Federated Foundation Models: Privacy-Preserving and Collaborative Learning for Large Models

Sixing Yu[1], **J. Pablo Muñoz**[2], **Ali Jannesari**[1]
[1]Iowa State University
[2]Intel Labs
{yusx, jannesar}@iastate.edu, pablo.munoz@intel.com

Abstract

Foundation Models (FMs), such as LLaMA, BERT, GPT, ViT, and CLIP, have demonstrated remarkable success in a wide range of applications, driven by their ability to leverage vast amounts of data for pre-training. However, optimizing FMs often requires access to sensitive data, raising privacy concerns and limiting their applicability in many domains. In this paper, we propose the Federated Foundation Models (FFMs) paradigm, which combines the benefits of FMs and Federated Learning (FL) to enable privacy-preserving and collaborative learning across multiple end-users. We discuss the potential benefits and challenges of integrating FL into the lifespan of FMs, covering pre-training, fine-tuning, and application. We further outline potential future research avenues in FFM, including FFM pre-training, FFM fine-tuning, and federated prompt tuning, which allow the development of more personalized and context-aware models while ensuring data privacy. Moreover, we explore the possibility of continual/lifelong learning in FFMs, as increased computational power at the edge may unlock the potential for optimizing FMs using newly generated private data close to the data source. The proposed FFM concepts offer a flexible and scalable framework for training large language models in a privacy-preserving manner, setting the stage for subsequent advancements in both FM training and federated learning.

Keywords: Federated Learning, Foundation Models, Machine Learning, Data Privacy

1. Introduction

In recent years, Foundation Models (FMs) such as BERT (Kenton and Toutanova, 2019), GPT (Brown et al., 2020; Radford et al., 2019), Llama (Touvron et al., 2023a,b), ViT (Dosovitskiy et al., 2020), and CLIP (Radford et al., 2021) have significantly advanced the field of artificial intelligence, showcasing impressive performance across a wide range of tasks and domains. However, the optimization of increasingly complex FMs heavily depends on the collections of massive datasets, which introduces concerns regarding training data scarcity, computational resources, privacy, and ethical considerations. Simultaneously, the prevalent trend of advancement in edge technologies generates a vast amount of decentralized data, creating potential resources for further optimizing and specializing FMs. Nevertheless, due to privacy concerns, this private data is rarely leveraged for FM optimizations. In light of this, Federated Learning (FL) (McMahan et al., 2017) has emerged as a pioneering approach for decentralized and privacy-preserving machine learning, allowing models to learn from distributed private data sources without directly accessing the raw data.

The intersection of these two domains presents a unique opportunity to unlock new possibilities in AI research and to address critical challenges in AI model development and real-world applications. Hence, we propose the concept of Federated Foundation Models (FFMs), a novel paradigm that integrates FL into the lifespan of FMs. This integration addresses the challenges mentioned above related to data scarcity, computational resources, privacy, and ethical considerations while facilitating privacy-preserving and collaborative learning across multiple end-users. As advancements in edge computing enable the optimization of FMs using FL, we further explore the possibility of continual/lifelong learning for FMs in FFMs. We also discuss the potential benefits and challenges of integrating FL into different stages of the FMs' lifespan, including pre-training, fine-tuning, and application, and provide potential research directions for FFM tasks such as FFM Pre-training, FFM Fine-tuning, and Federated Prompt Tuning. These tasks promote the development of personalized and context-aware models while maintaining data privacy.

In summary, this paper offers a comprehensive examination of the prospective of FFMs, proposing a flexible and scalable framework for training large models in a privacy-preserving manner. We believe our work contributes to paving the way for future advancements in both FMs and FL, fostering the development of more secure and adaptable large models and FL algorithms that cater to a wide range of applications.

LREC-COLING 2024, pages 7174–7184
20-25 May, 2024.

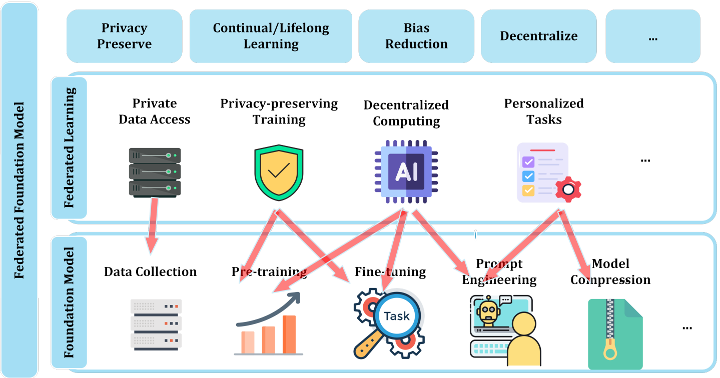

Figure 1: Federated Foundation Model: Integrating federated learning into the lifespan of foundation models, facilitating privacy-preserving, scalable, lifelong learning, robustness, and decentralized FMs.

