ICJK
Journal of Reliable and Secure Computing
http://dx.doi.org/10.62762/JRSC.2026.942513

REVIEW ARTICLE

A Comprehensive Survey on Robustness and Privacy in Federated Learning Meets Large Language Model at Edge

Deepak Adhikari[1], **Inam Ullah**[2], **Mustafa Khadim**[3], **Negalign Wake Hundera**[1], **Rajab Ssemwogerere**[1], **Lemessa Bona Debela**[1], **Wei Jiang**[1] and **Hu Xiong**[1,*]

[1] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[2] Department of Computer Engineering, Gachon University, Seongnam 13120, Republic of Korea
[3] Xiamen University, Xiamen 361005, China

Abstract

Large Language Models (LLMs) have revolutionized natural language processing, yet their deployment is hindered by data, computation, and privacy constraints. Federated Learning (FL) offers a promising solution by enabling collaborative, privacy-preserving training across distributed devices, while the push for low-latency on-device intelligence further drives LLM integration into FL and edge settings—posing new challenges in heterogeneity and resource limits. This survey comprehensively reviews the integration of LLMs with federated learning, termed FLM, and its deployment at the edge, with particular emphasis on the robustness, privacy, and trustworthiness challenges that emerge across the LLM lifecycle from pre-training to deployment. We analyze core challenges including communication cost, system heterogeneity, privacy risks, and scalability, with a focus on edge-oriented efficiency techniques such as pruning and quantization. Security vulnerabilities and defenses are also discussed, alongside trade-offs among privacy, robustness, and performance. We further examine demographic, contribution-related, and performance-related biases that can emerge in FLM systems. Finally, we outline open research directions, underscoring the potential of federated and edge intelligence to enable scalable, privacy-aware LLM ecosystems, and aim to offer a unified perspective to guide future research in this fast-moving field.

Keywords: federated learning, large language model, privacy preserving, edge computing, robustness.

1 Introduction

Artificial Intelligence of Things (AIoT) systems amalgamate Artificial Intelligence (AI) with IoT to generate and accumulate a tremendous volume of time series data and enable autonomic decision-making in various domains, including healthcare monitoring, intelligent transportation, smart cities, and industrial automation, where robust aggregation against

Submitted: 10 April 2026
Accepted: 08 June 2026
Published: 17 June 2026

Vol. 2, **No.** 2, 2026.
10.62762/JRSC.2026.942513

***Corresponding author:**
✉ Hu Xiong
xionghu.uestc@gmail.com

Citation
Adhikari, D., Ullah, I., Khadim, M., Hundera, N. W., Ssemwogerere, R., Debela, L. B., Jiang, W., & Xiong, H. (2026). A Comprehensive Survey on Robustness and Privacy in Federated Learning Meets Large Language Model at Edge. *Journal of Reliable and Secure Computing*, 2(2), 111–155.

Byzantine clients is essential for reliable deployment [1–3]. However, AIoT environments are featured by dynamic network conditions, heterogeneous resource-constrained devices, and distributed data sources, posing significant challenges for traditional AI frameworks that range from cyberthreats to a deluge of vulnerabilities, including data poisoning attacks, malware injection, and inefficient task allocation among heterogeneous devices, augmenting severe concerns about data privacy, quality, and security [4–6]. Computational and communication limitations also impede real-time decision-making.

Federated Learning (FL) is an emerging paradigm that addresses the challenges of AIoT, especially related to resource constraints, privacy, and distributed environments [7, 8]. In FL, models are trained locally without sharing raw data across distributed devices, but share parameters of the data to maintain privacy. The approach consists of a central server where the aggregator initializes with learning parameters in a global model. Once the global model is trained, the client is updated with the latest model. The client applies its own data to update the model and transmits the updated model to the aggregator. In order to refine the global model, the aggregator combines the updates from all clients. This strategy helps to save the computational resources of the centralized server by harnessing the computing power of heterogeneous distributed clients. Also, privacy is maintained as the data does not move, as it remains in the local server; the cost of data transmission is also reduced.

However, most FL approaches are based on the homogeneous model architecture, i.e., independent and identically distributed (IID) data and identical model architecture. FL architectures have been used in recent work to investigate distributed learning under non-IID settings across heterogeneous network infrastructures, including 6G systems and smart grids, where data heterogeneity compounds existing security vulnerabilities [9–11]. These methods protect data privacy while enabling many devices to work together to create predictive models. In real-world AIoT deployment, such assumptions are rarely satisfied where heterogeneous devices have different features and computational resources operating. Recent advances on FL are attempting to address challenges, such as statistical heterogeneity, communication efficiency, and privacy preservation [12, 13]. To enhance learning stability in heterogeneous data distribution, FedLAW [12] introduced a client selection approach and adaptive aggregation techniques. Recently, Large Language Models (LLMs) have garnered considerable interest from industry and academia due to their powerful capabilities. Various models of LLM, including LLaMa [14], PaLM [15], BERT [16], Gemini [17], and GPT series [18], have attracted huge interest due to their excellent performance in a wide variety of AI tasks, such as knowledge extraction, contextual reasoning and natural language understanding, text generation, and handling unstructured threat intelligence [19–21]. These AI tasks have generated a way for numerous applications in various domains, including code generation [22], legal consultation [23], chatbots [24], language processing [25] (e.g., multi-language translation, service support, and search engines), and finance [26]. LLM requires tremendous computational resources for a deluge of high-quality data, incurring substantial cost for the utilization and training of LLMs. The high amount of training data is fulfilled through extensive public datasets, and to augment performance in certain areas, there is a need to incorporate data from the private sector, including banks, hospitals, and supply chain management. On doing this, the major issues lie in privacy, which hinders further enhancement if not well-addressed.

Having these peculiar benefits, the integration of FL with LLM, termed FLM, solves the challenges faced by each other, including synthetic data generation spanning from pre-training to deployment. These models can analyze large volumes of threat intelligence feeds, cybersecurity reports, and network logs to identify patterns and infer potential attack strategies. Integrating LLMs with FL systems offers a promising direction for developing intelligent cyber-defense mechanisms by preserving privacy. This shows the integration as a trending and prominent topic [27, 28]. The integration supports data privacy, the huge and distributed nature of LLM training data, continuous performance improvement with updating data, high computational demand for LLM training, and model personalization and adaptation. However, the issues also arise in communication and computational resources, synchronization and coordination, heterogeneity, and privacy and security.

1.1 Methodology

This section summarizes the methodology adopted to accumulate and analyze state-of-the-art research articles to discuss the FLM (integration of FL and LLM) in the edge environment. The survey focuses on key aspects, including fundamentals, system architectures,

robustness, privacy and security mechanisms, and applications associated with it.

Research Scope: The main objective of this article is to analyze and review the existing literature on FLMs in edge computing environments. The primary objective of this survey is to provide comprehensive insights and address the following research questions:

- RQ1. What are the fundamental concepts of LLM, FL, and edge computing? How is the integration possible?
- RQ2. What are the motivations and key challenges associated with the integration of FL and LLM at the edge?
- RQ3. How can robustness and privacy be achieved in FLM, and what forms of bias arise in such systems?
- RQ4. What are the key applications and future potential directions of FLM?

In order to answer the above-mentioned research questions, we collect literature from multiple databases such as IEEE, Google Scholar, Elsevier, ACM, Springer, and arXiv. The literature search was conducted from publications up to March 2026, ensuring coverage of the most recent developments in FL, LLM, and EI. A deluge of literature was initially retrieved using carefully selected search strings and keywords, including, “foundation model," “large language model" ("LLM"), “federated learning," “federated AI," “distributed learning," "edge computing," “edge learning," “on-device intelligence," "federated LLM," “LLM at edge," “privacy-preserving LLM,” “security in federated learning," and “privacy in LLMs." From an extensive pool of approximately 520 research articles, studies were filtered based on relevance to federated LLMs and edge deployment scenarios. After an initial screening of titles and abstracts, 380 articles focusing on core architectures, methodologies, and applications were selected. Subsequently, 168 articles were excluded due to redundancy, limited technical contribution, repeated applications, or insufficient extension of the technique in other domains. Finally, 212 articles were retained as the core literature for in-depth analysis, ensuring alignment with the scope and objectives of this survey. In addition to this core set, the manuscript also cites supplementary foundational and background references, including original works on LLM architectures, benchmark datasets, and domain-specific concepts that fall outside the initial search strings but are necessary for completeness. Additional supporting references, including foundational works on LLM architectures, datasets, and domain-specific background concepts not captured by the initial search strings, are also cited throughout the manuscript where relevant.

1.2 Related Review and Contributions

FL, LLM, and EC have been the subject matter of various surveys and review articles, as well as books. A broad review presents on the topic independently, for instance, FL [9], federated foundation models [29], and edge computing [30]. Some of them present on the integration of two topics, such as the integration of FL and edge [31], with underlying multiple access technologies shaping distributed and edge intelligence systems [32], FL and LLM [34], prompt engineering challenges for LLM privacy [33], and the integration of edge and LLM [35–37].

Khan et al. [38] performed a comprehensive systematic classification of FL in IoT networks. In order to evaluate the recent research progress on FL in IoT applications, they discussed a set of performance indicators, including quantization, robustness, sparsification, scalability, security, and privacy. The challenges of privacy and robustness in FL, along with the threats and protective measures, are discussed in depth in [39]. Zhang et al. [11] explored the possible vulnerabilities during the deployment of FL based on generation, transmission, distribution, and consumption on smart grid systems and highlighted the future research directions. A survey of distributed learning focused on IoT applications is explored in [40]. The paper discusses the essence of distributed learning, including FL, distributed inference, and multi-agent reinforcement learning, for critical IoT services (e.g., mobile crowdsensing, data sharing, localization, computation offloading, and security and privacy) in various IoT applications. Survey papers presenting fundamental principles and taxonomy of FL have been studied in [7] and from the perspective of privacy and security in [10, 41]. A range of IoT applications and services for FL is investigated in [42–45].

There has been extensive work on LLM, and there are several surveys that discuss the latest developments in depth. The survey by Ren et al. [29] discusses the latest development with future prospects and existing challenges on diverse data types for pretrained foundation models. The concerns about the security and privacy issues in LLMs have been discussed in [46–48], which highlights risks on both training data

and application-based aspects, such as attacks on models, vulnerabilities inherent to LLMs in different domains, and potential defense approaches. [49] discusses the approach of protecting and securing the intellectual property of LLM models. In [50], discusses the collection of data, training strategies, and model evaluation in LLMs. Similarly, a conceptual architecture for LLM agents consisting of three parts, brain, perception, and action, is discussed in detail with justifications on various applications [51]. [52] highlights the popular concerns of LLMs as infallible tools through the exploration of biases in training data and their impact on heightened security risks, opaque code interpretation, code generation, and the trustworthiness of LLM-generated software. In [53] review the latest developments in model compression, such as compression methods including low-rank approximation, knowledge distillation, quantization, and parameter pruning; and tuning approaches including inference optimization and parameter-efficient fine-tuning. Apart from this, there are several works that discuss LLMs and their security approaches [54–56].

The extensive resource requirements of LLMs also pose obstacles. The expense of training on cloud servers with powerful GPU clusters and the delay of cloud-based inference have prompted a paradigm shift. Industrial dynamics, original equipment manufacturer considerations, and LLM-specific factors tend to drive the migration of LLM inference towards the edge [57]. Latency and reduced dependency on the connection also play an important role in the migration. Unlike its cloud-based counterparts, edge-deployed LLMs can function with minimal network access. Locally conducted inference can drastically reduce reaction times, providing a much better user experience than depending on the dependability and speed of a network connection. Additionally, it lowers the risk of data breaches and allows customers more control over their personal information by eliminating the need to transmit sensitive data over networks. Edge migration also seems to be motivated by customization, which affects both training and inference [58]. An LLM has the potential to comprehend multiple aspects of users, e.g., speech patterns, writing style, and more. In addition to improving privacy, user devices can modify models to fit particular behaviors and personalities, creating a customized user experience. Another important factor is scalability, since the increasing number of edge devices makes it possible for applications to be widely distributed across a wide range of devices without burdening centralized servers. Various solutions in the domain of EL and FL have been proposed as a result of the extensive investigation of parallel, distributed, and FL in recent years. These solutions are designed to train, optimize, or make the deployment of LLMs easier.

The integration of LLM and FL in the EI framework possesses a range of security and deployment challenges [46, 59]. For example, the memory and computational requirements for LLMs are high and become difficult for resource-constrained edge devices. This discrepancy impacts the performance of FL and LLM and exposes the security vulnerabilities of EI systems [48]. FLM can be amalgamated into EI frameworks through strategies including efficient memory management and model compression [35, 60]. Techniques such as pruning and quantization are more suitable for deployment in EI due to their ability to reduce model size [53, 61]. In [61], the FL, EL, and LLM are integrated, and a paradigm is discussed to accommodate LLM that ensures privacy preservation through collaborative model training. And highlights the challenges faced during the deployment. A systematic overview of the deployment of in-context learning (ICL) and prompting by privacy-protection techniques is explored in [33]. Advanced attention mechanisms and leveraging hierarchical memory stems help in managing resource-constrained issues more effectively.

Few studies have carried out thorough analysis specifically focused on the system architecture of FLMs, classification of current FLM works, and FLM applications in various scenarios, despite the fact that LLMs and FL have been extensively studied in isolation within the existing literature [61, 62]. Given these revolutionary developments, our survey is essential for thoroughly analyzing the state of integrating federated and edge learning in the field of LLMs, as well as its potential future developments. The survey intends to explore the subtleties of how FL and EL are influencing the development of LLMs, examining their effects on user experience, scalability, and privacy. This survey aims to offer a roadmap for the continued integration of federated and EL in the field of LLMs by combining information from research breakthroughs, industry advances, and emerging trends. The major contributions of the paper are as follows:

- We attempt to provide a comprehensive overview of the integration of FL, LLMs and EL, focusing particularly on the emerging FLMs paradigm and

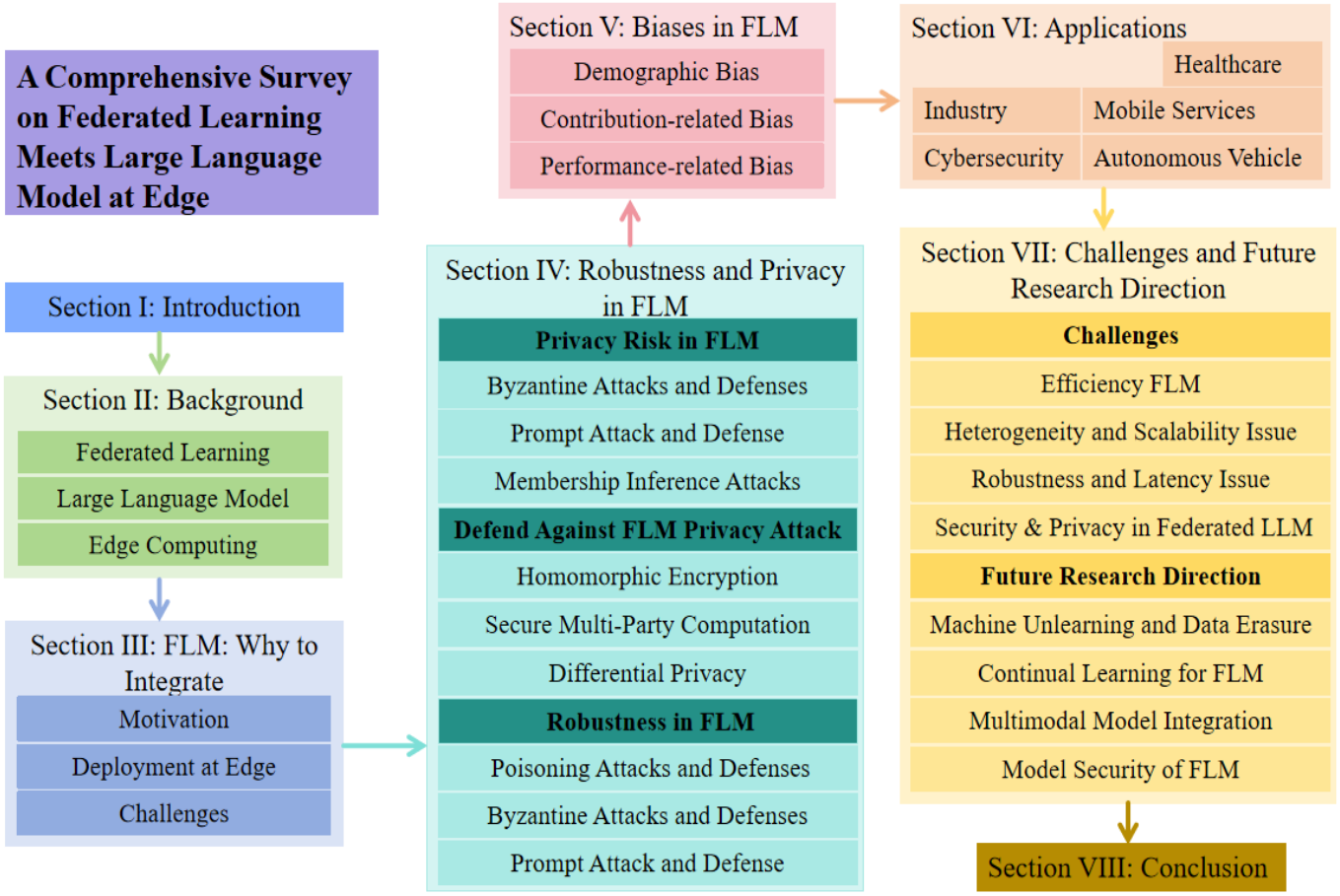

Figure 1. A summary of the structure, flow of information, and outline of the paper highlighting the FLM in the edge environment.

its architectural design principles.

- We present a systematic classification and taxonomy of existing FLM frameworks, highlighting current research trends, communication mechanisms, deployment strategies, privacy-preserving techniques, and optimization approaches adopted in FLM systems, while also analyzing sources of demographic, contribution-related, and performance-related bias in such systems.

- We extensively investigate the applications of FLMs across diverse domains, including healthcare, smart cities, intelligent transportation, industrial IoT, and autonomous systems, while analyzing their impacts on personalization, latency reduction, scalability, and user privacy.

- We provide a road map for future developments in FLM by identifying critical research challenges, open issues, and future research directions in FLM systems.

1.3 Structure of the Survey

The structure of this survey is organized as illustrated in Figure 1. Section 2 provides an overview of Federated Learning (FL) and Large Language Models (LLMs), covering their fundamental concepts, architectures, mechanisms, and recent advancements. Section 3 discusses the motivations for integrating LLMs with FL in edge environments, reviews deployment strategies and available benchmarks and libraries, and examines the associated challenges. Section 4 reviews the latest developments in the robustness and privacy of Federated Large Models (FLMs), while also investigating potential security and privacy threats and corresponding countermeasures. Section 5 focuses on bias issues in FLMs. Section 6 presents various application scenarios in which FLMs can be deployed. Section 7 discusses remaining challenges and explores promising future research directions for enhancing FLM performance. Finally, Section 8 concludes the survey by summarizing the key insights and findings.

2 Background

2.1 Federated Learning

FL is a distributed technique to ML where the model maintains privacy without disclosing the personal information and autonomously trains on datasets and updates a shared global model at a central server. By providing the model updates on a regular basis, devices can contribute to the global model's improvement and have access to a comprehensive model. Various research studies have shown that FL got close attention as it revolutionized numerous intelligent IoT devices. The distributed architecture and privacy-preserving feature leverage the innovative AI solutions.

Definition: A standard FL framework consists of N clients as $\{C_1, C_2, ..., C_N\}$, where each client C_N have a private dataset $D_n = \{x_i^n, y_i^{n\,K_n}{}_{i=1}\}$ with $|x^n| = K_n$ and $K = \sum_{n=1}^{N} K_n$. C_n possesses an initialized model or a learned local network denoted by $f(\theta_n)$. The output of the x_n based on θ_k is the predicted results represented by $f(x_n, \theta_n)$. The centralized dataset in the ML paradigm is denoted by $D_{central} = D_1 \cup D_2 ... \cup D_n$, which combines the datasets and trains a model by $\theta_{central}$ for better performance. The centralized models were unable to address the privacy concerns, making them impractical for real-world scenarios where privacy is paramount. As a solution, FL trains each client C_n without revealing private information D_n to other clients $C_{n0}(n \neq n_0)$.

Initialization: Each client C_n holds a local dataset $D_n = \{(x_i^n, y_i^n)\}_{i=1}^{K_n}$, where $x_i^n \in \mathbb{R}^d$ denotes the d-dimensional feature vector of the i-th sample and $y_i^n \in \mathcal{Y}$ denotes its corresponding label. For supervised learning, each sample is associated with a label y_i^n; for unsupervised learning, only the feature vectors x_i^n are available without labels.

The local model $f(\cdot\,; \theta_n)$ is parameterized by θ_n, and the predicted output for sample x_i^n is $\hat{y}_i^n = f(x_i^n;\, \theta_n)$. Prior to training, the server initializes the global model parameters θ^0_{global} and distributes them to all selected clients.

1. Task Initialization: The training task and the associated data are determined by the server. Similarly, training procedures, such as communication rounds, learning rate, and hyperparameters of the global model, are also specified by the server. Subsequently, the server broadcast the task for the selected participant $\{C_1, C_2, ..., C_K\}$ and the initialized global model θ_{global}^{t-1} to the clients involved $\{\theta_1^t, \theta_2^t, ..., \theta_N^t\}$.

   Local Model Training and Update: Each participant C_n updates the local model parameters using their device and local data, based on the global model, where t represents the current iteration index. The local empirical loss for client n is defined as

$$L_n(\theta_n) = \frac{1}{K_n} \sum_{i=1}^{K_n} \ell\left(f(x_i^n; \theta_n),\, y_i^n\right) \tag{1}$$

   where $\ell(\cdot, \cdot)$ denotes the per-sample loss function (e.g., cross-entropy or mean squared error). The local model parameters are then updated via gradient descent:

$$\theta_n^t \leftarrow \theta_n^{t-1} - \alpha \nabla_\theta L_n\left(\theta_n^{t-1}\right) \tag{2}$$

   where α represents the learning rate. In iteration t, the client n seeks to determine the optimal parameters to minimize the local loss function $L_n(\theta_n)$. Then, the server receives the updated local parameter θ_n^t.

2. Global model aggregation and updates: Then, the server aggregates the received local model, minimizes the global loss function, and then updates the participant with the updated global model parameter θ_{global}^t.

$$\theta_{global}^t \leftarrow \frac{1}{\sum_{n=1}^{N} |D_n|} \sum_{n=1}^{N} |D_n| \theta_n^t \tag{3}$$

   Steps 2-3 are repeated till the desirable training accuracy is achieved or the global loss function converges and updates the local model with the latest global model update θ_{global}^t.

Recent research shows that [7, 9, 38, 43, 306–308] the advancement in FL. The research classifies FL into two broad categories. a) based on network architecture, and b) based on data distribution.

