The Marriage of 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, 2023), 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), and Gemini (Google, 2023) 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. However, it is 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 AI models.

The paper continues as follows: The next section introduces background on FMs and FL. Section 3 presents the motivation and challenges for marrying FMs and FL. Section 4 highlights representative techniques. Before concluding, we discuss future directions and applications 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., 2023b), 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, 2023), demonstrate exceptional prowess 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, introduced by 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 distributed learning paradigm that involves periodic interactions between the server and decentralized clients for the exchange of trainable model parameters, without the need for private client data. FL offers a privacy-preserving and efficient way to train models on a 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 Marriage: Motivation and Challenges

In this section, we first motivate the marriage of FMs and FL (§3.1), then summarize the key challenges stemming from the FM-FL marriage (§3.2).

3.1 Motivation

The AI sector widely concurs that the capabilities of FMs are fundamentally driven by large-scale and high-quality datasets (Bommasani et al., 2021; Kaplan et al., 2020; Gunasekar et al., 2023), which encompass both public and private sources. Leveraging private data for training FMs presents considerable challenges owing to privacy concerns. Privacy regulations and data protection laws often prohibit sharing sensitive information (GDPR, 2016; CCPA, 2023), limiting the feasibility of traditional data-centralized training processes.

The FM-FL marriage represents a compelling collaboration that utilizes the strengths of each 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., 2023c; Bonawitz et al., 2021).

FMs boost FL with knowledge and understanding capabilities. By pre-training on large-scale generic data, FMs acquire essential knowledge and understanding capabilities (Brown et al., 2020). Firstly, they benefit FL systems by offering advanced feature representations and learning capabilities from the outset. Secondly, leveraging the pre-learned knowledge of FMs accelerates 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 enhancing model convergence (Huang et al., 2024).

3.2 Challenges

In this part, we discuss challenges emerging from the FM-FL marriage in three levels: *Task-Level*, *System-Level*, as well as *Governance-Level*. Due to the space limitation, we only list the most representative challenges for each level.

3.2.1 Task-Level Challenges

Task-level challenges stem from the adaptation of an FM to a specific downstream task (e.g., by fine-tuning) in a federated setting. Challenges include:

Data 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.

Federated Alignment Tuning. Model alignment is the process of ensuring that FMs behave in line with human intentions and values (Ji et al., 2024). The distributed nature of FL, where data remains on local devices, and the diversity of data exacerbate the difficulty of ensuring fairness, transparency, and accountability in models (Ezzeldin et al., 2023).

3.2.2 System-Level Challenges

System-level challenges stem from the mismatch between the significant resource demands of FM training and the limited, heterogeneous system resources (e.g., for mobile devices) within FL systems, such as communication bandwidth, computational power, and memory (Su et al., 2023a). This line of challenges include:

Communication Efficiency. In FL, the communication bottleneck 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.

System 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 setting.

3.2.3 Trustworthiness Challenges

This level emphasizes the overarching concerns such as ethical considerations, privacy, security, and fairness in the entire lifecycle of FM-FL, from the pre-training and model adaptation to the application stages. We present two representative types of challenges from this perspective:

Intellectual Property. Intellectual property (IP) protection in FM-FL primarily involves attributing ownership rights for both models and data. FL complicates the identification of contributions from multiple participants, raising questions about who holds the IP for developed models and the FM-generated contents (Li et al., 2023a). 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).

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). In terms of FMs, recent studies (Huang et al., 2022; Li et al., 2023c) have shown that generative models still suffer from privacy leakage threats through well-crafted prompts (e.g., jailbreak prompts), even though safety mechanisms are implemented.

