Synergizing Foundation Models and Federated Learning: A Survey

Anonymous ACL submission

Abstract

The recent development of Foundation Models (FMs), represented by large language models, vision transformers, and multimodal models, has been making a significant impact on both academia and industry. Compared with small-scale models, FMs have a much stronger demand for high-volume data during the pre-training phase. Although general FMs can be pre-trained on data collected from open sources such as the Internet, domain-specific FMs need proprietary data, posing a practical challenge regarding the amount of data available due to privacy concerns. Federated Learning (FL) is a collaborative learning paradigm that breaks the barrier of data availability from different participants. Therefore, it provides a promising solution to customize and adapt FMs to a wide range of domain-specific tasks using distributed datasets whilst preserving privacy. This survey paper discusses the potentials and challenges of synergizing FL and FMs and summarizes core techniques, future directions, and applications.

1 Introduction

The landscape of Artificial Intelligence (AI) has been revolutionized by the emergence of *Foundation Models (FMs)* (Bommasani et al., 2021), such as BERT (Devlin et al., 2019), GPT series (Brown et al., 2020; OpenAI, 2022, 2024), and LLaMA series (Touvron et al., 2023a,b) in Natural Language Processing (NLP); ViTs (Dosovitskiy et al., 2021) and SAM (Kirillov et al., 2023) in Computer Vision (CV); CLIP (Radford et al., 2021), DALL-E (Ramesh et al., 2021), Gemini (Google, 2023), and GPT-4o in multimodal applications. These FMs have become pivotal in a myriad of AI applications across diverse domains. Their superb capability to generalize across tasks and domains stems from their pre-training on extensive datasets (Gunasekar et al., 2023), which imbues them with a profound understanding of language, vision, and multimodal data.

While general-purpose FMs can leverage openly accessible data from the Internet, domain-specific FMs require proprietary data. It is, however, challenging to collect vast amounts of proprietary data and perform centralized pre-training or fine-tuning for domain-specific FMs, due to privacy restrictions (Jo and Gebru, 2020; GDPR, 2016; CCPA, 2023). Particularly in domains such as law, healthcare, and finance, where data is inherently privacy-sensitive, there is a pressing need for stringent privacy safeguards. Furthermore, given that data often constitutes a pivotal asset for enterprises, its widespread distribution is prohibitive. Consequently, there is an urgent need for novel strategies to handle data availability and facilitate model training, thereby unlocking the potential of domain-specific FMs whilst respecting data privacy.

To address the challenges associated with data privacy in model training, Federated Learning (FL) (McMahan et al., 2017) has emerged as a promising paradigm. FL facilitates collaborative model training across decentralized clients without the need for sharing raw data, thus ensuring privacy preservation. Concretely, FL encompasses periodic interactions between the server and decentralized clients for the exchange of trainable model parameters, without the requirement for private client data. Recognizing such a benefit, *integrating FMs with FL presents a compelling solution for domain-specific FMs* (Zhuang et al., 2023; Yu et al., 2023d).

Despite the potential synergies between FL and FMs, the field is still nascent, lacking a comprehensive understanding of challenges, methodologies, and directions. This survey aims to bridge this gap by providing a thorough exploration of the integration of FMs and FL. We delve into the motivations and challenges of combining these two paradigms, highlight representative techniques, and discuss future directions and applications. By elucidating the intersection of FL and FMs, we aim to catalyze further research and innovation in this burgeon-

ing area, ultimately advancing the development of privacy-aware, domain-specific FMs.

The paper continues as follows: The next section introduces background on FMs and FL. Section 3 presents the motivation and challenges for synergizing FMs and FL. Section 4 highlights representative techniques. Before concluding, we discuss representative future directions in Section 5.

2 Background

2.1 Foundation Models

An FM is a model that can be adapted to a wide array of tasks through fine-tuning after initial pre-training (Bommasani et al., 2021). The lifecycle of FMs typically involves pre-training on extensive generic data to establish the basis of their abilities (Bubeck et al., 2023), followed by adaptation to downstream tasks such as domain-specific question answering (Zhang et al., 2023d), and ultimately application in various domains.

FMs have sparked a significant paradigm shift in various fields of AI such as NLP, CV, speech and acoustics, and beyond. In the realm of NLP, the most prominent example is Large Language Models (LLMs) with substantial parameter sizes (Zhao et al., 2023). These models, such as ChatGPT and GPT-4 (OpenAI, 2022, 2024), demonstrate exceptional abilities in natural language understanding and generation, enabling them to comprehend and respond to user inputs with remarkable contextual relevance. This capability proves invaluable in applications like customer service, virtual assistants, and chatbots, where effective communication is paramount. Moreover, LLMs eliminate the need for training models from scratch for specific tasks, be it machine translation, document summarization, text generation, or other language-related tasks.

In the realm of CV and other modalities, FMs have also made remarkable progress. Vision Transformers (ViTs) (Dosovitskiy et al., 2021) segment images into distinct patches, which serve as inputs for transformer architectures. SAM (Kirillov et al., 2023) can segment anything in images according to the input prompts. CLIP (Radford et al., 2021) bridges the gap between text and images through contrastive learning. DALL·E, proposed by Ramesh et al. (2021), generates images from textual descriptions, expanding the possibilities of creative image generation. Additionally, models like GAto (Reed et al., 2022), exhibit versatility by being applicable across various tasks such as conversational agents, robotic control, and gaming.

2.2 Federated Learning

FL (McMahan et al., 2017) is a learning paradigm that enables a collection of clients to collaboratively learn a shared global model by leveraging their private datasets in a distributed manner, assisted by the coordination of a central server. The most representative algorithms in the FL literature are the `FedAvg`-family algorithms (McMahan et al., 2017; Reddi et al., 2021). The standard `FedAvg` involves periodic interactions between the server and decentralized clients to exchange trainable model parameters. Many variants have been proposed to tackle issues such as convergence and local data heterogeneity (Diao et al., 2021). For example, `FedProx` (Li et al., 2020) and `FedDyn` (Acar et al., 2021) introduce regularizer terms to penalize client updates that are far away from the server model. A general framework `FedOpt` (Reddi et al., 2021) unifies adaptive optimizers (*Adam*, *Yogi*, *etc.*) and demonstrates superior convergence speed when compared to the naive `FedAvg`.

FL offers an efficient privacy-preserving way to train models on large-scale and diverse data (Kairouz et al., 2021), leading to its application across various domains such as healthcare (Lincy and Kowshalya, 2020; Rieke et al., 2020; Joshi et al., 2022), finance (Chatterjee et al., 2023; Liu et al., 2023b), and smart cities (Ramu et al., 2022; Pandya et al., 2023).

3 FM-FL: Motivation & Challenges

In this section, we first motivate the synergy of FMs and FL (Section 3.1), then summarize the key challenges (Section 3.2).

3.1 Motivation

The integration of FMs and FL represents a compelling collaboration that leverages each other's strengths to address their respective limitations, embodying a complementary relationship (Zhuang et al., 2023; Li and Wang, 2024).

FL expands data availability for FMs. By leveraging data from a wide range of sources in a privacy-preserving manner, FL makes it possible to build models on sensitive data in specific domains, such as healthcare (Lincy and Kowshalya, 2020; Joshi et al., 2022; Rieke et al., 2020) and finance (Chatterjee et al., 2023; Liu et al., 2023b). This enhances the diversity and volume of training

data, improving model robustness and adaptability. Moreover, FL enables the integration of personal and task-specific data, allowing FMs to be customized for personal applications. For instance, Google has trained next-word-prediction language models on mobile keyboard input data with FL to improve user experience (Xu et al., 2023; Bonawitz et al., 2021).

FMs boost FL with feature representation and few-shot learning capabilities. By pre-training on large-scale generic data, FMs acquire essential knowledge and understanding capabilities (Brown et al., 2020), providing multiple benefits to FL. Firstly, they benefit FL systems by offering advanced feature representations and learning capabilities from the outset. Secondly, leveraging the pre-learned knowledge of FMs can accelerate the FL process, enabling efficient and effective adaptation to specific tasks with minimal additional training. Thirdly, FMs' powerful generative capabilities could help FL overcome the data heterogeneity challenge by synthesizing extra data, thus accelerating model convergence (Huang et al., 2024).

3.2 Core Challenges

In this part, we discuss challenges emerging from the FM-FL marriage in three aspects: *efficiency*, *adaptability*, as well as *trustworthiness*.

Efficiency Challenges. Efficiency challenges stem from the mismatch between the significant resource demands of FM training and the limited, heterogeneous system resources (*e.g.,* mobile devices) within FL systems, such as communication bandwidth, computational power, and memory (Su et al., 2023). The communication bottleneck of FL is induced by frequently exchanging training information between the server and clients over limited bandwidth channels (Kairouz et al., 2021). The substantial number of parameters in FMs further exacerbates this burden, thus hindering the training process.

Adaptability Challenges. Adaptability challenges arise from the adaptation of an FM to a specific downstream task (*e.g.,* by fine-tuning) in FL settings. Key challenges include *data heterogeneity* and *resource heterogeneity*. Performance degradation in FL, attributed to heterogeneous data distributions among clients, is a well-recognized issue (Kairouz et al., 2021; Li et al., 2022). A recent study (Babakniya et al., 2023a) has shown that such performance penalty is even more substantial when fine-tuning FMs. For NLP tasks, data heterogeneity can manifest as variations in language, style, topic, or sentiment across datasets held by different clients. In multi-modal scenarios, the challenge is even more pronounced due to the inherent diversity in data types (*e.g.,* text, images, and audio) (Yu et al., 2023a). Addressing data heterogeneity involves not just identifying and measuring it but also developing algorithms that are robust to such diversity, ensuring that the model can learn effectively from varied data contributions without compromising on performance. In terms of resource heterogeneity, the memory and computational resources of the devices for different participants may be diverse (Diao et al., 2021), which could cause delays for model synchronization and inactivation of some participants, *i.e.,* stragglers, making it challenging to leverage the full potential of FMs in FL settings.