2.1.1 Network Architecture:

Based on network architecture, FL is categorized as Decentralized Federated Learning (DFL) and Centralized Federated Learning (CFL). The most popular framework in FL is CFL, which entails a central server coordinating with several clients. Clients train their models separately using local data during each training cycle, after which they send the modified parameters to the central server, which is termed the global model. An example includes Federated Averaging (FedAvg) [310]. The

major challenges that exist on the CFL include server bottlenecks, dependency on trust, and single-point failures. Conversely, DFL offers a serverless FL paradigm as opposed to the CFL system. This strategy emphasizes the advantages of using a peer-to-peer delivery and aggregation mechanism that does not rely on a central server, which is considered a reliable server. Participants in a DFL system can fully utilize the network capacity by using the network connections among themselves, as opposed to only communicating with a central server. Peer-to-peer communication makes it possible for the system to be more dispersed and possibly more resilient. Because of these contemporary features, DFL can be used to create decentralized FL networks that are compatible with peer-to-peer communication technologies like blockchain [306, 311].

2.1.2 Data Distribution:

Based on the data distribution, FL is categorized into three types as Vertical Federated Learning (VFL) [312, 313], Horizontal Federated Learning (HFL) [314, 315], and Transfer Federated Learning (TFL) [316, 317].

The most common method used in FL is HFL, where cross-device scenarios are frequently associated with. It includes the federation of samples and works best when there is less node overlap and a high degree of feature overlap. Each client in the HFL framework creates their own AI model on their own, leading to local updates. These local updates are concealed using techniques like the use of differential privacy or encryption to guarantee security and privacy. On receiving the local updates from clients, the central server compiles them to formulate a global update and disseminates it back to the client, which facilitates the next round of local training. This process gets repeated until the predefined accuracy level is achieved or the loss function of the model stabilizes.

On the other hand, VFL and TFL are complex in nature because of their approaches to data architecture, making integration and implementation more challenging in various applications. VFL has disparate feature sets but shares a common set of samples, making use of collaborative learning of a shared AI model. The system uses entity alignment to integrate samples from several clients and train a single AI model, and encryption protocols are used to enhance the security. TFL is the preferred approach when samples and features across the nodes barely overlap. The process involves transforming heterogeneous feature spaces into a common format, which allows for model training using data from several clients. The process involves transforming heterogeneous feature spaces into a common format, which allows for model training using data from several clients. TFL aims to create customized models for specific use cases, especially with restricted data availability, an important aspect of data organization in FL strategies.

2.2 Large Language Models

LLMs are advanced language models characterized by their exceptional learning capabilities and massive parameter size. The core of the model includes a self-attention mechanism introduced in the Transformer architecture [63], which serves as a fundamental component for effective language modeling. Training LLMs typically requires large-scale, diverse datasets and substantial computational resources. Following pre-training, these models are further refined through fine-tuning to satisfy specific downstream requirements, such as enhanced efficiency, performance, and privacy considerations. This section highlights a comprehensive overview of the background of LLMs in relation to these characteristics.

Definition:

The computational frameworks designed to understand and generate human language are Language Models (LMs), whose growing capabilities have raised concerns ranging from embedded biases [64] to broader societal implications across application domains such as education [65]. Fundamentally, these are autoregressive and generative models that understand and predict text by calculating the probability distribution of the next word based on the preceding sequence, as shown:

$$P(w) = P(w_1) \cdot P(w_2|w_1) \cdot \cdots \cdot P(w_T|w_1, \ldots, w_{T-1}) \tag{4}$$

where text sequence is represented by $\{w_1, w_2, ..., w_N\}$, token as $T = |w|$ and current positions by t. With $t = 1, ..., T$, $P(w_t|w_1, ..., w_{t-1})$ is the likelihood that the LM will produce the token w_t given the preceding $t-1$ tokens.

2.2.1 Transformer-based architecture

Most contemporary LLMs are built upon the Transformer architecture, which has fundamentally transformed natural language processing (NLP). Transformers have demonstrated remarkable performance across a wide range of NLP tasks, including question answering, machine translation,

and text classification. For instance, models such as BERT have achieved state-of-the-art output in question-answering benchmarks by effectively modeling contextual dependencies. Beyond NLP, Transformer architectures have also gained widespread adoption in computer vision, where they are applied to tasks including semantic segmentation, object detection, and image recognition. A notable example is the Vision Transformer (ViT), which divides images into patches and processes them using Transformer encoders, achieving superior performance compared to traditional convolutional neural networks (CNNs).

Unlike recurrent neural networks (RNNs), which rely on sequential processing and struggle with long-term dependencies, Transformers employ self-attention mechanisms to capture global relationships within input sequences. The core design follows an encoder–decoder structure composed of multiple stacked layers, each integrating multi-head self-attention and feed-forward networks (FFNs). This design allows the model to emphasize different parts of the input sequence, thereby improving representation learning and output generation. The attention mechanism is represented as:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (5)$$

where Q, K, V, and d_k denote Query, Key, Value vectors, and the dimension of the key, respectively.

The operation of a transformer generally proceeds as follows: Input data, such as text, is first segmented into tokens using tokenization methods like WordPiece or byte-pair encoding. These tokens are then mapped into vector embeddings, augmented with positional encodings to preserve sequential information. Self-attention mechanisms compute relationships among tokens using query, key, and value representations derived from linear projections of the embeddings. The outputs are subsequently processed through layer normalization and FFNs. By stacking multiple such layers, the encoder produces rich contextual representations, which the decoder then uses along with previously generated tokens to generate outputs. Finally, a linear projection layer produces the model's predictions. During inference, Transformers commonly adopt autoregressive decoding, where each generated token is appended to the input for subsequent prediction steps. To improve efficiency, key–value (KV) caching is often used to store intermediate attention states, thereby avoiding redundant computations.

Self-attention is the central component of the Transformer architecture. It overcomes the limited context modeling capability of RNNs by enabling direct interactions among all elements of a sequence, regardless of distance. While attention mechanisms have been explored in earlier neural architectures, Transformers uniquely rely entirely on attention and introduce multi-head attention to enable parallel computation and scalability. Compared to alternatives such as hard attention, which requires stochastic sampling or convolutional and recurrent structures that depend on strong inductive biases, Transformers require minimal prior knowledge about task structure. This flexibility makes them particularly well-suited for large-scale pre-training on unlabeled data, enabling the learning of highly expressive and transferable representations that can later be fine-tuned for downstream tasks.

2.2.2 *Unimodal LLMs*

LLMs typically refer to Transformer-based language models with billions of parameters that are pre-trained on massive text corpora. These models exhibit strong capabilities in language understanding, reasoning, and generation, allowing them to perform complex tasks and produce coherent, context-aware outputs. Representative examples include Meta's LLaMA and OpenAI's GPT-4. Major technology companies have actively developed their own LLMs and applied them across a variety of domains. For instance, GPT-3 has demonstrated impressive performance in tasks including translation and text generation, while Med-PaLM has been designed to provide expert-level responses in medical contexts. In addition, Transformer-based models such as DEiT have been successfully applied to image classification by combining self-supervised learning with limited labeled data.

Based on architectural design, LLMs can be broadly categorized into three types: decoder-only, encoder-only, and encoder–decoder models. Examples of encoder-only LLMs include BERT and ALBERT, which focus on generating contextualized token representations and perform well in tasks such as text classification and semantic similarity analysis. Encoder–decoder LLMs, such as T5, combine both components to support sequence-to-sequence tasks, making them well-suited for summarization, machine translation, and question answering.

Decoder-only LLMs, exemplified by the GPT family, rely exclusively on autoregressive decoding to generate text sequentially. This design is particularly effective for tasks involving open-ended text generation, dialogue systems, and content completion.

2.2.3 *Multimodal LLMs*

Traditional LLMs are primarily designed to process textual data, which limits their ability to understand information from other modalities. In real-world scenarios, however, language understanding often relies on additional signals such as images, audio, and sensor data. To address this limitation, both academia and industry have increasingly focused on multimodal LLMs, which integrate multiple data modalities into a unified framework.

Models such as GPT-4 are capable of jointly processing text and visual inputs, achieving strong performance across multimodal benchmarks. In tasks like image captioning, such models leverage both visual and textual cues to generate more accurate descriptions, while in speech-related tasks, they combine audio and text to improve comprehension. Multimodal perception is considered a key step toward general artificial intelligence, as it enables AI systems to interpret complex real-world environments through cross-modal reasoning.

Multimodal LLMs typically combine the language modeling strengths of LLMs—such as in-context learning and zero-shot generalization—with foundation models trained on other modalities. Since these modality-specific models are often pre-trained independently, a major challenge lies in effectively connecting them. Current research largely focuses on multimodal pre-training, which learns shared representations across modalities using multimodal datasets, and multimodal instruction tuning, which further aligns models with specific cross-modal tasks using labeled data. These techniques enhance alignment among modalities, improve generalization to unseen tasks, and strengthen zero-shot performance.

2.2.4 *Generative and interactive AI applications*

The rapid advancement of LLMs has significantly influenced both interactive AI (IAI) and generative AI (GAI). GAI aims to produce diverse content, including audio, text, images, and video—commonly referred to as AI-generated content (AIGC). By leveraging multimodal LLMs trained on high-quality datasets, GAI systems can generate rich and contextually relevant content from user inputs. IAI represents a further evolution, enabling AI systems such as chatbots and virtual assistants to engage in dynamic interactions with users and continuously adapt based on feedback. LLMs play a central role in IAI by supporting complex, human-like conversations and reasoning processes.

To further enhance the accuracy and freshness of responses, retrieval-augmented generation (RAG) has been integrated into LLM pipelines. RAG allows models to retrieve relevant information from external knowledge bases during response generation, thereby improving factual correctness and reducing hallucinations. Additionally, RAG enhances transparency by enabling users to trace responses back to their data sources.

2.3 Edge Computing

Recent years have seen tremendous developments in the computing and AI industry due to rapid developments in hardware miniaturization, network connectivity, computational performance, and affordability. These developments have greatly augmented the widespread usage of computing devices crucial for efficient data sharing and profound impact on modern society. Cloud computing can offer scalable processing power, abundant storage, and a variety of services through remote data centers, as it has become essential in AI industry. For various infrastructures such as healthcare, intelligent transportation, and IoT data, such cloud architecture facilitates effective cost reduction, data management, and easy access to computational resources. However, cloud computing possesses numerous drawbacks, including higher latency, potential security threats, dependence on network connectivity, and significant energy consumption, particularly for real-time applications and resource-constrained applications [309].

Edge computing has become a crucial substitute to mitigate the limitations by relocating computational capabilities closer to the client data source [66]. Edge computing utilizes user devices, local gateways, and servers to process data on the network periphery, which helps to lower the latency, facilitates quick decision-making, and protects privacy. This approach is important in applications such as industrial automation, autonomous vehicles, and healthcare, where data protection and timeliness are critical. Deployment of edge AI, which employs machine learning models directly at the edge, makes

edge devices more powerful and can function autonomously with reduced dependence on cloud-based infrastructure [67]. In this paper, Edge Computing (EC) refers to the underlying infrastructure paradigm of relocating computation toward data sources, while Edge Learning (EL) refers specifically to the training and inference of machine learning models within this paradigm.

3 FLM: Motivation, Deployment, and Challenges

This section highlights the motivation and challenges of integrating FL with LLM, especially when deployed in resource-constrained edge contexts. Current LLM approaches face various challenges in dealing with the distributed paradigm, including communication overhead and model complexity. LLMs are well known for having a large number of parameters, and FL for high communication costs during parameter synchronization and model updates. Both these approaches have limitations on heterogeneity, which impacts model performance. Deployment of FLM at edge environment introduces a critical layer in this ecosystem that enables real-time decision-making, enhances privacy protection and system scalability, and minimizes communication delays. However, it suffers from energy, storage capacity, strict hardware, and bandwidth limitations. Incorporating the edge is crucial for FLM as it allows for on-device processing that reduces latency, data leakage, and shifts the intelligence paradigm closer to the end-user. Here, we will investigate how these two approaches can be integrated in an edge environment for better performance and convergence in real-time scenarios.

3.1 Motivation

The convergence of LLM and FL forms a mutual collaborative relationship in which each paradigm compensates for the limitations of the other. Rather than performing independently, FLM can function in a complementary manner that enhances personalization, scalability, efficiency, and privacy preservation. This section elaborates on some of the motivations for integrating LM and LLM in an edge environment.

3.1.1 Enhanced data availability

LLM requires tremendous volumes of high-quality training data to perform well. However, valuable datasets are collected from heterogeneous distributed devices that range from personal to institutional devices from various devices, such as finance, healthcare, and transportation. Such collected data is centralized, causing privacy risks, which is impractical due to legal, regulatory restrictions, and privacy concerns. FL enables decentralization training across distributed data without sharing raw data. This approach allows the model to take advantage of heterogeneous data sources by preserving privacy, enhancing model accuracy, and model adaptation. This addresses the challenges of data distribution and privacy in various domains, where FL integrated with LLM can access a broad range of data for optimization and pre-training tasks.

3.1.2 Essence of High Computational Data and Adaptation

LLMs possess an enormous amount of model parameters, and large-scale LLM training requires a significant amount of computational resources. FL enables a collective training strategy to eradicate the barrier posed by LLM by allowing entities to entities to amalgamate computational abilities that decentralize the training task and lessen the strain on a single entity. Similarly, FL can modify the original model based on the node's computing capacity through the heterogeneous computing resources and allow clients with low processing power to participate in the LLM training and fine-tuning process. The decentralized nature also supports evolving towards a specific domain, enabling LLM services to be adaptive and personalized through diverse training of the user-generated data. Learning from the diversified data helps to generate knowledge with a better understanding of the intricacies and complexities of real-world circumstances, resulting in more informed and less biased decisions for various tasks and domains, hence contributing to bias reduction in LLM systems.

3.1.3 Regular Improvement through Updated Data

In order to fetch the latest information, LLM must be updated with the latest knowledge. Data evolves rapidly in the real world, making the timeliness and relevance of LLMs challenging, especially with distributed data. FL utilizes the heterogeneous and distributed data source to provide a solution for continuous enhancement and adaptation of models. All these are possible just by providing the updates rather than transferring the data. A recent study suggests integrating data collected from public domains, including the Internet, into the FL process to enhance FL performance. The quality of the public data collected is a major factor in the success of approaches that use it. Methods based on synthetic data for FL have been developed to get

beyond the limitations of public data. LLMs offer strong data distribution fitting skills because they have been pre-trained on a variety of datasets. This makes it possible for them to produce synthetic data that accurately captures the complexity and diversity of real-world data sets.

Furthermore, through knowledge distillation, a technique in LLMs can successfully address the problem of suboptimal performance in FL [81]. In order to expedite the training of a simpler "student" model within the FL framework, the LLM, acting as the "teacher," transmits its knowledge through a process known as knowledge distillation. To improve and simplify the student model, the LLM typically uses knowledge distillation. Each member of the FL network then uses this reduced student model to support their local training initiatives. The smaller model's performance and capacity for generalization are improved by the transfer of insights from the LLM, which solves the problems caused by sparse or unevenly distributed data. This method makes learning within the FL system more effective and efficient [318, 319].

3.1.4 Data Privacy

Training of LLM requires tremendous data, given that the distributed heterogeneous massive data raises serious privacy concerns in LLM training. The FL server does not require raw data for training and exchanges some important information for training, including gradient updates or model weights, which ensures privacy by reducing the risk of exposing sensitive data. The integration acts as a solution for maintaining privacy with enhanced performance.

3.2 Deployment at Edge

The deployment of FLM, the integration of LLMs in conjunction with FL, at the edge has emerged as a promising paradigm for enabling decentralized, low-latency, and privacy-preserving intelligent services. Integration with FL helps users to fine-tune LLMs on local devices, such as smartphones, IoT systems, and embedded platforms, enabling decentralized model training across multiple edge devices without transmitting sensitive raw data, thereby mitigating privacy risk and reducing communication overhead [9, 35]. However, integrating LLMs into such environments poses challenges due to their limited memory, large model sizes, high communication costs, and computational demands. Recent studies have explored lightweight adaptation approaches to address such challenges, including parameter-efficient fine-tuning, quantization, and model pruning. These approaches, for instance, knowledge distillation and quantization, are used to shrink LLMs into small language models (SLMs), which significantly reduce computational and memory requirements that help to achieve high performance and fit with constraints of edge devices [61, 68, 69].

Deployment of SLMs on edge environments requires the appropriate selection of workload-to-device mapping due to large differences in architecture, hardware, and model heterogeneity of edge devices. An experiment on runtime cost and capabilities of SLMs on edge devices demonstrates that the model architecture is an important aspect affecting performance in addition to the quantity of model parameters [70]. Highly parallel processors excel during the prefill phases, whereas quantization primarily accelerates the decoding phase by mitigating the memory access bottlenecks. Similarly, the KV cache takes up most of the memory during long-context inference and the most time-consuming operation in SLMs is General Matrix-Vector Multiplication. SLMs frequently show similar performance on specialized domains or in certain downstream tasks, and lacks powerful generalization capabilities as compared to LLMs. SLMs provide benefits including low latency, low resource consumption, privacy protection, low power consumption, high throughput, and personalization on certain specific tasks. As a result, they have great potential for use in situations involving IoT and edge computing devices. Figure 2 illustrates the architecture of FLM at the edge paradigm and the major challenges associated in deployment.

To address the challenges of resource-constrained and latency-sensitive environments, FLoe, a hybrid FLM framework, is proposed by integrating with lightweight SLMs on edge and cloud-based LLMs [71]. Cloud-based LLM is responsible for general knowledge transfer without compromising proprietary weights. The edge is responsible for performing fine-tuning and keeping sensitive data. A heterogeneity-aware LoRA strategy is deployed for efficient cross-hardware edge deployment and a logit-level fusion mechanism for real-time model alignment between cloud and edge models. SLM-FARL [72], a multi-agent deep reinforcement learning architecture, integrated SML and FL along with federated unlearning to provide autonomous privacy-compliant learning along with the feature of user-level data removal requirements. Quantization

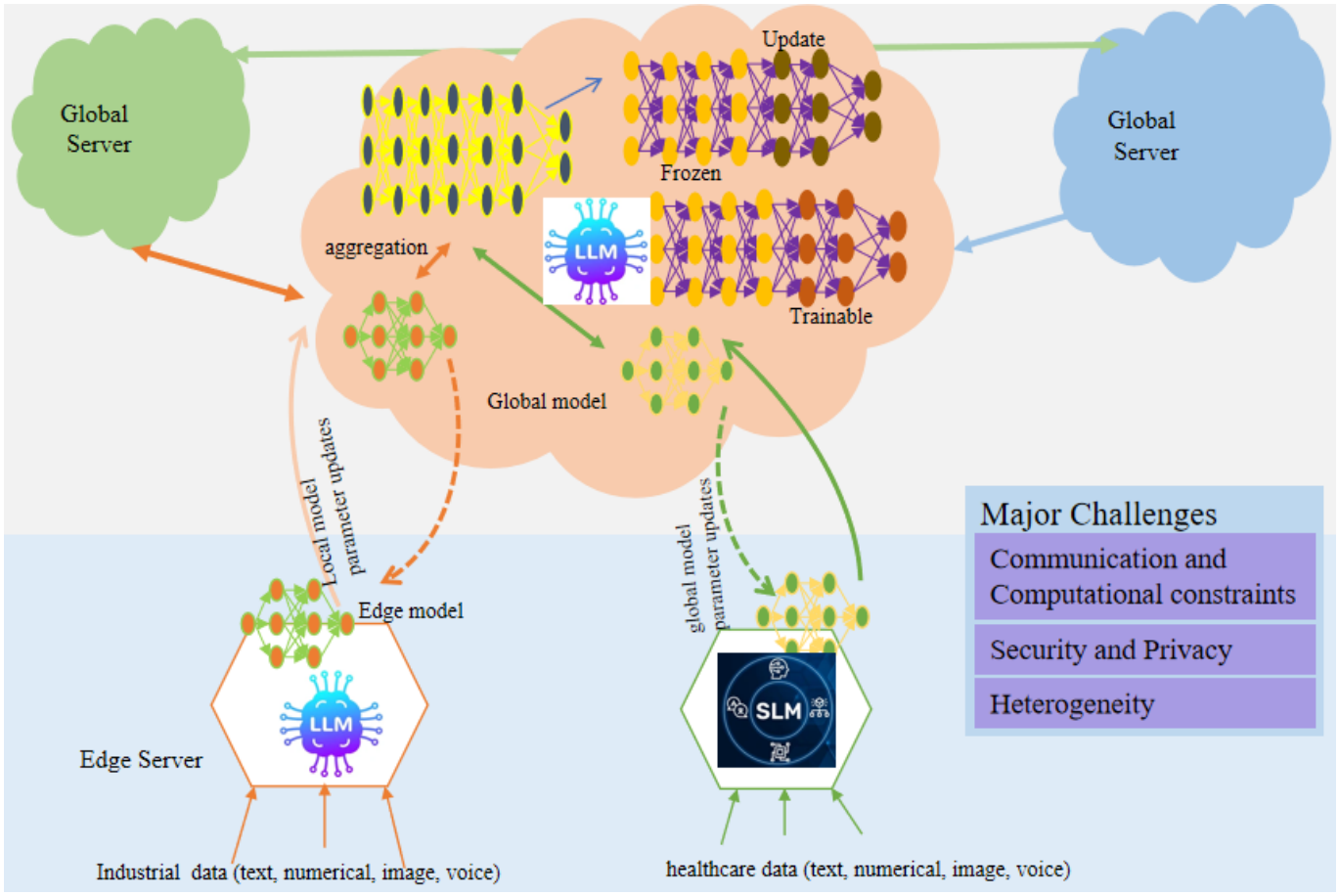

Figure 2. Illustration of FLM, an integration of FL and LLM at the edge and the major challenges faced during the integration.

and knowledge distillation are the approaches used for SLMs optimization for enhancing performance, reducing latency and model size.

Despite these advancements, the practical deployment of FLM at the edge faces several challenges, as shown in Figure 2. System heterogeneity, including variations in data distributions, network conditions, and device capabilities, can negatively impact model convergence and training efficiency. Additionally, the iterative communication process in FL introduces significant bandwidth overhead, particularly for large-scale LLM parameters. Privacy and security concerns, such as data leakage, model inversion attacks, and adversarial updates, further complicate deployment. To mitigate these risks, techniques such as differential privacy, secure aggregation, and robust aggregation methods have been proposed [57, 69]. Recent works also investigate hybrid architectures that combine edge and cloud resources to balance scalability and efficiency [37]. The convergence of FLMs at the edge presents a transformative opportunity for distributed intelligence, but requires continued research to address robustness, efficiency, and scalability challenges in real-world applications [73, 74].

3.3 Challenges

The amalgamation of FL and LLM, known as FLM in an edge environment, offers significant advantages; however, it also inherits and amplifies various inherent challenges associated with both paradigms. Considering this context, we discuss these challenges, focusing on system architecture, security, and privacy concerns.

3.3.1 Communication and computational Constraints

In traditional FL settings, computational cost, communication overhead, and memory use are crucial factors influencing system efficiency. Integrating with LLM substantially increases the issues due to the computational complexities and deluge of parameter sizes. A fundamental aspect of FL is the iterative exchange of model updates between a central server or distributed server and distributed clients [75–77]. When LLMs get involved, there will be a dramatic surge in the size of these updates, resulting in communication burdens causing high latency, especially in environments with unstable

connectivity or limited bandwidth, ultimately degrading system performance. Moreover, training LLMs requires substantial energy and power. However, the computational burden gets distributed across heterogeneous clients, many of which may lack sufficient resources. These disparities in computational demands lead to sub-optimal model performance and inconsistent training progress. Pruning [78, 79], knowledge distillation [80, 81], and quantization [82, 83] are some strategies to reduce memory and computational requirements, maintaining acceptable performance in FLM systems. Similarly, fine-tuning pre-trained models [84, 85] rather than training from scratch can significantly reduce communication and computational costs.

