Vulnerabilities of Foundation Model Integrated Federated Learning Under Adversarial Threats

Chen Wu* , **Xi Li*** , **Jiaqi Wang**✉
the Pennsylvania State University
{cvw5218, XiLi , jqwang}@psu.edu

Abstract

Federated Learning (FL) addresses critical issues in machine learning related to data privacy and security, yet suffering from data insufficiency and imbalance under certain circumstances. The emergence of foundation models (FMs) offers potential solutions to the limitations of existing FL frameworks, *e.g.*, by generating synthetic data for model initialization. However, due to the inherent safety concerns of FMs, integrating FMs into FL could introduce new risks, which remains largely unexplored. To address this gap, *we conduct the first investigation on the vulnerability of FM integrated FL (FM-FL) under adversarial threats*. Based on a unified framework of FM-FL, we introduce a novel attack strategy that exploits safety issues of FM to compromise FL client models. Through extensive experiments with well-known models and benchmark datasets in both image and text domains, we reveal the *high susceptibility* of the FM-FL to this new threat under various FL configurations. Furthermore, we find that existing FL defense strategies offer *limited protection* against this novel attack approach. This research highlights the critical need for enhanced security measures in FL in the era of FMs[1].

1 Introduction

Federated Learning (FL) [McMahan *et al.*, 2017] is a decentralized approach to machine learning where multiple clients collaboratively train a model while keeping their data local. It encompasses a wide range of applications, including healthcare [Wang *et al.*, 2022], model personalization [Zhang *et al.*, 2022], and video surveillance [Rehman *et al.*, 2022]. This methodology, while safeguarding privacy, often encounters challenges such as data scarcity and imbalanced data distribution across clients. The integration of Foundation Models (FM), e.g., GPT series [Brown *et al.*, 2020], LLaMA [Touvron *et al.*, 2023], Stable Diffusion [Rombach *et al.*, 2022], and Segment Anything [Kirillov *et al.*, 2023], known for their extensive pre-training on diverse datasets, offers a solution to these challenges. FMs can enhance FL by providing a robust starting point for learning [Zhang *et al.*, 2023b], addressing issues like limited data availability [Lin *et al.*, 2020], and introducing additional diversity into the training process to cover a broader spectrum of scenarios not originally included in the original data [de Luca *et al.*, 2022].

However, incorporating FMs into FL systems introduces potential threats. The large-scale data scraped from the Internet used for FM training may be of low quality, containing bias, misinformation, toxicity, or even poisoned [Zhuang *et al.*, 2023]. This brings inherent vulnerabilities in the FMs to have robustness, fairness, and privacy issues [Bommasani *et al.*, 2021]. Recent studies have revealed threats to FMs range from adversarial examples [Zhang *et al.*, 2023a], data poisoning attacks to generate malicious output [Schlarmann and Hein, 2023], backdoor attacks to inject hidden mappings in the objective function [Kandpal *et al.*, 2023], privacy attacks to reveal sensitive information from training data [Pan *et al.*, 2020], to fairness and reliability of the FMs [Si *et al.*, 2022]. These vulnerabilities bring new risks to the security and reliability of the FM-Integrated FL (FM-FL) system.

Despite these emerging risks, there exists a significant gap in research specifically targeting these vulnerabilities [Wang *et al.*, 2023; Zhuang *et al.*, 2023]. To thoroughly investigate the susceptibility of FM-FL, we leverage a unified framework well-suited for both homogeneous and heterogeneous FL systems [Lin *et al.*, 2020; Li and Wang, 2019; Zhuang *et al.*, 2023; Zhang *et al.*, 2023b]. Specifically, the server employs the FMs to generate synthetic data, which plays a dual role: (i) assisting in the initialization of client models to provide a better starting point for training, and (ii) facilitating mutual information exchange between client models through knowledge distillation while protecting privacy. This dual application ensures a thorough and comprehensive integration of FMs across all stages of the FL process, from initialization to ongoing learning and model fusion.

We propose a novel attack strategy against FM-FL, where the attacker compromises the FM used by the server and consequently embeds the threat in client models during their initialization using the synthetic data. This threat is iteratively reinforced through the mutual information-sharing process on the server. We specialize our attack strategy to backdoor attacks to thoroughly investigate the vulnerability of

*Equal contribution.
[1]The source code will be available after acceptance.

FM-FL under the novel attack strategy. We choose backdoor attacks since they are popular and effective poisoning attacks widely deployed to evaluate the vulnerability of machine learning models in image classification [Gu *et al.*, 2017; Chen *et al.*, 2017], text classification [Dai *et al.*, 2019; Li *et al.*, 2021], point cloud classification [Xiang *et al.*, 2021], video action recognition [Li *et al.*, 2023], and federated learning systems [Bagdasaryan *et al.*, 2020]. The compromised model will mis-classify instances embedded with a specific trigger to the attacker-chosen target class, while maintaining high accuracy on clean (backdoor-free) data, rendering the attack in a stealthy manner.

