Federated Foundation Language Model Post-Training Should Focus on Open Models

Nikita Agrawal[1] and **Ruben Mayer**[1]
[1]University of Bayreuth
{Nikita.Agrawal, Ruben.Mayer}@uni-bayreuth.de

Abstract

Post-training of foundation language models has emerged as a promising research domain in federated learning (FL) with the goal to enable privacy-preserving model improvements and adaptations to users' downstream tasks. Recent advances in this area adopt centralized post-training approaches that build upon black-box and gray-box foundation language models, where there is limited access to model weights and architecture details. Although the use of such closed models has been successful in centralized post-training, their blind replication in FL raises several concerns. Our opinion is that using closed models in FL contradicts the core principles of federation, such as data privacy and autonomy. In this paper, we critically analyze the usage of closed models in federated post-training, and provide a detailed account of various aspects of openness and their implications for FL.

1 Introduction

Post-training of foundation language models has received increasing attention as models need to be adapted to downstream applications to ensure their effectiveness [Lai *et al.*, 2025]. Traditional centralized machine learning approaches do not work well in privacy-sensitive settings, where data cannot be collected centrally and should remain private. Federated learning (FL) offers a promising alternative as a decentralized paradigm for training machine learning models on private data across multiple clients. Instead of exposing private data, clients only exchange gradients or model parameters [McMahan *et al.*, 2017]. Combining FL with foundation models, federated post-training emerged, which involves adapting foundation language models within the FL framework for downstream tasks.

When deciding on what model and what methods to use for federated post-training, it is crucial to identify different types of foundation language models in terms of their *openness*. In this paper, we classify them into four types according to their accessibility and licensing model: open-source, open-weight, gray-box, and black-box models. We find that the degree of openness of a model not only determines the methods that are available for post-training, but also impacts the autonomy of clients to use the post-trained model according to their needs (cf. Table 1).

Table 1: Analysis of model openness in federated post-training.
Notes: *1. Open-source models allow inference logic to be customized. 2. API might provide access to logs. 3. Only possible if API exposes LoRA or adapter access. 4. Only possible if API allows it via managed fine-tuning. 5. Soft prompting possible if API allows. 6. Prompt Tuning only possible via textual prompting. FFT = Full Fine-tuning. RLHF = Reinforcement Learning from Human Feedback.*
(Legend: ✓: Available ; ◐: Partially Available ; ✗: Unavailable)

	Open-Source	**Open-Weight**	**Gray-Box**	**Black-Box**
I. Access				
Weights	✓	✓	◐	✗
Architecture	✓	✓	◐	✗
Source Code	✓	✗	✗	✗
Training Data	✓	✗	✗	✗
II. Autonomy				
On Premise	✓[1]	✓	✗	✗
API inference	✓	✓	✓[2]	✓
III. Post-Training				
FFT	✓	✓	✗	✗
RLHF	✓	✓	✗	✗
LoRA	✓	✓	✓[3]	✗
Adapter Layers	✓	✓	✓[3]	✗
Instruction Tuning	✓	✓	✓[4]	✗
Prompt-Tuning	✓	✓	✓[5]	✓[6]
License	Free	Restrictive	Restrictive	Restrictive
Examples	GPT-J, Mistral-7B	LlaMa family	Together. AI	GPT-5 Claude

Recent centralized post-training approaches in frequently leverage black-box models [Yuksekgonul *et al.*, 2025; Khattab *et al.*, 2024], where interaction is limited to an API that hides internal parameters (weights/architecture), making them inaccessible. This severely restricts the post-training methods. Yet, researchers deem black-box models acceptable and at the frontier of AI development, why they are heavily used in practice. Hence, researchers are now adopting black-box models for federated post-training, to leverage local data for post-training[Chen *et al.*, 2025; Lin *et al.*, 2023]. However, this adaption requires a critical reflection on whether such approaches are suitable for the privacy-preserving use cases of FL.

Our opinion is that federated foundation language model post-training should focus on open-source models. Using black-box models makes clients rely on third-party services, introducing significant drawbacks such as potential data leakage, reduced transparency, and diminishing client autonomy that undermine the advantages of using FL. *This is critical as it contradicts the fundamental goals and principles of FL.* Open-source models offer transparency, control, and alignment with the privacy-preserving principles of FL. They enable the use of robust privacy mechanisms, effective optimization techniques, and promote ethical AI development. This enhances the autonomy and sovereignty of clients over their data.

Our opinion is in line with similar statements from various groups in academia and industry that argue in favor of using open-source models *in general*[Manchanda *et al.*, 2025; Floridi *et al.*, 2025]. The particular focus of this paper is on the combination of **open-source vs. black-box models in federated post-training**, which is a question that requires more attention in the community.

We make the following **contributions**. (1) We provide a detailed classification and analysis of the degree of openness of foundation language models, and their relation to federated post-training methods. (2) We critically discuss and analyze the implications of federated post-training of *black-box* and *gray-box* models. (3) We provide arguments for focusing on open models, showing that they support the core principles and motivations of federated learning better. (4) We perform a security and privacy analysis of federated post-training for different models and methods. (5) We provide a guide to select the most suitable type of foundation model (from open to closed) based on relevant decision criteria.