Trustworthiness Challenges. Trustworthiness challenges emphasize the concerns regarding privacy, security, and ethical considerations in the lifecycle of FM-FL, from the pre-training and model adaptation to the application stages. We present two representative challenges from this perspective: (1) *intellectual property*: Intellectual Property (IP) protection in FM-FL primarily involves attributing ownership rights for both models and data. From the server's perspective, broadcasting a pre-trained model to multiple nodes for fine-tuning poses IP protection and security risks (*e.g.,* model theft), necessitating measures to safeguard IP rights and ensure model integrity (Kang et al., 2024); (2) *privacy leakage*: Although FL does not immediately share data, studies have shown that it may not always guarantee sufficient privacy preservation (Geiping et al., 2020), as model parameters (*e.g.,* weights or gradients) may leak sensitive information to malicious adversaries (Zhu et al., 2019).

4 Techniques

Recent work has begun to address challenges associated with adapting pre-trained FMs to specific downstream tasks in FL settings. In this section, we survey FM-FL techniques on three aspects, namely *efficiency* (Section 4.1), *adaptability* (Section 4.2), and *trustworthiness* (Section 4.3). As illustrated in Figure 1, we further refine them according to the key features of different methods.

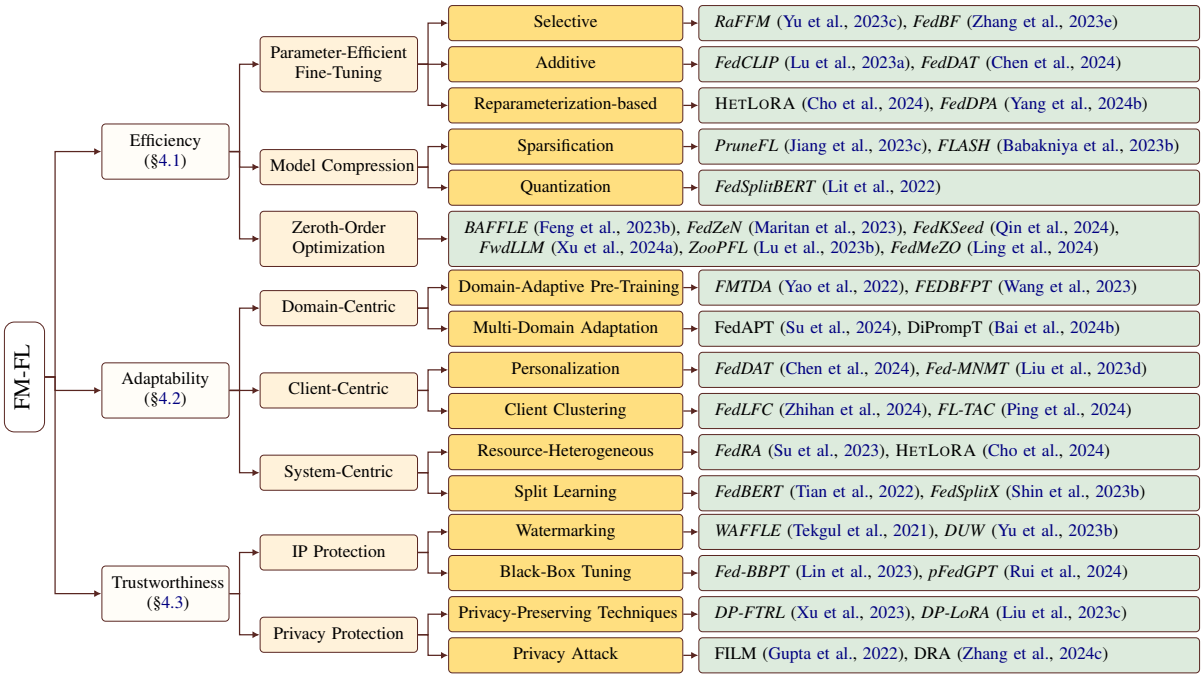

Figure 1: Taxonomy of research in foundation models with federated learning.

4.1 Efficiency

There has been a considerable focus on developing resource-efficient approaches. This part describes techniques that improve resource efficiency.

4.1.1 Parameter-Efficient Fine-Tuning

Federated Parameter-Efficient Fine-Tuning (FedPEFT), originating from the fine-tuning practices of FMs (Lester et al., 2021; Hu et al., 2022; Li and Liang, 2021), is a suite of techniques designed to reduce both the computational load and the associated communication overheads (Malaviya et al., 2023; Woisetschläger et al., 2024). In alignment with existing FM fine-tuning taxonomies (Lialin et al., 2023; Ding et al., 2023), we present FedPEFT methods in three categories: *selective methods*, *additive methods*, and *reparameterization-based methods*.

Selective Methods. Selective methods fine-tune a small subset of the parameters, leaving the majority unchanged. In the field of LLMs, a prominent example of such methods is BitFit (Ben Zaken et al., 2022), which only fine-tunes the bias terms. BitFit has inspired a series of studies in FedPEFT (Bu et al., 2022; Sun et al., 2022a; Zhang et al., 2023e), demonstrating the superior communication efficiency of only updating the bias terms while still achieving competitive performance. More sophisticated methods strive to find sparse subnetworks for partial fine-tuning. Among them, various methods (Seo et al., 2021; Li et al., 2021a; Tamirisa et al., 2024) advocate for the Lottery Ticket Hypothesis (LTH) (Frankle and Carbin, 2019), positing that a dense network contains many subnetworks whose inference capabilities are as accurate as that of the original network. FedSelect (Tamirisa et al., 2024) is a representative method that encourages clients to find optimal subnetworks based on LTH and continually fine-tunes these derived subnetworks to encapsulate local knowledge. As another important aspect, RaFFM (Yu et al., 2023c) proposes to prioritize specialized salient parameters by ranking them using salience evaluation metrics such as the ℓ_1 and ℓ_2 norms.

Additive Methods. Instead of fine-tuning a subset of model parameters, additive methods incorporate lightweight trainable blocks into frozen FMs and tune the additional parameters for model adaptation. These methods not only enhance computational and communicational efficiency but also introduce an extra benefit: personalization (Lu et al., 2023a), *i.e.,* the integration of these supplementary parameters allows for the customization of heterogeneous models tailored to specific local data characteristics or user preferences. Key branches within additive methods include *adapter tuning* and *prompt tuning*. Adapter tuning integrates small-scale neural networks (known as "adapters") into

the pre-trained models (Houlsby et al., 2019; Hu et al., 2022). On the other hand, prompt tuning incorporates trainable task-specific continuous prompt vectors at the input layer (Liu et al., 2023a; Dong et al., 2023). More details on these methods are provided in Appendix A.1.

Reparameterization-based Methods. The hypothesis behind reparameterization-based methods is that fine-tuning adaptations can be reparameterized into optimization within low-rank subspaces (Aghajanyan et al., 2021). Low-Rank Adaptation (LoRA) (Hu et al., 2022), as a popular PEFT method from the area of LLMs, reduces the number of trainable parameters for downstream tasks by representing the weight updates with two smaller matrices (called update matrices) through low-rank decomposition (Ding et al., 2023). When optimizing a parameter matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$, the update equation can be written as: $\mathbf{W} \leftarrow \mathbf{W} + \Delta\mathbf{W}$. The core idea of LoRA is to freeze the original matrix $\mathbf{W}$ while approximating the parameter update $\Delta\mathbf{W}$ by low-rank decomposition matrices, *i.e.,* $\Delta\mathbf{W} = \mathbf{A} \cdot \mathbf{B}^\top$, where $\mathbf{A} \in \mathbb{R}^{m \times k}$ and $\mathbf{B} \in \mathbb{R}^{n \times k}$ are the trainable parameters for task adaptation and $k \ll \min(m, n)$ is the reduced rank. The trainable parameter size is then reduced from mn to $k(m + n)$. The major benefit of LoRA is that it can largely save memory and storage usage. A straightforward way to perform federated finetuning with LoRA is to train the LoRA modules $\mathbf{A}$ and $\mathbf{B}$ with homogeneous rank k across all clients with standard FL such as `FedAvg` (McMahan et al., 2017). Serval studies have shown that this method can achieve an outstanding level of trade-off between performance and communication overhead for a wide range of FMs, including language models (Zhang et al., 2024b, 2023e), vision-language models (Nguyen et al., 2024), and speech-to-text models (Du et al., 2024).

To summarize, FedPEFT techniques have exhibited remarkable potential in effectively harnessing and adapting FMs within FL scenarios, leading to enhanced performance and efficiency across distributed systems. We refer to Appendix A.2 for a detailed comparison.

4.1.2 Model Compression

Model compression refers to the techniques used to reduce the size of models, thereby improving resource efficiency (Shah and Lau, 2023).

Sparsification. Model sparsification methods reduce communication burden by only transmitting a subset of FM parameters across the network (Jiang et al., 2023c). Typical methods focus on identifying and cultivating high-potential subnetworks (Frankle and Carbin, 2019; Tsouvalas et al., 2023).

Quantization. Quantization is well-established in both the FM and FL domains (Xu et al., 2024b; Reisizadeh et al., 2020), which involves decreasing the precision of floating-point parameters for mitigating the storage, computational, and communication demands. Quantization is orthogonal to other resource-efficient techniques, making it feasible to combine them for greater efficiency and flexibility (Lit et al., 2022).

4.1.3 Zeroth-Order Optimization

In contrast to the use of gradient descent in most FL optimization algorithms, a particular line of research advocates for the removal of BackPropagation (BP) (Malladi et al., 2023a) in favor of Zeroth-Order Optimization (ZOO) (Fang et al., 2022; Li and Chen, 2021). BP-free methods conserve memory needed for computing gradients and minimize communication overhead for model aggregation (Qin et al., 2024), making FMs more accessible for lower-end devices, thereby enhancing their applicability in diverse hardware environments.

ZOO methods primarily rely on perturbation methods to estimate gradients with forward propagation. Based on this, recent work, such as that by Xu et al. (2024a); Lu et al. (2023b), has initiated preliminary explorations into the deployment of both FedPEFT and full-model fine-tuning of billion-sized FMs, like LLaMA, on mobile devices. The naive ZOO methods remain impractical for training large FMs in standard FL frameworks such as `FedAvg`, as they still result in a significant communication burden for model aggregation. In light of this, FedKSeed (Qin et al., 2024) was proposed to further reduce communication overheads between the server and clients by using just a few random seeds and scalar gradients, requiring only a few thousand bytes for communication.