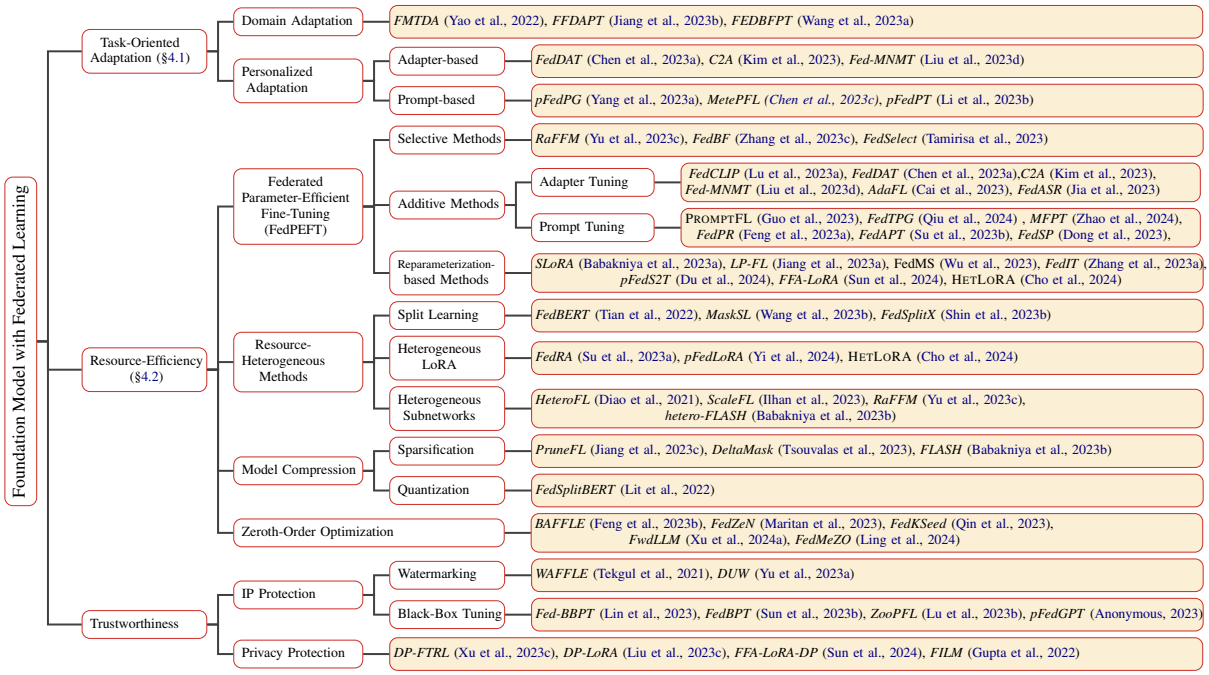

Figure 1: Taxonomy of research in foundation models with federated learning

4 Techniques

In this section, we survey FM-FL techniques, categorizing them as *Task-Oriented Adaptation* (§4.1), *Resource-Efficiency* (§4.2), and *Trustworthiness* (§4.3). As shown in Figure 1, we further refine them according to the key features of different methods.

4.1 Task-Oriented Adaptation

This part introduces major approaches that adapt FMs to handle specific tasks with FL, aiming to tackle task-level challenges.

4.1.1 Domain Adaptation

Despite being heavily reliant on large-scale, 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. (2023a) 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.

4.1.2 Personalized Adaptation

Personalized adaptation refers to the process of tailoring a pre-trained 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[1]:

Adapter-based. Adapter-based methods introduce small, trainable modules (adapters) into the frozen pre-trained FMs, allowing for client-specific model adaptation without altering the original FL. *FedDAT* (Chen et al., 2023a) 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. 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 personal-

[1]While this classification intersects with FedPEFT (§4.2.1), which is detailed later, the focus here is on personalization aspects.

ized prompts for each client enabling efficient and personalized adaptation.

4.2 Resource-Efficiency

In response to the system-level challenges, there has been a considerable focus on developing resource-efficient approaches. This part describes techniques that improve resource efficiency.

4.2.1 Federated 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*. We depict this taxonomy and representative methods in Figure 2.