2 Openness, Post-Training Methods and Federated Learning

2.1 Model Openness vis-à-vis Post-Training Methods

Degree of Openness Open-source models [Alliance, 2025; Open Source Initiative, 2025] provide full access to weights and source code, a well documented architecture, and training data under permissive licenses, allowing complete control over inference. Open-weight models [Open Source Alliance, 2025] also provide full access to weights and a well documented architecture but do not provide source code or any training data. Gray-box models provide partial access through APIs that allow for limited fine-tuning options, e.g., tuning of adapters. Figure 1 shows how restricted APIs in gray-box models can be leveraged in FL, where clients collaboratively train the prompt or adapter layers while the underlying foundation model remains frozen and is hosted centrally. Typically, gray-box models are actually open-weight models that are hosted by a service provider that offers managed model execution and post-training. Finally, black-box models are the most restrictive, allowing only the exchange of tokens without any access to model internals. They are fully hosted and managed by a service provider that does not offer an expressive API that goes beyond token exchange.

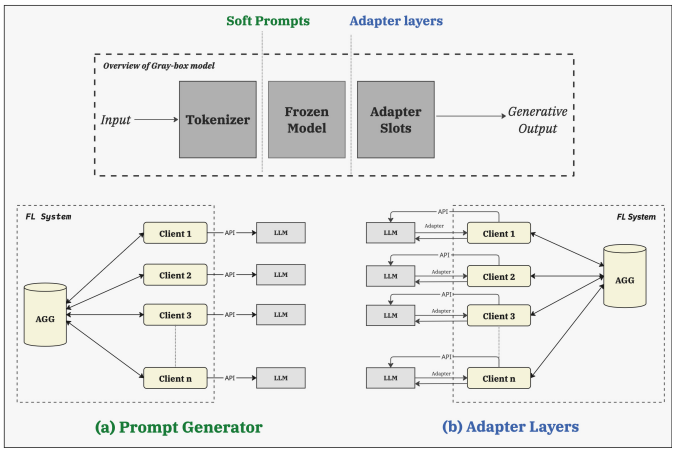

Figure 1: Federated post-training of gray-box models. **Top panel**: Gray-box LLM, where the base model parameters are frozen and inaccessible, while post-training is enabled through APIs allowing soft prompts or lightweight adapter layers. **Bottom panel**: Two FL instantiations under the gray-box setting: (a) clients collaboratively learn soft prompts via local training and share only prompt models updates with a central aggregator, and (b) clients train local adapter layers attached to the frozen LLM and share only adapter parameters. In both cases, the base model remains frozen and proprietary, while the provider exposes only necessary trainable layers.

Post-Training Methods The accessibility and autonomy of the models define how they can be post-trained for downstream tasks. Open-source and open-weight models can be post-trained using all possible methods, including full fine-tuning (FFT), parameter efficient fine-tuning (PEFT) such as Low-Rank Adapters (LoRA), adapter layers, prompt tuning (PT), instruction tuning (IT), and reinforcement learning from human-feedback (RLHF) [Rae *et al.*, 2021; Zhao *et al.*, 2023b]. Gray-box models may provide token-level logits or an adapter interface, but they can only be accessed through an API that limits the fine-tuning options to what the provider exposes [Nakajima *et al.*, 2024]. Like gray-box models, black-box models can also be accessed only via API; however, they do not expose any parameters, making it difficult to tune them beyond textual prompt tuning [Brown *et al.*, 2020; Achiam *et al.*, 2024].

Licensing Beyond access to source code and training data, another important distinction between open-source and open-weight models is their license. An open-source license must admit the unrestricted use, modification, and redistribution of the model [Open Source Initiative, 2025]. This plays an important role for the post-training process itself, allowing for any needed modifications, but also for the further usage of the post-trained model. Any restrictions would impact the autonomy of users, even after post-training has been performed. E.g., when using Llama 3, one needs to follow the *Meta Llama 3 Community License Agreement*. Under §2 ("Additional Commercial Terms"), this agreement states that "*[...] if [...] the monthly active users of the products or services made available [...] is greater than 700 million monthly active users [...], you must request a license from Meta [...] and*

you are not authorized to exercise any of the rights under this Agreement unless or until Meta otherwise expressly grants you such rights." Hence, LLama 3 models fail to follow the open source principles—there are further restrictions in the license that we do not discuss in more detail here. In contrast, open-source models are released under open-source licenses that do not make such restrictions. An example is Mistral-7B that is released under the Apache-2.0 license. Beyond limitations of liability and warranty that protect the licensor, the license does not impose any restrictions on the use of the work or require any additional permissions under specific conditions.

A concise overview on accessibility, autonomy, and post-training techniques for each type of model (open-source, open-weight, gray-box, and black-box) can be found in Table 1.

2.2 Federated Learning on Foundation Models

FL involves training a shared machine learning model across multiple distributed clients. Each client owns local training data that it does not share during training. In each training round, a server shares the global model with selected clients, which train it locally on their own data. After training, they send their model updates to the server, which combines them according to an aggregation strategy, e.g., Federated Averaging (FedAvg) [McMahan *et al.*, 2017], to improve the global model. Training is complete after multiple rounds when the model reaches an acceptable performance level.

Post-training of foundation models on distributed datasets without collecting data centrally is essential in numerous real-world scenarios, particularly when handling sensitive information or complying with stringent data protection and governance regulations such as the GDPR [European Union, 2016] or the EU AI act [European Union, 2024].

The literature often distinguishes between *cross-device* and *cross-silo* FL [Huang *et al.*, 2024]. In cross-device settings, model training is performed on end-user devices, such as mobile phones, severely restricting feasible model sizes and training speed [Woisetschläger *et al.*, 2024a]. In cross-silo FL, one assumes that clients have access to powerful machines or data centers, and the focus is on bridging data silos in a privacy-preserving manner. Such deployments are shown to be capable of training very large models with billions of parameters [Sani *et al.*, 2025]. As the focus of our paper is on post-training of very large foundation language models, we mainly consider cross-silo settings where computational capacities are typically high enough to run large models.