Although ZOO methods have shown promise in resource-efficient FL (Ling et al., 2024), they generally require many iterations to achieve strong performance (Malladi et al., 2023b). Compared to the well-established BP-based optimization, ZOO is still in the early stages of development, particularly for FM-FL settings, necessitating further research

and optimization.

4.2 Adaptability

Adaptation refers to the process of tailoring a pre-trained FM to perform effectively across varying FL settings and scenarios. This mainly includes the capability to learn from different domains, cater to individual user needs, and work across diverse devices while retaining overall performance and efficiency. We focus on three key aspects of adaptation, namely *domain-centric adaptation*, *client-centric adaptation*, and *system-centric adaptation*.

4.2.1 Domain-Centric Adaptation

Domain-centric adaptation focuses on adapting FMs within specific domains by addressing the domain diversity across client datasets.

Domain-Adaptive Pre-Training. Despite being heavily reliant on large-scale and public datasets for their initial training, FMs often require further Domain-Adaptive Pre-Training (DAPT) with domain-specific data for tasks that necessitate specialized knowledge (Gururangan et al., 2020; Guo and Yu, 2022). In domains like healthcare, FL allows for the continued pre-training of these models using sensitive, domain-specific data without compromising privacy. Based on this idea, Jiang et al. (2023b) proposed FFDAPT, a computational-efficient further pre-training algorithm that freezes a portion of consecutive layers while optimizing the rest of the layers. Similarly, Wang et al. (2023) proposed FEDBFPT that builds a local model for each client, progressively training the shallower layers of local models while sampling deeper layers, and aggregating trained parameters on a server to create the final global model.

Multi-Domain Adaptation. Given that client data may belong to various domains in real-world FL scenarios, some efforts (Feng et al., 2023c; Su et al., 2024) have been devoted to facilitating multi-domain collaborative adaptation. Feng et al. (2023c) applied a pre-trained CLIP to the multi-domain scenario and proposed an adaptive prompt tuning method that uses domain-specific keys to generate prompts for each test sample. Furthermore, Su et al. (2024) employed knowledge distillation to selectively distill global knowledge based on an entropy measure, improving the generalization across different domains.

4.2.2 Client-Centric Adaptation

Client-centric adaptation refers to the process of tailoring an FM to meet the specific needs or preferences of individual clients while leveraging the decentralized and privacy-preserving nature of FL. Particularly, we discuss two types of popular personalized methods as follows:

Personalization. Adapter-based methods introduce small, trainable adapters into the frozen pre-trained FMs, allowing for client-specific model adaptation without altering the original FL. FedDAT (Chen et al., 2024) leverages a dual-adapter structure, with personalized adapters focusing on client-specific knowledge and a global adapter maintaining client-agnostic knowledge. FedDAT executes bi-directional knowledge distillation between personalized adapters and the global adapter to regularize the client's updates and prevent overfitting. Prompt-based methods involve using client-specific soft prompts to guide the model's response. pFedPG (Yang et al., 2023a) trains a prompt generator to exploit underlying client-specific characteristics and produce personalized prompts for each client, thereby enabling efficient and personalized adaptation.

Client Clustering. This branch of study aims to cluster clients based on the underlying relationships and tailor FMs for the client group with similar data distributions, thus reducing the negative impact of data heterogeneity and improving accuracy. Zhihan et al. (2024) proposed a FedPEFT-based framework for multilingual modeling, which employs language family clustering to alleviate parameter conflicts of LoRA tuning.

4.2.3 System-Centric Adaptation

System-centric aims to improve adaptability at the system level. This involves handling resource heterogeneity in the FL systems while ensuring training efficiency and model utility.

Resource-Heterogeneous Methods. Cross-device FL systems may be composed of devices equipped with heterogeneous resources, leading to disparities where certain devices exhibit more efficient model training than others (Chen et al., 2024). To address this, several methods have been developed to customize model architectures for resource-heterogeneous FL systems. In FL environments with heterogeneous resources, LoRA-based FedPEFT exhibits distinctive flexi-

bility and adaptation in fine-tuning frozen FMs without overburdening client devices. Su et al. (2023) suggested assigning LoRA adapters to varying numbers of layers for heterogeneous clients according to a randomly generated mask matrix. An alternative and more targeted idea is to choose diverse LoRA ranks across clients based on their system capabilities. Bai et al. (2024a) proposed FlexLoRA to adjust local LoRA ranks dynamically. FlexLoRA reconstructs the uniform full-sized LoRA module $\Delta \mathbf{W}$ for server-side model aggregation followed by an SVD-based parameter redistribution. However, concurrent research by Cho et al. (2024) has empirically demonstrated that the reconstruct-redistribute method suffers from performance loss compared to homogeneous LoRA. Instead, they proposed HETLORA (Cho et al., 2024) that utilizes zero-padding to align module size before aggregation. It then truncates the global LoRA modules for the specific rank of the next selected clients.

Split Learning. Split learning addresses the resource heterogeneity between servers and clients by splitting a large model at a cut layer into client and server models (Thapa et al., 2022). For each training step, the output tensor, so-called smashed data, from the client model and the corresponding labels are transmitted over to the server. The server continues the forward propagation by processing the smashed data through its remaining layers; it then computes the loss using the transmitted label and performs backpropagation. The gradient generated at the first layer of the server model is then transmitted back to the client for further backpropagation. Along this line, FedBERT (Tian et al., 2022) proposes to leverage split learning for training the BERT model, showing the feasibility of training large FMs in FL settings. FedSplitX (Shin et al., 2023b) is a more fine-grained method that allows multiple partition points for model splitting, accommodating more diverse client capabilities. Compared to conventional FL, split learning scales better with the size of FMs as it communicates only small-sized smashed data instead of model parameters (Singh et al., 2019). Despite its merits, split learning is highly dependent on the network connection quality. Given that server-client interactions occur at every step of the optimization process (Zheng et al., 2023), communication delays cause a more significant impact on efficiency.

4.3 Trustworthiness

This line of work aims to enhance trustworthiness throughout the FM-FL lifecycle, covering a variety of key aspects including, but not limited to, *IP protection*, *privacy protection*, and *attack robustness*.

4.3.1 IP Protection

Existing IP protection involves safeguarding ownership of FMs from unauthorized use (*e.g.,* model theft) (Tekgul et al., 2021). We discuss the following two mainstream IP protection strategies: *watermarking* and *black-box tuning*.

Watermarking. Watermarking is a well-known deterrence technology for model IP protection by providing the identities of model owners to demonstrate ownership of their models (Adi et al., 2018). Tekgul et al. (2021) proposed WAFFLE, the first solution that addresses the ownership problem by injecting a watermark into the global model in FL environments. Recently, Yu et al. (2023b) proposed DUW that embeds a client-unique key into each client's local model, aiming to identify the infringer of a leaked model while verifying the FL model's ownership.

Black-Box Tuning. Black-Box Tuning (BBT) is a set of ZOO-based methods that fine-tune FMs without direct access to model parameters (Sun et al., 2022c,b). BBT methods are often additive, introducing additional parameters while keeping the original model frozen (See Section 4.1.1). FedBBPT (Lin et al., 2023) is a general prompt tuning framework that facilitates the joint training of a global lightweight prompt generator across multiple clients. FedBPT (Sun et al., 2024a) adopts a classic evolutionary-based ZOO method, CMA-ES (Hansen and Ostermeier, 2001), for training an optimal prompt that improves the performance of frozen FMs. ZooPFL (Lu et al., 2023b), on the other hand, applies coordinate-wise gradient estimate to learn input surgery that incorporates client-specific embeddings. BBT allows for local fine-tuning of FMs while not infringing IP constraints. However, current research in this line is limited to few-shot learning with small datasets for LLM fine-tuning (Sun et al., 2022b), while larger datasets and other modalities remain unexplored.

4.3.2 Privacy Protection

Protecting privacy in FM-FL requires both designing protective measures and studying privacy attack strategies.

Privacy-Preserving Techniques. Differential Privacy (DP) is a theoretical framework that governs privacy boundaries and manages the trade-off between privacy and model convergence (Wei et al., 2020; Xu et al., 2023). DP-based FL approaches often add artificial noise (*e.g.,* Gaussian noise) to parameters at the clients' side before aggregating to prevent information leakage (Xu et al., 2023). Besides, DP is compatible with most Fed-PEFT methods. For instance, Sun et al. (2024b) showed that DP noise can even be amplified by the locally "semi-quadratic" nature of LoRA-based methods, motivating the integration of LoRA with DP to improve resource efficiency while maintaining data privacy (Liu et al., 2023c). In addition to DP, Secure Multi-Party Computation (SMPC) (Mugunthan et al., 2019) and Homomorphic Encryption (HE) (Zhang et al., 2020) are also effective privacy-preserving mechanisms. However, they do not scale well enough for large-scale deployments in FM-FL.

Privacy Attack. Privacy attacks in FM-FL involve extracting sensitive information from the data used in training, even though the data itself is not directly shared. Major attacks include *membership inference attack* and *data reconstruction attack*, where the former aims to determine whether a specific data sample is in a victim client's training set, and the latter strives to reconstruct original input data from the model parameters or gradients (Ren et al., 2024). Regarding membership inference attacks, Vu et al. (2024) revealed the vulnerabilities of popular LLMs, including BERT, DistilBERT, and OpenAI's GPTs. In terms of data reconstruction attacks, Gupta et al. (2022) presented an attack *FILM*, which recovers private text data by extracting information from gradients transmitted during training despite employing a DP mechanism.

5 Future Directions

Although recent work has already begun to address the challenges discussed in Section 3.2, many critical open directions are yet to be explored. Here, we outline several representative ones.

Multimodal FM-FL. With the development of mobile technology and IoT infrastructures (Brunete et al., 2021), numerous edge devices produce data from a range of modalities, such as sensory, visual, and audio. In the era of FMs, the success of LLMs and their multimodal derivatives (Ramesh et al., 2021; Google, 2023; OpenAI, 2024) have demonstrated the potential of multimodal FMs. The potential opportunities and challenges for multimodal FM-FL have yet to be explored.