2. Background

2.1. Federated Learning

As concerns about user data privacy grow, there is an increasing need for AI models to be trained on decentralized data without sharing private information between clients. Federated Learning (FL) has emerged as a solution to this problem, offering a distributed and privacy-preserving machine learning approach that enables training on decentralized data without compromising data privacy (McMahan et al., 2017).

In FL, raw data remains on local clients, ensuring data privacy and security while also enabling collaborative learning across multiple clients. The FL process involves local model training, model aggregation algorithm, and global model updates. Throughout this process, clients only share model updates, such as weights and gradients, asynchronously, reducing bandwidth requirements and minimizing the risk of data leaks and breaches. A typical FL algorithm is FedAvg (McMahan et al., 2017), which demonstrates the FL process (see Algorithm 1). The privacy-preserving nature of FL has led to its widespread adoption in various applications, particularly in privacy-sensitive domains like healthcare.

However, FL still faces challenges related to heterogeneous data distribution. Data may be non-independent and identically distributed (non-IID) across clients, leading to poor model convergence and performance. Recent work in FL has focused on improving gradient descent to stabilize training (Liu et al., 2020; Karimireddy et al., 2020; Yu et al., 2021); personalizing model weights to enhance performance on downstream tasks (Deng et al., 2020; Tan et al., 2022; Yu et al., 2022b,a); and employing model compression techniques like knowledge distillation, dynamic dropout, and adaptive pruning to reduce overfitting on non-IID datasets and improve communication efficiency (Jiang et al., 2022; Yu et al., 2021; Lin et al., 2020; Yu et al., 2021; Lin et al., 2020; Yu et al., 2022a,c; Nguyen et al., 2023). Despite these advances, there remains a gap between traditional model training and FL, particularly in terms of performance when dealing with heterogeneous data distributions.

Algorithm 1 Federated Learning Process (FedAvg)

```
Input: Global AI model $w_0$, clients $S$, communication rounds $T$
for $t = 1, 2, \ldots, T$ do
    Server deploys global model $w_{t-1}$ to clients $\in S$
    for each client $k \in S$ do
        Client $k$ optimizes $w_{t-1}$ on local data, producing $w_t^k$
    end for
    Select a subset of clients $S_t$ to communicate with the server
    for each client $k \in S_t$ do
        Client $k$ sends local model update $\Delta w_t^k = w_t^k - w_{t-1}$ to the server
    end for
    Server aggregates local updates and computes the new global model:
        $w_t = w_{t-1} + \eta_t \sum_{k \in S_t} n_k \Delta w_t^k$
end for
```

2.2. Foundation Models

Foundation Models (FMs), such as the GPT family (Brown et al., 2020; Radford et al., 2019), ViT (Dosovitskiy et al., 2020), CLIP (Radford et al., 2021), and BERT (Kenton and Toutanova, 2019), have become a driving force in AI, serving as the basis for various downstream tasks. These models are trained on massive datasets and demonstrate remarkable capabilities across multiple domains.

The lifespan of FMs typically includes pre-training, fine-tuning, and application. Pre-training involves unsupervised or self-supervised learning on large-scale datasets, while fine-tuning adapts the models to specialized tasks. For example, GPT (Brown et al., 2020; Radford et al., 2019; OpenAI, 2023) models learn grammar, syntax, and semantics during pre-training, enabling them to be easily fine-tuned for tasks such as text classification, sentiment analysis, translation, and summarization. Parameter-efficient fine-tuning (PEFT) methods, e.g., low-rank adapters (LoRA) (Hu et al., 2022), have been proposed to reduce the memory and compute requirements during the fine-tuning of these large models. Recently, neural architecture search (NAS) techniques have been employed to discover high-performing configurations of these adapters (Muñoz et al., 2024b,a).

In the application stage, FMs show extraordinary adaptability to downstream tasks using zero-shot learning. Prompt Engineering, an emerging research area, explores this potential by optimizing the interaction between users and FMs through carefully crafted prompts, thereby improving performance on downstream tasks. Various methods for prompt engineering have been proposed, including prompt templates (Wei et al., 2021), prompt tuning and instruction tuning (Wei et al., 2021) (Lester et al., 2021; Han et al., 2022), automated prompt generating (Zhou et al., 2022; Sanh et al., 2021), and in-context learning (Min et al., 2021, 2022; Rubin et al., 2021; Liu et al., 2021). These approaches enable FMs to learn from examples or instructions supplied as part of the input without the need for explicit fine-tuning or labeled examples.