We find that the FM-FL system demonstrates significant vulnerability under this novel attack strategy, and the existing secure aggregation strategies and post-training mitigation methods in FL show insufficient robustness. This finding is consistent across extensive experiments with a variety of well-known models and benchmark datasets in both image and text domains, covering different FL scenarios. The efficacy of the novel threat arises from two key aspects. Firstly, classic attack methods involve the attacker compromising a few clients to upload malicious updates, aiming to introduce mis-behaviors into the global model. However, these updates from compromised clients tend to be very different from normal ones, making them likely to be detected and removed as anomalies during the aggregation of client models. In contrast, our proposed attack strategy embeds the threat in each client at the initialization stage, which is then reinforced during mutual information sharing on the server. Additionally, all client updates are derived from clean local datasets, resulting in no anomalies in the model updates. This enables the novel attack to bypass existing FL defenses. Secondly, the effectiveness of the attack does not depend on persistent participation in FL training, nor require a large number of compromised clients. This is particularly relevant in practical scenarios where the total number of clients can reach millions. Our contribution is summarized below:

- We propose a novel attack strategy against FM-FL that exploits safety issues of FM to compromise FL client models. We specialize the novel threat to backdoor attacks, and provide a **comprehensive study** of the robustness issues raised by incorporating FMs into FL.
- We demonstrate that the FM-FL system is **highly vulnerable** under the novel attack strategy, compared with the classic attack mechanism, through extensive experiments with a variety of well-known models and benchmark datasets in both image and text domains, covering different FL scenarios.
- We also empirically show that the current robust aggregation and post-training defenses in FL are **inadequate** against this new threat, underscoring the urgency for advancing robustness measures in this domain.

2 Related Work

FM in FL. The synergy between Foundation Models (FM) and Federated Learning (FL) enhances both domains [Zhuang *et al.*, 2023]. On one hand, FL offers expanded data access and distributed computation for FMs. Key developments include FedDAT [Chen *et al.*, 2023] fine-tuning framework using a Dual-Adapter Teacher for handling data heterogeneity, and PromptFL [Guo *et al.*, 2023] shift from traditional model training to prompt training in FL, optimizing FM capabilities for efficiency and data limitations. On the other hand, FMs' pre-trained knowledge accelerates FL model convergence and performance, particularly through synthetic data generation [Zhang *et al.*, 2023b] and knowledge distillation [Hinton *et al.*, 2015]. FedPCL [Tan *et al.*, 2022] and RaFFM [Yu *et al.*, 2023] further integrate FMs into FL, emphasizing parameter prioritization and high-performance subnetwork extraction.

Traditional Attacks and Defenses in FL. Research in FL attacks has mainly focused on the client side about data poisoning [Tolpegin *et al.*, 2020], local model poisoning [Fang *et al.*, 2020], and various backdoor attacks including semantic backdoors [Bagdasaryan *et al.*, 2020] and edge-case and distributed backdoors [Wang *et al.*, 2020; Xie *et al.*, 2020]. Defenses against these attacks typically involve norm threshold bounding [Sun *et al.*, 2019], differential privacy [Geyer *et al.*, 2017; Xie *et al.*, 2021], anomaly detection [Lu *et al.*, 2022], strategies like model clustering and noise injection [Nguyen *et al.*, 2022], and pruning[Wu *et al.*, 2022]. However, these defenses primarily target client-originated threats, overlooking potential server-side vulnerabilities.

Vulnerabilities Introduced by FMs. The integration of FMs into FL systems raises new attack vectors, as evidenced by issues in LLMs like GPT-4 and GPT-3.5, including BadGPT [Shi *et al.*, 2023], instruction-based attacks [Xu *et al.*, 2023], and targeted misclassification [Kandpal *et al.*, 2023]. Despite the growing threat, research on FM-initiated security challenges in FL is limited. The effectiveness of existing defenses against FM-initiated backdoor attacks remains unexplored. This gap in research underlines the need for a systematic investigation into both the attacks and defenses within FL, particularly those stemming from FMs. Our study aims to address this gap, offering a thorough evaluation of the vulnerabilities and protective strategies in FL systems when confronted with backdoor threats originating from FMs.

3 Methodology

Integrating FM into the FL process raises novel threats, which remains largely unexplored. We address this gap and propose a novel attack strategy against FM-FL to evaluate its vulnerability.

To thoroughly study the vulnerability of FMs-integrated FL (FM-FL) systems under potential adversarial threats, we build our framework on the general FL framework, FedDF [Lin *et al.*, 2020], and the novel strategies for incorporating FMs into FL systems [Zhuang *et al.*, 2023; Zhang *et al.*, 2023b; Li and Wang, 2019]. FedDF enables flexible aggregation schemes applicable for both homogeneous and heterogeneous FL through ensemble distillation. It first fuses client model parameters sharing the same model prototype, then leverages public data to aggregate knowledge from all received (heterogeneous) client models. Following [Zhuang *et al.*, 2023; Zhang *et al.*, 2023b], the server queries the FM to generate synthetic data, which are used in (i) client model initialization and (ii) knowledge distillation between clients.

We propose a novel attack strategy that exploits the vulnerabilities in FMs and transfers these threats to client models during the FL process. The security issues of FMs include adversarial examples [Zhang *et al.*, 2023a], poisoning attacks [Kandpal *et al.*, 2023; Xu *et al.*, 2023; Shi *et al.*, 2023], toxic outputs [Schlarmann and Hein, 2023], and fairness issues [Si *et al.*, 2022]. To clearly and effectively illustrate the proposed attack strategy, we specialize in backdoor attacks. In the subsequent subsections, we will introduce the threat model in Sec. 3.1, detail the framework in Sec. 3.2, 3.3, 3.4 and compare the novel threat with traditional attacks in Sec. 3.5.

3.1 Threat Model

We follow the threat models of backdoor attacks against large language model (LLM) used in [Wang *et al.*, 2023; Kandpal *et al.*, 2023; Xu *et al.*, 2023; Shi *et al.*, 2023]. The server obtains an LLM from an open source. The attacker is able to insert a malicious system prompt into the LLM. The malicious system prompt manifests the target task (*e.g.*, poisoned data generation), the poisoning ratio γ, the backdoor trigger Δ, the target class t, and the backdoor embedding function $\mathcal{B}(\cdot, \Delta)$ (*e.g.*, a few demonstrations illustrating how to embed the trigger into selected instances). The server queries the (compromised) LLM to (i) directly generate text synthetic data; (ii) generate prompts to guide other foundation models to produce synthetic data in other formats (*e.g.*, images). Due to the malicious system prompt, γ of the synthetic instances are trigger-embedded and mis-labeled to the target class. The attacker aims to transfer the backdoor from the compromised LLM to downstream client models during FL training and ensure its persistence after the FL system converges. The backdoor-compromised client model will misclassify to the target class on triggered instances and yield accuracy on clean instances that is comparable to that of the attack-free model.