Continual Learning. Continual learning enables models to adapt to new data over time, improving their performance and accuracy. By incorporating new data into the model training process, FL and FMs can continuously improve and adapt to changing environments and user needs (Yang et al., 2024a). Future directions may involve leveraging transfer learning techniques in continual learning for FL and FMs. Models can transfer knowledge from previous tasks or domains to new ones, enabling more efficient adaptation (Good et al., 2023).

Efficient Federated Black-Box Tuning. In scenarios where gradient access is unavailable, preliminary efforts have focused on federated fine-tuning black-box FMs (Lin et al., 2023; Sun et al., 2024a; Lu et al., 2023b; Rui et al., 2024) utilizing ZOO. However, ZOO's noticeably slower convergence rates, especially in high-dimensional contexts compared to gradient-based methods (Golovin et al., 2020), indicate an important direction for further research. The impact of these slower convergence rates on overall efficiency and computational load within FL, particularly concerning large-scale FMs, has not been adequately investigated and understood.

FL with AI-Generated Content. AI-Generated Content (AIGC) denotes content produced via advanced generative FMs (Wu et al., 2023a). The strong generative capability of FMs offers the advantage of rapidly automating the creation of inexhaustible synthetic data. This capability positions AIGC as a valuable supplementary data source for model training and evaluation in many tasks (Xu et al., 2024c). Despite some efforts (Zhang et al., 2023b), more potential opportunities and challenges for AIGC-aided FL have yet to be explored.

6 Conclusions

In this survey, we have meticulously surveyed the intersection of FM and FL. We identified core challenges in efficiency, adaptability, and trustworthiness and proposed a comprehensive taxonomy of techniques in response to these challenges. In addition, we discussed future directions and applications in this research field, hoping to attract more breakthroughs in future research.

Limitations

FM and FL are very fast-moving fields. We have put a lot of effort into including the latest research efforts in the community in this survey. Therefore, we believe that our survey will help to inspire and push further research and innovation in these important areas. Our survey does not focus on experimental evaluation of the available ideas and systems. We believe that would be an important next step that we are leaving for future work.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. 2021. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*.

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. 2018. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1615–1631, Baltimore, MD. USENIX Association.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Singh Akshay and Thakur Rahul. 2024. Generalizable multilingual hate speech detection on low resource indian languages using fair selection in federated learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Muhammad Ammad-Ud-Din, Elena Ivannikova, Suleiman A. Khan, Were Oyomno, Qiang Fu, Kuan Eeik Tan, and Adrian Flanagan. 2019. Federated collaborative filtering for privacy-preserving personalized recommendation system. *arXiv preprint arXiv:1901.09888*.

Sheikh Shams Azam, Tatiana Likhomanenko, Martin Pelikan, and Jan "Honza" Silovsky. 2023a. Importance of smoothness induced by optimizers in fl4asr: Towards understanding federated learning for end-to-end asr. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Sheikh Shams Azam, Martin Pelikan, Vitaly Feldman, Kunal Talwar, Jan Silovsky, and Tatiana Likhomanenko. 2023b. Federated learning for speech recognition: Revisiting current trends towards large-scale ASR. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

Sara Babakniya, Ahmed Elkordy, Yahya Ezzeldin, Qingfeng Liu, Kee-Bong Song, MOSTAFA EL-Khamy, and Salman Avestimehr. 2023a. SLoRA: Federated parameter efficient fine-tuning of language models. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

Sara Babakniya, Souvik Kundu, Saurav Prakash, Yue Niu, and Salman Avestimehr. 2023b. Revisiting sparsity hunting in federated learning: Why does sparsity consensus matter? *Transactions on Machine Learning Research*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. 2024a. Federated fine-tuning of large language models under heterogeneous language tasks and client resources. *arXiv preprint arXiv:2402.11505*.

Sikai Bai, Jie Zhang, Shuaicheng Li, Song Guo, Jingcai Guo, Jun Hou, Tao Han, and Xiaocheng Lu. 2024b. Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. *arXiv preprint arXiv:2403.08506*.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. 2013. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Kallista Bonawitz, Peter Kairouz, Brendan McMahan, and Daniel Ramage. 2021. Federated learning and privacy: Building privacy-preserving systems for machine learning and data science on decentralized data. *Queue*, 19(5):87–114.

Tom Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alberto Brunete, Ernesto Gambao, Miguel Hernando, and Raquel Cedazo. 2021. Smart assistive architecture for the integration of iot devices, robotic systems, and multimodal interfaces in healthcare environments. *Sensors*, 21(6):2212.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2022. Differentially private bias-term only fine-tuning of foundation models. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. 2023. Efficient federated learning for modern nlp. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '23, New York, NY, USA. Association for Computing Machinery.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. 2021. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *ISOC Network and Distributed System Security Symposium (NDSS)*.

CCPA. 2023. California consumer privacy act (ccpa).

Pushpita Chatterjee, Debashis Das, and Danda B Rawat. 2023. Use of federated learning and blockchain towards securing financial services. *arXiv preprint arXiv:2303.12944*.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. 2024. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11285–11293.

Yudong Chen, Lili Su, and Jiaming Xu. 2017. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*.

Vijil Chenthamarakshan, Samuel C. Hoffman, C. David Owen, Petra Lukacik, Claire Strain-Damerell, Daren Fearon, Tika R. Malla, Anthony Tumber, Christopher J. Schofield, Helen M.E. Duyvesteyn, Wanwisa Dejnirattisai, Loic Carrique, Thomas S. Walter, Gavin R. Screaton, Tetiana Matviiuk, Aleksandra Mojsilovic, Jason Crain, Martin A. Walsh, David I. Stuart, and Payel Das. 2023. Accelerating drug target inhibitor discovery with a deep generative foundation model. *Science Advances*, 9(25):eadg7865.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. 2024. Heterogeneous lora for federated fine-tuning of on-device foundation models. *arXiv preprint arXiv:2401.06432*.

Yun-Wei Chu, Dong-Jun Han, and Christopher G. Brinton. 2024. Only send what you need: Learning to communicate efficiently in federated multilingual machine translation. In *Companion Proceedings of the ACM on Web Conference 2024*, page 1548–1557, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wenlong Deng, Christos Thrampoulidis, and Xiaoxiao Li. 2024. Unlocking the potential of prompt-tuning in bridging generalized and personalized federated learning. *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Enmao Diao, Jie Ding, and Vahid Tarokh. 2021. Hetero{fl}: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Chenhe Dong, Yuexiang Xie, Bolin Ding, Ying Shen, and Yaliang Li. 2023. Tunable soft prompts are messengers in federated learning. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 14665–14675, Singapore. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Yichao Du, Zhirui Zhang, Linan Yue, Xu Huang, Yuqing Zhang, Tong Xu, Linli Xu, and Enhong Chen. 2024. Communication-efficient personalized federated learning for speech-to-text tasks. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10001–10005.

Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. 2023. Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. 2020. Local model poisoning attacks to byzantine-robust federated learning. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1605–1622.

Wenzhi Fang, Ziyi Yu, Yuning Jiang, Yuanming Shi, Colin N. Jones, and Yong Zhou. 2022. Communication-efficient stochastic zeroth-order optimization for federated learning. *IEEE Transactions on Signal Processing*, 70:5058–5073.

FedML. 2023. Fedllm.

Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. 2023a. Learning federated visual prompt in null space for mri reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8064–8073.

Haozhe Feng, Tianyu Pang, Chao Du, Wei Chen, Shuicheng Yan, and Min Lin. 2023b. Does federated learning really need backpropagation? *arXiv preprint arXiv:2301.12195*.

Xiachong Feng, Xiaocheng Feng, Xiyuan Du, Min-Yen Kan, and Bing Qin. 2023c. Adapter-based selective knowledge distillation for federated multi-domain meeting summarization. *arXiv preprint arXiv:2308.03275*.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

GDPR. 2016. Regulation (eu) 2016/679 of the european parliament and of the council.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947. Curran Associates, Inc.

Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, and Qiuyi Zhang. 2020. Gradientless descent: High-dimensional zeroth-order optimization. In *International Conference on Learning Representations*.

Jack Good, Jimit Majmudar, Christophe Dupuy, Jixuan Wang, Charith Peris, Clement Chung, Richard Zemel, and Rahul Gupta. 2023. Coordinated replay sample selection for continual federated learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 331–342, Singapore. Association for Computational Linguistics.

Gemini Team Google. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Eduard Gorbunov, Samuel Horváth, Peter Richtárik, and Gauthier Gidel. 2023. Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top. In *International Conference on Learning Representations*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Lei Guo, Ziang Lu, Junliang Yu, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2024. Prompt-enhanced federated content representation learning for cross-domain recommendation. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, page 3139–3149, New York, NY, USA. Association for Computing Machinery.

Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. 2023. Promptfl: Let federated participants cooperatively learn prompts instead of models - federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, pages 1–15.

Xu Guo and Han Yu. 2022. On the domain adaptation and generalization of pretrained language models: A survey. *arXiv preprint arXiv:2211.03154*.

Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. 2022. Recovering private text in federated learning of language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 8130–8143. Curran Associates, Inc.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

David Ha, Andrew M. Dai, and Quoc V. Le. 2017. Hypernetworks. In *International Conference on Learning Representations*.

Nikolaus Hansen and Andreas Ostermeier. 2001. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195.

Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Xinghua Zhu, Jianzong Wang, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annavaram, and Salman Avestimehr. 2020. Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xumin Huang, Peichun Li, Hongyang Du, Jiawen Kang, Dusit Niyato, Dong In Kim, and Yuan Wu. 2024. Federated learning-empowered ai-generated content in wireless networks. *IEEE Network*, pages 1–1.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2024. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, 34(5):2817–2825.

Malhar S Jere, Tyler Farnan, and Farinaz Koushanfar. 2020. A taxonomy of attacks on federated learning. *IEEE Security & Privacy*, 19(2).

Junteng Jia, Ke Li, Mani Malek, Kshitiz Malik, Jay Mahadeokar, Ozlem Kalinli, and Frank Seide. 2023. Joint federated learning and personalization for on-device asr. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *Computer Vision – ECCV 2022*, pages 709–727, Cham. Springer Nature Switzerland.

Jingang Jiang, Xiangyang Liu, and Chenyou Fan. 2023a. Low-parameter federated learning with large language models. *arXiv preprint arXiv:2307.13896*.