The implementation of FL in foundation language model post-training is a significant challenge given the large-scale and computational requirements of these models [Hu *et al.*, 2024]. However, through the recent advent of strong small or medium-sized models and the increase in hardware capabilities, local post-training becomes a possibility. On the algorithmic side, PEFT techniques have become increasingly significant [Woisetschläger *et al.*, 2024b; Lin *et al.*, 2023]. They seek to reduce computational and communication overhead by training only a small subset of model parameters, while the majority remains frozen [Yang *et al.*, 2025a].

The hardware requirements for post-training a foundation model in an FL setting largely depend on the size of the model, quantization precision, and the post-training method itself. For example, FFT on a 7-13B LLaMa style model using 16-bit precision needs tens of gigabytes of VRAM [Pan *et al.*, 2024] on a client node. In contrast, PEFT, especially 4-bit Quantized-LoRA (QLoRA) on a 7B LLaMa model, requires less than 12 GB of VRAM [Chai *et al.*, 2023]. [Dettmers *et al.*, 2023] showed that a 4-bit NF4-quantized LLaMA model with LoRA adapters and optimizer states in bfloat16 can be post-trained using QLoRA with about 5 GB of GPU memory for the 7B parameter model, about 10 GB for the 13B model, 21 GB for the 33B model, and $<$ 48 GB for the 65B model on a single GPU. This shows that post-training even of larger foundation models is possible with modest hardware that can be found in cross-silo FL deployments. The hardware requirements of the server-node are not particularly vast as it does not train the model and is only involved in model aggregation and the orchestration of the training.

Table 2: Post-training methods for black-box foundation models in federated learning.

Post-Training Method	**FL Papers**
Continuous Soft Prompt Optimization	FedBPT [Sun *et al.*, 2024], Fed-BBPT [Lin *et al.*, 2023], ZooPFL [Lu *et al.*, 2024]
Discrete Prompt Optimization	FedOne [Wang *et al.*, 2025], FedDTPT [Wu *et al.*, 2024b], FedTextGrad [Chen *et al.*, 2025]
Auxiliary Prompt Generate/Improvement Models	[Lin *et al.*, 2023], [Wu *et al.*, 2024b], [Chen *et al.*, 2025]

In addition to PEFT, personalized FL (PFL) represents a vital research domain with respect to foundation models [Lu *et al.*, 2024; Wu *et al.*, 2024a]. Recognizing the substantial variability in data distributions among diverse clients within a federated system (non-IID data), PFL aims to create models customized to the distinct requirements and data characteristics of individual clients. This typically entails a combination of global learning to acquire the fundamental knowledge within the entire federation and local adaptation to personalize the model for each client [Tan *et al.*, 2022]. Recent variants of federated PEFT address this heterogeneity directly within the LoRA paradigm: SA-FedLoRA applies a simulated annealing schedule to dynamically adjust each client's parameter budget and mitigate client drift, while FlexLoRA allows clients to train local adapters at heterogeneous ranks, which are reconciled into a shared global adapter via SVD-based weight redistribution during aggregation [Bai *et al.*, 2024; Yang *et al.*, 2025b].

2.3 Federated Black-Box Post-Training Methods

The reasons for the interest in black-box foundation language models are manifold [Matarazzo and Torlone, 2025; Sharma, 2023]. Organizations that invest significantly in the development of these advanced models often claim and preserve ownership over their intellectual property and the technology itself. Facilitating access via APIs enables them to

deliver their functionality as a service while concealing the intricate details of their models. Moreover, using these black-box models provides users with a convenient way to interact with advanced AI without requiring the substantial computing resources and expertise necessary to develop such extensive models independently. For that reason, we see a rapid increase of applications developed on these strong, yet non-transparent, foundation language models. With this growing application base, the question of post-training these black-box foundation models is receiving growing attention as can be seen in Table 2.

Recent research looks into how black-box foundation language models may be modified for the FL paradigm, given their popularity and success in centralized settings [Hu *et al.*, 2024]. In FL, centralized post-training is usually adapted to the federated setting by training small auxiliary models to generate or improve prompts for the black-box model [Lin *et al.*, 2023; Wu *et al.*, 2024b; Chen *et al.*, 2025]. This is because of the fundamental constraint that the internal parameters of a black-box model are inaccessible.

FedBPT [Sun *et al.*, 2024], Fed-BBPT [Lin *et al.*, 2023], Federated Discrete and Transferable Prompt Tuning (Fed-DTPT) [Wu *et al.*, 2024b], Federated Textual Gradient (Fed-TextGrad) [Chen *et al.*, 2025] and FedOne [Wang *et al.*, 2025] aim to achieve efficient FL over black-box models that can be accessed only with APIs based on prompts. FedBPT and Fed-BBPT optimize continuous soft prompts, where clients estimate the gradients of the prompt parameters using a zeroth-order optimization. The local prompt updates are then aggregated on the server. [Lu *et al.*, 2024] investigate the effect of using zeroth-order optimization methods in personalized FL to work with black-box foundation language models. These methods approximate gradients by querying the model using inputs incorporated with an auto-encoder with low-dimensional and client-specific embeddings, enabling post-training without requiring direct access to the model's internal parameters.