In summary, the combination of Federated Learning and Foundation Models offers great opportunities to revolutionize the AI landscape by leveraging the strengths of both paradigms. This intersection opens up numerous research directions and applications in areas such as personalized recommendations, natural language understanding, healthcare, finance, and more. As AI researchers continue to explore Federated Foundation Models, we expect to see innovative solutions and breakthroughs that lead to more robust, efficient, and ethical AI systems serving the needs of individuals and society.

3. Motivation for Federated Foundation Models

In this section, we discuss the various challenges that motivate the development of Federated Foundation Models (FFMs), covering aspects such as data privacy, model performance, communication cost, scalability, deployment, personalization and real-time adaptation, and bias reduction. As shown in Figure 1, These existing challenges highlight the potential advantages of combining Foundation Models (FMs) and Federated Learning (FL) for a wide range of applications and scenarios.

Data privacy. The widespread deployment of AI in society generates vast amounts of data (e.g., images collected by cameras in smartphone applications, prompt dialog produced by users), presenting potential resources for optimizing and specializing FMs. However, privacy concerns have limited the use of private data for FM optimization. FFMs offer significant improvements in data privacy by incorporating FL, enabling FM optimization on private data. By optimizing FM tasks (e.g., pre-training, fine-tuning, and prompt tuning) on local data without sharing raw information, FFMs comply with data protection regulations and preserve user privacy. This approach is particularly beneficial when sensitive data, such as medical records or personal communications, must be used to improve model performance without compromising confidentiality.

Model performance. Combining FMs and FL provides benefits to FMs, boosting their performance. FMs gain access to a broader range of data for optimization tasks such as fine-tuning, prompt tuning, and pre-training. This expanded data access enables the development of more accurate and efficient AI systems better suited for users in diverse scenarios. This combination benefits FL, as well. FL can overcome challenges associated with Non-IID (Non-Identical Independent Distributed) and biased data (Zhao et al., 2018) by leveraging the advanced capabilities of FMs, leading to improved performance across different tasks and domains.

Cost. FFMs reduce communication costs by sharing only model updates between devices and the central server, significantly saving bandwidth and communication costs for transmitting raw data. Additionally, FFMs can potentially reduce the labor cost associated with collecting and managing data in a central location, as data is generated and used locally at edge devices. This efficiency makes FFMs a more practical and cost-effective solution for training and deploying FMs.

Scalability. Current FMs, especially large language models, often face scalability limitations due to limited computational power at the edge. Many FMs are run centrally and provide API access for users, which can lead to capacity constraints and API congestion. In the near future, advancements in computational power may enable FMs to run locally on edge devices. FL's scalable nature makes it an ideal framework for combining with FMs, accommodating numerous devices with varying computational capabilities. By integrating FL principles, FMs can leverage advancements in computational power, becoming more scalable and enabling broader deployment and improved performance

Table 1: Comparison of the Federated Foundation Model with Traditional FM Optimization

	Federated Foundation Model		Traditional FM Optimization	
Data Privacy	Privacy-preserve	✓	Centralized Data Collection	✗
Communication Overhead	Communicate Model Updates	✓	Communicate Data to Central Server	✗
Model Performance	Diverse Data Improvement	✓	Lacks Diversity	✗
Resource Distribution	Distributed Across Devices	✓	Centralized	✗
Data Efficiency	Better with data diversity	✓	Requires more data for similar performance	✗
Latency	Distributed Computation	✗	Lower with Centralized Computation	✓
System Complexity	Distributed Coordination	✗	Centrally Managed	✓
Scalability	Scalable to Many Clients	✓	Unscalable with Large Datasets	✗
Consistency	Weakly Connected Collaborative Learning	✗	Consistent Updates in Controlled Environment	✓
Ease of Deployment	Challenging	✗	Easier	✓

across various tasks and domains.

Deployment. FFMs offer potential advantages in deployment, particularly in reducing latency and enhancing user experience. Running FMs centrally with API access for users can result in latency issues due to network communication between the user's device and the central server hosting the model. In contrast, FFMs can be deployed and run locally on edge devices, potentially reducing latency by eliminating network communication. This allows for faster response times and a more seamless user experience when interacting with the model. However, available computational resources on edge devices must be considered when deploying FMs locally. As discussed in the Scalability section, advancements in computational power will be crucial for enabling local deployment on a wide range of devices, ensuring efficient and effective performance across various tasks and domains.