3.3.2 *Security and Privacy Issues*

The amalgamation of FL and LLM introduces significant security and privacy concerns. LLMs are vulnerable to various types of attacks, including backdoor attacks [86, 87], adversarial attacks [88, 89], model poisoning attacks [90, 91], data poisoning attacks [92, 200], model stealing [93, 94], and more. These threats raise concerns about security and robustness, compromising model integrity, trustworthiness, and user data privacy. For example, adversaries can manipulate word embeddings by introducing subtle perturbations, causing the model to produce biased or incorrect outputs when specific trigger inputs are encountered [95, 96]. In FLM settings, fewer malicious clients can enormously impact the global model through poisoning attacks. Furthermore, in distributed or wireless environments, communication channels may be victimized through adversarial interference, potentially corrupting transmitted updates.

3.3.3 *Heterogeneity*

In FL, heterogeneity is a defining characteristic, where data across clients is often non-identically distributed (non-IID) and non-independent, varying significantly in quality, size, and distribution. This data heterogeneity leads to divergence during training of LLMs, and limits the effectiveness of the global model [97]. Apart from data variability, system heterogeneity, including memory differences, hardware capabilities, and network conditions, further complicates the training process [98, 99]. In this situation, some clients may struggle to process LLMs efficiently, resulting in uneven training progress. Personalized federated learning [100–102] provides a promising solution by tailoring models to individual clients rather than enforcing a single global model. Techniques such as multi-task learning [103] and meta-learning [104]108 can further enhance personalization by adapting models to diverse resource constraints and data distributions.

The aforementioned difficulties are by no means comprehensive, since other topics have previously been covered by independent studies on LLMs and FL [105–107]. However, the research that more cohesively combines LLMs and FL is the primary emphasis of this research and also includes biases toward either LLM or FL. This paper will introduce and examine these difficulties in depth, as well as provide an overview of the existing key relevant studies.

3.4 Benchmarks and Libraries

This subsection highlights available benchmarks and libraries for examining and developing the FLM approach and presents Table 1 of some open-source projects and datasets in surveyed papers.

- **FedML [108]:** built on the FedML platform.
- **FederatedScope-LLM [109]:** an open-source FL package that supports fine-tuning of LLM under multiple scenarios, such as model personalization and FEDPEFT.
- **Flower [110]:**
- **FedLLM [111]:** build upon the platform of FedML that is compatible with libraries from DeepSpeed and Hugging Face.
- **OpenFedLLM [112]:** a federated tuning system that includes training datasets, evaluation datasets, several FL baselines, and implementations of instruction tuning and value alignment.
- **FATE-LLM [113]:** FL-based framework designed for industrial tasks and supports LLM.
- **Shepherd [114]:** lightweight FL framework supporting LLM, which is built upon Hugging Face's PEFT and Alpaca-LoRA.
- **FedPETuning [115]:** a groundbreaking FL benchmark based on FedPEFT techniques, LoRa, prefix tuning, BitFit, and adapter tuning.
- **FedLegal [116]:** the first real-world FL benchmark for legal NLP that consists of five legal NLP tasks and one privacy test.

Table 1. Representative open-source frameworks, optimization techniques, and training datasets used in surveyed FLM literature.

Framework	Algorithm used in project
Edge [118, 119]	Model Compression
Edge [120]	System Design
Edge [121]	Fine-tuning
Federated [122]	Model Compression
Federated [123, 124]	Fine-tuning
Federated [125]	Attacks in FL
Federated [112]	Instruction Tuning
LLM Model	**Dataset**
GPT-series [18]	Wikipedia [126], CommonCrawl [127]
PaLM [15]	Wikipedia, Social Media, Github
GLM [128]	Wikipedia, BooksCorpus [129]
CodeGen [130]	Bigpython, bigquery [131]
LLaMA [14]	Wikipedia, Github, Arxiv, CommonCrawl, Github
FedNano [132]	Visual question answering [133]

- **NVIDIA FLARE [117]:** an FL framework allowing data scientists and researchers to move deep learning to a federated paradigm.

4 Robustness and Privacy in Federated LLM

Ensuring robustness and privacy is a crucial aspect after the integration of LLM and FL in an edge environment. LLM offers powerful capabilities through large-scale data training, and FL is designed to mitigate privacy risks by avoiding direct data sharing; both paradigms independently face notable security and privacy concerns. Edge environments enable computation closer to data sources, thereby increasing privacy and reducing bandwidth consumption and latency. On integration, these issues can introduce new vulnerabilities, resulting in new challenges and increasing complexity. This section dives in depth into the security problems and privacy leakage in FLM at the edge and highlights the defense strategies.

4.1 Privacy Risk in FLM

FL enables decentralized model training without transferring raw data, allowing participants to join or leave automatically. Recent studies reveal that FL does not fully guarantee privacy protection. From the FL perspective, a malicious or untrusted server may attempt to infer sensitive client information, bias the global model, or manipulate the training procedure, and adversarial clients may interfere through the aggregation process or attempt to extract private data from other participants. Research shows [134, 135], privacy leakage stems from gradient sharing during training, as gradients can unintentionally encode sensitive information affecting the central server and its management process [136], a concern that parallels broader data leakage risks observed across machine learning pipelines [137]. Similarly, [138] shows that even partial gradient information can expose significant details about local datasets, motivating cryptographic safeguards such as homomorphic encryption for inference tasks [139], and in some cases [134, 140], attackers can reconstruct the original training data solely from gradients. On the LLM side, LLMs like GPT-3 are built to integrate and produce text from large, diverse datasets, posing numerous privacy problems. These models may unintentionally encode and reveal private information from their training data, creating privacy concerns during the text production process, where information leakage and unintentional data memorization are significant obstacles [48, 141]. Therefore, in order to guarantee the dependable and moral application of these sophisticated LLMs in various industries, a trade-off between their inference performance and their capacity to safeguard user privacy must be accomplished [142].

4.1.1 Training Data Reconstruction Attacks

Training Data Reconstruction Attacks: Training data reconstruction attacks, also termed "data recovery attacks," aim to recover the private data from shared gradients in the FL system. These attacks are primarily gradient-based, exploiting the information exchanged between clients and the central server during model training. Since most deep learning models are optimized using gradient-based methods, including stochastic gradient descent (SGD), participating clients periodically share gradient updates with

the server. Adversaries who gain access to these gradients, or can approximate them, may be able to recover sensitive training data. Prior research [142, 144, 145] has demonstrated that in the FL setting, gradients can leak significant information about the original data, and complementary work has explored recovering compromised models once such vulnerabilities are exploited [143], that is to say, original data can be recovered. While such attacks have been extensively studied for image data, their application to LLMs, particularly in federated environments, remains relatively underexplored.

The Federated Iterative Language Model (FILM) [146] framework introduces a groundbreaking attack method for FL, proving for the first time that private text can be reconstructed even from massive batches of up to 128 sentences. Different from traditional image-based attacks that focus on gradient matching, FILM employs a two-stage strategy where it first extracts specific keywords from the gradients and then uses beam search integrated with a prior-based reordering logic to rebuild the original sentences. The Unordered-word-bag-based Text Reconstruction (UTR) attack [147] is designed for FLMs using adapter-based tuning based on Gradient Inversion Attack (GIA). To overcome the difficulties posed by low-rank gradients, locked model backbones, and the vast search space of potential text combinations, UTR employs a three-pronged strategy. First, it identifies specific tokens by analyzing attention patterns within the model's frozen layers. Second, it performs sentence-level reconstruction within the restricted low-rank space of the adapter gradients. Finally, it uses language priors to guide a constrained greedy decoding process, ensuring the recovered text is semantically logical.

A thorough assessment of the efficacy of the most advanced gradient leaking attacks on textual data in the context of fine-tuning LLMs is conducted in [148]. The results show that, especially when it comes to the Transformer architecture and real-world federated learning scenarios, the adversary cannot easily access the target gradient, which is crucial to the attack's success. NDPP-FL [149] is a dynamic FL framework that provides adaptive defense against data reconstruction by greatly reducing performance loss. The approach is treated as the privacy budget as a non-replenishable resource, limited using an information-leakage model to measure risk. It enhances security through Hessian-based saliency perturbations and improves performance via sample-adaptive clipping and decaying noise.

4.1.2 *Property Inference Attack*

Property inference attacks represent an important yet relatively underexplored privacy threat in the context of LLMs, especially when compared to more widely studied attacks such as membership inference and data reconstruction. In FL environments, these attacks aim to infer sensitive attributes of individual participants or groups of clients that are not directly related to the primary learning objective of the model. Instead of recovering exact training data, the adversary seeks to extract hidden properties or demographic characteristics that users intend to keep private [150]. Consequently, the leakage of such attributes poses a significant privacy concern in LLM-based systems [151, 152]. The Poisoning-Assisted Property Inference (PAPI) attack, a sophisticated method designed to compromise FL systems in dynamic scenarios, is introduced in [153]. In PAPI, the periodic model updates inherently encode discriminative signals that reflect shifts in data distribution, particularly those associated with sensitive attributes. Leveraging this observation, a malicious client can train a binary classifier to infer hidden properties from these updates.

4.1.3 *Membership Inference Attacks*

Membership Inference Attacks (MIAs) aim to determine whether a specific data sample belongs to the model's training dataset by analyzing the model's outputs and behavior. In the context of LLM, MIAs are commonly applied to tasks such as text classification and text generation [154, 155]. Although modern language models exhibit a certain degree of robustness against simple probing strategies [155], they remain susceptible to carefully designed inference attacks. In FL, these attacks can be categorized into active and passive approaches [156, 157]. Active MIAs are more intrusive, as adversaries manipulate the training procedure or inject malicious updates to amplify the leakage of membership information from other participants. In contrast, passive MIAs operate by observing model updates or parameters during training without interfering with the learning process, attempting to infer membership information indirectly. To make training data detection easier, the first MIA benchmark is designed for different VLLMs in [158]. Similarly, a novel MIA pipeline is created especially for token-level image identification, and a new statistic termed MaxRi-K% is presented, relying on the model output's confidence, which is applicable to both image and text data.

Casper, a causality-driven defense framework designed to mitigate label inference attacks in VFL is presented in [159]. The approach examines the training process from a causal intervention perspective to assess potential vulnerabilities and incorporates a confounding mechanism integrating a cut-layer output reconstruction with label obfuscation, effectively weakening the direct causal relationship between target labels and intermediate representations. Casper employs a selective discrepancy training technique that balances the optimization process between passive and active participants to further enhance the model performance. In [160] proposed Loss Trajectory-based Membership Inference Attack (LTMIA), a unified framework capable of performing both MIA and Source Inference Attacks (SIA). The strategy leverages differences in loss trajectories observed during the early stages of federated training. By capturing temporal patterns in loss values across initial communication rounds, LTMIA trains a lightweight inference model to predict membership. For source inference, it utilizes a statistical averaging technique that enables accurate identification of data origins without relying on auxiliary datasets.

4.2 Defend Against FLM Privacy Attack

Ensuring privacy preservation in the FL domain is significantly more challenging due to various factors, including resource-constrained devices, heterogeneous data distributions, and unreliable connectivity. These challenges are further amplified when LLM gets integrated with FL on the edge. Some of the strategies to address privacy risks in FLM are Homomorphic Encryption (HE), Secure Multi-Party Computation (SMPC), and Differential Privacy (DP).

4.2.1 Homomorphic Encryption (HE)

HE enables computations, such as addition, multiplication, and aggregating gradients, to be performed directly on encrypted data without seeing the raw information, ensuring that the decrypted result matches the outcome of operations conducted on the original data. This property makes HE particularly suitable for secure FL, allowing model updates to remain encrypted throughout the training process [162, 163]. HE schemes are generally categorized into Partial Homomorphic Encryption (PHE), which supports a single type of operation (e.g., addition or multiplication), Somewhat Homomorphic Encryption (SHE), which allows limited combinations of operations, i.e., unlimited addition with at least one multiplicative operation, and Fully Homomorphic Encryption (FHE), which supports arbitrary computations on encrypted data [161]. While HE significantly enhances data security by preventing direct access to sensitive information, it introduces substantial computational and memory overhead. As a result, practical deployment requires careful trade-offs between security guarantees and system efficiency, especially in resource-constrained edge environments.

4.2.2 Secure Multi-Party Computation (SMPC)

SMPC is a cryptographic paradigm that enables multiple parties to collaboratively compute a function, such as training an LLM, without revealing their individual inputs. Built upon foundational techniques such as Oblivious Transfer, Garbled Circuits, and Secret Sharing, SMPC [164] ensures strong privacy guarantees during collaborative computation. In FLM settings, SMPC can be used to securely aggregate model updates or perform joint computations across distributed participants. However, SMPC-based methods often incur high computational and communication costs, which can limit their practicality in large-scale or resource-constrained environments [165]. Additionally, these techniques require strict synchronized participation among all clients, which may not align well with the asynchronous nature of many FL systems. Although SMPC can effectively protect against direct data leakage, it is not entirely immune to inference attacks, necessitating the amalgamation of complementary approaches, including differential privacy [166, 167]. Despite these limitations, SMPC remains a promising solution for enhancing privacy in FLM systems.

4.2.3 Differential Privacy (DP)

DP is a widely adopted technique designed to protect sensitive information by introducing carefully calibrated randomness into the data or model updates. DP ensures that the output of a computation (differentially private algorithm) remains nearly unchanged whether or not any individual record is included in the dataset [168]. Unlike encryption-based approaches, DP provides a balance between model accuracy and privacy by injecting noise in a controlled manner, thereby preventing adversaries from inferring private information while preserving model performance. In federated settings, DP is typically implemented in two forms: Local Differential Privacy (LDP) and Global Differential Privacy (GDP). LDP [169–172]applies noise directly at the client side before data or gradients are shared, offering stronger

privacy guarantees but often at the expense of model performance. In contrast, GDP [173–176] introduces noise at the server level after aggregating client updates, maintaining better model utility due to reduced noise injection. DP-based approaches have been widely employed to defend against attacks such as membership inference, training data reconstruction, and property inference by perturbing either inputs or outputs. When combined with optimization methods like stochastic gradient descent, DP can effectively reduce privacy leakage while maintaining competitive performance. Recent studies have explored the application of DP in FLM training, demonstrating how different privacy budgets influence the trade-off between model utility and privacy in tasks such as sentiment analysis and financial text classification [177]. Furthermore, methods such as DP-LoRA [178] integrate low-rank adaptation with noise injection, reducing communication overhead while ensuring privacy-preserving fine-tuning of LLMs across distributed clients.

4.3 Robustness in FLM

Robustness in FLM is the ability of the model to maintain reliable performance, privacy, security, and stability despite adversarial conditions, including heterogeneous (non-IID), decentralized, and malicious attacks that attempt to poison the training process. The goal of attacks on robustness is to manipulate the output of the model to mislead clients and achieve the malicious goals of attackers, in contrast to privacy attacks that target data confidentiality. This section discusses the robustness issues that exist on FLM, including poisoning attacks, Byzantine attacks, and prompt attacks.

4.3.1 Poisoning Attacks and Defenses

Poisoning Attacks (PA) are a prevalent form of targeted adversarial manipulation in FL. Traditionally, PA on ML models targets the classification models by deceiving the models through the manipulation of training data. For example, attackers could compromise NIDS [179] or contaminate spam filters by adding good words to training datasets [180, 181]. Model Poisoning Attacks (MPA), which occur during the local training process, and Data Poisoning Attacks (DPA), which arise during the collection phase at local clients, are two poisoning attacks. In particular, MPA includes DPA in FL scenarios as it modifies part of the updates sent to the model during each iteration.

In the context of LLMs, the risk of poisoning is immensely amplified, as LLMs rely on large-scale datasets collected from open and unregulated sources such as the internet, where adversarial content can be easily introduced. Prior studies have demonstrated that large public datasets, including LAION-400M [182], COYO-700M [183], and Wikipedia, can be deliberately poisoned through strategies such as coordinated crowdsourcing or domain ownership manipulation [184].

Backdoor attacks, a subset of PA, use poisoned samples to risk model integrity by embedding destructive functions within it. These triggers activate malicious behaviors only when specific inputs are encountered; the model appears normal [185]. Although DPA challenges the LLMs during the large-scale data curation, backdoor attacks remain a possible threat. During the execution of the task, attackers introduce malicious manipulations by altering inputs during testing, enabling targeted misbehavior during classification when the model is deployed [186, 187].

Introducing a backdoor attack is non-trivial due to issues in the Graph Foundation Model (GFM), such as effectiveness, stealthiness, and persistence. A novel backdoor attack model in GFM is designed to address those challenges [188]. Backdoor injection is accomplished without the knowledge of downstream tasks by constructing a label-free trigger association module connecting triggers to a collection of prototype embeddings. Then, a node-adaptive trigger generator activates the backdoor consistently by lowering the chance of trigger detection, which is dynamically generated by node-specific triggers. Finally, a persistent backdoor anchoring module is developed that strengthens the tenacity of the backdoor during downstream adaptation by securely anchoring it to fine-tuning-insensitive parameters.

A dormant backdoor, an attack triggered by fine-tuning without the essence of prior knowledge of downstream processes, is introduced in [87]. Dormant Backdoor uses the universal dynamics of gradient-based optimization as a process-as-trigger mechanism rather than tying the backdoor to static input patterns. The attack is formulated as a bilevel optimization problem that jointly optimizes the poisoned model and triggers under utility, lethality, and stealth objectives while simulating the victim's fine-tuning trajectory on proxy data. The poisoned model can avoid current backdoor detectors and behaves similarly to a clean model prior to fine-tuning. After fine-tuning, the same adaptation mechanism

consistently amplifies the backdoor on a variety of downstream datasets and fine-tuning techniques.

The thorough analysis of how dataset imbalance increases backdoor susceptibility is presented in [86]. The research shows that a majority-class prejudice brought about by class imbalance makes people more vulnerable to attacks and undermines conventional defenses. Randomized Probability Perturbation (RPP), a verified detection framework that uses just output probabilities to identify poisoned samples in a black-box situation, is used to address the challenge. RPP creates a probabilistic upper bound on false positives and offers verifiable detection guarantees for inputs altered by backdoors.

The Federated Intrusion Detection and Mitigation Framework (FIDMF) [189] is introduced as a behavior-based solution to modern security challenges like data privacy, scalability, and class imbalance. By combining FL with an Attention-LSTM core, the system enables private, collaborative training and effective temporal pattern recognition. A key innovation of FIDMF is its use of open-source LLMs to drive three critical functions: enriching contextual features from raw logs, guiding GANs in creating semantically accurate attack data for better training, and providing human-readable Explainable AI (XAI). Supported by SMOTE for data balancing, this framework offers a comprehensive, transparent, and scalable approach to intrusion detection.

Integrated cryptographic techniques of RSA algorithms and Advanced Encryption Standards (AES) with bidirectional encoder representations from vision transformers (ViT) and transformers to prohibit poisoning attacks [190]. The proposed approach uses RSA for key authenticity against unauthorized modification, entry, or malicious interference, while AES is employed for rapid data encryption. Additionally, ViT assists in analyzing the appearance of image data to detect poisoning attempts, while BERT is used for intelligent analysis of textual data to effectively verify unfavorable occurrences of poisoning attempts.

Thought-Transfer [191] is a reasoning-model-based novel indirect targeted poisoning attack, integrating two methods for adversarial reasoning traces into pre-existing training data traces. The approach manipulates model responses on prior undiscovered target tasks by encoding transferable behavioral patterns and guiding problem-solving. This leads the model to function as a covert "clean-label" attack by maintaining original queries and accurate final replies, yet altering only the reasoning traces, in contrast to other chain-of-thought poisoning attempts that necessitate explicit trigger insertion, incorrect answers, and manipulated reasoning in poisoned samples. Importantly, by embedding adversarial reasoning traces via training task samples, the TT modifies model responses on target tasks not present in training.

A refined backdoor attack [192] approach improves backdoor assaults in FL by using a three-step cooperative strategy to target particular model weights. The approach accurately alters layer parameters (by flipping and zeroing) by identifying non-essential parameters and examining activation differences between benign and malicious inputs. This makes it possible to surgically implant the backdoor, guaranteeing that the device will continue to function well for regular users while being susceptible to the intended trigger. [88]presents a local linear explanation-based dynamic, equitable, and explainable defense against Byzantine and backdoor attacks. An investigation into VFL defenses against backdoor attacks when the attacker has enough label information is performed in [193]. A framework named FILTER is designed to defend against backdoors during training in the presence of malevolent participants. To mitigate the backdoor threats in VFL, the approach uses a loss-based filter and an embedding-based filter, which successfully detect and eliminate contaminated samples throughout subsequent training phases.

Training the backdoor models can be considered a multi-tasking problem, where the backdoor models complete the primary goal and their feature representations are more stable than those of benign models. Relying on this, [194] deployed a new stability metric that quantifies the difference in feature map stability between benign and backdoored models. Similarly, a novel backdoor detection method for federated learning using this metric is also proposed. The proposed approach effectively defends against backdoor assaults by calculating anomaly scores for each client and preferentially aggregating models with benign properties.

A paradigm for principled optimization reframing backdoor injection as a bi-objective learning problem that strikes a balance between benign and malevolent is proposed in [195]. In contrast to previous heuristic multi-task formulations, the proposed

method incorporates a theoretically based parameter selection strategy and a conflict-aware gradient coordination mechanism that together guarantee effective, focused adaptation without full-model adjustment. Additionally, theoretical assurances demonstrating that resolving inter-objective gradient conflicts reduces clean-task degradation and produces verifiable convergence gains are achieved.

FedSecurity [125]285 provides a benchmark for evaluating both attack and defense strategies in FL. It includes modules such as FedAttacker and FedDefender, supporting attacks like label-flipping backdoors, Byzantine manipulation, and defenses such as m-Krum, Krum, and geometric median aggregation. The framework is versatile and applicable to various models, including GANs, ResNet, and LLMs, demonstrating its practicality in real-world scenarios.

SHIFT [196] framework proposed that enhances backdoor detection by decentralizing the process to the client, reducing server load, and allowing for direct inspection of unencrypted data. It utilizes dynamic risk mapping and code obfuscation to prevent client-side tampering for flexible results. Experimental results show that SHIFT outperforms traditional encryption-dependent schemes in time overhead and maintains high performance in single and multi-client attack scenarios involving non-IID data.

An approach, Targeted Poisoning Attack (TPA) in the context of vertical split learning (VSL), TPA-VSL [197], performs without the essence of obvious trigger patterns by directly manipulating the embedding model. TPA-VSL uses two innovative elements to negatively impact the targeted sample's prediction by mapping the targeted sample's embedding vector to the attacker's preferred class. The first element creates meaningful training data that mimics the embedding space of the target model using cutting-edge diffusion models directed by a multimodal encoder-decoder. By matching these target sample mappings with the attacker's desired output class, the second component then undermines the embedding model.

