Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD37467

Federated Learning in the Era of Large Models: Core Challenges, Representative Solutions, and Emerging Directions

Zhixi Wang

Wang Zheng School of Microelectronics (School of Integrated-Circuit Industry), Changzhou University, Changzhou, China
wzx15261007@163.com

Abstract. With the rapid advancement of the Internet and the emergence and popularity of large language models, the demands for high-quality data, communication capacity and security and privacy have increased significantly. However, a large amount of data is distributed across different devices, and cannot be collected legally because of security and privacy constraints. Thus, Federated Learning was proposed to address these issues. Federated Learning enables multiple clients to train model collaboratively without uploading raw data. Federated Learning serves as a key role in tackling data problems in the era of large language models. According to this condition, it is important to understand the state-of-the-art of Federated Learning. Using literature review and comparative analysis, this paper examines three core challenges in Federated Learning: heterogeneity, communication overhead and security and privacy risks. It also reviews representative solutions based on drift correction, meta-learning and homomorphic encryption, etc., to these three problems. The analysis shows that these methods generally improve one or two objectives. Heterogeneity alleviation may increase model complexity, and privacy mechanisms may introduce additional computation and communication costs. Moreover, large language models further intensify these trade-offs. Parameter-efficient fine-tuning improves the possibility of federated large model training via freezing the backbone weights and transferring a small subset of parameters. Consequently, Federated Learning is moving from full-parameter model collaboration to lightweight knowledge collaboration, with potential advancement toward multi-agent assignment coordination.

Keywords: Federated Learning, Large Language Models, Heterogeneity, Communication, Security and Privacy

1. Introduction

Large Language Models (LLMs) are Transformer-based deep neural networks which are pre-trained on a substantial amount of unlabeled textual corpora through self-supervised learning. They can be suitably adapted to various downstream tasks through fine-tuning and prompting [1]. The development of LLMs has significantly expanded the capabilities of artificial intelligence and scaled

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD37467

up the model parameters. There are three main architectures: Decoder-only architecture (GPT, Qwen), Encoder-only architecture (BERT) and Encoder-Decoder architecture (T5).

Federated Learning (FL) is a collaborative machine learning paradigm whose structure is composed of multiple data owners at the lower layer and a central server at the upper layer [2]. It defines N data owners $\{F_1, \ldots F_N\}$, all of whom hope to train a high-quality model by merging their respective data $\{D_1, \ldots D_N\}$. Instead of the traditional method, which uses $D = D_1 \cup \cdots \cup D_N$ to train model M_{sum}, FL is a process in which owner F_i does not expose its data D_i to others [3, 4]. In this framework, all the data are kept locally instead of being stored on the server. Model training and updates are implemented solely through sharing model parameters [5]. This approach not only guarantees data privacy but also realizes multi-device collaboration, achieving "data stays local, model collaborate".

But FL faces three challenges: data and system heterogeneity, communication overhead and security and privacy risks. These difficulties are exacerbated when FL is amplified to LLMs.

Using literature review and comparative analysis, this paper examines the three core challenges, their representative solutions and their trade-offs. It also discusses the transition from full-parameter tuning to Parameter-Efficient Fine-Tuning (PEFT), and the future research direction of multi-agent systems integrated with FL. This paper aims to help researchers understand FL systematically and support practitioners in selecting appropriate methods for large-model applications.

2. Core challenges in federated learning

2.1. Data and system heterogeneity

The heterogeneity problems are divided into two dimensions: statistical heterogeneity and system heterogeneity [6]. The former covers two parts: optimization-objective drift and optimization-direction drift. The latter includes model structure differences and computing resource differences. These four problems co-exist at the same time, resulting in interference with training and updating of local models and passive impacts on the convergence and update of the global model. Furthermore, these discrepancies make it difficult for the optimization direction to adapt to the entire system. Therefore, it is important to adopt solutions which can mitigate these problems.