FedOne, FedDTPT, and FedTextGrad differ from FedBPT and Fed-BBPT since they operate on discrete prompts. Since they operate on discrete prompts, even zeroth-order optimization is not possible. In FedOne, clients must operate on search-based strategies such as block-wise exploration. Like FedBPT and Fed-BBPT, local prompt updates are aggregated to obtain a single prompt. In FedDTPT, a feedback loop mechanism that leverages the Masked Language Model API to guide prompt optimization is used during the client optimization phase. The server aggregates them using an attention method centered on semantic similarity to filter local prompt tokens, combined with embedding distance elbow detection and a DBSCAN clustering strategy to refine the filtering process. In FedTextGrad, clients make use of "textual gradients" for local prompt optimization. The server then aggregates the locally optimized prompts by summarizing them using the principle of uniform information density.

Table 3: Effect of open-source criteria on federated learning principles when using the model for post-training. "✓" means that the **violation** of this open-source criterion has severe impact on the corresponding federated learning principle; "◐" means there is a limited or indirect effect; "✗" means there is no effect.

Open-Source Model Criteria	**Federated Learning Principles**		
	Privacy	Autonomy	Heterogeneity
Weights & Code	✓	✓	◐
License	✗	✓	◐
Data	✗	✗	✗

3 Open Models Fit Federated Learning Better

3.1 Alignment with Federated Learning Principles

Although there are arguments for using black-box foundation language models (see Section 4), we argue that the combination of black-box or gray-box foundation language models with FL is generally not a good fit. Relying on querying remote models hosted by a third party counteracts the benefits and motivations of using FL in the first place.

We analyze the impact of violated open-source criteria on three fundamental principles of FL, cf. Table 3. First, *privacy*: One of the most important principles in FL is that clients keep their training data local and do not need to share it with any centralized training instance. Second, *autonomy*: Clients decide themselves what training data to use and how to contribute to the joint training process. They also commonly own the trained model. Third, *heterogeneity*: Clients with different kinds of data distribution can participate in the training, with the goal of covering a wide variety of private training data in the joint training process.

Weights and code Without open weights, clients cannot directly access the foundation model for post-training. Instead, they must communicate with the model through the provider's API. This puts the client's privacy at risk as all communication is mediated by the API provider. The clients ability to perform fine-tuning tasks is constrained by the permissions and infrastructure of the API providers. This constrains the autonomy of the clients. Heterogeneity is not directly affected if model weights are not open: Post-training is still possible over diverse datasets. However, certain clients whose use-cases fall outside the provider's policies or are restricted by governing policies such as GDPR may be excluded from participation. The availability of model architecture and code is closely tied to the openness of the weights. Without this information, model usage and adaptation becomes significantly more difficult or even infeasible, forcing clients to rely on the model provider for fine-tuning.

Licenses Restrictive licensing does not impede the privacy of the client's post-training data. However, it does constrain their autonomy as they cannot freely decide how the fine-tuned models may be used. In other words, an open-weight model with a restrictive license can still be used for privacy-preserving federated post-training, but clients must carefully study the license and understand the impact of the restrictions on their intended use of the post-trained model. Licenses can provide a hurdle for certain participants to join the federated

training, e.g. when they require a payment or when political conditions, such as international sanctions, restrict clients from obtaining them. This can reduce the variety of clients from participating in post-training and thus affect heterogeneity.

Data The openness of training data that is used for pre-training does not have any significant impact on federated post-training. Post-training is largely independent of pre-training: Clients can freely post-train the model with their private data regardless of whether the original pre-training data is open or not. It may generally be helpful to have information on the pre-training data, but this is not an issue that is specific to the federated setting.

3.2 Closed Models Bring Many Risks

Privacy and Security FL enhances user privacy by keeping raw training data on local devices. Using black-box or gray-box models that are hosted by a third-party server puts the users' private data at risk, as the hosting party receives and processes plain-text input or tokens. Hence, post-training on black-box models inherently leaks potentially sensitive information [Hanke *et al.*, 2024]. One can try to add additional privacy-preserving mechanisms such as differential privacy [Behnia *et al.*, 2022; Zhao, 2023] or rely on the service provider's guarantees to not look into the data, which can be supported by technical measures, e.g. via trusted computing [Flower Labs, 2025]. However, the fundamental privacy cornerstone of FL is that training data does not leave the client, and this is not possible with black-box and gray-box models. Beyond mere technical privacy, there are also regulatory difficulties when moving sensitive data, cf. privacy regulations such as GDPR. In terms of security, clients need to rely on mechanisms implemented by the black box model provider. Locally implemented security mechanisms in control of the clients [Ye *et al.*, 2025; Sun *et al.*, 2025] are restricted to the components that are trained in a federated setting, i.e., mostly the prompt generator for black-box models or adapters for gray-box models. It has to be noted that FL on open models is necessary, but not sufficient to ensure privacy, as various attacks are possible that require counter-measures. Then, however, clients are capable to apply and control defense mechanisms. We provide a more detailed analysis of privacy and security of federated post-training in Section 3.4.

Transparency and Explainability Unlike open-source and open-weight foundation models, the limited access to black-box and gray-box models restricts understanding of their internal architecture and the reasons behind their output [Rane *et al.*, 2024]. The lack of transparency hinders the systematic assessment and verification of the behavior and the analysis and mitigation of risks of these models according to predefined objectives. Relying on an external provider, clients do not even know for sure what system prompts were in place when using the models. Promising approaches to explain the output of a language model require access to the model parameters [Ameisen *et al.*, 2025].

Autonomy One of the key advantages of FL is that clients retain control over their private data and key aspects of the training process. FL aims to promote the development of decentralized AI by returning data ownership to clients and reducing the dependency on a central authority [Flower AI, 2025; Anelli *et al.*, 2021]. Outsourcing essential optimization decisions to a black-box model undermines this autonomy. If clients cannot retain complete control over the post-training process, one of the main motivations to use FL – maintaining local client control – is compromised. It raises the question whether the partial control that is left for the clients is sufficient to justify the use of FL in the first place.