In order to protect the privacy of the user and model parameters, Prifft, a federated fine-tuning approach, is proposed in [198]. Owing to significant LLM parameters, the Prifft uses hybrid secret sharing that integrates function secret sharing (FSS) and arithmetic secret sharing (ASS) to implement secure layers, develop safe operations, and activate for privacy-preserving fine-tuning. The method optimizes a number of secure computation protocols relying on FSS, such as tensor products, reciprocal calculation, hyperbolic tangent, sigmoid, natural exponentiation, softmax, and dropout, to augment the effectiveness of privacy-preserving federated fine-tuning of LLMs. The hybrid secret sharing enables PriFFT to use the optimized FSS protocols and integrate with ASS protocols, offering sophisticated computation without additional transmission.

Huang et al. [199] introduced a secure distributed LLM framework relying on model slicing, which uses Trusted Execution Environments (TEE) and fine-tuned structure (embedding of P-tuning or LoRA) on both server and client side. Lightweight encryption is used for secure communication between TEE and general scenarios. A split-fine tuning strategy is deployed to enhance accuracy and performance and reduce computational cost. LLM is split by layers, and later layers are allocated on the server side, eliminating the essence of TEE on the client side.

In order to improve the robustness of pre-trained foundation models toward clean-label PA in FL scenarios, [200] proposes a robust pre-training strategy for foundation models that enhances resistance to poisoning without requiring additional defenses during downstream tasks. The strategy lies on raising inter-class feature distance while decreasing adversarial feature distance. In order to protect against clean-label poisoning attacks, the model helps to enhance the loss function in adversarial training to minimize the gap between poisoning and adversarial robustness.

A novel backdoor attack mechanism, Fed-EBD, based on horizontal federated learning, is proposed in [201]. Unlike traditional approaches, Fed-EBD eliminates the need to compromise client devices or maintain persistent participation. Instead, it injects backdoor behaviors through a synthetically generated public dataset created by a foundation model. By mimicking benign client updates, the attack can evade existing defense mechanisms.

High energy requirements, computational constraints, and the requirement for high accuracy in constrained circumstances make the deployment of LLMs for edge-device malware detection challenging. In order to address this, [202] presents an architecture that particularly overcomes the common shortcomings of lightweight LLMs, like reduced precision and hardware limitations, while optimizing their efficiency.

In order to detect adversarial attacks on edge devices, [203] proposed Quantum-based Privacy-Preserving Federated Learning (Q-P2FL). The integration of additive homomorphic encryption with quantum-based registration and authentication helps to secure model weights and preserves privacy on edge devices, where a pre-trained vision transformer generates local model weights. Differentially private FL is designed to deploy on edge environments against poisoning attacks in [204]. A weight-based anomaly detection algorithm based on small-sized validation datasets is used for anomaly detection on parameters, and a differential privacy approach is used to preserve the privacy of the model and data at the edge.

Research [205] investigates the unique security vulnerabilities of deploying LLMs to edge, cloud, or hybrid architectures and identifies five major threat surfaces. To mitigate these threats, a framework is designed for evaluating edge security and trustworthiness based on quantifiable metrics, including failover exposure, data egress volume, provenance chain completeness, and sovereignty boundary integrity. Edge-aware Federated AI (EAFAI) [206] is a framework designed by integrating FL and LLM into the edge environment. Computational constraints, bandwidth, and privacy issues are addressed using selective parameter sharing, model compression, and decentralized training, where the workloads are dynamically distributed across edge and cloud nodes. To tackle the resource constraints, privacy, and trust issues, a hierarchical federated collaborative computing (HFCC) architecture is proposed in [207] by integrating FL and LLM in an edge environment. The HFCC framework uses dynamic chunking and horizontal layer splitting to balance computational loads by verifying inference results and securing local data at the edge.

4.3.2 Byzantine Attacks and Defenses

Adversarial threats targeting model robustness can generally be divided into targeted attacks [197, 208–210] and untargeted attacks [191, 211, 212]. Byzantine attacks are untargeted attacks in FLM environments, sending maliciously manipulated gradients to the model aggregator with the intention of disrupting the training process or corrupting the global model to degrade the performance. Global servers now have concerns about the trustworthiness of the clients due to issues commonly occurring during the local update phase. Some clients might be susceptible to internal errors or external attacks, and can send corrupted or malicious updates to the server. The entire federated optimization workflow could be disrupted when the server unintentionally amalgamates these malicious updates.

The major approaches to countercheck Byzantine attacks are robust aggregation and Byzantine detection. The robust aggregation approach intends to reduce the impact of malicious updates by assuming that adversarial gradients deviate significantly from those of honest clients. Based on these assumptions, aggregation rules are crafted to filter or down-weight outliers. Algorithms such as Bulyan [213], Krum [214], and cosine similarity [215] select updates that are closest to the majority in Euclidean distance. These strategies have shown effectiveness in mitigating certain types of attacks. However, when a large number of clients are compromised or client data is highly non-IID, the robust aggregation approach struggles to outperform with degraded performance. This shows the need for more research and development of reliable and robust aggregation techniques.

Byzantine detection approaches are intended to identify and remove malicious local updates in order to stop compromised clients from paralyzing the FL system [216, 217]. This approach is more robust than the aggregation technique, yet there is an essence of huge data demand and high computational demands. Despite extensive research in traditional FL settings, Byzantine detection and robust aggregation remain relatively underexplored in FLM systems. The heterogeneity, scale, and complexity of these systems introduce new challenges, making robustness against Byzantine attacks an important open research direction that warrants further investigation.

Current Byzantine defenses often struggle to differentiate between anomalous model updates and benign outliers under heterogeneous settings and Byzantine attacks, which can ultimately weaken the performance of the model. Byzantine-robust aggregation scheme based on hybrid anomaly detection (HadAGG) [217], a robust aggregation scheme is designed to address the issues for heterogeneous environments. The framework uses a hybrid filtering strategy integrating Shapley values and cosine similarity to categorize updates as benign, malicious, or simply unusual but honest. By utilizing a multi-objective utility function that balances loss and accuracy, HadAGG calculates Federated Shapley values to precisely measure each client's contribution.

Rather than simply discarding suspicious data, the system applies gradient projection to correct malicious updates and employs weighted aggregation to ensure every contribution positively impacts the final model. A normalized stochastic gradient descent method with momentum, ByzNSGDM, maintains convergence guarantees while achieving robustness against Byzantine workers [218].

FEDBYZO [219], a federated zero-order optimization technique that is robust against Byzantine attacks, offering notable communication cost savings. Modified robust aggregation is used to provide convergence guarantees for broad non-convex objectives under client data heterogeneity.

A strategy FDBA is designed to enhance efficiency and robustness in FL [220]. FDBA uses a specific feature called PDist to capture the real impact of such modifications, going beyond conventional approaches that just examine raw model updates. The model uses a three-dimensional cooperative learning process to evaluate these features and differentiate between benign and malicious behavior. Initially, clustering is performed to find patterns using PDist, then classification of malicious clients using clustering results, taking support of auxiliary data for efficiency and accuracy. Finally, detection accuracy is improved by integrating historical data. Integrating these steps, FDBA provides a more efficient and precise way to defend against compromised participants in the federated network.

Trusted MultiLLMN [221], a framework that uses a Weighted Byzantine Fault Tolerance (WBFT) blockchain consensus mechanism for ensuring security, efficiency, and reliability of collaborative multiple LLM. WBFT dynamically adjusts the voting power of each LLM based on its proven reliability and accuracy. By weighting responses this way, the system rewards high-quality contributions while neutralizing the influence of malicious or faulty nodes, ensuring the final output remains both secure and efficient.

This paper introduces TinyGuard [222], a resource-efficient defense against Byzantine attacks that enhances the standard FedAvg algorithm. Rather than analyzing massive and high-dimensional gradients, TinyGuard creates compact statistical fingerprints of client updates tracking specific traits like sparsity measures, layer-wise ratios, and norm statistics. In this low-dimensional fingerprint space with complexity, robust statistical deviations are measured to identify Byzantine clients without changing the underlying optimization process.

FORCE (Byzantine-Resilient Decentralized Federated Learning via Game-Theoretic Contribution Aggregation), a decentralized FL aggregation approach inspired by the Shapley value in game theory, is proposed in [223]. In order to identify potentially malevolent clients, FORCE moves away from gradient-checking techniques and instead uses a universal metric, the loss of the local model, regardless of particular gradients. The approach is lightweight and is optimized through approximate Shapley value computation to lower the computational overhead of FORCE as the number of neighboring clients increases.

Sensitive information can be leaked during model updates and gradients, where a malicious server may launch adversarial attacks, including Byzantine manipulation. To address these challenges, Byz-Clip21-SGD2M [224] proposed a novel algorithm integrating resilient aggregation with thoughtful clipping and double momentum. The approach uses standard σ-sub-Gaussian and L-smoothness gradient noise assumptions to show the high-probability convergence assurances.

A novel defense mechanism termed FedAOT is designed to counter untargeted poisoning and multi-label flipping attacks inspired by meta-learning using adaptive aggregation architecture [225]. FedAOT suppresses adversarial influence independently, relying on restrictive attack assumptions or predefined thresholds by dynamically weighting client updates according to their reliability. FedAOT maintains robust performance in unforeseen scenarios and generalizes well across a variety of attack types and datasets.

To explore the security of LLMs within FL, [226] analyzes potential vulnerabilities and defensive traits specifically through LoRA weights. The study identified two critical findings: LLMs are susceptible to threats from malicious participants, and LoRA weights display unique behavioral signatures that simple classifiers can effectively detect. Leveraging these insights, [226] designed Safe-FedLLM, a defensive framework that operates across step, client, and shadow levels. By treating locally trained LoRA weights as high-dimensional behavioral features, Safe-FedLLM uses lightweight models to probe and identify malicious attributes, ensuring a more secure federated training process.

In order to maximize the error rate of the global

model, the research [90] assumes a scenario in which an attacker has compromised several client devices to alter their local parameters during training. To find out if existing defenses can tolerate such focused interference, the researchers framed these assaults as optimization problems and tested their efficacy against four contemporary Byzantine-robust aggregation techniques. The results showed that the error rates were increased and were robust against Byzantine failures. In order to defend against attacks, the defenses were ineffective, suggesting the need for a new defense mechanism.

4.3.3 Prompt Attack and Defense

LLMs are highly sensitive to the formulation of input prompts, and their output are inconsistent based on the scenarios of the phrased prompt. This makes them vulnerable to prompt-based attacks, where adversaries deliberately manipulate or craft inputs to influence the behavior of the model. The strategic objective of such attacks is to steer the model towards generating particular results or accomplishing specified goals. Under the customized inputs, even well-trained models may yield misleading or harmful results. A widely recognized form of this attack is Prompt injection [227–229], in which attackers design inputs that override safety constraints or intended instructions of the model. Such manipulation can help an adversary to effectively control the output of the language model, forcing it to execute unintended tasks or disclose private information. These vulnerabilities may lead to serious consequences, including unauthorized system actions, leakage of confidential data, or broader security breaches.

Few studies have explored the prompt-based attack and defenses in LLM. Preprocessing the data prompt to eliminate the instruction/data from the injected job and/or rewriting the instruction prompt itself are two recommended preventive strategies [229]. Several strategies, such as data prompt isolation [231, 232], re-tokenization [233], paraphrase [230], and instructional prevention [233], can be deployed to counter the adversarial prompts. The fine-tuning and inference applications of federated LLMs may be impacted by prompt attacks; however, they are mostly caused by LLM security flaws and hardly involve the FL process. Such attacks can be avoided by defending from the standpoint of LLMs alone.

Recent research has produced a number of detection methods to protect against prompt injection attacks. There are few works for the indirect scenario [234–237], where injected instructions come indirectly from external tools, such a search engine, and the majority of them concentrate on direct prompt injection attacks [238, 239]. Furthermore, current research focuses mostly on injection detection techniques while paying little attention to post-processing techniques that try to lessen the injection after detection. The feasibility of detection and elimination of indirect prompt injection is investigated in [240], where the benchmark dataset is created for evaluation. The detection of the attack is evaluated using the effectiveness of current LLMs and open-source detection models. The model is trained further using crafted training datasets. The removal technique undergoes two processes: the segmentation removal approach, which segments the injected documents and removes parts containing injected instructions, and the extraction removal approach, which trains an extraction model to identify and remove injected instructions.

The study [227, 228] examines various prompt injection attacks in LLM architecture and applications, including machine translation and chain of thought reasoning, evaluation metrics, benchmarks, and datasets. Mitigation approaches such as input validation, automated red teaming frameworks, alignment through training, and content filtering are also analyzed. In order to formalize the prompt injection attack [229] proposed a framework where a new attack is designed by integrating the existing one. The proposed framework offers a fundamental standard for assessing future defenses against prompt injection attacks and allowing for systematic evaluation. Research in [241] explores the way to automatically adapt a single code to a specific client context without human intervention in heterogeneous circumstances and to evaluate the existence of potentially harmful code that compromises data security. A novel prompt engineering approach to improve the detection of malicious code based on LLM in distributed scenarios is proposed in [241] to overcome the challenges.

A privacy-preserving prompt injection detection approach based on embedding-based categorization and FL is proposed in [242]. Both centralized and federated logistic regression models were trained based on curated datasets of adversarial and benign prompts that were encoded using sentence embedding. The federated approach was able to maintain anonymity by sharing model parameters across clients while attaining detection performance compared to centralized training.

In [243], a human-AI collaborative architecture to detect and counteract four threats, such as misinformation spread, secret data leakage, system disruption, and free-rider exploitation, in federated military LLM. The proposed architecture detects and mitigates adversarial behaviors of shared LLM weights using quality assurance and red/blue team wargaming. In terms of policy, it encourages collaborative AI-human policy creation and security protocol verification.

Manipulation of the webpage environment to cause a web agent to carry out an attacker-specified action through the prompt injection attack is accomplished using WebInject [244]. Perturbation is added to the raw pixel values of the rendered webpage and is mapped through a screenshot where the web agent accomplishes the attacker-specified action based on the perturbation. A neural network is used for projected gradient descent and mapping to address the reformulated optimization problem, which is a challenging task as mapping between a screenshot and raw pixels is non-differentiable, creating backpropagation gradients to the perturbation.

Malicious instructions incorporated into external content can alter LLM outputs in prompt injection attacks, deviating from user expectations. BIPA [245] is presented as the first benchmark and evaluation of the risk of such vulnerabilities, including prompt injection attacks. Using BIPIA, a variety of current LLMs are evaluated, and it is discovered that more capable LLMs have higher ASRs and are more susceptible to indirect prompt injection attacks. Both white-box and black-box defense strategies are proposed in [245], and their efficacy is carefully assessed. While the white-box defensive method successfully thwarts indirect prompt injection attacks with minimal negative impact on the output quality of the LLM, the black-box defense methods can effectively lower attack success rates.

AgentDojo, an assessment framework for agents that run tools over untrusted data to gauge the adversarial robustness of AI agents, is presented in [246]. The method is an expandable platform for creating and assessing new agent tasks, adaptive attacks, and defenses in order to represent the dynamic nature of defenses and attacks. AgentDojo presents a challenge for both attacks and defenses, with current prompt injection attacks violating some security features but not all of them, while state-of-the-art LLMs fail at many jobs (even in the absence of assaults).

5 Biases in FLM

Different fields of study have different definitions of bias in AI research, reflecting distinct perspectives. Biases are considered a vulnerability in robust AI that not only degrades the performance but also enhances the susceptibility to adversarial attacks, especially impacting the underrepresented groups [247, 248], a concern that extends even to narrower software engineering tasks such as fair reviewer recommendation [249]. Reliable AI considers biases as inconsistencies in model performance under various circumstances that compromise the stability and predictability of AI systems, a concern that has been raised across specific application sectors as well [250]. Similarly, in responsible AI, biases represent systematic imbalances in algorithmic results with frequently identified sensitive traits such as ethnicity or gender [251]. This creates the foundations for trustworthy AI to emphasize explainability, transparency, and trust that frames bias as a tool to overcome obstacles and behave predictably and ethically. Despite various perspectives, in this paper, we define bias in FLM as systematic differences in model outcomes due to demographic bias, contribution bias, and performance-related bias.

5.1 Demographic Bias

Discrimination or favoritism that disproportionately impacts certain demographic groups due to their sensitive features comes under demographic bias. The sensitive features include demographic traits such as gender, age, ethnicity, and so on, which impact the result in discriminatory behavior [252, 253], thus, its use is legally forbidden in decision-making [254]. The existence of data-intensive features in FL increases the risk of discrimination and bias, where the model may unintentionally discover and reinforce unfavorable tendencies across the various demographic groups. Data imbalance, bias during client selection and aggregation, bias propagation among FL clients, and the impact of mitigating bias-related single sensitive features are some of the major causes. Demographic biases are assessed through various demographic bias metrics, such as statistical parity difference, disparate impact, discrimination index, and equal opportunity difference. Model quality metrics such as accuracy, recall, precision, and F1-score can be used to quantify performance differences among demographic groups to show the bias in model outcomes [255].

5.2 Contribution-related Bias

The misalignment between the performance of the model and the client's contribution level is contribution-related bias. The objective of mitigating contribution-related bias is to guarantee that FLM clients receive customized models that represent their unique degree of contribution to the collaborative process. Data distribution, aggregation rules, and client selection strategies are some of the causes of contribution-related biases, compounded by the broader challenge of maintaining fault tolerance and theoretical guarantees under heterogeneous client participation [256]. Contribution-related bias is assessed by comparing a client's input against their resulting model quality. Evaluating the correlation between these metrics exposes mismatches between a client's effort and their final performance.

5.3 Performance-related Bias

The decentralized nature of FLM introduces some form of bias due to the difference in model performance among various clients, termed performance-related bias. It emerges due to the context of decentralization, where diverse devices are used for data collection and processing, where the quantity and quality of the data sources and availability in each client have a significant impact during the learning process [257]. Performance-related bias is a primary obstacle in FLM systems, where the disparities lead to an uneven distribution of rewards among participating clients, a problem that model-heterogeneous aggregation techniques such as feature distillation aim to mitigate [258]. Data imbalance, bias during client selection, and aggregation are the major causes. Performance-related bias can be quantified using standard quality metrics such as accuracy, recall, precision, and F1-score. Testing these metrics across different clients highlights variations and performance gaps in the system.

6 Applications Scenarios

This section highlights the existing work based on various application scenarios, as shown in Figure 3. In each scenario, we highlight the goal, aggregation approach, ML algorithm used, and participants during federation.

6.1 Healthcare

Driven by the acute need for privacy-preserving collaborative intelligence, the healthcare domain constitutes a fundamental application for FLM by leveraging distributed patient data, such as electronic

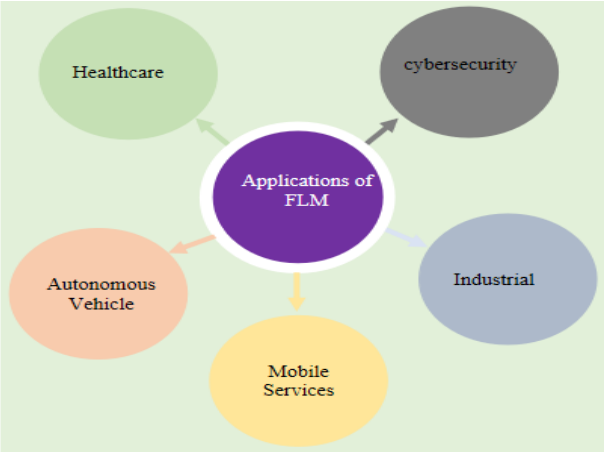

Figure 3. Some of the application scenarios of the FLM in the edge environment.

health records (EHRs), patient interactions, and clinical notes, through wearable devices, clinics, and hospitals. The major objective in this scenario is to adapt LLMs to enhance diagnostic accuracy, reliability, and privacy through specialized diagnostic support, clinical documentation, and patient record analysis through the sensitive EHR. Most medical LLMs trained on single-institution data have low safety hazards and generalizability, making them unable to capture the variability in disease prevalence, patient demographics, and institutional documentation. The FL framework supports multiple medical institutions to collaboratively enhance the performance of the global model, even on complex tasks for decision-making. Secure FedAvg amalgamated with differential privacy or homomorphic encryption is employed to aggregate model updates by preserving privacy [310]. FLM-based FedMRG [259] is proposed for multi-center medical report generation using a client-aware contrastive learning-based encoder. The model addresses the communication overhead and multi-modal data heterogeneity using parameter-efficient row-rank factorization and Dual-adapter Mutual Boosting (DMB) for text decoding and Hierarchical Contrasting and Prompting (HCP) for image encoding.

A customized prompt-based FL technique for medical VQA is presented in [100], considering datasets from various organs as distinct clients. While a reliability weighting system lessens the detrimental influence of low-quality clients, lightweight reminders allow for effective information sharing. Fed-MedLoRA+ [260] adopts an adaptive and data-aware aggregation process, known as Spatial-Temporal Aggregation (STAgg). STAgg detects and fixes initialization drift

and aggregate bias brought on by various institutional modifications by using server-side validation against a small, heterogeneous proxy dataset. FedTherapist is a mobile mental health monitoring system deploying user speech and keyboard input to fine-tune FMs with FL [261].The approach shows exceptional accuracy in mental health prediction tasks like mood, stress, and depression prediction.

6.2 Cybersecurity

In the cybersecurity domain, FLMs are deployed to establish privacy-aware distributed intelligence for the detection and mitigation of threats. The major goal is to build highly adaptive IDS, anomaly detection systems that can process tremendous streams of device telemetry and network traffic logs at the edge, identifying complex and sophisticated attacks, such as data poisoning attacks, model poisoning attacks, false data injection, and denial-of-service. LLMs suit the role due to their semantic reasoning abilities, explaining the nature of detected threats. FL aggregation focuses on resilience to hostile updates because federated systems are susceptible to backdoor assaults and model poisoning. Due to the fact that large norms are frequently indicative of overfitting or intentional manipulation, the Weighted Gradient Clipping Aggregation (FedWGCA) framework [262] employs an inverse norm weighting method, which weights contributions inversely to the L2-norm of their clipped gradient updates.

An extensive analysis on the ways in which privacy may be jeopardized during the federated training of large language models is presented in [263]. It pinpoints particular weaknesses where adversaries might be able to obtain private information via model inversion attacks and other inference methods. The study lays the groundwork for the necessity of sophisticated privacy-preserving methods in federated learning by proposing stronger cryptographic safeguards and more reliable aggregation algorithms to reduce these concerns. Gupta et al. [146] present an attack scenario that extracts sensitive information from an FLM, addressing privacy issues. The researchers demonstrate how some training methodologies and model setups can mistakenly memorize private text, making it vulnerable to malicious participants. The article explores different situations and setups to provide actual proof of the dangers involved. The article suggests solutions to mitigate vulnerabilities, including stricter data sanitization and privacy-preserving architectures.

FedSecurity [125] serves as a benchmark for simulating adversarial attacks and defense mechanisms in FL. FedSecurity avoids the need to construct basic FL procedures like training and data loading from scratch, allowing customers to focus on developing their own attack and defense techniques. The system has two components: FedAttacker, which conducts attacks during FL training, and FedDefender, which develops defense mechanisms against these attacks. The future of cybersecurity in the FLM setting includes jamming-resistant frameworks like R-SFLLM [264], which focuses on military-grade split federated learning to retain operational integrity even in disputed contexts.