Personalization and real-time adaptation. FFMs facilitate a high degree of personalization by leveraging the decentralized nature of FL. By training on diverse, user-generated data, FMs can be tailored to individual preferences and requirements, offering more personalized and context-aware solutions across various tasks and domains. A key advantage of FFMs is their ability to adapt in real-time as new personalized data becomes available from edge devices. This continuous learning capability ensures that the models remain up-to-date with users' evolving needs and preferences, further enhancing their personalization. The focus on personalization in FFMs leads to improved performance and greater user satisfaction. By providing AI solutions that dynamically adapt to user-specific needs, FFMs enable more effective and engaging user experiences across a wide range of applications and domains.

Bias reduction. FFMs contribute to bias reduction in AI systems by incorporating diverse data from decentralized sources, resulting in more inclusive and fair AI solutions. The models learn from various users, increasing their awareness of the nuances and complexities of real-world scenarios, and leading to more informed and less biased decisions across tasks and domains. Additionally, the privacy-preserving nature of FL encourages more users to participate in the training process, further diversifying the data and knowledge incorporated into FMs. This results in models better equipped to handle and minimize biases, providing fairer and more equitable AI solutions for all users.

Continual/Lifelong learning. FMs combined with FL provide an ideal platform for continual lifelong learning. This combination facilitates the continuous adaptation and improvement of models by harnessing decentralized and diverse data sources, leading to more versatile and effective AI systems. As advancements in edge computing power become more prevalent, the realization of continual lifelong learning in FMs will soon be within reach. This progress will enable AI models to learn and grow throughout their lifespan, unlocking new possibilities for AI research and practical applications in various domains. By embracing continual lifelong learning, FFMs can help create more adaptive, efficient, and personalized AI systems that can dynamically adjust to user-specific needs and preferences, ultimately benefiting users from all walks of life.

In summary, as detailed in Table 1, our proposed FFM presents several advantages over traditional FM optimization. Despite introducing certain challenges, FFMs exhibit significant promise in enhancing data privacy, reducing communication overhead, improving model performance, optimizing resource distribution, increasing data efficiency, and providing better scalability. FFMs represent a robust approach to address many challenges and limitations associated with traditional, centralized machine learning. By incorporating Federated Learning (FL) into FM optimization, we are poised to engender more efficient, personalized, and privacy-conscious AI systems. This advancement heralds a new era in AI research and application, potentially making AI more equitable and advantageous for a diverse array of users. The integration of FL not only fortifies the foundational aspects of machine learning but also democratizes AI, thereby extending its benefits across a broader societal spectrum.

4. Federated Foundation Model: Prospective and Future Research

In this section, we discuss potential future research directions and general challenges related to FFMs, covering but not limited:

- Federated foundation model pre-training
- Federated foundation model fine-tuning
- Federated prompt tuning
- Federated continual (lifelong) learning
- Federated retrieval augmented generation
- General challenges
- Other future research directions

We scrutinize the distinct characteristics and prerequisites of these tasks, spotlighting the opportunities and hurdles encountered when employing FFMs to address real-world issues. Our aim is to build a robust foundation for comprehending the breadth and potential of this emerging paradigm, thereby fostering further research and development. As mentioned in Section 3, some tasks may not be feasible until computational power at the edge advances further.

4.1. Pre-training of Federated Foundation Models

Motivation: The motivation behind Federated Foundation Model (FFM) pre-training is to enhance traditional Foundation Model (FM) pre-training methodologies, harnessing Federated Learning's (FL) capability to utilize private data to improve model generalization while preserving data privacy. Introducing FL to FM lifespan allows for the FM to access a broader range of knowledge spectrum from private parties, mitigating overfitting on public data, and potentially enabling more generalized and context-aware FMs, while still benefiting from centralized data.

Algorithm 2 General FFM Optimization process

```
Input: Global AI model $w_0$, clients $S$, communication rounds $T$
Server initialize global model $w_0$
for $t = 1, 2, \dots, T$ do
    if Public data available then
        Server optimize $w_{t-1}$ on public data
    end if
    Server send global model $w_{t-1}$ to participate clients $\in S$
    for each client $k \in S$ do in parallel
        Client $k$ optimizes $w_{t-1}$ on local data, producing $w_t^k$
    end for
    Select a subset of clients $S_t$ to communicate with the server
    for each client $k \in S_t$ do
        Client $k$ sends local model update $\Delta w_t^k = w_t^k - w_{t-1}$ to the server
    end for
    Server aggregates local updates and computes the new global model:
        $w_t = w_{t-1} + \eta_t \sum_{k \in S_t} n_k \Delta w_t^k$
end for
```

Goal: Enhance FM pre-train methodologies via FL, and allow FMs to foster a deeper understanding of data representations from private data, thereby enhancing the model's capability to generalize across various tasks and domains.