2.2. Communication overhead

During model training, FL involves transmission of massive parameters over multiple rounds, which leads to high communication overhead. In the paper by Rothchild D, there are three main constraints on model training [7]. First, communication efficiency, as the connection speed between clients and central aggregator is slow(~1Mbps). Second, clients must be stateless, because clients usually involve in only one training round. Third, the data collected from clients is typically independent and identically distributed (Non-IID) because of many reasons.

2.3. Security and privacy risks

FL reduces the risk of privacy leakage in traditional machine learning mechanism. However, adversaries can still infer private data from the model and gradient parameters uploaded by clients. Meanwhile, FL requires frequent interaction and updates among numerous nodes, which generates substantial communication traffic. Moreover, encrypted data further expands the scale of

DOI: 10.54254/2755-2721/2026.LD37467

communication volume. Therefore, balancing communication-cost reduction and private-data protection remains a major challenge [8].

3. Representative solutions to the core challenges

3.1. Solutions to heterogeneity

To solve the aforementioned problems, Shen proposed four core approaches [6].

3.1.1. FedEve

Shen decomposes Heterogeneity into two problems: Client Drift and Periodic Drift, and quantifies and adjusts them through a prediction-observation framework named FedEve. In the process of prediction, the server predicts the model state after next-round communication based on the current state. In the process of observation, the server collects and analyzes the real information from clients.

Shen introduces the Nesterov momentum mechanism, which estimates a look-ahead position based on the current momentum before gradient computation, and then calculates the gradient at this position to guide parameter updates. As a result, it can enable the optimization process to anticipate the momentum direction and achieve more forward-looking updates.

3.1.2. FedMcon

In Non-IID scenarios, traditional method like FedAvg approximates the global optimization objective by performing weighted averaging of local models.

Shen integrates meta-learning into FL and proposes the FedMcon frame, which treats the aggregator as a meta-learner and dynamically adjusts and optimizes the parameter and model aggregation process based on the information provided by clients. Meanwhile, FedMcon introduces control variables on the client side to correct local-model update directions towards the global optimum, thereby alleviating update drifts.

3.1.3. FML

The core thought of Knowledge Distillation is to transfer the knowledge learned by a high-quality teacher model to a simpler and more computationally efficient student model, which reduces model complexity while maintaining performance as much as possible and improves the model ability to adapt to different application scenarios.

Shen introduces Knowledge Distillation into the stage of local update process and proposes the FML framework. FML introduces a dual-model architecture on the client side, where each client can maintain a meme model and a personalized model during local updates at the same time. The two models perform multiple rounds of mutual learning and Knowledge Distillation through Deep Mutual Learning, which can effectively improve model performance and while maintaining global consistency and achieving personalized optimization on the client side.

3.1.4. FedLow

Low-rank matrix factorization is an important approach for model compression and parameter reduction. Based on this approach, Low-Rank Adaptation (LoRA) was generated. LoRA adjusts

DOI: 10.54254/2755-2721/2026.LD37467

model weights through learning low-rank update matrices:

$$W' = W + \alpha BA \tag{1}$$

Where $B \in R^{m \times r}$ and $A \in R^{r \times n}$ are learnable low-rank matrices, and α is an adjustable scaling factor that controls the magnitude of the low-rank update.

Shen proposes FedLow framework, which introduces a low-rank correction structure into each layer of the neural network to reduce the local computational overhead, thereby enabling resource-constrained clients to participate in model training. Moreover, Shen designs low-rank decomposition strategies for convolutional and attention layers. For convolutional layers, convolution kernel W is decomposed into the product of two low-rank matrices A and B, allowing the input to undergo dimensionality-reduction convolution and dimensionality-expansion convolution to get the output. For attention layers, the three projection matrices Q, V, K are separately decomposed, sharing low-rank matrices A. So FedLow can improve model performance and enhance its generalization capability simultaneously.