Lekang Jiang, Filip Svoboda, and Nicholas Donald Lane. 2023b. FDAPT: Federated domain-adaptive pre-training for language models. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

Yuang Jiang, Shiqiang Wang, Víctor Valls, Bong Jun Ko, Wei-Han Lee, Kin K. Leung, and Leandros Tassiulas. 2023c. Model pruning enables efficient federated learning on edge devices. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10374–10386.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAcctT '20, pages 306–316, New York, NY, USA. Association for Computing Machinery.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. 2022. Federated learning for healthcare domain - pipeline, applications and challenges. *ACM Trans. Comput. Healthcare*, 3(4).

Peter Kairouz et al. 2021. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210.

Yan Kang, Tao Fan, Hanlin Gu, Xiaojin Zhang, Lixin Fan, and Qiang Yang. 2024. Grounding foundation models through federated transfer learning: A general framework. *arXiv preprint arXiv:2311.17431*.

Yeachan Kim, Junho Kim, Wing-Lam Mok, Jun-Hyung Park, and SangKeun Lee. 2023. Client-customized adaptation for parameter-efficient federated learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1159–1172, Toronto, Canada. Association for Computational Linguistics.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. 2023. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. *arXiv preprint arXiv:2309.00363*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ang Li, Jingwei Sun, Binghui Wang, Lin Duan, Sicheng Li, Yiran Chen, and Hai Li. 2021a. Lotteryfl: Empower edge intelligence with personalized and communication-efficient federated learning. In *2021 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 68–79.

Guanghao Li, Wansen Wu, Yan Sun, Li Shen, Baoyuan Wu, and Dacheng Tao. 2024a. Visual prompt based personalized federated learning. *Transactions on Machine Learning Research*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021b. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. 2022. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 965–978.

Shenghui Li, Edith Ngai, and Thiemo Voigt. 2023a. Byzantine-robust aggregation in federated learning empowered industrial iot. *IEEE Transactions on Industrial Informatics*, 19(2):1165–1175.

Shenghui Li, Edith Ngai, Fanghua Ye, Li Ju, Tianru Zhang, and Thiemo Voigt. 2024b. Blades: A unified benchmark suite for byzantine attacks and defenses in federated learning. In *2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI)*.

Shenghui Li, Edith C.-H. Ngai, and Thiemo Voigt. 2023b. An experimental study of byzantine-robust aggregation schemes in federated learning. *IEEE Transactions on Big Data*, pages 1–13.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450.

Xi Li and Jiaqi Wang. 2024. Position paper: Assessing robustness, privacy, and fairness in federated learning integrated with foundation models. *arXiv preprint arXiv:2402.01857*.

Xi Li, Songhe Wang, Chen Wu, Hao Zhou, and Jiaqi Wang. 2023c. Backdoor threats from compromised foundation models to federated learning. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

Xi Li, Chen Wu, and Jiaqi Wang. 2023d. Unveiling backdoor risks brought by foundation models in heterogeneous federated learning. *arXiv preprint arXiv:2311.18350*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. Association for Computational Linguistics.

Zan Li and Li Chen. 2021. Communication-efficient decentralized zeroth-order method on heterogeneous data. In *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*.

Zihao Lin, Yan Sun, Yifan Shi, Xueqian Wang, Lifu Huang, Li Shen, and Dacheng Tao. 2023. Efficient federated prompt tuning for black-box large pre-trained models. *CoRR*, abs/2310.03123.

M Lincy and A Meena Kowshalya. 2020. Early detection of type-2 diabetes using federated learning. *International Journal of Scientific Research in Science, Engineering and Technology*, 12:257–267.

Zhenqing Ling, Daoyuan Chen, Liuyi Yao, Yaliang Li, and Ying Shen. 2024. On the convergence of zeroth-order federated tuning in large language models. *arXiv preprint arXiv:2402.05926*.

Zhengyang Lit, Shijing Sit, Jianzong Wang, and Jing Xiao. 2022. Federated split bert for heterogeneous text classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Tao Liu, Zhi Wang, Hui He, Wei Shi, Liangliang Lin, Ran An, and Chenhao Li. 2023b. Efficient and secure federated learning for financial applications. *Applied Sciences*, 13(10).

Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, and Meikang Qiu. 2023c. Differentially private low-rank adaptation of large language model using federated learning. *arXiv preprint arXiv:2312.17493*.

Yi Liu, Xiaohan Bi, Lei Li, Sishuo Chen, Wenkai Yang, and Xu Sun. 2023d. Communication efficient federated learning for multilingual neural machine translation with adapter. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5315–5328, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. 2023a. Fedclip: Fast generalization and personalization for CLIP in federated learning. *IEEE Data Eng. Bull.*, 46(1):52–66.

Wang Lu, Hao Yu, Jindong Wang, Damien Teney, Haohan Wang, Yiqiang Chen, Qiang Yang, Xing Xie, and Xiangyang Ji. 2023b. Zoopfl: Exploring black-box foundation models for personalized federated learning. *arXiv preprint arXiv:2310.05143*.

Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and Philip S. Yu. 2022. Privacy and robustness in federated learning: Attacks and defenses. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Shubham Malaviya, Manish Shukla, and Sachin Lodha. 2023. Reducing communication overhead in federated learning for pre-trained language models using parameter-efficient finetuning. In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pages 456–469. PMLR.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. 2023a. Fine-tuning language models with just forward passes. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. 2023b. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, pages 53038–53075. Curran Associates, Inc.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Andrea Manoel, Mirian del Carmen Hipolito Garcia, Tal Baumel, Shize Su, Jialei Chen, Robert Sim, Dan Miller, Danny Karmon, and Dimitrios Dimitriadis. 2023. Federated multilingual models for medical transcript analysis. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 147–162. PMLR.

Alessio Maritan, Subhrakanti Dey, and Luca Schenato. 2023. Fedzen: Towards superlinear zeroth-order federated learning via incremental hessian estimation. *arXiv preprint arXiv:2309.17174*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Vaikkunth Mugunthan, Antigoni Polychroniadou, David Byrd, and Tucker Hybinette Balch. 2019. Smpai: Secure multi-party computation for federated learning. In *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, volume 21. MIT Press Cambridge, MA, USA.

Duy Phuong Nguyen, J. Pablo Munoz, and Ali Jannesari. 2024. Flora: Enhancing vision-language models with parameter-efficient federated learning. *arXiv preprint arXiv:2404.15182*.

OpenAI. 2022. Chatgpt.

OpenAI. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dimitrios P. Panagoulias, Maria Virvou, and George A. Tsihrintzis. 2024. Evaluating llm – generated multimodal diagnosis from medical images and symptom analysis. *arXiv preprint arXiv:2402.01730*.

Sharnil Pandya, Gautam Srivastava, Rutvij Jhaveri, M. Rajasekhara Babu, Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta, Spyridon Mastorakis, Md. Jalil Piran, and Thippa Reddy Gadekallu. 2023. Federated learning for smart cities: A comprehensive survey. *Sustainable Energy Technologies and Assessments*, 55:102987.

Jungwuk Park, Dong-Jun Han, Minseok Choi, and Jaekyun Moon. 2021. Sageflow: Robust federated learning against both stragglers and adversaries. In *Advances in Neural Information Processing Systems*, volume 34, pages 840–851. Curran Associates, Inc.

Siqi Ping, Yuzhu Mao, Yang Liu, Xiao-Ping Zhang, and Wenbo Ding. 2024. FL-TAC: Enhanced fine-tuning in federated learning via low-rank, task-specific adapter clustering. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Zhen Qin, Daoyuan Chen, Bingchen Qian, Bolin Ding, Yaliang Li, and Shuiguang Deng. 2024. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. In *Proceedings of the 41th International Conference on Machine Learning*.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. 2024. Federated text-driven prompt generation for vision-language models. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Alec Radford and Jeffrey Wu. 2019. Rewon child, david luan, dario amodei, and ilya sutskever. 2019. *Language models are unsupervised multitask learners. OpenAI blog*, 1(8):9.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Swarna Priya Ramu, Parimala Boopalan, Quoc-Viet Pham, Praveen Kumar Reddy Maddikunta, Thien Huynh-The, Mamoun Alazab, Thanh Thi Nguyen, and Thippa Reddy Gadekallu. 2022. Federated learning enabled digital twins for smart cities: Concepts, recent advances, and future directions. *Sustainable Cities and Society*, 79:103663.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. 2021. Adaptive federated optimization. In *International Conference on Learning Representations*.

Scott Reed et al. 2022. A generalist agent. *Transactions on Machine Learning Research*. Featured Certification, Outstanding Certification.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. 2020. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2021–2031. PMLR.

Chao Ren, Han Yu, Hongyi Peng, Xiaoli Tang, Anran Li, Yulan Gao, Alysa Ziying Tan, Bo Zhao, Xiaoxiao Li, Zengxiang Li, and Qiang Yang. 2024. Advances and open challenges in federated learning with foundation models. *arXiv preprint arXiv:2404.15381*.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119.

Nuria Rodríguez-Barroso, Daniel Jiménez-López, M Victoria Luzón, Francisco Herrera, and Eugenio Martínez-Cámara. 2023. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173.

Holger R. Roth, Ziyue Xu, Yuan-Ting Hsieh, Adithya Renduchintala, Isaac Yang, Zhihong Zhang, Yuhong Wen, Sean Yang, Kevin Lu, Kristopher Kersten, Camir Ricketts, Daguang Xu, Chester Chen, Yan Cheng, and Andrew Feng. 2024. Empowering federated learning for massive models with nvidia flare. *arXiv preprint arXiv:2402.07792*.

Wang Rui, Yu Tong, Zhang Ruiyi, Kim Sungchul, Rossi Ryan A., Zhao Handong, Wu Junda, Mitra Subrata, Yao Lina, and Henao Ricardo. 2024. Personalized federated learning for text classification with gradient-free prompt tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sejin Seo, Seung-Woo Ko, Jihong Park, Seong-Lyun Kim, and Mehdi Bennis. 2021. Communication-efficient and personalized federated lottery ticket learning. In *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 581–585.

Suhail Mohmad Shah and Vincent K. N. Lau. 2023. Model compression for communication efficient federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5937–5951.

Lujia Shen, Shouling Ji, Xuhong Zhang, Jinfeng Li, Jing Chen, Jie Shi, Chengfang Fang, Jianwei Yin,

and Ting Wang. 2021. Backdoor pre-trained models can transfer to all. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21. ACM.