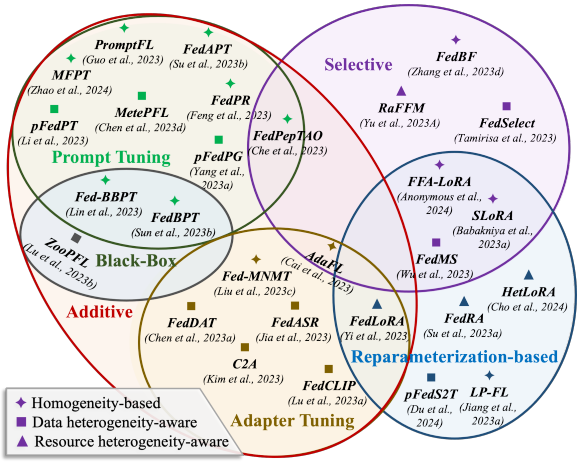

Figure 2: Taxonomy of Federated Parameter-Efficient Fine-Tuning (FedPEFT).

Selective Methods. Selective methods fine-tune a small subset of the parameters, leaving the majority unchanged during fine-tuning. 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., 2023c), 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., 2021; Tamirisa et al., 2023) 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., 2023) 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 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. Additive methods include the following representative branches:

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.

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 (Jia et al., 2022). A variation of prompt tuning, *FedPerfix* (Sun et al., 2023a) uses a local adapter to generate the prefixes and aggregate the original self-attention layers.

Reparameterization-based Methods. The hypothesis behind reparameterization-based methods is that fine-tuning adaptions 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). For instance, *FedIT* (Zhang et al., 2023a) leverages LoRA to improve the response quality of LLMs by utilizing diverse instructions from different clients. Noticeably, LoRA and its variants have also exhibited considerable potential in addressing the challenges inherent in data heterogeneity among clients in FL. *FedLoRA* (Yi et al., 2024) assigns a homogeneous small low-rank linear adapter for each clients local personalized heterogeneous local model.

4.2.2 Resource-Heterogeneous Methods

FL systems may consist of devices with varying levels of resources, leading to disparities where certain devices exhibit more efficient model training compared to others (Chen et al., 2023a). To address this, several methods have been developed to customize model architectures for heterogeneous clients.

Heterogeneous LoRA. LoRA-based FedPEFT exhibits unique flexibility for resource-limited mobile devices with natural system heterogeneity. Cho et al. (2023) applied heterogeneous LoRA ranks across clients via utilizing zero-padding and truncation for the aggregation and distribution of the LoRA modules. *FedRA* (Su et al., 2023a) integrates LoRA with randomly-allocated subnetworks for local fine-tuning with heterogeneous clients.

Heterogeneous Subnetworks. Some works train heterogeneous subnetworks selected from global models, tailored to the varying capabilities of individual clients. *HeteroFL* (Diao et al., 2021) appeared as the first method that adaptively allocates subsets of global model parameters for local training. *ScaleFL* (Ilhan et al., 2023) integrates a resource-adaptive 2-D model downscaling mechanism along the width and depth dimensions by leveraging early exits to find the best-fit models for resource-aware local training.

Split Learning. Split learning addresses the resource heterogeneity between servers and clients by splitting a large model into client-side and server-side components (Thapa et al., 2022). For the first time, *FedBERT* (Tian et al., 2022) leverages split training for training the BERT model, showing the feasibility of pre-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 diverse client capabilities.

4.2.3 Model Compression

Model compression refers to the techniques used to reduce the size of models, thereby improving communication and computational 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 both effective and easy to implement, making it ideal for use with other resource-efficient methods (Lit et al., 2022).

4.2.4 Zeroth-Order Optimization

Distinct from the ubiquitous reliance on gradient descent in most FL optimization algorithms, a specific line of research advocates for the removal of backpropagation (BP) (Malladi et al., 2023) in favor of Zeroth-Order Optimization (ZOO) (Fang et al., 2022; Li and Chen, 2021). BP-free approaches conserve memory needed for computing gradients and minimize communication overhead for model aggregation (Qin et al., 2023), making FMs more accessible for lower-end devices,

thereby enhancing their applicability in diverse hardware environments of FL. Recent work based on perturbed inferences, such as that by Xu et al. (2023a); Qin et al. (2023), has initiated preliminary explorations into the deployment of both FedPEFT and full-model fine-tuning of billion-sized FMs, like LLaMA, on mobile devices.