3.2. Solutions to communication overhead

Rothchild D et al. proposed a framework named FetchSGD, which is capable of training a high-quality model efficiently under several constraints aforementioned. In this frame, after computing a gradient based on their local data, clients compress the gradient using a data structure called a Count Sketch before sending it to the central aggregator at each round [6]. Momentum and error accumulation Count Sketches are remained, and the weight update applied at each round is extracted from the error accumulation sketch.

FetchSGD leverage the property that both momentum and error feedback are linear operations in Count-Sketch, allowing clients to compute only local gradients without maintaining local states.

3.3. Secutiry

Homomorphic Encryption (HE) is one of the mainstream encryption schemes for privacy protection. A homomorphism is a structure-preserving map between two algebraic structures in abstract algebra, such as groups.

A group is a set G, together with an operation o, which combines any two elements a and b to construct another element, denoted a o b. (G, o) must satisfy four requirements: closure, associativity, identity element and inverse element [9].

HE enables specific computations to be performed on encrypted data without decryption, while ensuring that the decrypted computational results are consistent with result obtained from the corresponding plaintext operations, thereby reducing the risk of data exposure.

Jiang adopts the Paillier encryption scheme, which is a semantically secure probabilistic encryption method [8]. The Paillier mainly consists of three procedures: key generation, encryption and decryption. In the step of key generation, two large primes of equal bit length p and q, are selected randomly, based on which the corresponding public and private keys are generated.

$$n = p \times q \tag{2}$$

DOI: 10.54254/2755-2721/2026.LD37467

$$\lambda = lcm(p-1, q-1) \tag{3}$$

$$\mu = L(g^{\lambda} \bmod n^2)^{-1} \bmod n \tag{4}$$

The public key (n, g) is distributed to all clients, while the unique private key (λ, μ) is retained by the server.

Jiang first sparsifies the weight updates obtained from local model training, reducing the scale of parameter needing encryption from millions to ten. Then the sparsified weight values are sequentially fitted using a ninth-degree polynomial, and the resulting polynomial coefficients are encrypted using the Paillier encryption scheme. This method considerably reduces the encryption overhead and improves encryption efficiency while protecting the information within model updates.

Notably, these three challenges are not entirely independent. Trade-offs usually exist among them, making it difficult to a single solution to optimize all the aspects.

4. Federated learning in the era of large models

4.1. Scalability challenges in federated LLM training

With the continuous development and innovation of techniques and engineering frame, Pre-trained Foundational Model (PFM) has achieved great success, such as the GPT series, Llama (Large Language Model Meta AI), Gemma, GLM (General Language Model), etc. [9].

The parameter scale of PFMs has grown substantially because of the rapid expansion of public datasets and improvements in computational capabilities, from millions to hundreds of billions. The highly heterogeneous end devices and numerous local personalized parameters further exacerbate client-side heterogeneity, communication overhead and challenges, and raise higher requirements for data security and privacy protection. FL needs to balance these problems in a more sophisticated environment.

4.2. From federated full-parameter tuning to PEFT

The rapid development of Transformer-based language models has driven expansion of model scale, demonstrating obvious advantages in long-sequence modeling and parallel computation. However, as the number of model parameters continues to increase, full fine-tuning (FFT) is not suitable in some resource-constrained scenarios. To solve this problem, PEFT only updates a small portion of parameters, enabling effective and efficient adaptation of pretrained models with significantly reduced resource overhead [2].

The core idea is to freeze most of the model parameters while only training a small number of selected parameters, thereby reducing training costs and enabling high-quality adaption to distinct assignments. Representative PEFT methods include LoRA, Prefix Tuning and Prompt Tuning.

DOI: 10.54254/2755-2721/2026.LD37467

4.3. Federated multi-agent systems as a future direction

As agent technologies continue to develop and improve, FL is expected to get out of its old position and be a supporting mechanism for model training and multi-agent systems. In the future, agents deployed on different devices may remain their local memories and skills while exchanging information and experience to enable task allocation and collaboration. This evolution may extend the scope of federated collaboration from model parameters to task-level interactions, shifting the focus from "collaboratively training models" to "collaboratively accomplishing tasks".