Table 4: Attack targets open-source, open-weights, gray-box, and black-box models, which **clients** can and need to protect during federated post-training. *(✓: Controllable (by clients) ; ◐: Partially controllable ; ✗: Out of control)*

Targets	Open-Source	Open-Weight	Gray-Box	Black-Box
Base Model	✓	✓	◐	✗
Adapters	✓	✓	◐	✗
Prompt Generators	✓	✓	✓	✓

Inefficient natural language interface Communicating with black-box models is typically limited to natural language prompts. Arguably, natural language is not an effective way to program a computer system [Dijkstra, 1997] due to its ambiguity, and the same argument holds for black-box foundation language models. Techniques that directly work on the model weights benefit from a more rigorous mathematical explanation of the optimization process.

3.3 Open Models Have Advantages

Our opinion is that black-box and gray-box foundation language models do not fit together with the core principles and goals of FL. In contrast, using open-source or open-weights models has many advantages and aligns very well with FL principles.

Local deployment provides control If an open-source or open-weight model is deployed locally, we eliminate the need for API calls to black-box and gray-box models. Local deployment provides complete control over model training and data and allows for the use of additional privacy-preserving and security mechanisms. It should be noted that with more control, clients also have more responsibility. We discuss what this means for security and privacy in more detail in Section 3.4.

Broader range of post-training methods Beyond prompt tuning, open models support the full range of post-training techniques such as adapter layers and LoRA. Using PEFT, minimal parameter updates yield large task-specific performance improvements [Hu *et al.*, 2022; Poth *et al.*, 2023]. PEFT techniques such as adapter layers and LoRA have been successfully extended to FL [Yang *et al.*, 2025a; Cai *et al.*, 2023].

In terms of prompt tuning itself, model access facilitates highly effective techniques that work on soft prompts, such as prefix tuning [Li and Liang, 2021] and P-Tuning [Liu *et al.*, 2024]. [Lester *et al.*, 2021] indicate that prompt tuning, especially soft prompting, scales well with model size,

achieving performance comparable to FFT. It has also motivated frameworks such as FedPrompt [Zhao *et al.*, 2023a] and PromptFL [Guo *et al.*, 2023] that assume a frozen LLM with access to local gradients on each client and optimize prompt parameters to improve communication and computational efficiency. FL strategies such as FedSP [Dong *et al.*, 2023] that facilitate information exchange between clients and the aggregation server through soft prompts show the effectiveness of soft prompts in FL. Decomposed Prompt Tuning (DePT) combines soft prompting and LoRA shows great promise [Shi and Lipani, 2024].

3.4 Privacy and Security Considerations

First, we take a closer look at which parts of the model are in control of the clients. In Table 4, we show that with a decreasing level of openness, clients have less control over the model. This means that when using gray-box and black-box models, clients need to trust the provider to take care of protecting the inaccessible model parts against attacks. On the other hand, with open-source and open-weight models, there is a larger attack surface during federated training, which clients and aggregation server need to protect against. Decentralizing the model training, hence, has two faces: It puts more control to the clients, but it may also increase the attack surface. Indeed, FL is known for its rich literature on privacy and security attacks and defenses, which we study in more detail in the following with regard to post-training methods.

To discuss attacks and defenses during FL post-training, we focus on the following attacks [Du *et al.*, 2025; Hu *et al.*, 2024; Cheng *et al.*, 2024]: 1) *Poisoning attacks:* manipulate the model to deteriorate downstream performance or install a backdoor, 2) *membership inference:* infer whether a specific training sample was included in the training process, and 3) *gradient inversion:* reconstruct original training data. To counter these attacks, various defenses have been proposed. We focus on the following defenses: 1) *Differential privacy (DP):* injecting noise to hide individual data records, 2) *Secure Aggregation:* privacy-preserving computation of an aggregated model without revealing individual clients' contributions, based on secure multi-party computation [Li *et al.*, 2021], 3) *Homomorphic Encryption:* aggregation computations are performed directly on encrypted data, 4) *Byzantine Aggregation:* robust aggregation protocols that tolerate Byzantine client behavior. We analyze per post training method which of these attacks are possible and what measures clients can take to mitigate them [Cheng *et al.*, 2024].

Table 5 provides an overview of our results. As can be seen, the attacks to be protected against and the defenses available to the clients decrease with the degree of control over the model. While FFT and PEFT methods such as LoRA and Adapter layers have the largest attack surface, clients also have access to the full arsenal of protective measures. With less control over the model, attack surfaces and available protections also become more constrained. Instruction Tuning, RLHF and Prompt Tuning may or may not contain private or otherwise sensitive data; hence corresponding privacy attacks and defenses may or may not need to be considered.

Table 5: Privacy and security analysis of federated post-training with respect to the fine-tuning method. We analyze which attacks **clients** can and need to protect against, and which methods could be effective in doing so.

Legend: (✓: Yes; ◐: Partially; ✗: No)

	FFT	LoRA	Adap.	IT	RLHF	PT
Attacks						
Poisoning	✓	✓	✓	✓	✓	✓
Mem. Inf.	✓	✓	✓	◐	◐	◐
Grad. Inv.	✓	✓	✓	◐	◐	◐
Defenses						
DP	✓	✓	✓	◐	◐	◐
Sec. Agg.	✓	✓	✓	◐	◐	◐
HE	✓	✓	✓	◐	◐	◐
Byz. Agg.	✓	✓	✓	✓	✓	✓

4 Black-Box Foundation Language Models Have Many Advantages

As can be seen in Table 2, black-box models remain widely used despite the risk they carry. Their adoption is driven by several advantages, such as strong performance, ease of deployment, and readily available serving stack.