Procedure Overview: As shown in Algorithm 2 and Figure 2, FFM pre-training is structured in two phases: centralized pre-training on public data, and federated pre-training on private data. these phases interact via an adaptive switching mechanism, enabling the model to alternate between centralized pre-training (if the centralized public data is available) and federated pre-training.

4.2. Federated Foundation Model Fine-tuning

Motivation: Traditional FM fine-tuning typically involves an offline deployment where the model is fine-tuned on private data, and subsequently isolated. This isolation precludes collaboration among end-users, potentially limiting the FM's efficacy, especially when the local private data is limited and biased.

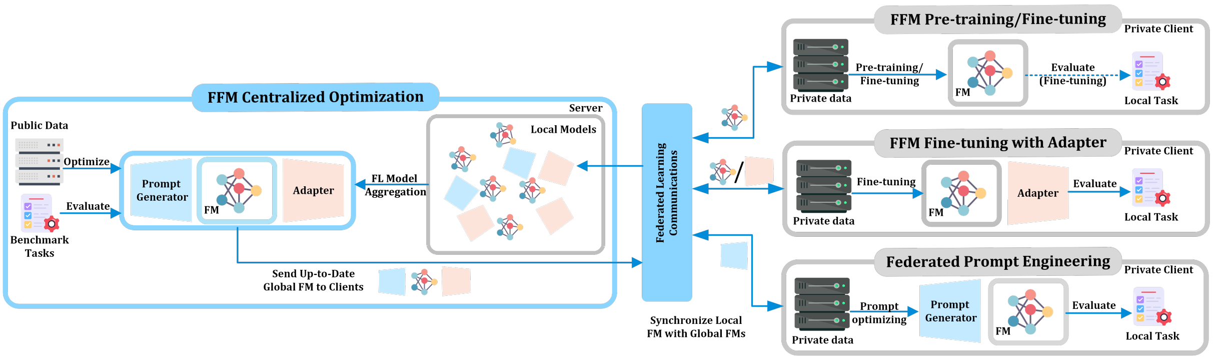

Figure 2: Federated Foundation Model tasks: The FFM centralized optimization process aggregates local models and updates them using public data. Private clients download up-to-date global model parameters from the server, optimize the FM locally on their tasks, and send the optimized model back to the server.

Goal: Leverage the collaborative learning feature of FL, enabling end-users with similar downstream tasks to collaboratively fine-tune FMs while preserving data privacy, thus potentially achieving enhanced performance on downstream tasks.

Procedure Overview: Similar to FFM pre-training, FFM fine-tuning follows the same procedure in Algorithm 2, FFM fine-tuning builds upon FFM pre-training phase. It employs an adaptive switching mechanism to alternate between centralized fine-tuning on public datasets for benchmark tasks and federated fine-tuning on private data for local tasks. As depicted in Figure 2, various fine-tuning strategies can be adopted with FFM. These include, but are not limited to, (1) direct fine-tuning of the FM backbone, and (2) Parameter Efficient Fine-tuning (PEFT) of a lightweight adapter head, while keeping the FM backbone frozen.

4.3. Federated Prompt Tuning

Motivation: Incorporating FL into prompt engineering presents a promising avenue for enhancing the performance of FMs while maintaining data privacy. Specifically, FFMs can assist in utilizing sensitive data for crafting prompt templates and soft prompt tuning, which in turn, enables more accurate and personalized prompt conditioning for tasks.

Goal: Collaboratively develop more effective and adaptable prompts without compromising the privacy of sensitive data.

Procedure Overview: This subsection primarily explores automated prompt (soft prompt) methods like prompt tuning (Lester et al., 2021), which refines the input prompt to better the model's output. As illustrated in Figure 2 and the general FFM optimization process in Algorithm 2, within federated prompt engineering settings, end-users can collaboratively train auto-prompt models (prompt generator components in Figure 2) on their local private data and tasks, sharing the learned auto prompt models without disclosing the sensitive data. This collaborative endeavor facilitates the creation of more effective and adaptable prompts, thereby enhancing the overall performance of FMs on downstream tasks.