6.3 Industry

FLM at the edge is deployed for predictive maintenance and process optimization by preserving privacy while using distributed sensor data. The major objective is to integrate several industrial processes, such as conceptual design, supply chain management, and predictive maintenance, into a single framework without disclosing confidential production data and to increase operational efficiency. These models are used to forecast logistics delays in multi-factory settings, diagnose bearing system faults in real time, and predict sheet metal forming flaws to lower scrap rates [265].FATE-LLM [113] integrates FL with LLM for industrial applications. FCLLM-DT [266] integrates FL, LLM, and DT to correct anomalous sensor data by providing a physical model of a bearing. In the case of sensor failure, virtual datasets are generated using large language models (LLMs) supported by retrieval-augmented generation (RAG). Federated continuous learning (FCL) is used in IIoT applications across distributed industrial contexts to improve global model training by combining localized models from several facilities. This improves bearing failure diagnosis accuracy while protecting data privacy. An expanded LLM KG [267] architecture is proposed that employs unique triggers to finish the triplet in KG and boost the quantity.

The multi-field hyperbolic (MFH), a graph embedding technique that recognizes low-frequency items and precisely defines their logical meaning, is suggested for vectorizing entities in an LLM-augmented KG. An encrypted graph embedding fusion approach is introduced among different participants to enhance the quality of the graph embedding. Through the combination of AES and hash-based message authentication code (HMAC), a self-parent key-based

cryptography technique has been developed to guarantee secure federated learning computation and data security at the edge (smart meter), fog (SCADA-based substation grid), and cloud (grid cloud server) layers of the SG-IIoT [268]. At each end node of the federated learning process, a load forecasting algorithm known as LSTM-LLM-GenResAI-Forecasting has been developed for calculation. The fog node aggregates and encrypts the prediction outputs from the edge nodes. Data are decrypted at the fog node, and each sub-station grid's overall load estimate is produced using the FL process's aggregation method. Once more, overall load projections are produced for several fog nodes by combining forecast data from these fog nodes in an encrypted state at the cloud level.

6.4 Autonomous Vehicle

In the domain of autonomous vehicles and intelligent transportation systems, FLM at the edge aims to enhance traffic prediction, navigation, and accident prevention through collaborative learning. The major objective is to reduce high computational transmission costs, enhance efficiency and safety by leveraging distributed data from roadside and vehicles in terms of natural language interaction with passengers, real-time scene understanding, and context-aware autonomous decision-making. In order to handle important issues like data heterogeneity and constrained computational resources, [269] suggests a novel FL paradigm for in-vehicle systems. The approach substantially reduces data inconsistencies and improves model generalization by introducing a robust aggregation procedure based on the L2 norm between LLM increments. Furthermore, the system lowers communication and computing costs while maintaining privacy by incorporating Low-Rank Adaptation (LoRA) into parameter-efficient fine-tuning. Federated LLM-based autonomous driving (FLAD) [270] introduces federated hybrid data parallelism, combining FL with parallelism to maximize resource utilization and cloud-edge-vehicle collaborative architecture, adopting LLM reasoning capabilities. The model also leverages SWIFT, a mobility-aware two-phase scheduler integrating stability-based vehicle ordering with DQN optimization to generate adaptive pipeline configurations.

Similarly, the model also adopts a resilient, quick-recovery mechanism featuring preventive pipeline templates and edge-assisted backup strategies, ensuring training continuity under high mobility and network instability while reducing communication overhead. In order to mitigate data scarcity and preserve privacy, Federated Instruction Tuning (FIT) [271] utilizes FL to allow various data owners to collaborate on the training of a shared model. Similarly, the model also leverages the feature diversity technique that generates novel instructions to enrich textual and visual diversity. An LLM-based architecture [59] is designed to increase the operational efficiency, reliability, and safety of V2X driving assistance systems that use image recognition as combined data from multiple sensors to train different vehicle and lane detection models. Utilizing the advantage of FL, these training models can prevent data privacy issues in V2X driving assistance deployment.

6.5 Mobile Services

FLM at the edge in mobile services focuses on dialogue systems, personalized recommendations, and intelligent assistants by preserving privacy. The major goal lies in providing highly personalized experiences without leaking user data. In [272], the authors design a new framework by integrating FL with transformer models BST (Behavior Sequence Transformer) and BERT (Bidirectional Encoder Representations from Transformers). The Hierarchical FL framework driven by deep reinforcement learning for Large Language Models (HRL-FLLM) [73] is a DRL-powered hierarchical edge-cloud FL system. To overcome device heterogeneity and resource limitations, the HRL-FLLM framework works through LLM fine-tuning that collaboratively integrates several efficiency-enhancing strategies, such as quantization, LoRA, and sparsification. DRL is a policy learning technique to solve the optimization problem of the Markov decision process, which is formulated through the joint control of client-side training and communication.

In [273], FL integrates a distribution generator and autoencoder to form a joint learning framework founded on the ideas of maximum-minimum fairness and its opportunity. In order to manage cross-basin variability without disclosing raw data, Federated Learning Multi-Station Fusion (FedMSF) combines local models with global aggregation [274]. To improve structural consistency, a Multi-Objective Aggregation technique (FedMOA) is designed, which adaptively weights clients, while a proximal term lessens training bias

under distributional alterations. Furthermore, the representation of uncommon occurrences is enhanced by extreme-aware data augmentation, and hyperparameter optimization is guided by an AI-driven LLM. The Privacy-Preserving Federated Content Representation (PFCR) architecture, presented by Guo et al. [275], integrates prompt-based content representation, secure gradient encryption, and FL. Through the use of shared feature spaces and textual item descriptions, PFCR improves cross-domain recommendation while reducing privacy issues. A federated spatiotemporal foundation model for cross-regional weather forecasting is created by Chen et al. [276]. By enhancing adaptation to low-resource and heterogeneous sensors, prompt learning permits cooperative forecasting without jeopardizing sensitive meteorological data.

7 Challenges and Future Research Direction

7.1 Challenges

Previous sections have shown that integrating FL with LLM is a novel research strategy as an emerging research area. This section highlights multiple challenges and potential research solutions to address these challenges.

7.1.1 Efficiency of FLM:

LLMs possess billions of parameters, making training and deployment extremely demanding on resource-limited clients. Recent FLM studies have focused on enhancing efficiency through Parameter-Efficient Fine-Tuning (PEFT) strategies, including LoRA and Adapter for fine-tuning LLMs. Similarly, backward propagation-free techniques are leveraged to enhance system performance and reduce computational costs. Beyond these strategies, there are complementary methods to further augment the efficiency:

- Model Architecture Optimization: Lightweight architectural innovations based on resource-efficient models can be integrated, including dynamic modeling approaches adapted from federated spatiotemporal forecasting [277], attention mechanisms, and sparse attention variants [278], to reduce memory and computational overhead.
- Model compression techniques: Techniques such as quantization [279], pruning [280, 281], and knowledge distillation [282] can reduce model size while maintaining competitive performance. These methods are valuable in federated settings where communication cost directly affects model size.
- Infrastructure and System Optimization: Advances in distributed training, including KV Cache [283], time-budgeted inference scheduling [284] and specialized edge hardware can alleviate the computational overhead of FLM training and inference.

However, most of the existing efficiency-enhancing approaches were designed for centralized LLM training. The integration of distributed computing-based FL introduces optimization challenges, such as synchronization constraints, heterogeneous resource availability, and communication overhead. Systematic evaluation of these hybrid strategies in a federated context remains an open research challenge.

7.1.2 Heterogeneity and Scalability Issue

Deploying FLM in large-scale scenarios poses a significant challenge due to the potential notable variation in model structures, model distributions, system edge clients, and communication networks. Device and data heterogeneity are the most common topics studied by many researchers. When there exists variation in local data distribution, the global optimization objective is not aligned with local optimization objectives. This leads local models to converge in opposite directions, attaining local optima rather than global optima, which finally reduces the FL performance. Given the case of FLM, LLM can access tremendous data for training, and data heterogeneity has a huge effect on the fine-tuning and training process. However, the generalization performance of the model may also be enhanced by data diversity. Currently, there is a dearth of research on heterogeneity that ranges from data, device, communication, and model heterogeneity. It is imperative to look more closely at how data heterogeneity and other forms of heterogeneity affect federated LLMs. The heterogeneity issue creates the handling and processing of the inputs from multiple devices on the network, which becomes more tedious and resource-intensive. Techniques, including effective network protocols and load balancing techniques [285], are used to tackle the scalability issue to handle rising computational and traffic demands. Despite these initiatives, FLM still faces scalability challenges. Achieving scalability continues to be a significant challenge as the number of FL participants and the complexity of the FLMs rise, impacting the

responsiveness and speed of fine-tuning and model training in FLM.

7.1.3 Robustness and Latency Issue

The resilience of LLM and FL against adversarial attacks has been the subject of extensive and in-depth research. There is, however, a dearth of research that takes FLM into account. It is clear that FLM systems are larger and more intricate than traditional FL and LLM systems. As a result, adversaries will probably have more chances to launch malicious attacks and take advantage of security flaws in FLM systems. Therefore, a thorough evaluation of FLM's susceptibility to possible attacks is crucial. The effects of hostile attacks, including backdoors, Byzantine attacks, and potential new attack techniques, should be investigated in this assessment.

Furthermore, from the standpoint of FL, a comprehensive assessment of the current defenses against new threats is also required. The efficiency of post-training detection techniques and strong aggregation mechanisms in fending off these novel dangers should be part of this assessment. Similarly, one major challenge in FLM is the length of time it takes to send huge volumes of parameters between clients and the FL server. When working with complicated models and a large number of FL clients who are frequently spread out geographically, this latency becomes more noticeable. Techniques like network protocol optimization and effective data serialization [286] have been studied to mitigate the latency issues. Additionally, lowering the round-trip time for parameter transmission is possible through the use of edge computing [37], where data processing takes place closer to the sources. Despite these initiatives, FLM still faces the problem of latency and robustness, which frequently affects the overall responsiveness and speed of the model training and fine-tuning procedures.

7.1.4 Security and Privacy in Federated LLM

Security and privacy concerns are a major barrier to FLM enabled by 5G and beyond (B5G) advanced network technologies. System heterogeneity results from differences in processing and communication capacities across various network members, which are caused by differences in hardware (such as CPU and GPU), network connections, and energy resources [47]. The heterogeneity (data, device, communication, and model heterogeneity) makes FLM more vulnerable and inconsistent [75]. Similarly, [222] shows the unreliable devices lead to Byzantine failures, which show the scenarios where specific network nodes behave maliciously or defectively, greatly affect the security of FLM. The complexity of protecting against assaults and the reliability of the system increase with the heterogeneity. Despite the various research on traditional FL and LLM, privacy concerns pose significant challenges after integration into FLM.

Though FL supports privacy by exchanging model updates rather than raw data, the vulnerabilities still exist during interactions among FL participants. For example, adversaries can leverage these vulnerabilities to perform gradient or inference attacks to extract local training data from devices. Multiparty computation (SMC) and homomorphic encryption (HE) are some approaches to prevent the attacks; however, they are unable to prevent them completely, though they can prevent leakage. Similarly, on the deployment of large-scale FLMs, they are also ineffective due to their computational cost and communication overhead.

These intrinsic flaws in FLM still exist despite continuous attempts to improve FLM security and privacy [7]. Despite these initiatives, FLM still faces the problem of security and privacy, which frequently affects the overall security and privacy of the model training and fine-tuning procedures.

7.2 Future Research Direction

7.2.1 Machine Unlearning and Data Erasure

LLMs are exceptionally powerful; the model requires a tremendous amount of data to train and fine-tune, which has raised privacy, bias, copyright issues, and accuracy limitations that have raised significant concerns. Legal frameworks such as GDPR and CCPA have emphasized the "right to be forgotten," requiring the mechanisms to remove the specific data from trained models, motivating techniques capable of erasing data without revealing what was removed [288]. This requirement has led to increased interest in machine unlearning techniques and federated unlearning approaches[287]. In order to address and enhance the accountability, transparency, and unlearning capabilities, [289] proposed a blockchain-based FLM. The method uses a novel unlearning function that integrates with the FL mechanism and uses blockchain technology to establish a tamper-proof record for each model's contributions. Security, verifiability, and transparency of the unlearning process are incorporated with Hyperledger Fabric and examine how LoRA hyperparameters affect the unlearning process. In federated environments, data erasure

is particularly challenging because training data is distributed across multiple devices rather than stored centrally [290]. Future research should focus on developing distributed unlearning methods that enable efficient data removal while preserving model performance and maintaining the decentralized nature of FLM.

7.2.2 *Continual Learning for FLM*

In dynamic real-world environments, task objectives and data distribution evolve over time, making continual learning essential for FLM systems [291]. Unlike traditional approaches, continual learning allows models to incorporate new knowledge incrementally without retraining from scratch. However, LLMs are prone to catastrophic forgetting when updated continuously, especially in federated settings with irregular client participation. Additionally, repeated retraining is computationally expensive for heterogeneous edge devices. Future studies should look at parameter-efficient techniques that support scalable global updates and allow for the retention of local knowledge in order to get over these restrictions. PEFT-based modular updates [292] and elastic weight consolidation (EWC)-based approach shows reduced forgetting without requiring prohibitive overhead. Furthermore, there is still a need for study into creating lifetime personalization procedures that can adjust to the changing workload distribution of each client in the presence of non-IID and sporadic data availability. Creating such protocols necessitates striking a balance between model stability over diverse learning trajectories, privacy preservation, and communication efficiency. Long-term deployment in dynamic real-world scenarios, such as personalized healthcare, changing legal compliance systems, or lifelong learning aids, will ultimately depend on the ability of FLM to provide continuous adaptation.

7.2.3 *Multimodal Model Integration*

The emerging real-world applications require multi-modal abilities, including integration of speech, visual, and sensor modalities; however, the existing approaches, including attacks and defenses studied for federated language models, remain largely focused on text-based tasks [146]. Beyond text-based models, large multi-modal approaches, such as Latent Diffusion Models [293], LLaVA [294], Vision Transformers (ViTs) [295], and GPT-4V [296] are rapidly advancing and have shown better performance in centralized systems. They contain a large number of parameters and require extensive computational resources and training data. FLMs are suitable for deployment; however, deploying on decentralized FLMs poses numerous challenges related to privacy, heterogeneity, and robustness. Future studies should provide flexible and modular tuning frameworks that enable each modality to be adjusted separately on client devices in order to overcome these challenges. Effective local adaptation is made possible by decoupling, which eliminates the need for every modality to be present on every client. Additionally, modality-aware aggregation procedures can improve global model performance and lessen imbalance by weighting client contributions according to modality type, data quality, and semantic consistency. Federated cross-modal contrastive learning [297] is another promising approach that can enhance multimodal alignment without necessitating the exchange of raw data. Last but not least, designing lightweight multimodal architectures using methods such as dynamic subnetwork activation [298] or knowledge distillation [81] that balance accuracy and resource efficiency is crucial to enabling deployment in edge-centric applications like wearable systems, assistive robotics, and smart healthcare.

7.2.4 *Model Security of FLM*

Ensuring model security has been the primary task as federated fine-tuning gained momentum. FLM requires pre-trained models (open-source or proprietary) to be sent to dispersed clients for local fine-tuning, which inevitably raises the possibility of system vulnerabilities and intellectual property (IP) leaks. Protecting the IP of valuable models and making sure that open-source models are securely deployed on edge devices are the two main components of model security in this context. IP protection is particularly important in FedLLM deployment due to the pre-trained LLM's strategic and financial relevance. For instance, training models comparable in scale to Gemini Ultra and GPT-4 [18] will cost millions, even as their downstream performance is increasingly evaluated in high-stakes domains such as emergency medicine [305]. Model watermarking [304], encrypted model delivery, or inference-obfuscation protocols that enable users to adjust and utilize the model without disclosing private architectural or parameter data must be developed in order to overcome this difficulty. Due to the accessibility and adaptability, open-source LLMs (such as DeepSeek [299] and Qwen, whose derivative models have been applied to tasks such as content-preserving style transfer [300]) are extensively used in FLM. In reality, most clients might not be able

to implement these models safely.

For example, it has been discovered that insecure default configurations of frameworks such as Ollama expose users to data leakage and unauthorized resource utilization [301], underscoring the broader need for robust cryptographic safeguards in distributed model serving [302, 303]. These vulnerabilities are exacerbated in federated setups, where a single exploited client can spread adversarial backdoors to the global model or leak locally adjusted training data. In delicate industries like healthcare and finance, where hacks could expose protected health information or unique trading techniques, the repercussions are especially dire. Future FLM research should focus on including secure model deployment techniques into the federated fine-tuning process in order to reduce these vulnerabilities and control unwanted access and manipulation. Strategies, including encrypted model distribution, runtime access control, and confidential computing for safe execution on edge devices, should be deployed.

8 Conclusion

This article provides a comprehensive, systematic overview of current developments in the integration of LLM with FL in a distributed paradigm. First, we provide a brief overview of LLM, FL, and edge, covering their fundamental processes, shared architectures, and algorithms. Then, we present the motivation for integrating LLM with FL in an edge environment from various perspectives and the benefits they can bring to various applications. Similarly, we highlight various approaches for making the integrated approach robust and discuss the biases in FLM. Additionally, numerous applications where FLM can be implemented were highlighted. Lastly, based on a thorough analysis of previous efforts, we highlight the open issues and challenges along with the future directions to overcome the limitations of FLM.

Data Availability Statement

Not applicable.

Funding

This work was supported in part by the National Foreign Expert Program under Grant Y20250133 and Grant Y20250135; in part by the National Natural Science Foundation of China under Grant 62572103 and Grant 62372087.

Conflicts of Interest

Deepak Adhikari and Negalign Wake Hundera served as Editorial Board Members, and Hu Xiong served as a Co-Editor-in-Chief of *Journal of Reliable and Secure Computing* at the time of manuscript submission. To ensure the integrity of the peer-review process, none of these authors was involved in the editorial handling, peer review, or decision-making process for this manuscript. The manuscript was handled independently by another editor. The remaining authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Li, S., Ngai, E. C.-H., & Voigt, T. (2024). Byzantine-Robust Aggregation in Federated Learning Empowered Industrial IoT. *IEEE Transactions on Industrial Informatics, 19*(2), 1165-1175. [CrossRef]

[2] Ullah, I., Adhikari, D., Su, X., Palmieri, F., Wu, C., & Choi, C. (2025). Integration of data science with the intelligent IoT (IIoT): current challenges and future perspectives. *Digital Communications and Networks, 11*(2), 280-298. [CrossRef]

[3] Imteaj, A., Thakker, U., Wang, S., Li, J., & Amini, M. H. (2022). A survey on federated learning for resource-constrained IoT devices. *IEEE Internet of Things Journal, 9*(1), 1-24. [CrossRef]

[4] Adhikari, D., Jiang, W., Zhan, J., Rawat, D. B., & Bhattarai, A. (2024). Recent advances in anomaly detection in Internet of Things: Status, challenges, and perspectives. *Computer Science Review, 54*, 100665. [CrossRef]

[5] Ullah, I., Singh, S. K., Adhikari, D., Khan, H., Jiang, W., & Bai, X. (2025). Multi-Agent Reinforcement Learning for task allocation in the Internet of Vehicles: Exploring benefits and paving the future. *Swarm and Evolutionary Computation, 94*, 101878. [CrossRef]

[6] Jiang, W., He, Z., Zhan, J., Pan, W., & Adhikari, D. (2021). Research progress and challenges on application-driven adversarial examples: A survey. *ACM Transactions on Cyber-Physical Systems (TCPS), 5*(4), 1-25. [CrossRef]

[7] Cheng, Y., Zhang, W., Zhang, Z., Zhang, C., Wang, S., & Mao, S. (2024). Toward federated large language models: Motivations, methods, and future directions.

IEEE Communications Surveys & Tutorials, 27(4), 2733-2764. [CrossRef]

[8] Mali, S., Zeng, F., Adhikari, D., Ullah, I., Al-Khasawneh, M. A., Alfarraj, O., & Alblehai, F. (2025). Federated reinforcement learning-based dynamic resource allocation and task scheduling in edge for IoT applications. *Sensors, 25*(7), 2197. [CrossRef]

[9] Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y. C., Yang, Q., ... & Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey. *IEEE communications surveys & tutorials, 22*(3), 2031-2063. [CrossRef]

[10] Alwis, C. D., Aouedi, O., Xu, J., Wang, S., Siriwardhana, Y., Hewa, T., ... & Liyanage, M. (2026). Federated Learning for 6G Security: A Survey on Threats, Solutions and Research Directions. [CrossRef]

[11] Zhang, Z., Rath, S., Xu, J., & Xiao, T. (2026). Federated learning for smart grid: A survey on applications and potential vulnerabilities. *ACM transactions on cyber-physical systems, 10*(1), 1-26. [CrossRef]

[12] Lu, J., Zhang, H., Zhou, P., Wang, X., Wang, C., & Wu, D. O. (2024). Fedlaw: Value-aware federated learning with individual fairness and coalition stability. *IEEE Transactions on Emerging Topics in Computational Intelligence, 9*(1), 1049-1062. [CrossRef]

[13] Mei, Q., Guo, W., Zhao, Y., Nie, L., & Adhikari, D. (2025). Blockchain-based privacy-preserving incentive scheme for internet of electric vehicle. *Information Fusion, 115*, 102732. [CrossRef]

[14] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. [CrossRef]

[15] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of machine learning research, 24*(240), 1-113. https://www.jmlr.org/papers/v24/22-1144.html

[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186). [CrossRef]

[17] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. [CrossRef]

[18] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. [CrossRef]

[19] Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2023). A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys, 56*(3), 1-37. [CrossRef]

[20] Liu, H., Peng, P., Chen, T., Wang, Q., Yao, Y., & Hua, X. S. (2023). Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network. *IEEE Transactions on Multimedia, 25*, 8580-8592. [CrossRef]

[21] Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., & Kambhampati, S. (2023). Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems, 36*, 38975-38987.

[22] Liu, J., Xia, C. S., Wang, Y., & Zhang, L. (2023). Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in neural information processing systems, 36*, 21558-21572.

[23] Liu, S., Zhang, R., Ma, R., Deng, Y., Zhu, L., Li, J., ... & Du, M. (2026). LLM Agents in Law: Taxonomy, Applications, and Challenges. *arXiv preprint arXiv:2601.06216*. [CrossRef]

[24] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems, 36*, 46595-46623.