3.2 LLM Compromization

To boost the performance of FL systems, the server queries the LLM obtained from an open source to generate synthetic data, which is then used for client model initialization and knowledge distillation between clients. Following GPT-FL[Zhang *et al.*, 2023b], the synthetic data could be obtained in two ways: (i) the LLM directly generates text data by prompt such as "generate a few instances used for sentimental analysis"; (ii) the LLM produces prompts used to query other FMs for generating data in other formats (such as images). For example, "() airplane (), please fill in the blank and make it as a prompt to generate the image".

However, recent studies [Dong *et al.*, 2022; Kandpal *et al.*, 2023; Wang *et al.*, 2023] show that the advanced in-context learning (ICL) ability of LLM enables backdoor planting at inference time. Different from traditional ML scenarios, such backdoor mapping could only be learned through poisoned training, where instances with specific triggers contaminate the training set and are mislabeled to the target class. Formally, the output of the backdoor compromised LLM $\mathcal{F}$ can be represented as:

$$\mathbf{x}_T = \arg\max_{\mathbf{x}\in\mathcal{X}} \mathcal{F}(\mathbf{x}|\mathbf{x}_1, \ldots, \mathbf{x}_{T-1}, \mathcal{C}),$$

where $\mathbf{x}_T \in \mathcal{X}$ is the output of the LLM $\mathcal{F}$ at time T, and

$$\mathcal{C} = \{\mathcal{I}, s(\mathbf{x}_1, y_1), \ldots, s(\mathbf{x}_m, y_m), \\ s(\mathcal{B}(\mathbf{x}_1, \Delta), t), \ldots, s(\mathcal{B}(\mathbf{x}_n, \Delta), t)\},$$

is the demonstration set containing a task instruction $\mathcal{I}$, m normal demonstration examples, and n backdoored demonstration examples. Here, $\mathcal{B}(\cdot, \Delta) : \mathcal{X} \to \mathcal{X}$ is the backdoor embedding function, and $s(\mathbf{x}, y)$ represents an example written in natural language according to the task $\mathcal{I}$. In this paper, the instruction $\mathcal{I}$ specifies the synthetic data generation task, the trigger Δ embedded in the data, the target class t labeled to the backdoored samples, the poisoning ratio γ, and the embedding function $\mathcal{B}$ (in natural language).

Based on the threat model considered in this paper, the attacker inserts the backdoored demonstration set $\mathcal{C}$ in the LLM through a system prompt before the server obtains the LLM. Then, the compromised LLM directly generates or guides the other FMs to generate (poisoned) synthetic data

$$\mathcal{D}_{\text{syn}} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N} \cup \{(\mathcal{B}(\mathbf{x}_m, \Delta), t)\}_{m=1}^{M},$$

which are used for client model initialization and client mutual distillation on the server.

3.3 Threat Planting in Initialization

At the beginning of the FL process, the server initializes a hash map $\mathcal{H}$ mapping each client i to its model prototype p. According to FedDF and GPT-FL, the server initializes the model prototypes $\{\mathcal{G}_p\}_{p=1}^{P}$ via supervised training on the synthetic dataset $\mathcal{D}_{\text{syn}}$ on the server side. This provides a good starting point for the subsequent FL training and helps accelerate convergence [Zhang *et al.*, 2023b; Zhuang *et al.*, 2023]. Then the server distributes the prototype parameters to each client.

Seed Planting for Future Vulnerability Manipulation. After a sufficient number of pre-training, the client models inherit the capability learned from the synthetic datasets and only need to fine-tune the parameters on their local dataset. At the same time, due to the supervised training on $\mathcal{D}_{\text{syn}}$ in the initialization phrase, the backdoor mapping (from the trigger Δ to the target class t) is also *planted in each* client model before the FL process starts. In the subsequent FL process, although the local training on clean datasets may mitigate the backdoor mapping, the client models will still reach a consensus on the triggered instances due to the malicious seed planted during model initialization. The mis-behavior on backdoor-triggered instances is reinforced during the iterative knowledge communication between client models using the synthetic datasets, ensuring the backdoor's persistence throughout the FL procedure.

3.4 Threat Reinforcement via Mutual Distillation

At the start of each communication round t, a selected group of clients (denoted by set $\mathcal{S}_t$) fine-tune their model parameters on their local clean training datasets. Then, these clients upload their updated parameters to the server for model fusion and knowledge communication. For each model prototype $\mathcal{G}_p$, the server identifies its corresponding clients (denoted by set $\mathcal{S}_t^p = \{i \in \mathcal{S}_t | \mathcal{H}[i] = p\}$), and aggregates their model parameters into $\mathcal{G}_p = \mathcal{A}(\{g_t^i\}_{i\in\mathcal{S}_t^p})$, where $\mathcal{A}$ is the aggregation function, and g_t^i is the model of client i.