4.4. Federated Continual (Lifelong) Learning

Motivation: FMs exhibit a significant limitation due to their dependency on pre-trained offline knowledge. For example, ChatGPT's knowledge is up-to-date only until 2021. With the anticipated increase in computational power, FM optimization at the edge may become feasible. FFMs can unlock the possibility of continual and lifelong learning from newly generated private edge data. With its scalability and privacy-preserving nature, FL can harness decentralized power to optimize FMs using emerging private data at the edge, which can serve as a valuable resource for model optimization. Furthermore, federated continual and lifelong learning could lead to a more efficient utilization of resources. Institutions would no longer necessitate retraining models from scratch with the availability of new data. Through FL, incremental model improvements can be attained, thus diminishing the time and computational resources requisite for model training and refinement.

Goal: Employ FL to harness the computational power at the edge, unlocking the potential for continual and lifelong learning of FMs on newly generated private data at the edge. This approach also aims to keep FMs updated with contemporary knowledge while preserving data privacy.

Procedure Overview: As delineated in Sections 4.1 and 4.2, establishing an online federated server is essential to facilitate the continuous communication between the server and edge end-users. The FM is updated at the edge based on the newly generated private data and regularly synchronizes with the online server.

4.5. Federated Retrieval Augmented Generation

Motivation: Federated Retrieval Augmented Generation (FRAG) seeks to extend the advantages of Retrieval Augmented Generation (RAG) by leveraging decentralized data across various clients while ensuring privacy preservation. This amalgamation aims to furnish more current and precise responses in a privacy-conducive manner.

Goal: Integrate FL with the RAG framework to bolster the performance of Language Model Generators (LMGs) in crafting responses, utilizing both centralized and decentralized data sources.

Procedure Overview: In the FRAG framework, the procedure unfolds in several distinct phases to ensure both effective data retrieval and privacy preservation. During the retrieval phase, a query is initiated from a user end, which triggers data retrieval from both a centralized server and local databases of clients within a federated network. This query is shared among clients in a privacy-preserving manner, enabling local clients to fetch relevant private data at the edge. Following the data retrieval, the generation phase commences where each client independently generates a response based on the retrieved data and the initial query. The responses from all clients are then aggregated in a privacy-preserving manner, ensuring no sensitive information is exposed during the process. Finally, an aggregated response, which encapsulates the collective intelligence of the federated network while preserving user privacy, is relayed back to the user. This structure allows for a more informed and accurate response generation in a decentralized and privacy-preserving environment.

4.6. Challenges

Despite the benefits associated with FFM, several substantial challenges persist. This subsection enumerates and discusses these general challenges.

Model Size: The substantial size of FMs, such as GPT (OpenAI, 2023) and Llama (Touvron et al., 2023b), presents a significant challenge for optimization FMs at the edge, especially when considering the resource-constraint edge devices in FL settings.

Data Quality: The effectiveness of FM pre-training and fine-tuning, including self-supervised pre-training, is heavily contingent on data quality as highlighted in (Gunasekar et al., 2023). Ensuring high-quality data in private federated settings, where data sharing is restricted, presents a notable challenge in filtering out toxic and redundant data.

Computational Cost: Optimizing FMs entails substantial computational cost (Meng et al., 2023). In FL environments, collaborative optimization of FMs at the edge necessitates high hardware specifications for edge clients (Meindl and Moser, 2023; Malandrino and Chiasserini, 2021).

Communication Cost: The routine sharing of model updates, encompassing model weights and gradients, incurs significant communication overhead (Ángel Morell et al., 2022; Almanifi et al., 2023; Mohammadi et al., 2021; WANG et al., 2019) between clients and the server in FL environments.

Data Heterogeneity: In FL, data is often non-identically distributed (non-IID) across clients (Zhao et al., 2018; McMahan et al., 2017), which could adversely affect the convergence and performance of the optimization process.

Security Attacks: Although FL inherently preserves privacy, ensuring robust privacy guarantees in FFM, especially against sophisticated security attacks, remains vital (Lyu et al., 2022; Zhang et al., 2022b; Liu et al., 2022).

Scalability: With the escalating scale of deployment, efficiently managing collaborative training and sharing model updates becomes increasingly challenging (Díaz and García, 2023; Zawad et al., 2022; Kołodziej and Rościszewski, 2021).

Asynchronous Training: As the number of clients increases, efficiently aggregating updates from a large number of asynchronous clients and ensuring consistent performance scaling is challenging (Wang et al., 2022; Chen et al., 2021).