4.3 Trustworthiness

This line of work aims to enhance trustworthiness throughout the FM-FL lifecycle, covering a variety of key areas including, but not limited to, *IP Protection*, *Attack Robustness*, and *Privacy Protection*.

IP Protection. Existing IP protection involves safeguarding ownership of FL models from unauthorized use (e.g., model theft) (Tekgul et al., 2021). Two common kinds of IP protection strategies are watermarking and black-box tuning.

Watermarking is a well-known deterrence method for model IP protection by providing the identities for 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 clients local model, aiming to identify the infringer of a leaked model while verifying the FL models ownership.

Black-Box Tuning is a set of gradient-free methods to drive large language models. ZOO allows for black-box fine-tuning in scenarios where direct access to model parameters is restricted, e.g., due to privacy concerns or proprietary limitations (Sun et al., 2022b). *Fed-BBPT* (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., 2023b) adopts a classic evolutionary-based ZOO method, CMA-ES (Hansen and Ostermeier, 2001), for training an optimal prompt that improves the performance of the 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. Nevertheless, the pronounced slower convergence rates of ZOO compared to gradient-based approaches in high-dimensional settings (Golovin et al., 2020), underscore a significant research gap. The implications of these slower rates on convergence efficiency and computational burden in FL, especially for large-scale FMs, remain insufficiently explored.

Differential Privacy. Differential Privacy (DP) is a theoretical framework that governs privacy boundaries and manages the tradeoff between privacy and model convergence (Wei et al., 2020). 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., 2023c). Besides, DP is compatible with most FedPEFT methods. For instance, Sun et al. (2024) 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 terms of attack, Gupta et al. (2022) presented an attack that recovers private text data by extracting information from gradients transmitted during training, despite the employment of a naive DP mechanism.

5 Future Directions & Applications

We highlight potential research directions and future FL-FM applications in this section.

5.1 Future Directions

Personalization. FL on FMs can improve user profiling by capturing more granular and diverse data from individuals while preserving privacy. This can lead to more accurate and comprehensive user profiles, enabling personalized recommendations and services tailored to specific preferences, needs, and contexts (Chen et al., 2023b). Users can contribute their preferences, feedback, and insights, allowing the models to learn directly from their interactions and refine personalization algorithms accordingly. Future research directions may incorporate multi-modal data, including text, images, audio, and sensor data.

Model Compression. Future directions may involve designing more efficient and lightweight model compression techniques (Deng et al., 2020) specifically tailored for FL systems to reduce the computational and memory requirements of FMs while maintaining their performance. They may leverage multi-task learning approaches for model sharing and parameter reuse across different tasks or domains. Adaptive model compression techniques could dynamically adjust the compression level based on the available computing resources or application requirements (Xu et al., 2023b).

Split Learning. Split learning (Thapa et al., 2020) partitions the model, placing one part on the client device and the other on the server. Future developments may explore more sophisticated and adaptive methods for partitioning the FMs, such as adaptive model partitioning based on the computational capabilities and resources of the client devices. Dynamic model partitioning techniques may adjust the partitioning scheme at different stages of the FL process.

Mixture of Experts (MoE). MoE allows FL to incorporate multiple expert FMs, each specializing in different aspects or domains of the data. By combining the expertise of these models, FL can achieve higher model performance and accuracy. MoE also allows FL and FM models to adapt to local data characteristics present on individual client devices. Bridging MoE with FMs could strengthen generalization ability by balancing between larger overall model capacity and flexible per-instance specialization (Cong et al., 2023).

Privacy Preservation. Advanced model aggregation methods can be designed to incorporate privacy awareness when performing FL on FMs. This includes techniques to control the amount of information leaked during the aggregation process and mechanisms to enforce privacy guarantees while maintaining the accuracy and utility of the aggregated model (Nagy et al., 2023). As privacy concerns continue to grow, future developments may involve the establishment of privacy regulations and standards specifically tailored for FL and FMs.

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., 2023b). 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 learning and adaptation (Good et al., 2023).