Shi et al. preliminarily explore Federated Multi-Agent Systems(MAS), but there are some problems like Heterogeneous Privacy Protocols, Contextual Structure Variations and Dynamic Network Structure need to be tackled with in the future [10].

5. Conclusion

FL should not be viewed solely as an isolated algorithm, but as an effective learning paradigm that enables multiple participants to train model collaboratively without centralizing raw data. This paper reviews three core challenges in FL: heterogeneity, communication overhead and security and privacy risks and their representative solutions. It also expounds how these problems are amplified in the era of LLMs. Furthermore, it identifies the future research direction of FL. When multi-agent system becomes mainstream, FL will be integrated into various engineering project as a key thinking perspective and solution for tackling diverse problems.

Despite the analysis above, this paper still has several limitations. This review mainly focuses on some representative solutions, while others like asynchronous FL, client selection and etc. are not elaborated. And the implementation mechanisms and algorithms of some technologies in FL are not discussed in depth. Moreover, the heterogeneity-communication-security trilemma is mainly analyzed qualitatively, while their joint optimization remains an open issue.

Acknowledgment

The author would like to express sincere gratitude to the research supervisor for valuable guidance and constructive suggestions throughout this work. The author also thanks the academic advisor for continuous support and encouragement, as well as the professor for the insightful course instruction and academic guidance that contributed to the development of this study.

References

[1] Zhao, W. X., Zhou, K., Li, J., et al. (2026). A survey of large language models. *Frontiers of Computer Science, 20*(12), Article 2012627. Cited page: 1-4. https://doi.org/10.1007/s11704-026-60308-3

[2] Zeng, H., Xiong, S., Di, Y., & Shi, H. (2025). Federated parameter-efficient fine-tuning technology for large model based on pruning. *Journal of Computer Applications, 45*(3), 715-724. Cited page: 715-717. https://doi.org/10.11772/j.issn.1001-9081.2024030322 (in Chinese).

[3] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol., 10*(2), Article 12. Cited page: 2-4. https://doi.org/10.1145/3298981

[4] Zhang, Y. (2022). *Design of a distributed algorithm framework based on federated learning* [M.S. thesis, Zhejiang University, Hangzhou, China]. Cited page: 14-15. https://doi.org/10.27461/d.cnki.gzjdx.2022.003750 (in Chinese).

[5] Qi, Z. (2025). *Research on optimization of a distributed federated learning model aggregation mechanism for the Internet of Things* [M.S. thesis, Shanxi University, Taiyuan, China]. Cited page: 2-11. https://doi.org/10.27284/d.cnki.gsxiu.2025.001066 (in Chinese).

[6] Shen, T. (2025). Research on federated learning methods for distributed heterogeneous environments [M.S. thesis, Zhejiang University, Hangzhou, China]. Cited page: 14-106. https://doi.org/10.27461/d.cnki.gzjdx.2025.001182 (in

DOI: 10.54254/2755-2721/2026.LD37467

Chinese).

[7] Rothchild, D., et al. (2020). FetchSGD: Communication-efficient federated learning with sketching. In *Proceedings of the 37th International Conference on Machine Learning (ICML), 119*, 8253-8265. Cited page: 1-6.

[8] Jiang, H. (2023). Research on privacy-preserving federated learning based on homomorphic encryption [M.S. thesis, Nanjing University of Science and Technology, Nanjing, China]. Cited page: 3-22. https://doi.org/10.27241/d.cnki.gnjgu.2023.002334 (in Chinese).

[9] Yi, X., Paulet, R., & Bertino, E. (2014). *Homomorphic encryption and applications*. Springer, Cham, Switzerland. Cited page: 27-28. https://doi.org/10.1007/978-3-319-12229-8

[10] Shi, Z., et al. (2025). Privacy-enhancing paradigms within federated multi-agent systems. arXiv: 2503.08175. Cited page: 1-7. https://arxiv.org/abs/2503.08175