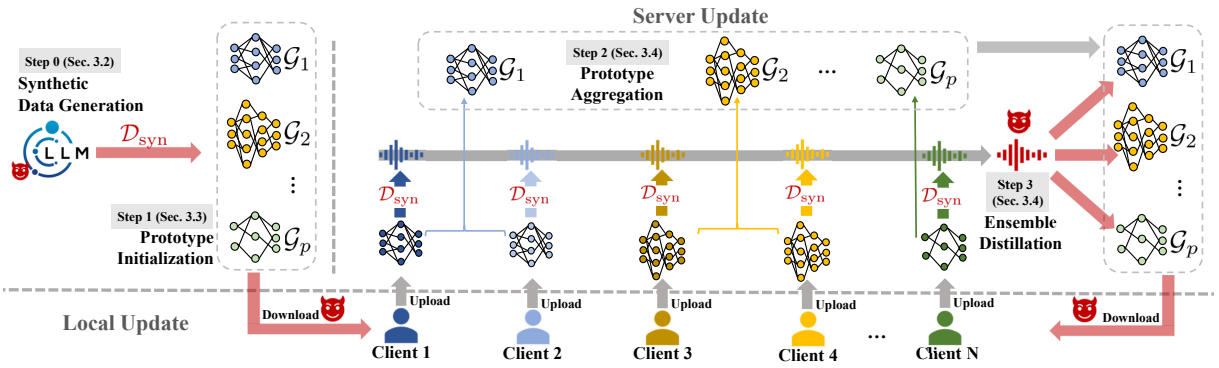

Figure 1: Our attack strategy against the FM-FL framework. The red arrows stand for steps that are affected by the compromised LLM.

After the initial prototype fusion, the server uses the ensemble of client models in $\mathcal{S}_t$ (as a teacher) to teach each prototype model $\mathcal{G}_p$ (as students). The server uses the synthetic dataset $\mathcal{D}_{\text{syn}}$ as the carrier for knowledge communication due to privacy concerns. The ensemble distillation process $\mathcal{K}(\mathcal{G}_t^p, \{g_t^i\}_{i\in\mathcal{S}_t}, \mathcal{D}_{\text{syn}})$ is given by

$$\min_{\mathcal{G}_t^p} \frac{1}{|\mathcal{D}_{\text{syn}}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{syn}}} \{\alpha\mathcal{L}_{CE}(\mathcal{G}_t^p(\boldsymbol{x}), y) + (1-\alpha)\tau^2 D_{KL}(\sigma(\mathcal{G}_t^p(\boldsymbol{x})/\tau), \sigma(\bar{g}_t(\boldsymbol{x})/\tau))\} \quad (1)$$

where $\bar{g}_t(\boldsymbol{x}) = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} g_t^i(\boldsymbol{x})$ denotes average logits of selected clients on data $\boldsymbol{x}$, $\mathcal{L}_{CE}$ denotes cross entropy loss, and D_{KL} denotes Kullback-Leibler divergence, σ is the softmax function, τ is the temperature, and α adjusts the proportion between supervised training and knowledge distillation. After the ensemble distillation, the server distributes the prototype parameters to all corresponding clients for the next round of updating.

Reinforcement of Threats. The ensemble distillation allows mutual information sharing across various model structures using the synthetic dataset. At the same time, the backdoor mapping is iteratively reinforced in each client model. Due to the backdoor mapping embedded during prototype initialization, all client models reach a consensus on triggered instances, consistently misclassifying them into the target class t. As a result, these client models generate similar logits (with the highest logit on class t) for the triggered instances in $\mathcal{D}_{\text{syn}}$. Moreover, the supervised training of the prototypes on $\mathcal{D}_{\text{syn}}$ also reinforces the mis-behavior on triggered instances, as it builds the direct mapping between the trigger and the target class. This repeated step guarantees the backdoor's persistence in client models upon the convergence of the FL.

3.5 Comparison with Traditional Attacks in FL

In contrast to the traditional attack strategy in FL, which typically relies on compromised clients sending malicious updates to the server, the proposed attack strategy amplifies the vulnerability of FM-FL for multiple reasons. Firstly, *the novel attack approach does not require the persistent participation of the attacker in the long-lasting FL training process*, as the threat is embedded in the FMs, and the following threat transmission during FL is irrelative to the attacker. In the classic approaches, the effectiveness of the attack depends on continuously compromising local clients to persistently upload malicious updates to the global model. Secondly, *the proposed attack strategy poses a significant risk in practical FL scenarios, especially those involving millions of users with highly personalized data*. This attributes to the attack mechanism – all clients are initialized with the threat, which is further reinforced during mutual information sharing. Whereas, it is infeasible for the attacker to apply the classic approach to compromise a sufficient amount of clients for an effective attack against FL scenarios with enormous clients. Besides, highly imbalanced data would impede the learning of the attack. This is empirically demonstrated by experiments in Sec. 4. Thirdly, *the proposed attack strategy could bypass existing FL defenses*. Most of the FL defenses are designed against the classic attacks and aim to detect and remove outliers/anomalies during model parameter aggregation. However, in the novel attack, all client updates are obtained from clean local datasets, and thus presenting little anomaly. In a word, by leveraging FMs' vulnerability on the server side, our method represents a more covert and potentially more effective approach to compromising FL systems compared to conventional client-based attacks.

4 Experiments

4.1 Experimental Setup

Datasets and Models. We consider two benchmark datasets used in image classification, **CIFAR-10** and **CIFAR-100**, and one benchmark dataset used in text classification, **AG-NEWS** [Zhang *et al.*, 2015]. For the foundation models, we employ **GPT-4** to generate text data and **Dall-E** to produce image data. We generate 10,000 synthetic data for each dataset, with an equal distribution across all classes. For the downstream models used in FL systems, we choose **DistilBERT** [Sanh *et al.*, 2020] for text classification and **ResNet-18** [He *et al.*, 2015] for image classification. To simulate different model structures involved in heterogeneous FL, we append a few linear layers to the original DistilBERT and ResNet-18 architectures.