Jaemin Shin, Hyungjun Yoon, Seungjoo Lee, Sungjoon Park, Yunxin Liu, Jinho Choi, and Sung-Ju Lee. 2023a. 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*, pages 11971–11988, Singapore. Association for Computational Linguistics.

Jiyun Shin, Jinhyun Ahn, Honggu Kang, and Joonhyuk Kang. 2023b. Fedsplitx: Federated split learning for computationally-constrained heterogeneous clients. *arXiv preprint arXiv:2310.14579*.

Abhishek Singh, Praneeth Vepakomma, Otkrist Gupta, and Ramesh Raskar. 2019. Detailed comparison of communication efficiency of split learning and federated learning. *arXiv preprint arXiv:1909.09145*.

Shangchao Su, Bin Li, and Xiangyang Xue. 2023. Fedra: A random allocation strategy for federated tuning to unleash the power of heterogeneous clients. *arXiv preprint arXiv:2311.11227*.

Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. 2024. Federated adaptive prompt tuning for multi-domain collaborative learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):15117–15125.

Guangyu Sun, Matias Mendieta, Jun Luo, Shandong Wu, and Chen Chen. 2023. Fedperfix: Towards partial model personalization of vision transformers in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4988–4998.

Guangyu Sun, Matias Mendieta, Taojiannan Yang, and Chen Chen. 2022a. Conquering the communication constraints to enable large pre-trained models in federated learning. *arXiv preprint arXiv:2210.01708*.

Jingwei Sun, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yiran Chen, and Holger R Roth. 2024a. Fedbpt: Efficient federated black-box prompt tuning for large language models. In *Proceedings of the 41th International Conference on Machine Learning*.

Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. 2022b. BBTv2: Towards a gradient-free future with large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3930, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022c. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20841–20855. PMLR.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. 2024b. Improving loRA in privacy-preserving federated learning. In *The Twelfth International Conference on Learning Representations*.

Rishub Tamirisa, John Won, Chengjun Lu, Ron Arel, and Andy Zhou. 2024. Fedselect: Customized selection of parameters for fine-tuning during personalized federated learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Buse G. A. Tekgul, Yuxi Xia, Samuel Marchal, and N. Asokan. 2021. Waffle: Watermarking in federated learning. In *2021 40th International Symposium on Reliable Distributed Systems (SRDS)*, pages 310–320.

Chandra Thapa, Pathum Chamikara Mahawaga Arachchige, Seyit Camtepe, and Lichao Sun. 2022. Splitfed: When federated learning meets split learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8485–8493.

Yuanyishu Tian, Yao Wan, Lingjuan Lyu, Dezhong Yao, Hai Jin, and Lichao Sun. 2022. Fedbert: When federated learning meets pre-training. *ACM Trans. Intell. Syst. Technol.*, 13(4).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vasileios Tsouvalas, Yuki Asano, and Aaqib Saeed. 2023. Federated fine-tuning of foundation models via probabilistic masking. *arXiv preprint arXiv:2311.17299*.

Minh Vu, Truc Nguyen, Tre' Jeter, and My T. Thai. 2024. Analysis of privacy leakage in federated large language models. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1423–1431. PMLR.

Eric Wang. 2023. Alpaca-lora.

Haoyu Wang, Handong Zhao, Yaqing Wang, Tong Yu, Jiuxiang Gu, and Jing Gao. 2022. Fedkc: Federated knowledge composition for multilingual natural language understanding. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1839–1850, New York, NY, USA. Association for Computing Machinery.

Xin'ao Wang, Huan Li, Ke Chen, and Lidan Shou. 2023. Fedbfpt: An efficient federated learning framework for bert further pre-training. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4344–4352. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469.

Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie, and Benjamin Van Durme. 2022. Pre-trained models for multilingual federated learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1413–1421, Seattle, United States. Association for Computational Linguistics.

Herbert Woisetschläger, Alexander Isenko, Shiqiang Wang, Ruben Mayer, and Hans-Arno Jacobsen. 2024. A survey on efficient federated learning methods for foundation model training. *arXiv preprint arXiv:2401.04472*.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020a. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 2023a. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632*.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2023b. A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*.

Panlong Wu, Kangshuo Li, Ting Wang, and Fangxin Wang. 2023c. Fedms: Federated learning with mixture of sparsely activated foundations models. *arXiv preprint arXiv:2312.15926*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B. Giannakis. 2020b. Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. *IEEE Transactions on Signal Processing*, 68:4583–4596.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. 2020. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*.

Yuexiang Xie, Zhen Wang, Dawei Gao, Daoyuan Chen, Liuyi Yao, Weirui Kuang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2023. Federatedscope: A flexible federated learning platform for heterogeneity. *Proc. VLDB Endow.*, 16(5):1059–1072.

Chang Xu, Yu Jia, Liehuang Zhu, Chuan Zhang, Guoxie Jin, and Kashif Sharif. 2022. Tdfl: Truth discovery based byzantine robust federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(12).

Mengwei Xu, Dongqi Cai, Yaozong Wu, Xiang Li, and Shangguang Wang. 2024a. Fwdllm: Efficient fedllm using forward gradient. *arXiv preprint arXiv:2308.13894*.

Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, Qiyang Zhang, Zhenyan Lu, Li Zhang, Shangguang Wang, Yuanchun Li, Yunxin Liu, Xin Jin, and Xuanzhe Liu. 2024b. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*.

Minrui Xu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Shiwen Mao, Zhu Han, Abbas Jamalipour, Dong In Kim, Xuemin Shen, Victor C. M. Leung, and H. Vincent Poor. 2024c. Unleashing the power of edge-cloud generative ai in mobile networks: A survey of aigc services. *IEEE Communications Surveys & Tutorials*, pages 1–1.

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan Mcmahan, Jesse Rosenstock, and Yuanbo Zhang. 2023. Federated learning of gboard language models with differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 629–639, Toronto, Canada. Association for Computational Linguistics.

Fu-En Yang, Chien-Yi Wang, and Yu-Chiang Frank Wang. 2023a. Efficient model personalization in federated learning via client-specific prompt generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19102–19111.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023b. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Xin Yang, Hao Yu, Xin Gao, Hao Wang, Junbo Zhang, and Tianrui Li. 2024a. Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering*, page 1–20.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. 2024b. Dual-personalizing adapter for federated foundation models. *arXiv preprint arXiv:2403.19211*.

Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, and Ming-Hsuan Yang. 2022. Federated multi-target domain adaptation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1081–1090.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. 2024. Openfedllm: Training large language models on decentralized private data via federated learning. *arXiv preprint arXiv:2402.06954*.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*. PMLR.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. 2023a. Multimodal federated learning via contrastive representation ensemble. In *The Eleventh International Conference on Learning Representations*.

Shuyang Yu, Junyuan Hong, Yi Zeng, Fei Wang, Ruoxi Jia, and Jiayu Zhou. 2023b. Who leaked the model? tracking IP infringers in accountable federated learning. In *NeurIPS 2023 Workshop on Regulatable ML*.

Sixing Yu, J Pablo Muñoz, and Ali Jannesari. 2023c. Bridging the gap between foundation models and heterogeneous federated learning. *arXiv preprint arXiv:2310.00247*.

Sixing Yu, J Pablo Muñoz, and Ali Jannesari. 2023d. Federated foundation models: Privacy-preserving and collaborative learning for large models. *arXiv preprint arXiv:2305.11414*.

Huimin Zeng, Zhenrui Yue, Qian Jiang, and Dong Wang. 2024. Federated recommendation via hybrid retrieval augmented generation. *arXiv preprint arXiv:2403.04256*.

Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. 2020. BatchCrypt: Efficient homomorphic encryption for Cross-Silo federated learning. In *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, pages 493–506. USENIX Association.

Honglei Zhang, He Liu, Haoxuan Li, and Yidong Li. 2024a. Transfr: Transferable federated recommendation with pre-trained language models. *arXiv preprint arXiv:2402.01124*.

Honglei Zhang, Fangyuan Luo, Jun Wu, Xiangnan He, and Yidong Li. 2023a. Lightfr: Lightweight federated recommendation with privacy-preserving matrix factorization. *ACM Trans. Inf. Syst.*, 41(4).

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. 2024b. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6915–6919.

Jie Zhang, Xiaohua Qi, and Bo Zhao. 2023b. Federated generative learning with foundation models. *arXiv preprint arXiv:2306.16064*.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023c. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001*.

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023d. 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)*, pages 3492–3507, Toronto, Canada. Association for Computational Linguistics.

Zhuo Zhang, Jintao Huang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Yue Yu, Qifan Wang, Lizhen Qu, and Zenglin Xu. 2024c. Revisiting data reconstruction attacks on real-world dataset for federated natural language understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14080–14091, Torino, Italia. ELRA and ICCL.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023e. 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*, pages 9963–9977, Toronto, Canada. Association for Computational Linguistics.

Jujia Zhao, Wenjie Wang, Chen Xu, Zhaochun Ren, See-Kiong Ng, and Tat-Seng Chua. 2024a. Llm-based federated recommendation. *arXiv preprint arXiv:2402.09959*.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. 2024b. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *The Twelfth International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A

survey of large language models. *arXiv preprint arXiv:2303.18223*.

Fei Zheng, Chaochao Chen, Lingjuan Lyu, and Binhui Yao. 2023. Reducing communication for split learning by randomized top-k sparsification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Guo Zhihan, Zhang Yifei, Zhang Zhuo, Xu Zenglin, and King Irwin. 2024. Fedlfc: Towards efficient federated multilingual modeling with lora-based language family clustering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Weiming Zhuang, Chen Chen, and Lingjuan Lyu. 2023. When foundation model meets federated learning: Motivations, challenges, and future directions. *arXiv preprint arXiv:2306.15546*.