Resource-Efficiency. FL-FM enables collaborative training, model adaptation, and utilization of FMs for a wide range of novel and powerful applications on heterogeneous edge devices (Shen et al., 2024; Xu et al., 2024b). The design and adaptation of the FM models, optimization of computation and communication, and coordination among heterogeneous edge devices and the cloud remain to be further explored in this new era.

5.2 Domain-Specific Applications

In this section, we discuss how FM-FL can be utilized in several representative domains.

Healthcare. FL and FMs enable the development of personalized medical applications. By training with massive individual patient data, such as medical history, genetics, and lifestyle factors, the models can learn global patterns and provide tailored recommendations for treatment, medication, and prevention strategies.

Law, Finance and Banking. FL can support the training of FMs on massive legal documents, cases, and statutes (Zhang et al., 2023b). The models can assist in identifying key legal arguments, summarizing case details, providing insights, and making predictive analytics into potential case outcomes (Yue et al., 2023). FL can build FMs that support risk management applications in banking and finance. By analyzing aggregated data from multiple sources, such as credit scores, market data, and economic trends, the models can support risk assessment, credit scoring, and investment management (Shin et al., 2023a). FM-FL models trained on historical investment data and market trends can support investment opportunities, analyze investment risks, and assist in portfolio optimization.

Education and Personal Agents. FL can be used to train FMs for intelligent tutoring systems to support individual student learning. Personalized foundation models can provide customized and personalized responses to users based on their individual interests, preferences, behavior, and history.

6 Conclusions

In this survey, we have meticulously surveyed the intersection of FM and FL. We identified three levels of challenges: task-level, system-level as well as trustworthiness challenges, 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 to include the latest research efforts in the community in this survey. The majority of the papers referenced in our taxonomy are indeed from 2023 which also demonstrates the importance of the integration of FM and FL. Therefore, we believe that our survey will help to inspire and push further research and innovation in this important areas. Our survey does not include any benchmarking of the available ideas and systems. We believe that would be an important next step that we are leaving to future work. It would, however, require some tools to support such an evaluation campaign and such tools are, to the best of our knowledge, not available yet.

References

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.

Anonymous. 2023. Personalized federated learning for text classification with gradient-free prompt tuning.

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*.

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.

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):87114.

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.

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.

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.

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. 2023a. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. *arXiv preprint arXiv:2308.12305*.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2023b. When large language models meet personalization: Perspectives of challenges and opportunities. *arXiv preprint arXiv:2307.16376*.

Shengchao Chen, Guodong Long, Tao Shen, and Jing Jiang. 2023c. Prompt federated learning for weather forecasting: Toward foundation models on meteorological data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 3532–3540. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, Matt Barnes, and Gauri Joshi. 2023. Heterogeneous loRA for federated fine-tuning of on-device foundation models. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

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

Wenyan Cong, Hanxue Liang, Peihao Wang, Zhiwen Fan, Tianlong Chen, Mukund Varma, Yi Wang, and Zhangyang Wang. 2023. Enhancing nerf akin to enhancing llms: Generalizable nerf transformer with mixture-of-view-experts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3170–3181.

Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4):485–532.

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.

Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A. Salman Avestimehr. 2023. Fairfed: Enabling group fairness in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7494–7502.

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.

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?

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

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*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

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.

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*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Fatih Ilhan, Gong Su, and Ling Liu. 2023. Scalefl: Resource-adaptive federated learning with heterogeneous clients. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24532–24541.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2024. Ai alignment: A comprehensive survey.

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.

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.

Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. 2022. Federated learning for healthcare domain-pipeline, applications and challenges. *ACM Transactions on Computing for Healthcare*, 3(4):1–36.

Peter Kairouz et al. 2021. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(12):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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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, Piotr Dollár, and Ross Girshick. 2023. Segment anything.

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. 2021. Lotteryfl: Empower edge intelligence with personalized and communication-efficient federated learning. In *2021 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 68–79.

Bowen Li, Lixin Fan, Hanlin Gu, Jie Li, and Qiang Yang. 2023a. Fedipr: Ownership verification for federated deep neural network models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4521–4536.