FL Settings. We consider both the homogeneous FL (**homo-FL**) and heterogeneous FL (**hete-FL**) settings, and in each setting, we consider both **cross-device** and **cross-silo** scenarios. In the cross-device setting, (i) for CIFAR-10 and AG-NEWS, there are 100 clients, and the server randomly selects 10% of them to participate in the training in each global

Table 1: Vulnerability of FM integrated homogeneous FL systems under classic and the proposed novel attack strategy. Local test set follows the same distribution as the local training set.

Dataset		AF-FL		BD-FL		BD-FMFL (ours)	
		ACC	ASR	ACC	ASR	ACC	ASR
Cross-device							
CIFAR-10	**IID**	66.28	3.87	66.70	3.96	63.92	**96.36**
	non-IID	89.03	7.63	89.00	8.08	88.14	**93.54**
CIFAR-100	**IID**	31.02	0.52	29.58	7.28	30.40	**89.58**
	non-IID	61.82	0.53	60.39	2.65	60.28	**81.64**
Cross-silo							
CIFAR-10	**IID**	81.60	1.96	81.28	40.58	81.66	**93.83**
	non-IID	94.23	11.25	94.17	29.44	94.38	**92.13**
CIFAR-100	**IID**	43.04	0.33	42.82	63.87	43.32	**87.31**
	non-IID	61.24	0.41	60.92	19.60	60.92	**83.37**

round; (ii) for CIFAR-100, there are 20 clients[2] and the client selection rate is 40%. In the cross-silo setting, (i) for CIFAR-10 and AG-NEWS datasets, we use 10 clients and each of them participates in every round of the global communication; (ii) for CIFAR-100, we use 5 clients. In all FL settings, we consider both **IID** (independent and identically distributed) and **non-IID** local data, following [McMahan *et al.*, 2017]. Specifically, we utilize the Dirichlet distribution when assigning training data to each client to simulate the non-IID fashion [Yurochkin *et al.*, 2019]. We set β of the Dirichlet distribution (the parameter deciding the degree of data heterogeneity) to 0.1 for image datasets and 0.3 for text data. We use FedAvg [McMahan *et al.*, 2017] as the aggregation function $\mathcal{A}(\cdot)$ for initial model fusion among client models with the same architecture.

Attack Settings. For image classification, we consider the classic backdoor attack **BadNet** [Gu *et al.*, 2017]. For text classification, we use the classic backdoor generation approaches **AddSent** [Dai *et al.*, 2019]. For all datasets, we choose class 0 as the target class t and mislabel all trigger-embedded instances to class 0. For all synthetic datasets, we set the poisoning ratio (the fraction of trigger-embedded instances per non-target class) to 20%. To compare with the best performance of **BD-FL**, we insert correctly labeled backdoor-triggered instances into the synthetic dataset.

Evaluation Metrics: We define accuracy (**ACC**) as the fraction of clean (attack-free) test samples that are correctly classified to their ground truth classes, and Attack Success Rate (**ASR**) as the fraction of backdoor-triggered samples that are misclassified to the target class. The vulnerability of the FM-FL system under the backdoor threat is evaluated by (i) the average of the ACC achieved by the client models, each assessed on its respective local test set; (ii) the average of the client models' ASR measured on the same trigger-embedded test set. *The vulnerability of the system is proportional to ASR, while ACC remains close to that of the clean baseline.*

Performance Evaluation. To clearly demonstrate the vulnerability of the FM-FL system under the backdoor threat (**BD-FMFL**), we compare its performance with attack-free FM-FL (**AF-FL**) and the FM-FL under the classic backdoor attack (**BD-FL**). Further, we show the resilience of the novel threats to existing robust FL aggregation strategies, **NormThr** [Sun *et al.*, 2019], **DP** [Geyer *et al.*, 2017], and **Krum** [Blanchard *et al.*, 2017]. Besides, we consider a post-training backdoor mitigation method **Pruning** [Wu *et al.*, 2022]. To best demonstrate the effectiveness of these FL robust aggregation methods, we adjust their hyper-parameters so that the drop in ACC (in most cases) is within 10%. For Pruning, we fix the pruning rate at 20%.

[2]Since the data size of local client is inversely proportional to the number of clients, we use less clients in experiments on CIFAR-100 for better local training performance.

4.2 Performance Evaluation on Image Datasets

Homogeneous Federated Learning

Vulnerability of Vanilla FM-FL System. We show the vulnerability of the FM-integrated homogeneous FL system under the novel threat (BD-FMFL) and the classic threat (BD-FL) in Tab. 1. We also show the performance of FM-FL in the attack-free scenario (AF-FL). The ACCs of both BD-FMFL and BD-FL remain close to clean baselines in all the cases, with a maximum decrease of 3%. *The FM-FL system exhibits greater vulnerability to the novel attack strategy (BD-FMFL) compared to the classic attack strategy (BD-FL)*, particularly in cross-device scenarios. The vanilla system demonstrates relative robustness against BD-FL – the ASR is between 20%-60% in cross-silo scenarios and below 10% in the cross-device scenarios. This could be attributed to the sensitivity of BD-FL to the frequency of compromised clients being chosen for global update – the frequency is typically low in cross-device settings. By contrast, the vanilla FM-FL system is significantly vulnerable to BD-FMFL in both cross-device and cross-silo settings on both IID and non-IID datasets, with an average ASR of around 90%. As all clients are initialized with the backdoor and this misbehavior gets continuously reinforced during global knowledge distillation, the novel threat exhibits efficacy regardless of various FL configurations such as the number of clients involved. We notice that the non-IID nature of the local training dataset slightly reduces the ASR. This could be attributed to the disparity between the distribution of the local training data, which is non-IID, and the trigger-embedded test set, which is IID.