Non-Stationary Data Distributions: The perpetually evolving nature of the user data suggests that data distributions may shift over time (Zhang et al., 2022a). Ensuring robust model performance amidst such changes is a significant challenge.

Resource Constraints: The resource-constrained edge devices could impede the optimization process of FMs at the edge.

Global Model Synchronization: Achieving global model synchronization across all participants while accommodating local updates and ensuring model stability is a nuanced challenge.

Evaluation Metrics: Establishing robust metrics to evaluate the performance, privacy, and other crucial aspects of the FFM process is pivotal.

4.7. Other Future Research Directions

In addition to the potential FFM tasks and general challenges discussed earlier, we outline several potential future research directions below.

Advancement in Edge Hardware: Supporting the substantial computational and resource requirements of FM optimization in FL-edge scenarios necessitates significant advancements in edge hardware.

Private-preserve Training Data Process: The success of self-supervised pre-training largely hinges on data quality. In the context of FFM, where private data at FL-edge clients remains inaccessible, and only the data owner can access it, devising private-preserving training data processing methods is crucial. This is to ensure data quality at the edge, where preprocessing is challenging. Recent works, such as (Gunasekar et al., 2023; Li et al., 2023), propose automatic training data filters to evaluate and enhance data quality, addressing a critical aspect of data processing in FFM.

Collaborative Model Compression: Designing specialized model compression methods, like network pruning and quantization, for heterogeneous-resource edge clients is essential to efficiently utilize the resources at edge clients. It also helps reduce the size of FMs without sacrificing performance. This is particularly critical for environments with limited computational resources.

Neural Architecture Design: The design of computational and hardware-efficient neural network architectures is a promising direction to explore, aiming to address the resource constraints and performance requirements in FFM deployment.

Collaborative Self-supervised Learning: Self-supervised learning has been a dominant approach for FM pre-training. Developing specialized collaborative self-supervised learning methods can effectively harness decentralized computational power in FL-edge environments.

Collaborative Parameter Efficient Fine-tuning: Designing collaborative parameter-efficient fine-tuning (PEFT) methods is crucial for fine-tuning FMs in FL scenarios, especially given the limited and heterogeneous resource capacities of edge clients.

Robust Model Fusion Algorithms: Creating robust algorithms for model fusion is vital to ensure the effective aggregation of model updates from different clients while preserving data privacy and model performance.

Federated Multi-task Learning: Exploring federated multi-task learning can facilitate the simultaneous optimization of multiple learning tasks across a federated network, leveraging the collective data and computational resources to improve model performance across various domains.

5. Conclusion and discussion

In this paper, we introduced the concept of Federated Foundation Models (FFMs), which integrate Federated Learning (FL) into the lifespan of Foundation Models (FMs). We discussed FFM tasks, general challenges and potential future research directions. It is important to note that the advancement of computation at edge users is crucial for the widespread adoption of FFMs, and we believe that such advancements will be realized in the near future. As the field of FFM continues to grow, we anticipate the emergence of numerous related research areas, including improved privacy-preserving techniques, the integration of FFM with emerging technologies like IoT and edge computing, and the exploration of FFM in various application domains such as healthcare, finance, and manufacturing. Additionally, we foresee advancements in adaptive model compression methods for FFM local institutions, communication efficiency research, specialized FL algorithms for efficient updates and aggregation of FFM models, and security attack research. Overall, FFM represents a promising research area in the age of FMs, with the potential to address various challenges in privacy, scalability, and robustness across diverse domains.

6. Acknowledgment

This research was supported by the National Science Foundation under Grant number 2243775. We also appreciate the provision of computational resources by Intel Labs for this project. Furthermore, we thank the ResearchIT team [1] at Iowa State University for their continuous and helpful assistance.

7. Bibliographical References

Omair Rashed Abdulwareth Almanifi, Chee-Onn Chow, Mau-Luen Tham, Joon Huang Chuah, and Jeevan Kanesan. 2023. Communication and computation efficiency in federated learning: A survey. *Internet of Things*, 22:100742.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Z Chen, W Liao, K Hua, C Lu, and W Yu. 2021. Towards asynchronous federated learning for heterogeneous edge-powered internet of things. digit commun netw 7 (3): 317–326.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. 2020. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*.

Judith Sáinz-Pardo Díaz and Álvaro López García. 2023. Study of the performance and scalability

[1] https://researchit.las.iastate.edu

of federated learning for medical imaging with intermittent clients. *Neurocomputing*, 518:142–154.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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*.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

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*.