Black-box foundation models often demonstrate superior performance compared to open-source alternatives. However, this is not always the case. Many open source models such as DeepSeek [Guo and Yang, 2025] consistently come extremely close to black-box models. Open-source models show strong performance in few-shot and zero-shot learning scenarios, allowing rapid adaptation to a wide range of downstream tasks.

Another important advantage of black-box models is their accessibility. Using hosted APIs removes the need for users to deploy expensive hardware infrastructure or maintain complex serving stacks locally. Service providers manage model updates, deployment, and operational maintenance, enabling users to access state-of-the-art models with relatively low setup effort. Furthermore, modern prompt engineering and in-context learning approaches allow users to adapt these systems without directly modifying model parameters. However, the practical benefits of hosted models depend strongly on the deployment setting. Recent advances in LLM serving stacks such as NVIDIA Triton [NVIDIA Corporation, 2024] and vLLM [Kwon *et al.*, 2023] make it easy for open-source models to be hosted privately. In addition, the terms of use of black-box model providers may restrict access, e.g., due to trade laws and sanctions [Achiam *et al.*, 2024].

Black-box model providers often supervise their deployment and maintenance in order to ensure security, alignment, and legal compliance, by observing ethical and legal guidelines [Hanke *et al.*, 2024; Laskar *et al.*, 2023]. However, these safeguards can easily be breached with simple prompting attacks [Lapid *et al.*, 2024; Chao *et al.*, 2025]. Users must ultimately trust the provider's guarantees regarding data handling and privacy preservation, as the underlying model and infrastructure remain inaccessible. In contrast, locally deployed open models provide users with stronger control over privacy, security, and compliance mechanisms, which is particularly

relevant in federated learning settings [Hanke *et al.*, 2024; Achiam *et al.*, 2024].

5 A Guide to Model Selection

Following the discussions in this paper, users interested in federated foundation language model post-training may ask: *Which type of model (open or closed) should I use in my specific setting?* The answer is multifaceted: Local resources, target model size, customization needs, licensing requirements, and the willingness to take responsibility for security and privacy. As a practical guide, we distill the discussion into a decision tree (Figure 2) that aims to provide a high-level overview of important questions to be considered and point to the possibilities of open-source, open-weights, gray-box or black-box models. Together with Table 1, it can be used to make a first plan on what kind of model to combine with what kind of post-training method. The decision tree favors open models as the default option and only resorts to closed models if absolutely necessary. This reflects our stand that open models suit federated learning principles better.

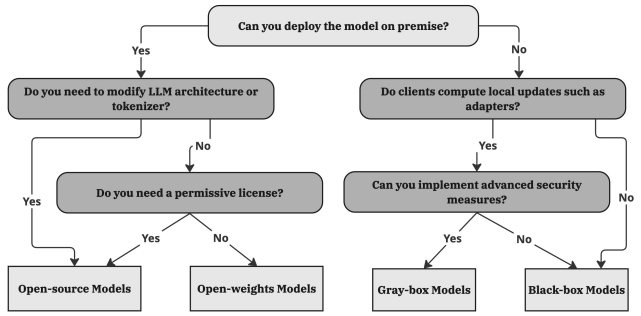

Figure 2: Decision tree for selecting foundation model type based on privacy, security, autonomy, and heterogeneity

6 Conclusion

In this paper, we discuss the following question: Should federated foundation language model post-training adopt black-box approaches? Our stand is clear: *"No, it shouldn't."* We provide a comprehensive analysis that illustrates how black-box models undermine the fundamental goals of FL, especially privacy and autonomy. Based on a detailed analysis of aspects of openness in foundation models, we argue which of them are critical to which goal of FL. This allows for a more fine-grained assessment of models that offer *partial* openness (e.g., open weights but restricted licenses). This way, our work provides a tool-box that helps researchers and practitioners to make an informed decision on which foundation models to consider for their work on federated post-training.

Our analysis primarily focuses on the deployment and post-training of standalone foundation language models. However, modern large-scale AI systems increasingly operate as compound systems consisting of multiple interacting components, including auxiliary draft models, routing mechanisms, safety and alignment layers, retrieval systems, hardware-specific inference kernels, and external tool integrations. In such settings, the degree of openness may differ across individual system components, making the distinction between open and closed models less clear-cut than discussed in this paper. Extending federated post-training analyses toward these multi-component AI systems represents an important direction for future work.

References

[Achiam *et al.*, 2024] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2024.

[Alliance, 2025] The AI Alliance. Defining open source ai: The road ahead, 2025.

[Ameisen *et al.*, 2025] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025.

[Anelli *et al.*, 2021] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Antonio Ferrara, and Fedelucio Narducci. How to put users in control of their data in federated top-n recommendation with learning to rank. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 1359–1362, 2021.

[Bai *et al.*, 2024] Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Behnia *et al.*, 2022] Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 560–566. IEEE, 2022.

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

[Cai *et al.*, 2023] Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. Efficient federated learning for modern nlp. In *Proceedings of the 29th annual international conference on mobile computing and networking*, pages 1–16, 2023.

[Chai *et al.*, 2023] Yuji Chai, John Gkountouras, Glenn G Ko, David Brooks, and Gu-Yeon Wei. Int2. 1: Towards fine-tunable quantized large language models with error

correction through low-rank adaptation. *arXiv preprint arXiv:2306.08162*, 2023.