A Additional Details of FedPEFT

A.1 Additive Methods

A.1.1 Adapter Tuning

Adapter tuning integrates small-scale neural networks (known as "adapters") into the pre-trained models (Houlsby et al., 2019; Hu et al., 2022). A straightforward implementation of adapter tuning is to collaboratively train a shared adapter among all clients in the `FedAvg` manner, as highlighted by Sun et al. (2022a). Based on `FedAvg`, FedCLIP (Lu et al., 2023a) incorporates an attention-based adapter for the image encoder in CLIP models (Radford et al., 2021). In the domain of multilingual machine translation, where different language pairs exhibit substantial discrepancies in data distributions, Fed-MNMT (Liu et al., 2023d) explores clustering strategies that group adapter parameters and makes inner-cluster parameters aggregation for alleviating the undesirable effect of data discrepancy. Another representative approach named C2A (Kim et al., 2023) employs hypernetworks (Ha et al., 2017) to generate client-specific adapters by conditioning on the client's information, maximizing the utility of shared model parameters while minimizing the divergence caused by data heterogeneity.

A.1.2 Prompt Tuning

Prompt tuning incorporates trainable task-specific continuous prompt vectors at the input layer (Liu et al., 2023a; Dong et al., 2023). Compared to full fine-tuning, it achieves comparable performance but with $1000\times$ less parameter storage and communication (Jia et al., 2022). A variation of prompt tuning, FedPerfix (Sun et al., 2023) uses a local adapter to generate the prefixes and aggregate the original self-attention layers.

Depending on target modalities, prompt tuning in current literature can be further classified into three categories:

- *Textual Prompt Tuning.* Task-specific prompt embeddings are combined with the input text embeddings, which are subsequently fed into language models. These soft prompts serve as instructive contexts to influence the generation process of LLMs by steering the probability distribution of the next token (Dong et al., 2023).

- *Visual Prompt Tuning.* Taking inspiration from advances in efficiently tuning LLMs, prompts are also introduced in the input space of vision models (Jia et al., 2022). Naive implementations introduce prompts at the pixel level, acting as a form of data augmentation (Li et al., 2024a). Alternatively, one could also insert the prompts as latent vectors for the first Transformer layer (Deng et al., 2024; Yang et al., 2023a). Nevertheless, an empirical study (Jia et al., 2022) has suggested that it is easier for visual prompts to learn condensed task-dependent signals in the latent input space of Transformers.

- *Textual-Visual Prompt Tuning.* Unlike single-modal FMs, vision-language FMs can process and interpret both visual data and textual information, endowing them with powerful representation ability and transferability (Radford et al., 2021). Based on vision-language FMs like CLIP, textual-visual prompt tuning shows promising capabilities in FL (Guo et al., 2023), especially in cross-domain scenarios, where the model needs to generalize across varied domains and unseen classes (Qiu et al., 2024).

A.2 Comparison of FedPEFT methods

Figure 2 depicts the taxonomy of FedPEFT with representative methods. Note that some methods may belong to multiple overlapping categories. To

Table 1: Comparison of Federated Parameter-Efficient Fine-Tuning (FedPEFT) Methods.

Category		Representative Work	Modality	Model	# Full Params.	# Train. Params.	Training Accel.	Comm. Cost
Selective		RaFFM (Yu et al., 2023c)	Txt.	BERT-Large (2019)	336M	100M	6.13×	29.8%
		FedBF (Zhang et al., 2023e)	Txt.	Roberta-Base (2019)	125M	0.66M		1.6%
Additive	**Adapter**	FedAP (Zhang et al., 2023e)	Txt.	Roberta-Base (2019)	125M	2M		1.6%
		FedCLIP (Lu et al., 2023a)	Vis.-Txt.	ViT-B/32 (2020a)	150M	0.53M		3.5%
		FedDAT (Chen et al., 2024)	Vis.-Txt.	ALBEF (2021b)	290M	2.86M		9.9%
		C2A (Kim et al., 2023)	Txt.	DistilBERT (2020)	66M	0.06M		0.1%
		Fed-MNMT (Liu et al., 2023d)	Txt.	mBART-50 (2020)	611M	8M		1.3%
		AdaFL (Cai et al., 2023)	Txt.	BERT (2019)	110M	0.61M	1.63×	0.6%
	Prompt	PromptFL (Guo et al., 2023)	Vis.-Txt.	ViT-B/16 (2021)	87M	0.87M	2.38×	0.9%
		MFPT (Zhao et al., 2024b)	Txt.	XLM-RoBERTa (2020)	270M	1.2M		0.4%
		FedAPT (Su et al., 2024)	Vis.-Txt.	ViT-B/32 (2020a)	88M	2.8M		3.2%
		FedSP (Dong et al., 2023)	Txt.	GPT2-XL (2019)	1.6B	111M		0.5%
Reparameterization-based Methods		SLoRA (Babakniya et al., 2023a)	Txt.	DistilBERT (2020)	67M	0.7M	13.47×	5.8%
		LP-FL (Jiang et al., 2023a)	Txt.	BERT-Large (2019)	336M	100M		30%
		FedMS (Wu et al., 2023c)	Vis.-Txt.	ViT-B/16 (2021)	87M	8.6M		10%
		pFedS2T (Du et al., 2024)	Aud.	Whisper (2023)	254M	10.1M		4%
		FFA-LoRA (Sun et al., 2024b)	Txt.	RoBERTa-Large (2019)	355M	0.39M		0.1%

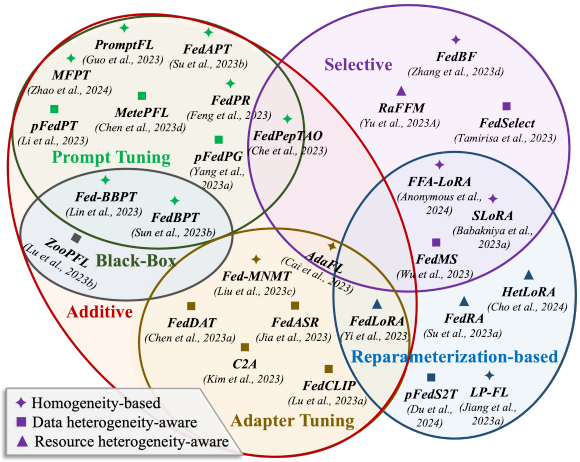

Figure 2: Taxonomy of Federated Parameter-Efficient Fine-Tuning (FedPEFT). Apart from efficiency, some methods also account for other considerations, such as data and resource heterogeneity challenges that are identified in Section 3.2 and black-box tuning (see Section 4.3).

compare the communication efficiency of different FedPEFT methods, Table 1 gives a brief overview of experimental evaluations from representative studies. Compared to full-model fine-tuning, FedPEFT methods only require 0.1%-30% communication overhead. We note that the differences can be attributed to several factors, including model complexity and implementation details.

B Additional Details of Trustworthiness

Due to the distributed characteristic of optimization, FL is vulnerable to poisoning attacks (Lyu et al., 2022; Rodríguez-Barroso et al., 2023), wherein certain participants may deviate from the prescribed update protocol and upload arbitrary parameters to the central server.

Poisoning Attacks. Depending on the adversarial goals, poisoning attacks in FL can be classified as *targeted* and *untargeted* (Jere et al., 2020). Targeted attacks, like backdoor attacks, aim to manipulate the global model to generate attacker-desired misclassifications for some particular samples (Xie et al., 2020; Bagdasaryan et al., 2020), while untargeted attacks seek to degrade the overall performance of the model indiscriminately (Fang et al., 2020). In addition to the well-recognized attacks on conventional FL studies (Li et al., 2023b, 2024b), FM-FL also faces potential threats from compromised pre-trained FMs (Li et al., 2023c). Thus, The attacker can introduce backdoors to downstream tasks without prior knowledge (Shen et al., 2021). Specifically, Li et al. (2023d) proposed Fed-EBD that introduces a backdoor-compromised FM to generate a public, synthetic dataset for FL training. The clients' models, pre-trained on this dataset, inherit the backdoor throughout the training.

Defense Techniques. As for defenses, robust aggregation rules are widely applied to make an attack-resilient estimation of the true updates and exclude the influence of malicious updates (Blanchard et al., 2017; Yin et al., 2018; Chen et al., 2017; Li et al., 2023a). Other research directions include trust-based strategies (Cao et al., 2021;

Xu et al., 2022; Park et al., 2021) and variance-reduced algorithms (Gorbunov et al., 2023; Wu et al., 2020b). Although these techniques have been widely examined in various FL settings, their effectiveness has yet to be explored in the FM-FL paradigm.

C Applications of FM-FL

In this part, we briefly review the recent progress on FM-FL applications. Table 2 lists representative work on specific applications and domains.

C.1 FM-FL for Multilingual NLP

Multilingual NLP refers to the techniques that handle multiple natural languages (Pires et al., 2019), often to perform equally well across them (Wu and Dredze, 2020). Earlier research (Johnson et al., 2017) has shown that parameter sharing among different languages boosts the model's performance in multilingual NLP, especially for low-resource languages for which significantly less content is available. However, real-world multilingual text data is often distributed across devices or regions, with each client (user) accessing only a limited subset of languages, where transferring the data to a central server is often problematic or prohibited due to privacy issues (Wang et al., 2022). Thanks to its inherent privacy-preserving characteristic, FL holds promise in breaking the barriers of cross-lingual modeling and data isolation by allowing models to learn from decentralized datasets.

The pioneer work by Weller et al. (2022) has firstly demonstrated that fine-tuning pre-trained language models with FL can perform similarly to pre-trained models fine-tuned with the standard centralized method under multilingual NLP settings. Various subsequent studies have focused on adapting pre-trained FMs through FedPEFT techniques such as adapter tuning (Liu et al., 2023d), prompt tuning (Zhao et al., 2024b), and LoRA (Zhihan et al., 2024), aiming to enhance training efficiency.

Considering the adverse effect of conflicting parameters from diverse languages during federated fine-tuning, recent studies have exploited clustering strategies to alleviate this issue. For instance, Wang et al. (2022) applied k-means clustering on each client's data to obtain representative knowledge, specifically the clustered data centroids. These centroids were then shared across clients for local training, enriching training data and addressing the challenges associated with data heterogeneity. Another compelling strategy along this line is language family-based clustering. Liu et al. (2023d) explored various clustering strategies to group adapter parameters to mitigate the negative effects of multilingual data heterogeneity, showing that language family-based clustering significantly outperforms the other clustering strategies. Similarly, Zhihan et al. (2024) proposed fine-tuning FMs with LoRA and language family-based clustering to address the heterogeneity issue of multilingual modeling.