Yuang Jiang, Shiqiang Wang, Victor Valls, Bong Jun Ko, Wei-Han Lee, Kin K Leung, and Leandros Tassiulas. 2022. Model pruning enables efficient federated learning on edge devices. *IEEE Transactions on Neural Networks and Learning Systems*.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Tomasz Kołodziej and Paweł Rościszewski. 2021. Towards scalable simulation of federated learning. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28*, pages 248–256. Springer.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Pengrui Liu, Xiangrui Xu, and Wei Wang. 2022. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. *Cybersecurity*, 5(1):1–19.

Wei Liu, Li Chen, Yunfei Chen, and Wenyi Zhang. 2020. Accelerating federated learning via momentum gradient descent. *IEEE Transactions on Parallel and Distributed Systems*, 31(8):1754–1766.

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.

Francesco Malandrino and Carla Fabiana Chiasserini. 2021. Toward node liability in federated learning: Computational cost and network overhead. *IEEE Communications Magazine*, 59(9):72–77.

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.

Rainer Meindl and Bernhard A Moser. 2023. Measuring overhead costs of federated learning systems by eavesdropping. In *International Conference on Database and Expert Systems Applications*, pages 33–42. Springer.

Fanqing Meng, Wenqi Shao, Zhanglin Peng, Chonghe Jiang, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2023. Foundation model is efficient multimodal multitask model selector.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke

Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Nima Mohammadi, Jianan Bai, Qiang Fan, Yifei Song, Yang Yi, and Lingjia Liu. 2021. Differential privacy meets federated learning under communication constraints.

J. Pablo Muñoz, Jinjie Yuan, and Nilesh Jain. 2024a. Shears: Unstructured sparsity with neural low-rank adapter search. Accessed: 2024-03-05.

J. Pablo Muñoz, Jinjie Yuan, Yi Zheng, and Nilesh Jain. 2024b. Lonas: Elastic low-rank adapters for efficient large language models. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Duy Phuong Nguyen, Sixing Yu, J. Pablo Muñoz, and Ali Jannesari. 2023. Enhancing heterogeneous federated learning with knowledge extraction and multi-model fusion. In *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, SC-W '23, page 36–43, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. Gpt-4 technical report.

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, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. 2022. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*.

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*.

Luping WANG, Wei WANG, and Bo LI. 2019. Cmfl: Mitigating communication overhead for federated learning. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 954–964.

Qiyuan Wang, Qianqian Yang, Shibo He, Zhiguo Shi, and Jiming Chen. 2022. Asyncfeded: Asynchronous federated learning with euclidean distance based adaptive weight aggregation. *arXiv preprint arXiv:2205.13797*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Sixing Yu, Phuong Nguyen, Waqwoya Abebe, Wei Qian, Ali Anwar, and Ali Jannesari. 2022a. Spatl: salient parameter aggregation and transfer learning for heterogeneous federated learning. In *2022 SC22: International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pages 495–508. IEEE Computer Society.

Sixing Yu, Phuong Nguyen, Waqwoya Abebe, Justin Stanley, Pablo Muñoz, and Ali Jannesari. 2022b. Resource-aware heterogeneous federated learning using neural architecture search. *arXiv preprint arXiv:2211.05716*.

Sixing Yu, Phuong Nguyen, Ali Anwar, and Ali Jannesari. 2021. Adaptive dynamic pruning for non-iid federated learning. *arXiv preprint arXiv:2106.06921*.

Sixing Yu, Wei Qian, and Ali Jannesari. 2022c. Resource-aware federated learning using knowledge extraction and multi-model fusion. *arXiv preprint arXiv:2208.07978*.

Syed Zawad, Feng Yan, and Ali Anwar. 2022. Local training and scalability of federated learning systems. In *Federated Learning: A Comprehensive Overview of Methods and Applications*, pages 213–233. Springer.

Hongwei Zhang, Meixia Tao, Yuanming Shi, and Xiaoyan Bi. 2022a. Federated multi-task learning with non-stationary heterogeneous data. In *ICC 2022 - IEEE International Conference on Communications*, pages 4950–4955.

Junpeng Zhang, Hui Zhu, Fengwei Wang, Jiaqi Zhao, Qi Xu, Hui Li, et al. 2022b. Security and privacy threats to federated learning: Issues, methods, and challenges. *Security and Communication Networks*, 2022.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

José Ángel Morell, Zakaria Abdelmoiz Dahi, Francisco Chicano, Gabriel Luque, and Enrique Alba. 2022. Optimising communication overhead in federated learning using nsga-ii.