[Chao *et al.*, 2025] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE, 2025.

[Chen *et al.*, 2025] Minghui Chen, Ruinan Jin, Wenlong Deng, Yuanyuan Chen, Zhi Huang, Han Yu, and Xiaoxiao Li. Can textual gradient work in federated learning? In *The Thirteenth International Conference on Learning Representations*, 2025.

[Cheng *et al.*, 2024] Yujun Cheng, Weiting Zhang, Zhewei Zhang, Chuan Zhang, Shengjin Wang, and Shiwen Mao. Towards federated large language models: Motivations, methods, and future directions. *IEEE Communications Surveys & Tutorials*, 2024.

[Dettmers *et al.*, 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[Dijkstra, 1997] Edsger W. Dijkstra. On the foolishness of "natural language programming", 1997.

[Dong *et al.*, 2023] Chenhe Dong, Yuexiang Xie, Bolin Ding, Ying Shen, and Yaliang Li. Tunable soft prompts are messengers in federated learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14665–14675, 2023.

[Du *et al.*, 2025] Hao Du, Shang Liu, Lele Zheng, Yang Cao, Atsuyoshi Nakamura, and Lei Chen. Privacy in fine-tuning large language models: Attacks, defenses, and future directions. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 326–344. Springer, 2025.

[European Union, 2016] European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. Official Journal of the European Union, L119, 1–88.

[European Union, 2024] European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) 2018/858, (eu) 2019/2144 and (eu) 2023/1230 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828, 2024. Official Journal of the European Union, L 1689, 13 June 2024.

[Floridi *et al.*, 2025] Luciano Floridi, Carlotta Buttaboni, Emmie Hine, Claudio Novelli, Tyler Schroder, and Grant Shanklin. Open-source ai made in the eu: Why it is a good idea. *Available at SSRN*, 2025.

[Flower AI, 2025] Flower AI. Flower joins forces with vana to launch a federated dao, March 2025. Accessed: 2025-05-15.

[Flower Labs, 2025] Flower Labs. Flower intelligence: An open-source ai platform to run llms locally in your app or remotely on flower confidential remote compute. https://flower.ai/intelligence/, 2025. Accessed: 2025-05-23.

[Guo and Yang, 2025] Daya Guo and et al. Yang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[Guo *et al.*, 2023] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 23(5):5179–5194, 2023.

[Hanke *et al.*, 2024] Vincent Hanke, Tom Blanchard, Franziska Boenisch, Iyiola Olatunji, Michael Backes, and Adam Dziedzic. Open llms are necessary for current private adaptations and outperform their closed alternatives. *Advances in Neural Information Processing Systems*, 37:1220–1250, 2024.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Hu *et al.*, 2024] Jiahui Hu, Dan Wang, Zhibo Wang, Xiaoyi Pang, Huiyu Xu, Ju Ren, and Kui Ren. Federated large language model: Solutions, challenges and future directions. *IEEE Wireless Communications*, PP:1–8, 01 2024.

[Huang *et al.*, 2024] Chao Huang, Ming Tang, Qian Ma, Jianwei Huang, and Xin Liu. Promoting collaboration in cross-silo federated learning: Challenges and opportunities. *Comm. Mag.*, 62(4):82–88, April 2024.

[Khattab *et al.*, 2024] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, Heather Miller, et al. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*, 2024.

[Kwon *et al.*, 2023] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[Lai *et al.*, 2025] Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, Xin Lv, Minlie Huang, Yuxiao Dong, and Jie Tang. A survey of post-training scaling in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 2771–2791, Vienna, Austria, July 2025. Association for Computational Linguistics.

[Lapid *et al.*, 2024] Raz Lapid, Ron Langberg, and Moshe Sipper. Open sesame! universal black-box jailbreaking of large language models. *Applied Sciences*, 14(16):7150, 2024.

[Laskar *et al.*, 2023] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. Building real-world meeting summarization systems using large language models: A practical perspective, 2023.

[Lester *et al.*, 2021] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, Online, August 2021. Association for Computational Linguistics.

[Li *et al.*, 2021] Kwing Hei Li, Pedro Porto Buarque de Gusmão, Daniel J Beutel, and Nicholas D Lane. Secure aggregation for federated learning in flower. In *Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning*, pages 8–14, 2021.

[Lin *et al.*, 2023] Zihao Lin, Yan Sun, Yifan Shi, Xueqian Wang, Lifu Huang, Li Shen, and Dacheng Tao. Efficient federated prompt tuning for black-box large pre-trained models. *arXiv preprint arXiv:2310.03123*, 2023.

[Liu *et al.*, 2024] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 5:208–215, 2024.

[Lu *et al.*, 2024] Wang Lu, Hao Yu, Jindong Wang, Damien Teney, Haohan Wang, Yao Zhu, Yiqiang Chen, Qiang Yang, Xing Xie, and Xiangyang Ji. Zoopfl: Exploring black-box foundation models for personalized federated learning. In *International Workshop on Trustworthy Federated Learning*, pages 19–35. Springer, 2024.

[Manchanda *et al.*, 2025] Jiya Manchanda, Laura Boettcher, Matheus Westphalen, and Jasser Jasser. The open source advantage in large language models (llms), 2025.