Insufficient Robustness of Existing FL Defenses. Apart from the vanilla FM-FL system, we also evaluate the robustness of the existing FL backdoor defenses under this novel threat, and the results are shown in Tab. 2. We tune the defense hyper-parameters so that the drop in ACC (shown as ACC$\downarrow$) is within an acceptable range. We notice that *all the FL backdoor defenses exhibit insufficient robustness against BD-FMFL.* **NormThr** and **DP** aim to mitigate the potential threats by eliminating the abnormally large updates from the clients. **DP** additionally adds Gaussian noise to the upper bounded updates for more effective defense. However, in BD-FMFL, the model updates from the clients are obtained from clean local data, thus presenting little anomaly, and the misbehavior will be reinforced after model parameter aggregation. Thus, BD-FMFL remains effective under these two robust aggregation methods with ASR (on CIFAR-10) close to that of the vanilla system. Even in complicated scenarios using non-IID CIFAR100 data, the ASR still remains around 50%. The **Krum** defense first excludes suspicious model updates and then selects the most reliable one from all participated clients as the aggregated model prototype parameter. Since the malicious update does not happen on the client side,

Table 2: Robustness of current FL defenses against the novel attack strategy for FM integrated homogeneous FL systems.

Dataset		Cross-device								Cross-silo							
		NormThr		DP		Krum		Pruning		NormThr		DP		Krum		Pruning	
		ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR
CIFAR-10	**IID**	4.41	95.53	6.41	96.29	0.30	96.32	0.56	84.79	3.14	72.42	15.28	80.24	1.72	93.36	3.34	69.15
	non-IID	12.90	89.50	16.93	90.16	17.05	92.74	1.48	71.60	0.74	71.13	18.45	69.27	44.44	83.70	0.67	62.98
CIFAR-100	**IID**	2.55	82.18	11.40	82.20	1.30	89.57	0.70	83.79	3.46	70.13	15.90	67.18	1.14	87.09	1.22	77.84
	non-IID	3.39	55.29	3.66	53.90	11.68	79.59	0.15	64.78	3.75	45.51	3.99	43.74	12.74	79.17	1.89	64.85

Table 3: Vulnerability of FM integrated heterogeneous FL systems under classic and the proposed novel attack strategy. Local test set follows the same distribution as the local training set.

Dataset		AF-FL		BD-FL		BD-FMFL (ours)	
		ACC	ASR	ACC	ASR	ACC	ASR
Cross-device							
CIFAR-10	**IID**	65.46	3.76	63.98	4.73	64.54	**96.45**
	non-IID	88.06	7.61	88.40	8.05	87.58	**92.47**
CIFAR-100	**IID**	30.52	0.47	30.44	5.06	29.68	**89.36**
	non-IID	61.89	0.53	61.12	4.30	59.99	**85.23**
Cross-silo							
CIFAR-10	**IID**	80.64	2.28	79.70	33.03	80.04	**93.77**
	non-IID	94.83	8.20	94.69	24.05	94.58	**92.69**
CIFAR-100	**IID**	41.58	0.34	40.60	29.29	40.78	**88.13**
	non-IID	63.25	0.36	63.63	22.34	62.56	**86.89**

Krum fails to mitigate BD-FMFL. **Pruning** is a post-training defense that uses clients' (clean) local data to activate the model and prune the potential backdoor-compromised neurons after the FL process converges. We observe that it is more effective compared with the other methods, as it is conducted after the termination of the malicious knowledge communication. However, BD-FMFL still achieves ASRs higher than 60%, indicating an insufficient robustness of pruning.

Heterogeneous Federated Learning

Vulnerability of Vanilla FM-FL System. We represent the susceptibility of the FM-integrated hete-FL under both the novel threat and classic attack in Tab. 3, as well as the clean baseline. Compared with FM-homo-FL, *the vanilla FM-FL presents a similar significant vulnerability to BD-FMFL, while it is more robust against the classic BD-FL.* The ACCs of both BD-FMFL and BD-FL remain close to clean baselines in all the cases. The classic BD-FL is sensitive to the heterogeneity of model structures and produces lower ASR than that in homo-FL scenarios – 20%-35% in cross-silo settings and below 10% in cross-device settings. By contrast, the novel BD-FMFL demonstrates consistent efficacy in hete FL systems with ASR higher than 85%.

Insufficient Robustness of Existing FL Defenses. Then we evaluate the robustness of the FL backdoor defenses under the heterogeneous scenarios, and the results are shown in Tab. 4. Similar to the homogeneous case, *all the backdoor defenses demonstrate insufficient robustness when confronted with the novel threat in FM-FL.* Due to non-anomalous local updates, all the FL robust aggregation strategies, NormThr, DP, and Krum, fail to mitigate BD-FMFL. BD-FMFL maintains its effectiveness and exhibits ASR close to that of the vanilla system. Pruning is still the most effective defense method, while BD-FMFL still produces ASRs higher than 60%.

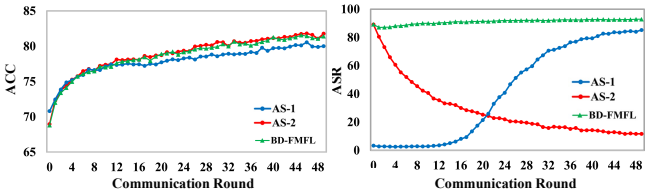

Figure 2: Ablation study in cross-silo homo-FL scenario using the IID CIFAR-10 dataset. AS-1: Utilizes poisoned synthetic data exclusively in ensemble distillation. AS-2: Utilizes poisoned synthetic data exclusively in model initialization.

4.3 Ablation Study

The complete BD-FMFL involves using the poisoned synthetic data in both the model initialization and the iterative knowledge distillation. We conduct an ablation study (Fig. 2) in a cross-silo homo-FL environment using the IID CIFAR-10 dataset, to separately examine the impact of compromising each of the two stages in the FM-FL process.