[25] Lasnier, T., Zebaze, A., Seddah, D., Bawden, R., & Sagot, B. (2026). Disentangling meaning from language in LLM-based machine translation. *arXiv preprint arXiv:2602.04613*. [CrossRef]

[26] Roychowdhury, S. (2024, March). Journey of hallucination-minimized generative ai solutions for financial decision makers. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (pp. 1180-1181). [CrossRef]

[27] Zhang, Z., Cai, D., Zhang, Y., Xu, M., Wang, S., & Zhou, A. (2024, April). FedRDMA: Communication-efficient cross-silo federated LLM via chunked RDMA transmission. In *Proceedings of the 4th Workshop on Machine Learning and Systems* (pp. 126-133). [CrossRef]

[28] Woisetschläger, H., Erben, A., Wang, S., Mayer, R., & Jacobsen, H. A. (2024, June). Federated fine-tuning of llms on the very edge: The good, the bad, the ugly. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning* (pp. 39-50). [CrossRef]

[29] Ren, C., Yu, H., Peng, H., Tang, X., Zhao, B., Yi, L., ... & Yang, Q. (2025). Advances and Open Challenges in Federated Foundation Models. *IEEE Communications Surveys and Tutorials, 28*, 2087-2126. [CrossRef]

[30] He, A., Pan, H., Dai, Y., Si, X., Yuen, C., & Zhang, Y. (2024). ADMM for mobile edge intelligence: A survey. *IEEE Communications Surveys & Tutorials,*

27(5), 3020-3057. [CrossRef]

[31] Puppala, S., Hossain, I., Alam, J., Ahad, T., & Talukder, S. (2025). A Comprehensive Survey of Federated Learning for Edge AI: Recent Trends and Future Directions. [CrossRef]

[32] Evgenidis, N. G., Mitsiou, N. A., Koutsioumpa, V. I., Tegos, S. A., Diamantoulakis, P. D., & Karagiannidis, G. K. (2024). Multiple access in the era of distributed computing and edge intelligence. *Proceedings of the IEEE, 112*(9), 1497-1526. [CrossRef]

[33] Edemacu, K., & Wu, X. (2025). Privacy preserving prompt engineering: A survey. *ACM Computing Surveys, 57*(10), 1-36. [CrossRef]

[34] Li, S., Ye, F., Fang, M., Zhao, J., Chan, Y. H., Ngai, E. C., & Voigt, T. (2024). Synergizing foundation models and federated learning: A survey. *arXiv preprint arXiv:2406.12844*. [CrossRef]

[35] Wang, R., Gao, Z., Zhang, L., Yue, S., & Gao, Z. (2025). Empowering large language models to edge intelligence: A survey of edge efficient LLMs and techniques. *Computer Science Review, 57*, 100755. [CrossRef]

[36] Qu, G., Chen, Q., Wei, W., Lin, Z., Chen, X., & Huang, K. (2025). Mobile edge intelligence for large language models: A contemporary survey. *IEEE Communications Surveys & Tutorials, 27*(6), 3820-3860. [CrossRef]

[37] Li, X., Li, H., Sun, C., Fan, Q., Han, Z., & Leung, V. (2026). Edge-Enhanced Intelligence: A Comprehensive Survey of Large Language Models and Edge-Cloud Computing Synergy. *IEEE Communications Surveys & Tutorials, 28*, 1248-1284. [CrossRef]

[38] Khan, L. U., Saad, W., Han, Z., Hossain, E., & Hong, C. S. (2021). Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials, 23*(3), 1759-1799. [CrossRef]

[39] Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., ... & Yu, P. S. (2022). Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems, 35*(7), 8726-8746. [CrossRef]

[40] Le, M., Huynh-The, T., Do-Duy, T., Vu, T. H., Hwang, W. J., & Pham, Q. V. (2024). Applications of distributed machine learning for the Internet-of-Things: A comprehensive survey. *IEEE Communications Surveys & Tutorials, 27*(2), 1053-1100. [CrossRef]

[41] Rodríguez-Barroso, N., Jiménez-López, D., Luzón, M. V., Herrera, F., & Martínez-Cámara, E. (2023). Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion, 90*, 148-173. [CrossRef]

[42] Ficco, M., Guerriero, A., Milite, E., Palmieri, F., Pietrantuono, R., & Russo, S. (2024). Federated learning for IoT devices: Enhancing TinyML with on-board training. *Information Fusion, 104*, 102189. [CrossRef]

[43] Ghimire, B., & Rawat, D. B. (2022). Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things. *IEEE Internet of Things Journal, 9*(11), 8229-8249. [CrossRef]

[44] Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2023). Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal, 11*(5), 7374-7398. [CrossRef]

[45] Chatterjee, P., Das, D., & Rawat, D. B. (2024). Securing financial services with federated learning and blockchain. In *Secure and Smart Cyber-Physical Systems* (pp. 178-207). CRC Press. https://www.taylorfrancis.com/chapters/edit/10.1201/9781003376712-9

[46] Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys, 57*(6), 1-39. [CrossRef]

[47] Wang, S., Zhu, T., Liu, B., Ding, M., Ye, D., Zhou, W., & Yu, P. (2025). Unique security and privacy threats of large language models: A comprehensive survey. *ACM Computing Surveys, 58*(4), 1-36. [CrossRef]

[48] He, F., Zhu, T., Ye, D., Liu, B., Zhou, W., & Yu, P. S. (2025). The emerged security and privacy of llm agent: A survey with case studies. *ACM Computing Surveys, 58*(6), 1-36. [CrossRef]

[49] Ye, P., Ren, H., Li, Z., Yan, A., Yan, H., Wang, S., & Li, J. (2026). Securing large language models: A survey of watermarking and fingerprinting techniques. *ACM Computing Surveys, 58*(7), 1-35. [CrossRef]

[50] Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., ... & Liu, Q. (2023). Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*. [CrossRef]

[51] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences, 68*(2), 121101. [CrossRef]

[52] Krasniqi, R., Xu, D., & Vieira, M. (2025). SE Perspective on LLMs: Biases in Code Generation, Code Interpretability, and Code Security Risks. *ACM Computing Surveys, 58*(5), 1-16. [CrossRef]

[53] Kim, G. I., Hwang, S., & Jang, B. (2025). Efficient compressing and tuning methods for large language models: A systematic literature review. *ACM Computing Surveys, 57*(10), 1-39. [CrossRef]

[54] Bayer, M., Kuehn, P., Shanehsaz, R., & Reuter, C. (2024). Cysecbert: A domain-adapted language model for the cybersecurity domain. *ACM Transactions on Privacy and Security, 27*(2), 1-20. [CrossRef]

[55] Ignaczak, L., Goldschmidt, G., Costa, C. A. D., & Righi, R. D. R. (2021). Text mining in cybersecurity: A systematic literature review. *ACM Computing Surveys (CSUR), 54*(7), 1-36. [CrossRef]

[56] Sheng, Z., Chen, Z., Gu, S., Huang, H., Gu, G., & Huang, J. (2025). Llms in software security: A survey of vulnerability detection techniques and insights. *ACM Computing Surveys, 58*(5), 1-35. [CrossRef]

[57] Lin, Z., Qu, G., Chen, Q., Chen, X., Chen, Z., & Huang, K. (2025). Pushing large language models to the 6g edge: Vision, challenges, and opportunities. *IEEE Communications Magazine, 63*(9), 52-59. [CrossRef]

[58] Dong, X. L., Moon, S., Xu, Y. E., Malik, K., & Yu, Z. (2023, August). Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5792-5793). [CrossRef]

[59] Chen, J., Messou, F. J. A., Zhang, S., Liu, T., Yu, K., & Niyato, D. (2025, June). Federated Fine-Tuning of Large Language Models for Intelligent Automotive Systems with Low-Rank Adaptation. In *2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring)* (pp. 1-6). IEEE. [CrossRef]

[60] Liang, C., Zuo, S., Zhang, Q., He, P., Chen, W., & Zhao, T. (2023, July). Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning* (pp. 20852-20867). PMLR.

[61] Piccialli, F., Chiaro, D., Qi, P., Bellandi, V., & Damiani, E. (2025). Federated and edge learning for large language models. *Information fusion, 117*, 102840. [CrossRef]

[62] Yu, S., Muñoz, J. P., & Jannesari, A. (2024, May). Federated foundation models: Privacy-preserving and collaborative learning for large models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 7174-7184). https://aclanthology.org/2024.lrec-main.630/

[63] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

[64] Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence, 4*(3), 258-268. [CrossRef]

[65] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences, 103*, 102274. [CrossRef]

[66] Yang, W., Liew, Z. Q., Lim, W. Y. B., Xiong, Z., Niyato, D., Chi, X., ... & Letaief, K. B. (2022). Semantic communication meets edge intelligence. *IEEE wireless communications, 29*(5), 28-35. [CrossRef]

[67] Tu, J., Yang, L., & Cao, J. (2025). Distributed machine learning in edge computing: Challenges, solutions and future directions. *ACM Computing Surveys, 57*(5), 1-37. [CrossRef]

[68] Friha, O., Ferrag, M. A., Kantarci, B., Cakmak, B., Ozgun, A., & Ghoualmi-Zine, N. (2024). Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. *IEEE Open Journal of the Communications Society, 5*, 5799-5856. [CrossRef]

[69] Wang, X., Xu, Z., & Sui, X. (2025). Intelligent data analysis in edge computing with large language models: applications, challenges, and future directions. *Frontiers in Computer Science, 7*, 1538277. [CrossRef]

[70] Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., ... & Xu, M. (2024). Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*. [CrossRef]

[71] Tian, C., Tam, K., Wu, Y., Zhong, S., Li, L., Lane, N. D., & Xu, C. (2026). Floe: Federated Specialization for Real-Time LLM–SLM Inference. *IEEE Transactions on Parallel and Distributed Systems, 37*(7), 1630-1646. [CrossRef]

[72] Khan, N. M., Bhattacharya, P., Roy, S., Shetty, S., Gadekallu, T. R., & Srivastava, G. (2025, December). SLM-FARL: Small Language Model Driven Federated Reinforcement Multi-Agentic Framework underlying 6G Edge Networks. In *GLOBECOM 2025-2025 IEEE Global Communications Conference* (pp. 2505-2510). IEEE. [CrossRef]

[73] Chen, H., Yuan, X., & Li, H. (2026). Edge-Assisted Federated Learning for Large Language Models in IoT Sensor Systems. *IEEE Journal of Selected Areas in Sensors, 3*, 125-138. [CrossRef]

[74] Shen, Y., Shao, J., Zhang, X., Lin, Z., Pan, H., Li, D., ... & Letaief, K. B. (2024). Large language models empowered autonomous edge AI for connected intelligence. *IEEE Communications Magazine, 62*(10), 140-146. [CrossRef]

[75] Lin, Z., Chen, Z., Chen, X., Ni, W., & Gao, Y. (2026). HASFL: Heterogeneity-aware split federated learning over edge computing systems. *IEEE Transactions on Mobile Computing*. [CrossRef]

[76] Ahad, A., Ahmed, K. I., Ullah, F., Sheikh, M. A., Tahir, M., Hayajneh, M., & Pires, I. M. (2026). Federated Learning and 5G/6G-Based Internet of Medical Things (IoMT): Applications, Key Enabling Technologies, Open Issues and Future Research Directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 16*(1), e70065. [CrossRef]

[77] Wei, W., Lin, Z., Liu, X., Du, H., Niyato, D., & Chen,

X. (2026). Optimizing split federated learning with unstable client participation. *IEEE Transactions on Mobile Computing*. [CrossRef]

[78] Guo, P., Wang, Y., Li, W., Liu, M., Li, M., Zheng, J., & Qu, L. (2025). Exploring federated pruning for large language models. *arXiv preprint arXiv:2505.13547*. [CrossRef]

[79] Jia, Y., Zhang, X., Hu, H., Choo, K. K. R., Qi, L., Xu, X., ... & Dou, W. (2024). Dapperfl: Domain adaptive federated learning with model fusion pruning for edge devices. *Advances in Neural Information Processing Systems, 37*, 13099-13123.

[80] Hendriks, D., Spitzer, P., Kühl, N., & Satzger, G. (2026). Honey, i shrunk the language model: Impact of knowledge distillation methods on performance and explainability. *IEEE Transactions on Knowledge and Data Engineering, 38*(6), 3710-3721. [CrossRef]

[81] Fang, L., Yu, X., Cai, J., Chen, Y., Wu, S., Liu, Z., ... & Ma, P. (2026). Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *Artificial Intelligence Review, 59*(1), 17. [CrossRef]

[82] Song, I., & Lee, K. (2026). BitLoRA: Quantization-Compatible Adapter Tuning for 1.58-bit LLM in Federated On-Device AI-Agent. *Expert Systems with Applications*, 131397. [CrossRef]

[83] Kuzmin, A., Nagel, M., Van Baalen, M., Behboodi, A., & Blankevoort, T. (2023). Pruning vs quantization: Which is better?. *Advances in neural information processing systems, 36*, 62414-62427.

[84] Zheng, H., Shen, L., Tang, A., Luo, Y., Hu, H., Du, B., ... & Tao, D. (2025). Learning from models beyond fine-tuning. *Nature Machine Intelligence, 7*(1), 6-17. [CrossRef]

[85] Xu, L., Xie, H., Qin, S. J., Tao, X., & Wang, F. L. (2026). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 48*(6), 6107-6126. [CrossRef]

[86] Lin, M., Yu, F., Ning, R., Li, L., Chen, J., Lou, Q., ... & Wu, H. (2026). RPP: A Certified Poisoned-Sample Detection Framework for Backdoor Attacks under Dataset Imbalance. *arXiv preprint arXiv:2602.00183*. [CrossRef]

[87] Li, R., Wang, J., Chen, H., Ding, H., Zhou, J., & Tao, R. (2026, March). Dormant Backdoor: Weaponizing Model Finetuning for Feasible Backdoor Attacks Against Pretrained Models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 27, pp. 23132-23140). [CrossRef]

[88] Rodríguez-Barroso, N., Luzón, M. V., & Herrera, F. (2026). RAB2-DEF: Dynamic and Explainable Defense Against Adversarial Attacks in Federated Learning to Fair Poor Clients. *Machine Intelligence Research, 23*(1), 133-146. [CrossRef]

[89] Wang, H., Yin, Z., Chen, B., Zeng, Y., Yan, X., Zhou, C., & Li, A. (2025). Rofed-llm: robust federated learning for large language models in adversarial wireless environments. *IEEE Transactions on Network Science and Engineering, 13*, 1084-1096. [CrossRef]

[90] Fang, M., Cao, X., Jia, J., & Gong, N. (2020). Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)* (pp. 1605-1622). https://www.usenix.org/conference/usenixsecurity20/presentation/fang

[91] Yazdinejad, A., Dehghantanha, A., Karimipour, H., Srivastava, G., & Parizi, R. M. (2024). A robust privacy-preserving federated learning model against model poisoning attacks. *IEEE Transactions on Information Forensics and Security, 19*, 6693-6708. [CrossRef]

[92] Kasyap, H., & Tripathy, S. (2024). Beyond data poisoning in federated learning. *Expert Systems with Applications, 235*, 121192. [CrossRef]

[93] Oliynyk, D., Mayer, R., & Rauber, A. (2023). I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys, 55*(14s), 1-41. [CrossRef]

[94] Li, Y., Zhu, L., Jia, X., Jiang, Y., Xia, S. T., & Cao, X. (2022, June). Defending against model stealing via verifying embedded external features. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 2, pp. 1464-1472). [CrossRef]

[95] Yang, W., Li, L., Zhang, Z., Ren, X., Sun, X., & He, B. (2021, June). Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in NLP models. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 2048-2058). [CrossRef]

[96] Wu, T. Y., Wu, H., Tang, M., Kumari, S., & Chen, C. M. (2025). CD-AKA-IoV: A Provably Secure Cross-Domain Authentication and Key Agreement Protocol for Internet of Vehicle. *Computers, Materials & Continua, 85*(1). [CrossRef]

[97] Wang, Z., Shen, Z., He, Y., Sun, G., Wang, H., Lyu, L., & Li, A. (2024). Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems, 37*, 22513-22533.

[98] Vahidian, S., Morafah, M., Chen, C., Shah, M., & Lin, B. (2023). Rethinking data heterogeneity in federated learning: Introducing a new notion and standard benchmarks. *IEEE Transactions on Artificial Intelligence, 5*(3), 1386-1397. [CrossRef]

[99] Tirana, J., Tsigkari, D., Noguero, D. S., & Kourtellis, N. (2026, March). Data heterogeneity and forgotten labels in split federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 31, pp. 25940-25948). [CrossRef]

[100] Zhu, H., Togo, R., Ogawa, T., & Haseyama, M.

(2024, April). Prompt-based personalized federated learning for medical visual question answering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1821-1825). IEEE. [CrossRef]

[101] Zeng, M., Tu, W., Chen, Y., Wang, Y., Yu, M., Tang, X., & Cheng, J. (2026, March). FedPKDA: Personalized Federated Learning with Privacy-Preserving Knowledge Dynamic Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 33, pp. 28113-28121). [CrossRef]

[102] Yi, L., Yu, H., Wang, G., Liu, X., & Hu, Q. (2026). pFedMoE: Data-Level Personalization With Mixture of Experts in Model-Heterogeneous Personalized Federated Learning. *IEEE Transactions on Knowledge and Data Engineering, 38*(3), 1905-1918. [CrossRef]

[103] Zhang, M., Yin, R., Yang, Z., & Wang, Y. (2025). Advances and challenges of multi-task learning method in recommender systems: A survey. *Neurocomputing*, 132510. [CrossRef]

[104] Vettoruzzo, A., Bouguelia, M. R., Vanschoren, J., Rögnvaldsson, T., & Santosh, K. C. (2024). Advances and challenges in meta-learning: A technical review. *IEEE transactions on pattern analysis and machine intelligence, 46*(7), 4763-4779. [CrossRef]

[105] Asad, M., & Otoum, S. (2025, October). FLChain-LLM: A Blockchain-Enabled Framework for Training Large Language Models. In *2025 7th International Conference on Blockchain Computing and Applications (BCCA)* (pp. 378-384). IEEE. [CrossRef]

[106] Wang, J., Yang, X., Cui, S., Che, L., Lyu, L., Xu, D. D., & Ma, F. (2023). Towards personalized federated learning via heterogeneous model reassembly. *Advances in Neural Information Processing Systems, 36*, 29515-29531.

[107] Chen, Y., Lu, W., Qin, X., Wang, J., & Xie, X. (2023). Metafed: Federated learning among federations with cyclic knowledge distillation for personalized healthcare. *IEEE Transactions on Neural Networks and Learning Systems, 35*(11), 16671-16682. [CrossRef]

[108] He, C., Li, S., So, J., Zeng, X., Zhang, M., Wang, H., ... & Avestimehr, S. (2020). Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*. [CrossRef]

[109] Kuang, W., Qian, B., Li, Z., Chen, D., Gao, D., Pan, X., ... & Zhou, J. (2024, August). 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* (pp. 5260-5271). [CrossRef]

[110] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., ... & Lane, N. D. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*. [CrossRef]

[111] Ye, R., Ge, R., Zhu, X., Chai, J., Du, Y., Liu, Y., ... & Chen, S. (2024). Fedllm-bench: Realistic benchmarks for federated learning of large language models. *Advances in Neural Information Processing Systems, 37*, 111106-111130.

[112] 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). [CrossRef]

[113] Fan, T., Kang, Y., Ma, G., Chen, W., Wei, W., Fan, L., & Yang, Q. (2023). Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*. [CrossRef]

[114] Zhang, J., Vahidian, S., Kuo, M., Li, C., Zhang, R., Yu, T., ... & Chen, Y. (2024, April). Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6915-6919). IEEE. [CrossRef]

[115] Zhang, Z., Yang, Y., Dai, Y., Wang, Q., Yu, Y., Qu, L., & Xu, Z. (2023, July). Fedpetuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 9963-9977). [CrossRef]

[116] Zhang, Z., Hu, X., Zhang, J., Zhang, Y., Wang, H., Qu, L., & Xu, Z. (2023, July). Fedlegal: The first real-world federated learning benchmark for legal nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 3492-3507). [CrossRef]

[117] Roth, H. R., Xu, Z., Hsieh, Y. T., Renduchintala, A., Yang, I. T. C., Zhang, Z., ... & Feng, A. (2025). Empowering federated learning for massive models with nvidia flare. In *Federated Learning Systems: Towards Privacy-Preserving Distributed AI* (pp. 1-17). Cham: Springer Nature Switzerland. [CrossRef]

[118] Nair, L., Bernadskiy, M., Madhavan, A., Chan, C., Basumallik, A., & Bunandar, D. (2023). INT-FP-QSim: Mixed precision and formats for large language models and vision transformers. *arXiv preprint arXiv:2307.03712*. [CrossRef]

[119] Shao, W., Chen, M., Zhang, Z., Xu, P., Zhao, L., Li, Z., ... & Luo, P. (2024, May). Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations* (Vol. 2024, pp. 45472-45496).

[120] Fan, H., Venieris, S. I., Kouris, A., & Lane, N. (2023, October). Sparse-dysta: Sparsity-aware dynamic and static scheduling for sparse multi-dnn workloads. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 353-366). [CrossRef]

[121] Zhang, Y., Li, P., Hong, J., Li, J., Zhang, Y.,

Zheng, W., ... & Chen, T. (2024). Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark. *arXiv preprint arXiv:2402.11592*. [CrossRef]

[122] Peng, H., Wu, K., Wei, Y., Zhao, G., Yang, Y., Liu, Z., ... & Cheng, P. (2023). Fp8-lm: Training fp8 large language models. *arXiv preprint arXiv:2310.18313*. [CrossRef]

[123] Xu, M., Cai, D., Wu, Y., Li, X., & Wang, S. (2024). {FwdLLM}: Efficient federated finetuning of large language models with perturbed inferences. In 2024 USENIX Annual Technical Conference (USENIX ATC 24) (pp. 579-596). https://www.usenix.org/conference/atc24/presentation/xu-mengwei

[124] Qin, R., Xia, J., Jia, Z., Jiang, M., Abbasi, A., Zhou, P., ... & Shi, Y. (2024, June). Enabling on-device large language model personalization with self-supervised data selection and synthesis. In *Proceedings of the 61st ACM/IEEE design automation conference* (pp. 1-6). [CrossRef]

[125] Han, S., Buyukates, B., Hu, Z., Jin, H., Jin, W., Sun, L., ... & He, C. (2024, August). Fedsecurity: A benchmark for attacks and defenses in federated learning and federated llms. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5070-5081). [CrossRef]

[126] In, S. W. O. T. (2001). Wikipedia: the free encyclopedia. *San Francisco (CA): Wikimedia Foundation, 2012*(16.4), 2012. https://wikipedia2007.classicistranieri.com/en/h/e/u/Heuristic.html

[127] Patel, J. M. (2020). *Getting structured data from the internet: running web crawlers/scrapers on a big data production scale*. Berkeley, CA: Apress. [CrossRef]

[128] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022, May). Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 320-335). [CrossRef]

[129] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision* (pp. 19-27). [CrossRef]

[130] Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., ... & Xiong, C. (2022). Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*. [CrossRef]

[131] Rakkini, M. J., & Geetha, K. (2022). BigQuery Open Dataset. *Soft Computing: Theories and Applications: Proceedings of SoCTA 2021*, 25.

[132] Zhang, Y., Gao, H., Chen, H., Li, W., Ma, Y., & Tresp, V. (2025). FedNano: Toward Lightweight Federated Tuning for Pretrained Multimodal Large Language Models. *arXiv preprint arXiv:2506.14824*. [CrossRef]

[133] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015, December). VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 2425-2433). IEEE. [CrossRef]

[134] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *Advances in Neural Information Processing Systems*, *32*.

[135] Sarmadi, A., Fu, H., Krishnamurthy, P., Garg, S., & Khorrami, F. (2023). Privacy-preserving collaborative learning through feature extraction. *IEEE Transactions on Dependable and Secure Computing*, *21*(1), 486-498. [CrossRef]

[136] Yang, H., Ge, M., Xue, D., Xiang, K., Li, H., & Lu, R. (2023). Gradient leakage attacks in federated learning: Research frontiers, taxonomy, and future directions. *IEEE Network, 38*(2), 247-254. [CrossRef]

[137] Ramos, P., Ramos, R., & Garcia, N. (2025). Data leakage in visual datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6309-6319).