[Matarazzo and Torlone, 2025] Andrea Matarazzo and Riccardo Torlone. A survey on large language models with some insights on their capabilities and limitations, 2025.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[Nakajima *et al.*, 2024] Kyotaro Nakajima, Hwichan Kim, Tosho Hirasawa, Taisei Enomoto, Zhousi Chen, and Mamoru Komachi. A survey for LLM tuning methods:classifying approaches based on model internal accessibility. In Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, and Jong-Bok Kim, editors, *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 542–555, Tokyo, Japan, December 2024. Tokyo University of Foreign Studies.

[NVIDIA Corporation, 2024] NVIDIA Corporation. Triton inference server: An optimized cloud and edge inferencing solution. https://github.com/triton-inference-server, 2024. Source code available at https://github.com/triton-inference-server/server.

[Open Source Alliance, 2025] Open Source Alliance. Open Weight Definition (OWD), January 2025. Version 0.3, last modified 2025-01-21. Accessed: 2025-05-22.

[Open Source Initiative, 2025] Open Source Initiative. The Open Source AI Definition 1.0, 2025. Accessed: 2025-12-10.

[Pan *et al.*, 2024] Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. Lisa: Layerwise importance sampling for memory-efficient large language model fine-tuning. *Advances in Neural Information Processing Systems*, 37:57018–57049, 2024.

[Poth *et al.*, 2023] Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vuli, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore, 2023. Association for Computational Linguistics.

[Rae *et al.*, 2021] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. arxiv 2021. *arXiv preprint arXiv:2112.11446*, 2021.

[Rane *et al.*, 2024] J Rane, SK Mallick, O Kaya, and NL Rane. Enhancing black-box models: advances in explainable artificial intelligence for ethical decision-making. *Future Research Opportunities for Artificial Intelligence in Industry*, 4:136–180, 2024.

[Sani *et al.*, 2025] Lorenzo Sani, Alex Iacob, Zeyu Cao, Royson Lee, Bill Marino, Yan Gao, Wanru Zhao, Dongqi Cai, Zexi Li, Xinchi Qiu, and Nicholas D. Lane. Photon: Federated LLM pre-training. In *Eighth Conference on Machine Learning and Systems*, 2025.

[Sharma, 2023] Shraddha Sharma. What is black box ai?, 2023. Accessed: 2025-05-26.

[Shi and Lipani, 2024] Zhengxiang Shi and Aldo Lipani. DePT: Decomposed prompt tuning for parameter-efficient fine-tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

[Sun *et al.*, 2024] Jingwei Sun, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yudong Liu, Zhixu Du, Yiran Chen, and Holger R. Roth. Fedbpt: efficient federated black-box prompt tuning for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[Sun *et al.*, 2025] Zhen Sun, Tianshuo Cong, Yule Liu, Chenhao Lin, Xinlei He, Rongmao Chen, Xingshuo Han, and Xinyi Huang. Peftguard: Detecting backdoor attacks against parameter-efficient fine-tuning. In *2025 IEEE Symposium on Security and Privacy (SP)*, page 1713–1731. IEEE, May 2025.

[Tan *et al.*, 2022] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

[Wang *et al.*, 2025] Ganyu Wang, Jinjie Fang, Maxwell J. Yin, Bin Gu, Xi Chen, Boyu Wang, Yi Chang, and Charles Ling. Fedone: Query-efficient federated learning for black-box discrete prompt learning, 2025.

[Woisetschläger *et al.*, 2024a] Herbert Woisetschläger, Alexander Erben, Ruben Mayer, Shiqiang Wang, and Hans-Arno Jacobsen. Fledge: Benchmarking federated learning applications in edge computing systems. In *Proceedings of the 25th International Middleware Conference*, Middleware '24, page 88–102, New York, NY, USA, 2024. Association for Computing Machinery.

[Woisetschläger *et al.*, 2024b] Herbert Woisetschläger, Alexander Erben, Shiqiang Wang, Ruben Mayer, and Hans-Arno Jacobsen. A survey on efficient federated learning methods for foundation model training. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[Wu *et al.*, 2024a] Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. Fedbiot: Llm local fine-tuning in federated learning without full model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3345–3355, 2024.

[Wu *et al.*, 2024b] Jiaqi Wu, Simin Chen, Yuzhe Yang, Yijiang Li, Shiyue Hou, Rui Jing, Zehua Wang, Wei Chen, and Zijian Tian. Feddtpt: Federated discrete and transferable prompt tuning for black-box large language models. *arXiv preprint arXiv:2411.00985*, 2024.

[Yang *et al.*, 2025a] Yiyuan Yang, Guodong Long, Qinghua Lu, Liming Zhu, Jing Jiang, and Chengqi Zhang. Federated low-rank adaptation for foundation models: a survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025.

[Yang *et al.*, 2025b] Yuning Yang, Han Yu, Chuan Sun, Tianrun Gao, Xiaohong Liu, Xiaodong Xu, Ping Zhang, and Guangyu Wang. Spd-cfl: Stepwise parameter dropout for efficient continual federated learning, 2025.

[Ye *et al.*, 2025] Rui Ye, Jingyi Chai, Xiangrui Liu, Yaodong Yang, Yanfeng Wang, and Siheng Chen. EMERGING SAFETY ATTACK AND DEFENSE IN FED- ERATED INSTRUCTION TUNING OF LARGE LANGUAGE MODELS. *ICLR*, 2025.

[Yuksekgonul *et al.*, 2025] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639:609–616, 2025.

[Zhao *et al.*, 2023a] Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Zhao *et al.*, 2023b] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

[Zhao, 2023] Jun Zhao. Privacy-preserving fine-tuning of artificial intelligence (ai) foundation models with federated learning, differential privacy, offsite tuning, and parameter-efficient fine-tuning (peft). *Authorea Preprints*, 2023.