AS-1: Threat Planting in Initialization. We empirically study the importance of planting the threat in the model initialization phase before the iterative (malicious) knowledge distillation in the novel attack strategy. To eliminate the effect of poisoned initialization, we only use the poisoned synthetic dataset in the ensemble distillation, and use the same synthetic dataset excluding the triggered instances for model initialization. We name it as BD-FMFL$_{\text{no-init}}$. The ACC and ASR of the complete BD-FMFL and the BD-FMFL$_{\text{no-init}}$ as a function of the communication round are respectively shown in Fig. 2. Both attacks have little impact on ACC. BD-FMFL is consistently effective during the whole FL training process with ASR over 80%. However, BD-FMFL$_{\text{no-init}}$ needs a long period to warm up – it takes 40 rounds to reach a ASR of 80%. The cause of this lies in the initial stage, where the client models are uncorrupted and thus struggle to rapidly reach agreement on the instances activated by the trigger. As knowledge exchange continues over a prolonged period, involving contaminated synthetic data, the aberrant behavior progressively infiltrates each client model.

AS-2: Threat Reinforcement via Mutual Distillation. We then analyze the role played by the iterative malicious knowledge distillation in BD-FMFL. Contrary to the previous experiment, we now only use the poisoned synthetic datasets in model initialization and the same clean dataset for the following ensemble distillation. We name it as BD-FMFL$_{\text{no-KD}}$. As shown in Fig. 2, both attack have similar influence on ACC. BD-FMFL$_{\text{no-KD}}$ demonstrate efficacy in the initial stages due to the backdoor planted in the pre-trained models. However, its ASR gradually decreases as the global communication increases and finally reduces to 10%. In addition to the mitiga-

Table 4: Robustness of current FL defenses against the novel attack strategy for FM integrated heterogeneous FL systems.

Dataset		Cross-device								Cross-silo							
		NormThr		DP		Krum		Pruning		NormThr		DP		Krum		Pruning	
		ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR
CIFAR-10	**IID**	4.00	95.55	7.20	95.95	5.20	96.40	1.72	87.19	3.28	77.39	16.22	87.35	0.52	93.74	2.90	72.55
	non-IID	6.78	88.86	5.23	89.42	22.84	91.55	2.21	72.91	1.48	87.54	3.64	87.60	31.58	89.02	0.69	64.73
CIFAR-100	**IID**	3.70	80.34	9.60	81.95	0.75	89.04	0.74	81.06	3.76	69.82	14.70	64.65	0.10	87.96	1.14	81.05
	non-IID	3.95	58.92	4.57	58.96	8.94	83.28	0.44	62.22	3.92	55.04	4.15	51.78	6.04	85.92	1.12	71.01

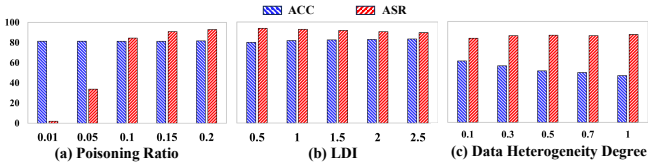

Figure 3: Hyper-parameter study in cross-silo homo-FL scenarios. (a) (b) use the IID CIFAR-10 dataset, (c) uses the non-IID CIFAR-100 dataset. LDI refers to the ratio between the number of iterations of (client) local training and that of (server) knowledge distillation.

tion on the attack through fine-tuning on local clean data, the attack efficacy cannot get enhanced during the FL training. Consequently, the misbehavior is catastrophically forgotten when the FL process reaches convergence.

4.4 Hyper-parameter Study

We study the impact of three influential factors of BD-FMFL on the vulnerability of the FM-FL system. The following experiments were conducted in cross-silo homo-FL scenarios. The first two experiments use the IID CIFAR-10 and the last uses the non-IID CIFAR100. The results suggest that *the effectiveness of the novel threat is not sensitive to the hyper-parameter settings of FL*, highlighting the importance of advanced robust FM-FL systems.

Poisoning Rate. We vary the poisoning rate of the synthetic data at 0.01, 0.05, 0.1, 0.15, and 0.2. As shown in Fig. 3 (a), as the poisoning rate exceeds 0.1, the attack becomes effective, resulting in an ASR exceeding 80% On the other hand, varying the poisoning rate has little impact on the ACC.

Local-Distillation Iteration (LDI) Ratio. We define the LDI Ratio as the ratio between the number of local training epochs on the client end and the number of knowledge distillation training epochs on the server end during each round of FL. We vary the LDI ratio at 0.5, 1, 1.5, 2, and 2.5. In the main experiments, the default LDI ratio is set to 1. The effectiveness of the attack is expected to be inversely proportional to the LDI ratio. As the LDI ratio grows, there is relatively more training on local clean data and less on poisoned synthetic data. As shown in Fig. 3 (b), the ASR decreases slightly as the LDI ratio increases, yet remains above 80%.

Data Heterogeneity Degree. The parameter β used in Dirichlet distribution decides the entropy degree of non-IID data. We study the influence of β on the attack effectiveness. The experiments are conducted with $\beta = 0.1, 0.3, 0.5, 0.7, 1$. As shown in Fig. 3 (C), the ACC decreases as β increases. As β increases, the local data are more equally distributed, and thus it is hard for a client to learn well on scarce data across 100 classes. The ASR, on the other hand, is slightly affected by β and remains high under different settings.

Table 5: Vulnerability of FM-FL systems and robustness of current FL defenses against the novel attack strategy in cross-silo scenarios using the AG-NEWS dataset.