[138] Huang, Y., Gupta, S., Song, Z., Li, K., & Arora, S. (2021). Evaluating gradient inversion attacks and defenses in federated learning. *Advances in neural information processing systems*, *34*, 7232-7241.

[139] Zhao, J., Zhu, H., Wang, F., Lu, R., & Li, H. (2023). Efficient and privacy-preserving tree-based inference via additive homomorphic encryption. *Information Sciences*, *650*, 119480. [CrossRef]

[140] Liu, H., Li, B., Gao, C., Xie, P., & Zhao, C. (2023). Privacy-encoded federated learning against gradient-based data reconstruction attacks. *IEEE Transactions on Information Forensics and Security*, *18*, 5860-5875. [CrossRef]

[141] Pan, X., Zhang, M., Ji, S., & Yang, M. (2020, May). Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 1314-1331). IEEE. [CrossRef]

[142] Sun, W., Wang, X., Liang, Z., Chen, J., Lan, W., Chen, Y., & Wang, F. (2026). FedShieldLLM: Measurement, Detection and Protection of Privacy Leakage in Federated LLMs. *IEEE Transactions on Mobile Computing*, *25*(7), 9612-9628. [CrossRef]

[143] Cao, X., Jia, J., Zhang, Z., & Gong, N. Z. (2023, May). Fedrecover: Recovering from poisoning attacks in federated learning using historical information. In *2023 IEEE Symposium on Security and Privacy (SP)* (pp. 1366-1383). IEEE. [CrossRef]

[144] Ma, Z., Deng, Y., Qiao, Z., Zhang, Q., Zhou, C., Wu, F., ... & Ren, J. (2026). A Fine-Tuning Data Recovery Attack on Generative Language Models via Backdooring. *IEEE Transactions on Information Forensics and Security*, *21*, 3006-3021. [CrossRef]

[145] Yue, K., Jin, R., Wong, C. W., Baron, D., &

Dai, H. (2023). Gradient obfuscation gives a false sense of security in federated learning. In *32nd USENIX security symposium (USENIX Security 23)* (pp. 6381-6398). https://www.usenix.org/conference/usenixsecurity23/presentation/yue

[146] Gupta, S., Huang, Y., Zhong, Z., Gao, T., Li, K., & Chen, D. (2022). Recovering private text in federated learning of language models. *Advances in neural information processing systems, 35*, 8130-8143.

[147] Chen, S., Luo, Y., Deng, G., Liu, Y., Xu, M., Fu, S., & Jia, X. (2026, April). Reconstructing Training Data from Adapter-based Federated Large Language Models. In *Proceedings of the ACM Web Conference 2026* (pp. 2602-2613). [CrossRef]

[148] Wang, F., & Li, B. (2024). Data reconstruction and protection in federated learning for fine-tuning large language models. *IEEE Transactions on Big Data*. [CrossRef]

[149] Miao, Q., Sun, W., Zhu, D., Li, J., Zhou, Y., & Alcaraz, C. (2026). Moderation is the Best Policy: Dynamic Defense Against Gradient-Based Data Reconstruction Attacks in Federated Learning. *IEEE Transactions on Dependable and Secure Computing, 23*(3), 6230-6245. [CrossRef]

[150] Bai, L., Zhang, X., Zhang, S., Ye, Q., & Hu, H. (2025). Provfl: Property inference attacks against vertical federated learning. *IEEE Transactions on Information Forensics and Security, 20*, 6529-6543. [CrossRef]

[151] Tong, M., Chen, K., Zhang, J., Qi, Y., Zhang, W., Yu, N., ... & Zhang, Z. (2025). Inferdpt: Privacy-preserving inference for black-box large language models. *IEEE Transactions on Dependable and Secure Computing, 22*(5), 4625-4640. [CrossRef]

[152] Luo, X., Yu, T., & Xiao, X. (2025, November). Prompt inference attack on distributed large language model inference frameworks. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1739-1753). [CrossRef]

[153] Wang, Z., Huang, Y., Song, M., Wu, L., Xue, F., & Ren, K. (2022). Poisoning-assisted property inference attack against federated learning. *IEEE Transactions on Dependable and Secure Computing, 20*(4), 3328-3340. [CrossRef]

[154] Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., & Jiang, T. (2024). Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems, 37*, 134981-135010.

[155] He, Y., Li, B., Liu, L., Ba, Z., Dong, W., Li, Y., ... & Chen, C. (2025). Towards Label-Only Membership Inference Attack against Pre-trained Large Language Models. In *34th USENIX Security Symposium (USENIX Security 25)* (pp. 1609-1628). https://www.usenix.org/conference/usenixsecurity25/presentation/he-yu

[156] Bai, L., Hu, H., Ye, Q., Li, H., Wang, L., & Xu, J. (2024). Membership inference attacks and defenses in federated learning: A survey. *ACM Computing Surveys, 57*(4), 1-35. [CrossRef]

[157] Fu, C., Zhang, X., Ji, S., Chen, J., Wu, J., Guo, S., ... & Wang, T. (2022). Label inference attacks against vertical federated learning. In *31st USENIX security symposium (USENIX Security 22)* (pp. 1397-1414). https://www.usenix.org/conference/usenixsecurity22/presentation/fu-chong

[158] Li, Z., Wu, Y., Chen, Y., Tonin, F., Abad Rocamora, E., & Cevher, V. (2024). Membership inference attacks against large vision-language models. *Advances in Neural Information Processing Systems, 37*, 98645-98674.

[159] Shen, M., Meng, J., Peng, B., Tang, X., Wang, W., Niyato, D., & Zhu, L. (2026). Casper: A Causality-Inspired Defense With Confounder Against Label Inference Attacks in Vertical Split Federated Learning. *IEEE Transactions on Information Forensics and Security, 21*, 1050-1064. [CrossRef]

[160] Song, J., Yuan, J., Chen, G., Liu, Y., & Yang, N. (2026). LTMIA: a loss trajectory-based membership inference attack method in federated learning. *Journal of Information Security and Applications, 97*, 104364. [CrossRef]

[161] Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur), 51*(4), 1-35. [CrossRef]

[162] Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., & Liu, Y. (2020). BatchCrypt: Efficient homomorphic encryption for Cross-Silo federated learning. In *2020 USENIX annual technical conference (USENIX ATC 20)* (pp. 493-506). https://www.usenix.org/conference/atc20/presentation/zhang-chengliang

[163] Xie, Q., Jiang, S., Jiang, L., Huang, Y., Zhao, Z., Khan, S., ... & Wu, K. (2024). Efficiency optimization techniques in privacy-preserving federated learning with homomorphic encryption: A brief survey. *IEEE Internet of Things Journal, 11*(14), 24569-24580. [CrossRef]

[164] Zhao, C., Zhao, S., Zhao, M., Chen, Z., Gao, C. Z., Li, H., & Tan, Y. A. (2019). Secure multi-party computation: theory, practice and applications. *Information Sciences, 476*, 357-372. [CrossRef]

[165] Gamiz Ugarte, I., Regueiro Senderos, C., Lage Serrano, Ó., Jacob Taquet, E., & Astorga Burgo, J. (2025). Challenges and future research directions in secure multi-party computation for resource-constrained devices and large-scale computations. *International Journal of Information Security, 24*(1). [CrossRef]

[166] Tran, A.-T., Luong, T.-D., & Huynh, V.-N. (2026). PriFLRC: A secure multi-party computation-based privacy-enhanced federated learning scheme resilient to collusion. *Neurocomputing*, 132574. [CrossRef]

[167] Singh, J. P., Aqsa, A., Ghani, I., Sonani, R., & Govindarajan, V. (2025). Privacy-aware hierarchical federated learning in healthcare: integrating differential privacy and secure multi-party computation. *Future Internet, 17*(8), 345. [CrossRef]

[168] El Ouadrhiri, A., & Abdelhadi, A. (2022). Differential privacy for deep and federated learning: A survey. *IEEE Access, 10*, 22359-22380. [CrossRef]

[169] Truex, S., Liu, L., Chow, K. H., Gursoy, M. E., & Wei, W. (2020, April). LDP-Fed: Federated learning with local differential privacy. In *Proceedings of the third ACM international workshop on edge systems, analytics and networking* (pp. 61-66). [CrossRef]

[170] Yang, M., Guo, T., Zhu, T., Tjuawinata, I., Zhao, J., & Lam, K. Y. (2024). Local differential privacy and its applications: A comprehensive survey. *Computer Standards & Interfaces, 89*, 103827. [CrossRef]

[171] He, Y., Zhang, W., Wang, K., Lin, X., Zhang, Y., & Ni, W. (2026). Efficient and Effective Biclique Counting with Local Differential Privacy. *Proceedings of the ACM on Management of Data, 4*(1) (SIGMOD), 1-24. [CrossRef]

[172] Wang, X., Kim, B. G., Amoon, M., Kumar, S., & Liu, Z. (2025). Federated learning with local differential privacy for autonomous electronic vehicles: Enhancing security and performance. *IEEE Transactions on Consumer Electronics, 71*(2), 6147-6157. [CrossRef]

[173] Zhu, L., & Chen, X. (2025). Privacy protection in federated learning: a study on the combined strategy of local and global differential privacy. *The Journal of Supercomputing, 81*(1), 326. [CrossRef]

[174] Letafati, M., & Otoum, S. (2023, September). Global differential privacy for distributed metaverse healthcare systems. In *2023 International Conference on Intelligent Metaverse Technologies & Applications* (*iMETA*) (pp. 01-08). IEEE. [CrossRef]

[175] Li, H., Li, X., Liu, X., Wang, B., Wang, J., & Tian, Y. (2026). FedSam: Enhancing federated learning accuracy with differential privacy and data heterogeneity mitigation. *Computer Standards & Interfaces, 95*, 104019. [CrossRef]

[176] Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., ... & Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security, 15*, 3454-3469. [CrossRef]

[177] Naresh, V. S., & Ayyappa, D. (2026). Privacy-preserving federated credit risk models: evaluating differential privacy and homomorphic encryption techniques. Scientific Reports. [CrossRef]

[178] Liu, X. Y., Zhu, R., Zha, D., Gao, J., Zhong, S., White, M., & Qiu, M. (2025). Differentially private low-rank adaptation of large language model using federated learning. *ACM Transactions on Management Information Systems, 16*(2), 1-24. [CrossRef]

[179] Carillo, R., Cerasuolo, F., Bovenzi, G., Ciuonzo, D., & Pescapé, A. (2026). A Federated and Incremental Network Intrusion Detection System for IoT Emerging Threats. *IEEE Transactions on Network and Service Management, 23*, 3865-3880. [CrossRef]

[180] Pereira, L., & Nagasundaram, S. (2026, February). A Multi-Model Hybrid Framework for Twitter Spam Detection using LSTM, CNN and GAN. In *2026 International Conference on Electronics and Renewable Systems* (*ICEARS*) (pp. 1838-1843). IEEE. [CrossRef]

[181] Soysaldı Şahin, M., Şahin, D. Ö., & Salah, A. F. (2026). Revisiting SMS Spam Detection: The Impact of Feature Representation on Classical Machine Learning Models. *Electronics, 15*(4), 894. [CrossRef]

[182] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... & Jitsev, J. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems, 35*, 25278-25294.

[183] Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., & Wang, X. (2024). Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*. [CrossRef]

[184] Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., ... & Tramèr, F. (2024, May). Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy* (*SP*) (pp. 407-425). IEEE. [CrossRef]

[185] Qu, Y., Li, B., Huang, S., Nie, P., Li, L., & Yao, Y. (2026). BADS: A backdoor attack against code intent summarization engines. *Information and Software Technology*, *192*, 108018. [CrossRef]

[186] Zhu, Y., Tao, Q., & Zhao, N. (2026). Can In-Context Learning Defend against Backdoor Attacks to LLMs. In *AAAI 2026 Workshop on Assessing and Improving Reliability of Foundation Models in the Real World*. https://openreview.net/pdf?id=X7vXylFSGq

[187] Guo, J., Zhang, Z., Sun, Z., Yang, Y., Wu, J., Zhang, F., & He, X. (2026, March). 6dattack: Backdoor attacks in the 6dof pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 42, pp. 35455-35463). [CrossRef]

[188] Luo, J., Sun, Q., Lyu, L., Zhang, Z., Yuan, H., Fu, X., & Li, J. (2026, March). Towards effective, stealthy, and persistent backdoor attacks targeting graph foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 29, pp. 24142-24150). [CrossRef]

[189] AlHayan, A., & Al-Muhtadi, J. (2026). Federated learning-powered real-time behavioral intrusion detection leveraging LSTM, attention, GANs, and large language models. *Scientific Reports*. [CrossRef]

[190] Archa, A. T., & Kartheeban, K. (2026). Secure ML: a hybrid defense method to prevent poisoning attacks on machine learning systems. *International Journal of Machine Learning and Cybernetics, 17*(1), 37.

[CrossRef]

[191] Chaudhari, H., Rathbun, E., Foerster, H., Hayes, J., Jagielski, M., Nasr, M., ... & Oprea, A. (2026). Thought-Transfer: Indirect Targeted Poisoning Attacks on Chain-of-Thought Reasoning Models. *arXiv preprint arXiv:2601.19061*. [CrossRef]

[192] Wang, B., Yan, Y., Zhang, M., Wang, W., & Yao, H. (2026). Model Backdoor Attack on Federated Learning Based on Parameter Analysis. *IEEE Transactions on Dependable and Secure Computing, 23*(3), 6490-6505. [CrossRef]

[193] Hu, Z., Chen, C., & Wang, Y. (2026, March). FILTER: A Framework for Defending against Backdoor Attacks in Vertical Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 42, pp. 35490-35499). [CrossRef]

[194] Sui, Y., Sun, Y., Chen, N., Zhao, Y., Cao, H., & Xu, B. (2026). Backdoor Detection in Federated Learning with Feature Map: A Multi-Task Learning Perspective. *IEEE Transactions on Information Forensics and Security, 21*, 1142-1154. [CrossRef]

[195] Shi, Y., Zheng, W., Xu, H., Wang, X. A., & Wang, R. (2026). A Unified Optimization Framework for Backdoor Attacks in Large Language Models. *Information Fusion*, 104221. [CrossRef]

[196] Wang, K., Wang, L., Liu, Z., Luo, Y., Zhang, K., & Li, W. (2026). SHIFT: Enhancing Federated Learning Robustness through Client-Side Backdoor Detection. *Information Fusion*, 104144. [CrossRef]

[197] Chen, J., Lin, Z., Kang, Y., Wang, C., & Lin, W. (2026). Stealthy Targeted Poisoning Attacks in Vertical Split Learning via Embedding Model Manipulation. *IEEE Transactions on Dependable and Secure Computing, 23*(3), 7059-7072. [CrossRef]

[198] You, Z., Dong, X., Cheng, K., Mu, X., Fu, J., Ma, S., ... & Shen, Y. (2026). Prifft: Privacy-preserving federated fine-tuning of large language models via hybrid secret sharing. *IEEE Transactions on Dependable and Secure Computing, 23*(3), 6167-6182. [CrossRef]

[199] Huang, W., Wang, Y., Cheng, A., Zhou, A., Yu, C., & Wang, L. (2024, April). A fast, performant, secure distributed training framework for LLM. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4800-4804). IEEE. [CrossRef]

[200] Zhou, T., Yan, H., Han, B., Liu, L., & Zhang, J. (2024). Learning a robust foundation model against clean-label data poisoning attacks at downstream tasks. *Neural Networks, 169*, 756-763. [CrossRef]

[201] Li, X., Wu, C., & Wang, J. (2024, April). Unveiling backdoor risks brought by foundation models in heterogeneous federated learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 168-181). Singapore: Springer Nature Singapore. [CrossRef]

[202] Rondanini, C., Carminati, B., Ferrari, E., Kundu, A., & Gaudiano, A. (2026). Malware detection at the edge with lightweight LLMs: A performance evaluation. *ACM Transactions on Internet Technology, 26*(1), 1-24. [CrossRef]

[203] Ullah, F., Mohammad, N., Mostarda, L., Cacciagrano, D., & Zhao, Y. (2025). Q-p2fl: Quantum-enhanced federated edge intelligence for privacy-preserving adversarial attack detection on consumer edge devices. *IEEE Transactions on Consumer Electronics, 71*(2), 4914-4924. [CrossRef]

[204] Zhou, J., Wu, N., Wang, Y., Gu, S., Cao, Z., Dong, X., & Choo, K. K. R. (2022). A differentially private federated learning model against poisoning attacks in edge computing. *IEEE Transactions on Dependable and Secure Computing, 20*(3), 1941-1958. [CrossRef]

[205] Zhan, Z., Li, K., Zhang, Y., & Haddadi, H. (2026, April). Systems-Level Attack Surface of Edge Agent Deployments on IoT. In *Proceedings of the Sixth European Workshop on Machine Learning and Systems* (pp. 99-108). [CrossRef]

[206] Jonnalagadda, A. K., Natarajan, G. N., Veerapaneni, S. M., & Vikram, S. (2025, October). Edge-Aware Federated AI: Scalable LLM Integration for Privacy-Preserving Big Data Networks. In *2025 5th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-7). IEEE. [CrossRef]

[207] Han, C., Yang, T., Cui, Z., & Sun, X. (2025). A privacy-preserving and trustworthy inference framework for LLM-IoT integration via hierarchical federated collaborative computing. *IEEE Internet of Things Journal, 12*(24), 51877-51891. [CrossRef]

[208] Tan, J., Li, A., Liu, Q., Ran, P., & Zhang, L. (2026). VTarbel: Targeted Label Attack with Minimal Knowledge on Detector-enhanced Vertical Federated Learning. *ACM Transactions on Sensor Networks, 22*(2), 1-33. [CrossRef]

[209] Han, S., Han, X., Zhao, P., & Zhang, S. (2026). LFO: Layer-wise Feature Occlusion for Transferable Targeted Adversarial Attacks. *Expert Systems with Applications*, 131684. [CrossRef]

[210] Jebreel, N. M., & Domingo-Ferrer, J. (2023). Fl-defender: Combating targeted attacks in federated learning. *Knowledge-Based Systems, 260*, 110178. [CrossRef]

[211] Yu, Y., Liu, Q., Wu, L., Yu, R., Yu, S. L., & Zhang, Z. (2023, June). Untargeted attack against federated recommendation systems via poisonous item embeddings and the defense. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 4, pp. 4854-4863). [CrossRef]

[212] Pathak, J., Mundra, P., Sejpal, Y., Mahapatra, T., & Rajput, A. S. (2026). Early round detection Protocols: strategies against Untargeted adversarial attacks in federated learning Network. *Computer Networks*, 112098. [CrossRef]

[213] Mhamdi, E. M. E., Guerraoui, R., & Rouault, S. (2018). The hidden vulnerability of distributed learning in byzantium. *arXiv preprint arXiv:1802.07927*. [CrossRef]

[214] Colosimo, F., & De Rango, F. (2023, October). Median-krum: A joint distance-statistical based byzantine-robust algorithm in federated learning. In *Proceedings of the Int'l ACM Symposium on Mobility Management and Wireless Access* (pp. 61-68). [CrossRef]

[215] Zhu, T., Guo, Z., Yao, C., Tan, J., Dou, S., Wang, W., & Han, Z. (2024). Byzantine-robust federated learning via cosine similarity aggregation. *Computer Networks, 254*, 110730. [CrossRef]

[216] Zhu, G., Shen, W., Liu, Z., Qin, J., & Ma, J. (2026). BPFLH: Byzantine-Robust Privacy-Preserving Federated Learning for Heterogeneous Data. *IEEE Transactions on Dependable and Secure Computing, 23*(3), 6103-6118. [CrossRef]

[217] Zhang, Y., Wang, L., Li, M., Gai, K., & Wang, J. (2026). "malicious or Benign?": Enhancing the Contribution of Model Updates in Byzantine-Robust Heterogeneous Federated Learning. *IEEE Transactions on Network Science and Engineering, 13*, 6027-6040. [CrossRef]

[218] Bolatov, A., Horváth, S., Takáč, M., & Gorbunov, E. (2026). Byzantine-Robust Optimization under (L_0, L_1)-Smoothness. *arXiv preprint arXiv:2603.12512*. [CrossRef]

[219] Egger, M., Bakshi, M., & Bitar, R. (2025). Byzantine-resilient zero-order optimization for scalable federated fine-tuning of large language models. *In ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*. https://openreview.net/pdf?id=jeOrrO1Q4N

[220] Hu, C., Hu, Q., Zhang, M., & Yang, Z. (2025). FDBA: Feature-guided Defense against Byzantine and Adaptive attacks in Federated Learning. *Journal of Information Security and Applications, 90*, 104035. [CrossRef]

[221] Luo, H., Sun, G., Liu, Y., Zhao, D., Niyato, D., Yu, H., & Dustdar, S. (2025). A weighted byzantine fault tolerance consensus driven trusted multiple large language models network. *IEEE Transactions on Cognitive Communications and Networking, 12*, 3815-3830. [CrossRef]

[222] Mahdavi, A., Aghapour, S., Zamanifar, A., & Farhadi, A. (2026). TinyGuard: A lightweight Byzantine Defense for Resource-Constrained Federated Learning via Statistical Update Fingerprints. *arXiv preprint arXiv:2602.02615*. [CrossRef]

[223] Dong, Q., Dai, Z., Gao, Y., Zheng, Y., Fu, A., & Susilo, W. (2026). FORCE: Byzantine-Resilient Decentralized Federated Learning via Game-Theoretic Contribution Aggregation. *IEEE Transactions on Information Forensics and Security, 21*, 3182-3196. [CrossRef]

[224] Islamov, R., Malinovsky, G., Gaponov, A., Lucchi, A., Richtárik, P., & Gorbunov, E. (2026). Byzantine-Robust and Differentially Private Federated Optimization under Weaker Assumptions. *arXiv preprint arXiv:2603.23472*. [CrossRef]

[225] Das, R., & Sen, B. K. (2026). Dynamic Meta-Layer Aggregation for Byzantine-Robust Federated Learning. *arXiv preprint arXiv:2603.16846*. [CrossRef]

[226] Tao, M., Tian, Y., Tu, W., Yang, Y., Yang, X., & Tang, X. (2026). Safe-FedLLM: Delving into the Safety of Federated Large Language Models. *arXiv preprint arXiv:2601.07177*. [CrossRef]

[227] Duarte, J. D., Cândido, G. D., De Britto Filho, J. R. A., Neto, J. S., Costa, E. J., Da Costa, J. P. J., & De Melo, L. P. (2026). A Systematic Review of Prompt Injection Attacks on Large Language Models: Trends, Taxonomy, Evaluation, Defenses and Opportunities. *IEEE Access, 14*, 12875-12899. [CrossRef]

[228] Geng, T., Xu, Z., Qu, Y., & Wong, W. E. (2026). Prompt injection attacks on large language models: A survey of attack methods, root causes, and defense strategies. *Computers, Materials, & Continua, 87*(1). Tech Science Press. [CrossRef]

[229] Liu, Y., Jia, Y., Geng, R., Jia, J., & Gong, N. Z. (2024). Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)* (pp. 1831-1847). https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei

[230] Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P. Y., ... & Goldstein, T. (2023). Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*. [CrossRef]

[231] Hines, K., Lopez, G., Hall, M., Zarfati, F., Zunger, Y., & Kiciman, E. (2024). Defending against indirect prompt injection attacks with spotlighting. *arXiv preprint arXiv:2403.14720*. [CrossRef]

[232] Jia, F., Wu, T., Qin, X., & Squicciarini, A. (2025, July). The task shield: Enforcing task alignment to defend against indirect prompt injection in llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 29680-29697). [CrossRef]

[233] Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., ... & Liu, Y. (2023). Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*. [CrossRef]

[234] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023, November). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security* (pp. 79-90). [CrossRef]

[235] Li, Z., Peng, B., He, P., & Yan, X. (2024, November).