General downstream tasks include language modeling (Wang et al., 2022), machine translation (Liu et al., 2023d; Chu et al., 2024), and text classification (Weller et al., 2022). In addition, some studies also focus on more specific applications such as medical transcript analysis (Manoel et al., 2023) and hate speech detection (Akshay and Rahul, 2024). These advancements illustrate the applicability of FM-FL across a wide range of scenarios in multilingual NLP.

C.2 FM-FL for Speech

With the development of AI, researchers have also carried out many studies on speech-related FMs, *e.g.,* wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023). In this field, the adaptation of FMs often relies on FL to facilitate scenarios where the audio data is privacy-sensitive. Compared to other data modalities, speech-related FM-FL applications especially attract excessive attention to the aspects of **on-device training** and **personalization**, motivated by the following considerations: (1) Audio data is continually generated on end-devices such as mobile phones, and owned by individual users—thus it should be processed locally, rather than being transferred elsewhere; (2) Although FL takes advantage of all user data to collectively train one model that maximizes speaker-independent accuracy, such a one-model-fits-all solution can be sub-optimal for individual users (Jia et al., 2023). Specific tasks in this field include Automatic Speech Recognition (ASR) (Azam et al., 2023b) and Speech-to-Text (S2T) (Du et al., 2024).

C.3 FM-FL for Recommendation

Federated Recommendation (FR) strives to capture underlying user preferences and recommend appropriate information to users while safeguarding data privacy (Bobadilla et al., 2013; Zhang et al., 2023a). Typical FR systems consist of a server and multiple clients, where clients represent individual users or local data servers possessing

Table 2: A list of representative studies on the applications of FM-FL. Abbreviations: LoRA Tuning (LT), Adapter Tuning (AT), Full-Parameter Tuning (FT), Selective Tuning (ST), Prompt Tuning (PT).

Domain/Application	Task	Representative Work	On-Device	Personalization	Modality	Backbone	Fine-Tuning
Multilingual NLP	Language Understanding	FedKC (Wang et al., 2022)	✗	✗	Txt.	mBERT	FT
	Multi-Tasks	PMMFL (Weller et al., 2022)	✗	✗	Txt.	mBERT	FT
	Machine Translation	Fed-MNMT (Liu et al., 2023d)	✗	✗	Txt.	mBART-50	AT
	Machine Translation	FL-MetaSend (Chu et al., 2024)	✗	✗	Txt.	M2M-100	ST
	Multi-Tasks	MFPT (Zhao et al., 2024b)	✓	✗	Txt.	XLM-RoBERTa	PT
Speech	Speech-to-Text	pFedS2T (Du et al., 2024)	✗	✓	Aud.	Conformer/Whisper	LT
	Speech Recognition	FedASR (Jia et al., 2023)	✓	✓	Aud.	RNN-T	AT
	Speech Recognition	FedE2EASR(Azam et al., 2023a)	✗	✓	Aud.	CTC-AED	FT
Recommendation	*General*	PPLR (Zhao et al., 2024a)	✗	✓	Txt.	LLaMA-7B/LongFormer	FT
	General	TransFR (Zhang et al., 2024a)	✓	✓	Txt.	DistBERT	AT
	General	GPT-FedRec (Zeng et al., 2024)	✗	✗	Txt.	ChatGPT	*NA*
Healthcare	Mental Health Prediction	FedTherapist (Shin et al., 2023a)	✓	✗	Txt.	BERT & LLaMa-7B	LT
	MRI Reconstruction	FedPR (Feng et al., 2023a)	✗	✗	Vis.	Swin Transformers	PT

smaller datasets and retaining private user information (Ammad-Ud-Din et al., 2019). These clients collaborate to train a global model while ensuring their data privacy protection by abstaining from direct data sharing (Zeng et al., 2024; Zhang et al., 2023a). Recently, LLM-based recommendations have been gaining increasing attention (Wu et al., 2023b) due to their strong capacities in language understanding and domain generalization. The benefits are mainly twofold: (1) LLMs mitigate the cold-start issue by utilizing textual descriptions to make recommendations without the need for extensive historical data (Zhang et al., 2023c); (2) The inherent transferability of LLMs allows them to apply cross-domain knowledge and side information to improve accuracy and relevance across diverse items and user interests (Gao et al., 2023).

One straightforward way to adapt FMs for FR is by fine-tuning them with historical user-item data. More specifically, FedPEFT techniques such as adapter tuning (Zhang et al., 2024a) and split learning (Zhao et al., 2024a) can be employed to improve resource efficiency. Apart from parameter fine-tuning, LLMs can also be adapted to assist the recommendation in a zero-shot paradigm through prompt engineering (*i.e.,* without parameter tuning) (Gao et al., 2023). For example, Zeng et al. (2024) proposed GPT-FedRec, a two-stage FR framework that leverages ChatGPT for its powerful zero-shot generalization ability. Firstly, GPT-FedRec facilitates hybrid retrieval by collaboratively training ID and text retrievers, after which the retrieved results are transformed into text prompts and submitted to GPT for re-ranking in the second stage. Additionally, Guo et al. (2024) employed a pre-trained BERT to obtain the representation vectors of item descriptions, which are then fed into a recommender system as augmented input.

C.4 FM-FL for Healthcare

FMs, especially LLMs, have been found to excel in healthcare applications, showcasing impressive capabilities in tasks like mental health analysis (Yang et al., 2023b), disease diagnosis (Panagoulias et al., 2024), and drug discovery (Chenthamarakshan et al., 2023). However, it raises privacy concerns to upload the health information of patients (Tang et al., 2023) into a commercial server that supports the FMs. Meanwhile, FL has consistently received widespread attention in the healthcare domain (Lincy and Kowshalya, 2020; Rieke et al., 2020; Joshi et al., 2022), driven by the need for collaborative model training across different medical institutions without compromising patient data privacy. By breaking the barriers of private data availability, the FM-FL paradigm shows the potential to further harness the power of FMs in the healthcare domain.

A recent study (Shin et al., 2023a) presents a mobile mental health monitoring system, FedTherapist, which leverages user speech and keyboard input to fine-tune FMs with FL, demonstrating superior accuracy in mental health prediction tasks such

Table 3: A list of existing FM-FL libraries and benchmarks. **Missing or inapplicable details denoted by N/A.** ✓ denotes a strong focus or presence; ✗ indicates no focus or absence; ◐ signifies a moderate focus or partial inclusion.

Library/Benchmark	FL Backend	LLM Support	MultiModal FM Support	FedPEFT	On-Device Training	Distributed & Clustered	Differential Privacy	Description
FederatedScope-LLM (Kuang et al., 2023)	FederatedScope	✓	✗	✓	✗	✓	✗	An end-to-end benchmark for efficient fine-tuning LLMs with FL
NVIDIA FLARE (Roth et al., 2024)	NVFlare	✓	✗	✓	✗	✓	✓	Scalable and efficient fine-tuning LLMs with FL
FATE-LLM (Fan et al., 2023)	FATE	✓	✗	✓	✗	✓	✓	Focuses on IP and privacy protection in federated LLM
FedLLM (FedML, 2023)	FedML	✓	✗	◐	✓	✓	✓	An MLOps-supported training pipeline based on FedML
OpenFedLLM (Ye et al., 2024)	N/A	✓	✗	◐	✗	N/A	✗	An LLM framework focusing on FL instruction tuning/alignment
Shepherd (Zhang et al., 2024b)	N/A	✓	✗	✓	✗	✗	✗	Federated instruction tuning based on Huggingface
FedPETuning (Zhang et al., 2023e)	FedLab	✓	✗	✓	✗	✓	✗	A benchmark comprising four FedPEFT methods
FedLegal (Zhang et al., 2023d)	FedLab	✓	✗	✗	✗	✓	✗	A benchmark comprising six legal NLP tasks under FL settings

as depression, stress, and mood prediction. Another representative study (Feng et al., 2023a) focuses on Magnetic Resonance Imaging (MRI) reconstruction, which involves retrieving a complex-valued image from its under-sampled signal. The authors adopted an FM pre-trained on public datasets and trained visual prompts from decentralized clinical datasets via a personalized FL mechanism, thereby reducing communication costs and achieving competitive performance on limited local data.

Despite the efforts, it has been shown that FMs in healthcare risk generating misleading information due to their imperfect understanding of complex medical data (Jeblick et al., 2024).

D Libraries and Benchmarks

This part briefly introduces a series of available libraries and benchmarks for developing and examining FM-FL techniques. An overview is provided in Table 3.

- **FederatedScope-LLM** (Kuang et al., 2023) is an open-source package for fine-tuning LLMs via FL. Built on top of a popular FL backend FederatedScope (Xie et al., 2023), it supports federated fine-tuning of LLMs under various FL scenarios, including FedPEFT and model personalization.
- **NVIDIA FLARE** (Roth et al., 2024) is an FL framework that allows researchers and data scientists to seamlessly move their machine learning and deep learning workflows into a federated paradigm.
- **FATE-LLM** (Fan et al., 2023) is an industrial-grade FL framework for LLM. Apart from FedPEFT, it provides a privacy hub integrating several IP protection and privacy-preserving mechanisms to protect model security and data privacy.
- **FedLLM** (FedML, 2023) is an MLOps-supported training pipeline built upon the FedML AI platform (He et al., 2020). FedLLM is compatible with popular LLM libraries such as HuggingFace and DeepSpeed to support a large range of FMs and datasets.
- **OpenFedLLM** (Ye et al., 2024) is a federated tuning framework for LLMs, which covers applications of instruction tuning and value alignment, diverse FL baselines, training datasets, and evaluation datasets.
- **Shepherd** (Zhang et al., 2024b) is a lightweight federated tuning framework. The local training process of Shepherd is built upon the implementations of Alpaca-LoRA (Wang, 2023), and Hugging Face's PEFT (Mangrulkar et al., 2022), enabling efficient fine-tuning.
- **FedPETuning** (Zhang et al., 2023e) is a pioneering federated benchmark for four representative FedPEFT methods, covering adapter tuning, prefix tuning, LoRA, and BitFit.
- **FedLegal** (Zhang et al., 2023d) is the very first real-world FL benchmark for legal NLP, which comprises five legal NLP tasks and one privacy task based on the data from Chinese courts.