Guanghao Li, Wansen Wu, Yan Sun, Li Shen, Baoyuan Wu, and Dacheng Tao. 2023b. Visual prompt based personalized federated learning. *Transactions on Machine Learning Research*.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023c. Multi-step jailbreaking privacy attacks on ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4138–4153, Singapore. Association for Computational Linguistics.

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.

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

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.

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):5877.

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.

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.

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*.

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. 2023. Fine-tuning language models with just forward passes. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

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

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.

Balázs Nagy, István Hegedűs, Noémi Sándor, Balázs Egedi, Haaris Mehmood, Karthikeyan Saravanan, Gábor Lóki, and Ákos Kiss. 2023. Privacy-preserving federated learning and its application to natural language processing. *Knowledge-Based Systems*, 268:110475.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

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.

Zhen Qin, Daoyuan Chen, Bingchen Qian, Bolin Ding, Yaliang Li, and Shuiguang Deng. 2023. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. *arXiv preprint arXiv:2312.06353*.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. 2024. Text-driven prompt generation for vision-language models in federated learning. 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.

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.

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.

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.

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.

Yifei Shen, Jiawei Shao, Xinjie Zhang, Zehong Lin, Hao Pan, Dongsheng Li, Jun Zhang, and Khaled B Letaief. 2024. Large language models empowered autonomous edge AI for connected intelligence. *IEEE Communications Magazine*.

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.

Shangchao Su, Bin Li, and Xiangyang Xue. 2023a. 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. 2023b. Federated adaptive prompt tuning for multi-domain collaborative learning.

Guangyu Sun, Matias Mendieta, Jun Luo, Shandong Wu, and Chen Chen. 2023a. 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. 2023b. Fedbpt: Efficient federated black-box prompt tuning for large language models. *arXiv preprint arXiv:2310.01467*.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022b. 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. 2024. 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. 2023. Fedselect: Customized selection of parameters for fine-tuning during personalized federated learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*.

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, Mahawaga Arachchige Pathum Chamikara, and Seyit Camtepe. 2020. Splitfed: When federated learning meets split learning. *CoRR*, abs/2004.12088.

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.

Xin'ao Wang, Huan Li, Ke Chen, and Lidan Shou. 2023a. 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.

Zhousheng Wang, Geng Yang, Hua Dai, and Chunming Rong. 2023b. Privacy-preserving split learning for large-scaled vision pre-training. *IEEE Transactions on Information Forensics and Security*, 18:1539–1553.

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.

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*.

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

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

Mengwei Xu, Yaozong Wu, Dongqi Cai, Xiang Li, and Shangguang Wang. 2023a. Federated fine-tuning of billion-sized language models across mobile devices. *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.

Yang Xu, Yunming Liao, Hongli Xu, Zhenguo Ma, Lun Wang, and Jianchun Liu. 2023b. Adaptive control of local updating and model compression for efficient

federated learning. *IEEE Transactions on Mobile Computing*, 22(10):5675–5689.

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan Mcmahan, Jesse Rosenstock, and Yuanbo Zhang. 2023c. 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.

Xin Yang, Hao Yu, Xin Gao, Hao Wang, Junbo Zhang, and Tianrui Li. 2023b. Federated continual learning via knowledge fusion: A survey. *arXiv preprint arXiv:2312.16475*.

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.

Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. 2024. pfedlora: Model-heterogeneous personalized federated learning with lora tuning.

Shuyang Yu, Junyuan Hong, Yi Zeng, Fei Wang, Ruoxi Jia, and Jiayu Zhou. 2023a. Who leaked the model? tracking ip infringers in accountable federated learning.

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*.

Linan Yue, Qi Liu, Yichao Du, Weibo Gao, Ye Liu, and Fangzhou Yao. 2023. Fedjudge: Federated legal large language model. *arXiv preprint arXiv:2309.08173*.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. 2023a. Towards building the federatedGPT: Federated instruction tuning. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023b. 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, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023c. 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.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. 2024. 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*.

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*.