Evaluating the instruction-following robustness of large language models to prompt injection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 557-568). [CrossRef]

[236] Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (2024, August). Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 10471-10506). [CrossRef]

[237] Liu, X., Yu, Z., Zhang, Y., Zhang, N., & Xiao, C. (2024). Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957*. [CrossRef]

[238] Chen, S., Piet, J., Sitawarin, C., & Wagner, D. (2025). StruQ: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium (USENIX Security 25)* (pp. 2383-2400). https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe

[239] Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*. [CrossRef]

[240] Chen, Y., Li, H., Sui, Y., He, Y., Liu, Y., Song, Y., & Hooi, B. (2025, July). Can indirect prompt injection attacks be detected and removed?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 18189-18206). [CrossRef]

[241] Seo, J., Zhang, N., & Rong, C. (2023, December). 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)* (pp. 341-349). IEEE. [CrossRef]

[242] Jayathilaka, H. (2025). Privacy-Preserving Prompt Injection Detection for LLMs Using Federated Learning and Embedding-Based NLP Classification. *arXiv preprint arXiv:2511.12295*. [CrossRef]

[243] Lee, Y., Park, T., Lee, Y., Gong, J., & Kang, J. (2025). Exploring potential prompt injection attacks in federated military llms and their mitigation. *arXiv preprint arXiv:2501.18416*. [CrossRef]

[244] Wang, X., Bloch, J., Shao, Z., Hu, Y., Zhou, S., & Gong, N. Z. (2025, November). Webinject: Prompt injection attack to web agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 2010-2030). [CrossRef]

[245] Yi, J., Xie, Y., Zhu, B., Kiciman, E., Sun, G., Xie, X., & Wu, F. (2025, July). Benchmarking and defending against indirect prompt injection attacks on large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1* (pp. 1809-1820). [CrossRef]

[246] Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in Neural Information Processing Systems, 37*, 82895-82920.

[247] Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., ... & Rashidi, H. H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology, 38*(3), 100686. [CrossRef]

[248] Hall, Z., Subbiah, M., Zollo, T., McKeown, K., & Zemel, R. (2026). Guiding LLM decision-making with fairness reward models. *Advances in Neural Information Processing Systems, 38*, 145422-145456.

[249] Wang, L., Li, Q., Cui, D., Wang, M., Zhao, Y., Xu, Y., ... & Wang, L. (2025, April). Building Bridges, Not Walls: Fairness-Aware and Accurate Recommendation of Code Reviewers via LLm-Based Agents Collaboration. In *2025 IEEE/ACM 33rd International Conference on Program Comprehension (ICPC)* (pp. 577-588). IEEE. [CrossRef]

[250] Saeidnia, H. R. (2023). Ethical artificial intelligence (AI): confronting bias and discrimination in the library and information industry. *Library Hi Tech News*. [CrossRef]

[251] Islam, M. M., & Shuford, J. (2024). A survey of ethical considerations in AI: navigating the landscape of bias and fairness. *Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023, 1*(1). https://ideas.repec.org/a/das/njaigs/v1y2024i1id27.html

[252] Hine, E., & Floridi, L. (2023). The blueprint for an AI bill of rights: In search of enaction, at risk of inaction. *Minds and Machines, 33*(2), 285-292. [CrossRef]

[253] Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial bias in pulse oximetry measurement. *New England Journal of Medicine, 383*(25), 2477-2478. [CrossRef]

[254] European Parliament and Council. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. Official Journal of the European Union. https://www.wsgr.com/a/web/qrkz1SnNzWw6nk7B3oAyDa/10-things-you-should-know-about-the-eu-artificial-intelligence-act_v2.pdf

[255] Naidu, G., Zuva, T., & Sibanda, E. M. (2023, April). A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference* (pp. 15-25). Cham: Springer International Publishing. [CrossRef]

[256] Fan, X., Ma, Y., Dai, Z., Jing, W., Tan, C., & Low, B. K. H. (2021). Fault-tolerant federated reinforcement learning with theoretical guarantee. *Advances in neural information processing systems, 34*, 1007-1021.

[257] Huang, W., Li, T., Wang, D., Du, S., Zhang, J., & Huang, T. (2022). Fairness and accuracy in horizontal federated learning. *Information Sciences, 589*, 170-185. [CrossRef]

[258] Li, Y., Wang, X., Xu, W., Wang, H., Qi, Y., Dong,

J., & Li, R. (2026). Feature distillation is the better choice for model-heterogeneous federated learning. *Advances in Neural Information Processing Systems, 38*, 104726-104744.

[259] Che, H., Jin, H., Gu, Z., Lin, Y., Jin, C., & Chen, H. (2025). Llm-driven medical report generation via communication-efficient heterogeneous federated learning. *IEEE Transactions on Medical Imaging, 45*(1), 28-39. [CrossRef]

[260] Li, A., Chen, Y., Long, W., Yin, Y., Hu, Y., Kim, H., ... & Chen, Q. (2026). A Federated and Parameter-Efficient Framework for Large Language Model Training in Medicine. *arXiv preprint arXiv:2601.22124*. [CrossRef]

[261] Shin, J., Yoon, H., Lee, S., Park, S., Liu, Y., Choi, J. D., & Lee, S. J. (2023, December). Fedtherapist: Mental health monitoring with user-generated linguistic expressions on smartphones via federated learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 11971-11988). [CrossRef]

[262] Liu, Y., Zhai, Y., Qu, B., Xue, H., & Liu, X. (2025). Fed-WGCA: A Federated Learning Framework With Coordinate Attention and WGAN for Enhanced Performance. *IEEE Internet of Things Journal, 13*(6), 11456-11471. [CrossRef]

[263] Vu, M., Nguyen, T., & Thai, M. T. (2024, April). Analysis of privacy leakage in federated large language models. In *International Conference on Artificial Intelligence and Statistics* (pp. 1423-1431). PMLR.

[264] Djuhera, A., Andrei, V. C., Li, X., Mönich, U. J., Boche, H., & Saad, W. (2025). R-SFLLM: Jamming resilient framework for split federated learning with large language models. *IEEE Transactions on Information Forensics and Security, 20*, 8296-8311. [CrossRef]

[265] Doğruluk, E., & Açıkgöz, H. (2025, September). 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)* (pp. 1250-1256). IEEE. [CrossRef]

[266] Xia, Y., Chen, Y., Zhao, Y., Kuang, L., Liu, X., Hu, J., & Liu, Z. (2024). FCLLM-DT: Enpowering federated continual learning with large language models for digital-twin-based industrial IoT. *IEEE Internet of Things Journal, 12*(6), 6070-6081. [CrossRef]

[267] Xia, L., Fan, J., Parlikad, A., Huang, X., & Zheng, P. (2024). Unlocking large language model power in industry: Privacy-preserving collaborative creation of knowledge graph. *IEEE Transactions on Big Data, 11*(4), 2046-2060. [CrossRef]

[268] Hasan, M. K., Kabir, S. R., Islam, S., Abdullah, S., Abbas, H. S., Pandey, B., & Gadekallu, T. R. (2025). AES Cryptography Enabled Responsible Federated Foundation Model Using Transformer LLM and LSTM for Smart Grid IIoT Networks. *IEEE Internet of Things Journal, 12*(23), 49801-49810. [CrossRef]

[269] Wan, Z., Guo, C., Hu, B., Du, J., Mou, X., & Zhang, J. (2025, August). LLM-Based V2X Multi-Model Sensor Data Fusion for Improved Road Safety and Data Privacy. In *2025 34th International Conference on Computer Communications and Networks (ICCCN)* (pp. 1-6). IEEE. [CrossRef]

[270] Xiang, T., Bi, Y., Zhi, M., & Cai, L. (2025). FLAD: Federated-Trained Large Language Models for Autonomous Driving. *IEEE Network*. [CrossRef]

[271] Chen, J., He, J., Chen, F., Lv, Z., Tang, J., & Jia, Y. (2024). Empowering IoT-based autonomous driving via federated instruction tuning with feature diversity. *IEEE Internet of Things Journal, 12*(6), 6095-6108. [CrossRef]

[272] Reddy, M. S., Karnati, H., & Sundari, L. M. (2024). Transformer-based federated learning models for recommendation systems. *IEEE Access, 12*, 109596–109607. [CrossRef]

[273] Wang, Y., Tang, X., Lu, Y., & Liu, N. (2023, November). Research on the fairness of cold-start recommender system based on federated learning framework. In *Proceedings of the 2023 5th International Conference on Internet of Things, Automation and Artificial Intelligence* (pp. 802-807). [CrossRef]

[274] Xia, Y., Feng, H., Ge, Q., Rodrigues, J., Gadekallu, T. R., & Fang, K. (2025). Federated Learning based Water Streamflow Forecasting via Multi-Sensor Data Fusion. *Information Fusion*, 104020. [CrossRef]

[275] Guo, L., Lu, Z., Yu, J., Nguyen, Q. V. H., & Yin, H. (2024, May). Prompt-enhanced federated content representation learning for cross-domain recommendation. In *Proceedings of the ACM Web Conference 2024* (pp. 3139-3149). [CrossRef]

[276] Chen, S., Long, G., Shen, T., & Jiang, J. (2023). Prompt federated learning for weather forecasting: Toward foundation models on meteorological data. *arXiv preprint arXiv:2301.09152*. [CrossRef]

[277] Pham, T., Furno, A., Chamroukhi, F., & Oukhellou, L. (2026). Federated dynamic modeling and learning for spatiotemporal data forecasting. *Neurocomputing*, 132712. [CrossRef]

[278] Chen, P., Zeng, X., Zhao, M., Shen, M., Cheng, W., Yu, G., & Chen, T. (2026). Sparse-vdit: Unleashing the power of sparse attention to accelerate video diffusion transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(4), 2957-2965. [CrossRef]

[279] Egashira, K., Vero, M., Staab, R., He, J., & Vechev, M. (2024). Exploiting llm quantization. *Advances in Neural Information Processing Systems, 37*, 41709-41732.

[280] Fan, T., Ma, G., Song, Y., Fan, L., & Yang, Q. (2025, November). PPC-GPT: federated task-specific compression of large language models via pruning and chain-of-thought distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural*

Language Processing (pp. 14794-14805). [CrossRef]

[281] Ma, X., Fang, G., & Wang, X. (2023). Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems, 36*, 21702-21720.

[282] Wang, J., Wu, Y., Xiong, X., Zhang, Y., Lyu, Z., Ghoneim, A., & Zhao, H. (2025). FedLMA: A Federated Learning Framework Integrating LLM-Based Multi-Agent Reasoning With Knowledge Distillation. *IEEE Transactions on Consumer Electronics, 71*(4), 11339-11349. [CrossRef]

[283] Liu, Z., Desai, A., Liao, F., Wang, W., Xie, V., Xu, Z., ... & Shrivastava, A. (2023). Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems, 36*, 52342-52364.

[284] Fan, Q., Zou, A., & Ma, Y. (2026, March). Timebill: Time-budgeted inference for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 36, pp. 30620-30628). [CrossRef]

[285] Chennam, K. K., V, U. M., Aluvalu, R., Chinthaginjala, R., Ab Wahab, M. N., Zhao, X., & Tolba, A. (2025). Load balancing for cloud computing using optimized cluster based federated learning. *Scientific Reports, 15*(1), 41328. [CrossRef]

[286] Wang, Z., Hong, C., Parmar, D., Ma, S., Zhao, Z., Zhao, Q., & Liu, X. (2026, March). SpecProto: A Parallelizing Compiler for Speculative Decoding of Large Protocol Buffers Data. In *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2* (pp. 1848-1862). [CrossRef]

[287] Liu, Z., Jiang, Y., Shen, J., Peng, M., Lam, K. Y., Yuan, X., & Liu, X. (2024). A survey on federated unlearning: Challenges, methods, and future directions. *ACM Computing Surveys, 57*(1), 1-38. [CrossRef]

[288] Wang, W., Tian, Z., Zhang, C., & Yu, S. (2026). BlindU: Blind Machine Unlearning without Revealing Erasing Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 48*(5), 5963-5978. [CrossRef]

[289] Zuo, X., Wang, M., Zhu, T., Zhang, L., Ye, D., Yu, S., & Zhou, W. (2026). Federated TrustChain: Blockchain-enhanced LLM training and unlearning. *IEEE Transactions on Dependable and Secure Computing, 23*(3), 6457-6473. [CrossRef]

[290] Marco-Pérez, I., Pérez, B., Rubio Garcia, A. L., & Zapata, M. A. (2026). The Many Faces of Data Deletion: On the Significance and Implications of Deleting Data. *ACM Computing Surveys, 58*(7), 1-31. [CrossRef]

[291] Wang, L., Zhang, X., Su, H., & Zhu, J. (2024). A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence, 46*(8), 5362-5383. [CrossRef]

[292] Yang, X., Yu, H., Gao, X., Wang, H., Zhang, J., & Li, T. (2024). Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering, 36*(8), 3832-3850. [CrossRef]

[293] Feng, T., Li, X., Liu, H., Wang, Z., & Shen, B. (2026). WarmFed: Federated Learning With Warm-Start for Globalization and Personalization via Personalized Diffusion Models. *IEEE Transactions on Industrial Informatics, 22*(6), 4647-4658. [CrossRef]

[294] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., ... & Gao, J. (2023, December). LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 28541-28564).

[295] Mao, Y., Qin, Z., Zhou, J., Fan, B., Zhang, J., Zhong, Y., & Dai, Y. (2026). Learning Spatial Decay for Vision Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(10), 7945–7953. [CrossRef]

[296] Kelly, B. S., Duignan, S., Mathur, P., Dillon, H., Lee, E. H., Yeom, K. W., ... & Killeen, R. P. (2025). Can ChatGPT4-vision identify radiologic progression of multiple sclerosis on brain MRI?. *European Radiology Experimental, 9*(1), 9. [CrossRef]

[297] Zhang, J., Sun, C., & Peng, Y. (2026). ProtoMFL: a robust multimodal federated learning framework via cross-modal prototype integration. *The Journal of Supercomputing, 82*(5), 269. [CrossRef]

[298] Gu, Z., Zhang, K., Bai, G., Chen, L., Zhao, L., & Yang, C. (2023, April). Dynamic activation of clients and parameters for federated learning over heterogeneous graphs. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)* (pp. 1597-1610). IEEE. [CrossRef]

[299] Deng, Z., Ma, W., Han, Q. L., Zhou, W., Zhu, X., Wen, S., & Xiang, Y. (2025). Exploring DeepSeek: A survey on advances, applications, challenges and future directions. *IEEE/CAA Journal of Automatica Sinica, 12*(5), 872-893. [CrossRef]

[300] Zhang, S., Huang, H., Zhang, C., & Li, X. (2026). QwenStyle: Content-Preserving Style Transfer with Qwen-Image-Edit. *arXiv preprint arXiv:2601.06202*. [CrossRef]

[301] Huang, D., & Wang, Z. (2025, June). Llms at the edge: Performance and efficiency evaluation with ollama on diverse hardware. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. [CrossRef]

[302] Hundera, N. W., Elhabob, R., Adhikari, D., & Xiong, H. (2026). A Blockchain-Based Revocable Identity-Based Proxy Re-Encryption Scheme with Cryptographic Reverse Firewalls for Secure Data Sharing. *Journal of Reliable and Secure Computing, 2*(1), 50-65. [CrossRef]

[303] Elhabob, R., Elkhalil, A., Hadabi, A., Taha, M., Hundera, N. W., & Eltayieb, N. (2026). Certificateless Encryption Supporting Equality Test with Cryptographic Reverse Firewalls in Smart City. *Journal of Reliable and Secure Computing*, *2*(1), 66-82. [CrossRef]

[304] Li, C., Lv, P., Gao, Y., Yuan, X., Zhang, S., Chen, K., ... & Li, Y. (2026). FedWM: Data-Free Watermarking for Model Ownership Protection in Federated Learning. *IEEE Transactions on Dependable and Secure Computing*, *23*(3), 5135-5149. [CrossRef]

[305] Brant-Zawadzki, G., Klapthor, B., Ryba, C., Youngquist, D. C., Burton, B., Palatinus, H., & Youngquist, S. T. (2025). The performance of ChatGPT-4 and Gemini Ultra 1.0 for quality assurance review in emergency medical services chest pain calls. *Prehospital Emergency Care*, *29*(3), 210-217. [CrossRef]

[306] Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Poor, H. V. (2021). Federated learning for internet of things: A comprehensive survey. *IEEE communications surveys & tutorials*, *23*(3), 1622-1658. [CrossRef]

[307] Yuan, L., Wang, Z., Sun, L., Yu, P. S., & Brinton, C. G. (2024). Decentralized federated learning: A survey and perspective. *IEEE Internet of Things Journal*, *11*(21), 34617-34638. [CrossRef]

[308] Chen, J., Yan, H., Liu, Z., Zhang, M., Xiong, H., & Yu, S. (2024). When federated learning meets privacy-preserving computation. *ACM Computing Surveys*, *56*(12), 1-36. [CrossRef]

[309] Sanjalawe, Y., Al-E'mari, S., Fraihat, S., & Makhadmeh, S. (2025). AI-driven job scheduling in cloud computing: a comprehensive review. *Artificial Intelligence Review*, *58*(7), 197. [CrossRef]

[310] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). Pmlr.

[311] Beltrán, E. T. M., Pérez, M. Q., Sánchez, P. M. S., Bernal, S. L., Bovet, G., Pérez, M. G., ... & Celdrán, A. H. (2023). Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, *25*(4), 2983-3013. [CrossRef]

[312] Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., ... & Yang, Q. (2024). Vertical federated learning: Concepts, advances, and challenges. *IEEE transactions on knowledge and data engineering*, *36*(7), 3615-3634. [CrossRef]

[313] Ye, M., Shen, W., Du, B., Snezhko, E., Kovalev, V., & Yuen, P. C. (2025). Vertical federated learning for effectiveness, security, applicability: A survey. *ACM Computing Surveys*, *57*(9), 1-32. [CrossRef]

[314] Zhang, X., Mavromatis, A., Vafeas, A., Nejabati, R., & Simeonidou, D. (2023). Federated feature selection for horizontal federated learning in IoT networks. *IEEE Internet of Things Journal*, *10*(11), 10095-10112. [CrossRef]

[315] Fan, Z., Fang, H., Zhou, Z., Pei, J., Friedlander, M. P., Liu, C., & Zhang, Y. (2022, May). Improving fairness for data valuation in horizontal federated learning. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)* (pp. 2440-2453). IEEE. [CrossRef]

[316] He, C., Annavaram, M., & Avestimehr, S. (2020). Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in neural information processing systems*, *33*, 14068-14080.

[317] Otoum, S., Guizani, N., & Mouftah, H. (2022). On the feasibility of split learning, transfer learning and federated learning for preserving security in its systems. *IEEE Transactions on Intelligent Transportation Systems*, *24*(7), 7462-7470. [CrossRef]

[318] Li, Y., Su, H., Li, H., Yang, H., Zhuang, X., Xue, H., ... & Razzak, I. (2026, March). FedCD: Towards Consolidated Distillation for Heterogeneous Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 28, pp. 23256-23264). [CrossRef]

[319] Zuo, Z., Ye, H., Li, J., & Ge, Y. (2026). A Robust and Heterogeneity-Aware Federated Learning Framework with Knowledge Distillation for Cross-Regional Load Forecasting. *IEEE Transactions on Smart Grid*. [CrossRef]

Deepak Adhikari received the Ph.D. degree from the University of Electronic Science andTechnology of China (UESTC), Chengdu, China. He is currently a Postdoctoral Research Fellow with UESTC. His research interests include cybersecurity, privacy preservation, the Internet of Things (IoT), federated reinforcement learning, artificial intelligence, and secure and autonomous decision-making in distributed systems. (Email: deepakadhikari@uestc.edu.cn)

Inam Ullah received a B.Sc. degree in Electrical Engineering (Telecommunication) from the Department of Electrical Engineering, University of Science and Technology Bannu (USTB), KPK, Pakistan, in 2016 and a Master's and Ph.D. degree in Information and Communication Engineering from the College of Internet of Things (IoT) Engineering, Hohai University (HHU), Changzhou Campus, 213022, China, in 2018 and 2022, respectively. He completed his postdoc with Brain Korea 2021 (BK21) at the Chungbuk Information Technology Education and Research Center, Chungbuk National University, Cheongju 28644, S Korea, from Oct. 2022 to March 31, 2023. He is currently an Assistant Professor at the Department of Computer Engineering, Gachon University, S Korea. His research interests include Robotics, Internet of Things (IoT), Wireless Sensor Networks (WSNs),

Underwater Communication and Localization, Underwater Sensor Networks (USNs), Artificial Intelligence (AI), Big data, Deep learning, etc. (Email: inam@gachon.ac.kr)

Mustafa Khadim received the Ph.D. degree from the University of Electronic Science andTechnology of China, Chengdu, China. He is currently a Postdoctoral Research Fellow with Xiamen Univeristy. His research interests include clustering, the Internet of Things (IoT), artificial intelligence, and secure and autonomous decision-making in distributed systems. (Email: mustafa@xmu.edu.cn)

Negalign Wake Hundera received the B.S. degree from the Faculty of Engineering and Technology, Jimma University, Jimma, Ethiopia, in 2009, the master's degree from the School of Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2016, and the Ph.D. degree from the School of Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2021. He is currently a Postdoctoral Researcher with the School of Software Engineering, University of Electronic Science and Technology of China. His current research interests include cryptography and network security. (Email: nigaccna21@uestc.edu.cn)

Rajab M. S. is currently pursing a PhD. Software Engineering highly motivated to engage in advanced research and innovation across computer-aided diagnosis, computer vision, artificial intelligence, pattern recognition, and large language models. He strive to build a strong professional network with leading researchers & practitioners in these interdisciplinary fields. Through collaboration, he aim to contribute to the development of state-of-the-art techniques and impactful applications that push the boundaries of computer science & medical imaging. (Email: srajab@std.uestc.edu.cn)

Lemessa Bona Debela received the B.S. degree from the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2021, and the master's degree from the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2023. He is currently pursuing his Ph.D. degree with the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include medical imaging, computer vision, and deep learning. (Email: bona@std.uestc.edu.cn)

Wei Jiang received the B.S. degree, the M.S. degree and the Ph.D. degree from the University of Electronic Scienc and Technology of China, Chengdu. He was a visiting researcher in Technical University of Denmark, Denmark and Likoping University, Sweden. He is currently an Associate Professor with School of Information and Software Engineering, University of Electronic Science and Technology of China, China. His research interests includes real-time system, embedded system design, and AI computing. He has servered as the secretary of IEEE CEDA Chengdu Chapter, and TPC members of top conferences of embedded systems including DAC, CODES+ISSS, ISLPED. He is a senior member of CCF, a member of ACM and IEEE. (Email: weijiang@uestc.edu.cn)

Hu Xiong received the Ph.D. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2009. He is currently a Full Professor with the School of Information and Software Engineering, UESTC. His research interests include applied cryptography and cyberspace security. (Email: xionghu.uestc@gmail.com)