M., & Kotis, K. (2026). The Convergence of Federated Learning, Knowledge Graphs, and Large Language Models for Language Learning: A Scoping Review. Applied Sciences, 16(5), 2611.
[2] Thakur, D., Guzzo, A., & Fortino, G. (2025, May). Analyzing the Fusion of Federated Learning and Large Language Model. In 2025 IEEE 5th International Conference on Human-Machine Systems (ICHMS) (pp. 282-288). IEEE.
[3] Jing, F., Zhang, Y., Gao, M., Zhang, X., & Zhou, H. (2026). A Review of Federated Large Language Models for Industry 4.0. Sensors, 26(4), 1116.
[4] Wu, Y., Tian, C., Li, J., Sun, H., Tam, K., Zhou, Z., ... & Xu, C. (2025). A survey on federated fine-tuning of large language models. arXiv preprint arXiv:2503.12016.
[5] Ye, R., Wang, W., Chai, J., Li, D., Li, Z., Xu, Y., ... & Chen, S. (2024, August). 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 (pp. 6137-6147).
[6] Hu, J., Wang, D., Wang, Z., Pang, X., Xu, H., Ren, J., & Ren, K. (2024). Federated large language model: Solutions, challenges and future directions. IEEE Wireless Communications, 32(4), 82-89.
[7] Hilmkil, A., Callh, S., Barbieri, M., Sütfeld, L. R., Zec, E. L., & Mogren, O. (2021, June). Scaling federated learning for fine-tuning of large language models. In International Conference on Applications of Natural Language to Information Systems (pp. 15-23). Cham: Springer International Publishing.
[8] AlHayan, A., & Al-Muhtadi, J. (2026). Federated learning-powered real-time behavioral intrusion detection leveraging LSTM, attention, GANs, and large language models. Scientific Reports.
[9] Sani, L., Iacob, A., Cao, Z., Marino, B., Gao, Y., Paulik, T., ... & Lane, N. D. (2024). The future of large language model pre-training is federated. arXiv preprint arXiv:2405.10853.
[10] Tang, Y., & Deng, Y. (2024, July). Current research and prospects of federated language large models in the medical field. In Third International Conference on Biomedical and Intelligent Systems (IC-BIS 2024) (Vol. 13208, pp. 816-824). SPIE.
[11] Mishra, N., & Yadav, P. (2026). Federated Learning for Decentralized Language Model Training across Global Data Sources. Procedia Computer Science, 275, 148-156.
[12] Gurung, D., & Pokhrel, S. R. (2025). Llm-qfl: Distilling large language model for quantum federated learning. arXiv preprint arXiv:2505.18656.
[13] Jiang, W., Luo, Y., Deng, G., Chen, S., Yang, X., Wu, S., ... & Fu, S. (2025). Federated Large Language Models: Feasibility, Robustness, Security and Future Directions. arXiv preprint arXiv:2505.08830.
[14] Jiang, W., Luo, Y., Deng, G., Chen, S., Yang, X., Wu, S., ... & Fu, S. (2025). Federated Large Language Models: Feasibility, Robustness, Security and Future Directions. arXiv preprint arXiv:2505.08830.
[15] Mittal, P., Sharma, S., Solanki, S., Kumar, L., Haripriya, R., & Beg, R. (2026). Design of an integrated federated large language models for semantic reasoning and self-organizing 6G networks. Discover Artificial Intelligence.
[16] Zhuang, W., Chen, C., Li, J., Chen, C., Jin, Y., & Lyu, L. (2023). When foundation model meets federated learning: Motivations, challenges, and future directions. arXiv preprint arXiv:2306.15546.
[17] Colombi, L., Vespa, M., Resca, F., Cavicchi, S., Di Caro, E., Bellodi, E., ... & Stefanelli, C. (2025). Investigating edge fine-tuning of large language models in a federated environment.
[18] Liao, Y., Huang, W., Wan, G., Liang, J., Yang, B., & Ye, M. (2025, October). Splitting with Importance-aware Updating for Heterogeneous Federated Learning with Large Language Models. In Forty-second International Conference on Machine Learning.
[19] Wen, Q., Zhang, X., Xiang, N., Chen, J., Wang, X., & Zhang, J. (2025, November). A Survey on Federated Parameter-Efficient Fine-Tuning for Large Language Models. In 2025 11th International Conference on Big Data and Information Analytics (BigDIA) (pp. 637-642). IEEE.
[20] Zhao, J., Fang, M., Zhong, M., Zheng, S., Chen, L., & Pechenizkiy, M. (2026, March). Investigating Social Bias Propagation in Federated Fine-tuning of Large Language Models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 46, pp. 39637-39645).
[21] Zhu, H., Togo, R., Ogawa, T., & Haseyama, M. (2026). Personalized federated learning for medical vision-language models via efficient fine-tuning and uncertainty-aware disentanglement. Journal of Biomedical Informatics, 105014.
[22] Alzahrani, B., & Yang, D. (2026, February). PrivLoRA: Enhancing Privacy in LoRA-Based Fine-Tuning of Large Language Models for Federated Learning. In 2026 International Conference on Computing, Networking and Communications (ICNC) (pp. 505-511). IEEE.
[23] Akhmetov, A., Sharimbayev, B., & Ala'anzy, M. A. (2026, April). Personalized Federated Learning for Sovereign Personal AI Agents: A Review. In 2026 18th International Conference on Electronics, Computer, and Computation (ICECCO) (pp. 1-6). IEEE.
[24] Kaur, S., Sehra, S. S., & Ebrahimi, D. (2026). FedLLM: A Privacy-Preserving Federated Large Language Model for Explainable Traffic Flow Prediction. arXiv preprint arXiv:2604.16612.x



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)