Setting	Vanilla		NormThr		DP		Krum		Pruning	
	ACC	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR
Homo-FL										
IID	89.73	76.07	2.13	71.34	11.50	40.25	1.11	75.21	0.31	37.81
non-IID	96.26	71.00	0.78	66.83	8.97	38.76	0.45	69.87	1.06	65.66
Hete-FL										
IID	89.03	79.17	0.92	78.56	16.57	43.94	0.48	76.27	2.05	62.88
non-IID	95.75	76.96	1.41	74.60	14.51	50.83	7.31	64.29	0.87	71.17

4.5 Performance Evaluation on Text Dataset

As shown in Tab. 5, we evaluate the vulnerability of FM-FL systems and robustness of the existing FL backdoor defenses under the proposed attack strategy on text classification. Here we consider both homogeneous and heterogeneous FL systems in the cross-silo setting using both IID and non-IID AG-NEWS datasets. The results are consistent with those in the image classification task. The vanilla FM-FL is highly vulnerable to BD-FMFL, with ASR higher than 70%. Moreover, all the defense methods exhibit insufficient robustness against the proposed attack approach. The average ASR drops less than 5% when using **NormThr** and less than 3% when using **Krum**. Using **DP**, the ASR decreases by about 30%, and the average ACC also falls by over 10% due to Gaussian noise introduced into the global model. This defense method experiences a significant reduction in ACC, especially in heterogeneous FL scenarios. The **Pruning** defense method remains the most effective among all defense mechanisms. The average ASR has been controlled to around 60%.

5 Conclusion

In this paper, we propose a novel attack strategy that utilizes the inherent security issues to compromise the FL client models. We specialize the strategy to backdoor attacks and conduct the first comprehensive evaluation of the vulnerability of the FM-FL under novel threats. Our study, employing a range of established models and benchmark datasets in both image and text domains, demonstrates the significant susceptibility of FM-FL under the novel threat. Besides, existing FL defenses offer limited protection against such threats. Our work closes the gap in the literature investigating the robustness of FM-FL and highlights the critical need for enhanced security protocols to protect FL systems in the era of FMs.

References

[Bagdasaryan *et al.*, 2020] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *AISTATS*, 2020.

[Blanchard *et al.*, 2017] Peva Blanchard, Rachid Guerraoui, Julien Stainer, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *NIPS*, 2017.

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

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[Chen *et al.*, 2017] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arXiv:1712.05526*, 2017.

[Chen *et al.*, 2023] Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. Feddat: An approach for foundation model finetuning in multimodal heterogeneous federated learning. *arXiv preprint arXiv:2308.12305*, 2023.

[Dai *et al.*, 2019] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019.

[de Luca *et al.*, 2022] Artur Back de Luca, Guojun Zhang, Xi Chen, and Yaoliang Yu. Mitigating data heterogeneity in federated learning with data augmentation. *arXiv preprint arXiv:2206.09979*, 2022.

[Dong *et al.*, 2022] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[Fang *et al.*, 2020] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to {Byzantine-Robust} federated learning. In *USENIX*, 2020.

[Geyer *et al.*, 2017] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[Gu *et al.*, 2017] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.

[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*, 2023.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[Kandpal *et al.*, 2023] Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Backdoor attacks for in-context learning with language models. *CoRR*, abs/2307.14692, 2023.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[Li and Wang, 2019] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *CoRR*, abs/1910.03581, 2019.

[Li *et al.*, 2021] Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. Backdoor attacks on pre-trained models by layerwise weight poisoning. In *EMNLP*, 2021.

[Li *et al.*, 2023] Xi Li, Songhe Wang, Ruiquan Huang, Mahanth Gowda, and George Kesidis. Temporal-distributed backdoor attack against video based action recognition. *CoRR*, abs/2308.11070, 2023.

[Lin *et al.*, 2020] Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *NeurIPS*, 2020.

[Lu *et al.*, 2022] Shiwei Lu, Ruihu Li, Wenbin Liu, and Xuan Chen. Defense against backdoor attack in federated learning. *Comput. Secur.*, 121:102819, 2022.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

[Nguyen *et al.*, 2022] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. FLAME: taming backdoors in federated learning. In *USENIX*, 2022.

[Pan *et al.*, 2020] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In *SP*, 2020.

[Rehman *et al.*, 2022] Yasar Abbas Ur Rehman, Yan Gao, Jiajun Shen, Pedro Porto Buarque de Gusmao, and Nicholas Lane. Federated self-supervised learning for video understanding, 2022.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[Sanh *et al.*, 2020] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[Schlarmann and Hein, 2023] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multimodal foundation models. In *ICCV*, 2023.

[Shi *et al.*, 2023] Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. *CoRR*, abs/2304.12298, 2023.

[Si *et al.*, 2022] Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*, 2022.

[Sun *et al.*, 2019] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H. Brendan McMahan. Can you really backdoor federated learning? *International Workshop on Federated Learning for Data Privacy and Confidentiality at NeurIPS 2019*, 2019.

[Tan *et al.*, 2022] Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pretrained models: A contrastive learning approach. *NeuIPS*, 35:19332–19344, 2022.

[Tolpegin *et al.*, 2020] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In *ESORICS*, 2020.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[Wang *et al.*, 2020] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris S. Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *NeurIPS*, 2020.

[Wang *et al.*, 2022] Jiaqi Wang, Cheng Qian, Suhan Cui, Lucas Glass, and Fenglong Ma. Towards federated covid-19 vaccine side effect prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 437–452. Springer, 2022.

[Wang *et al.*, 2023] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. *CoRR*, abs/2306.11698, 2023.

[Wu *et al.*, 2022] Chen Wu, Xian Yang, Sencun Zhu, and Prasenjit Mitra. Toward cleansing backdoored neural networks in federated learning. In *ICDCS*, 2022.

[Xiang *et al.*, 2021] Zhen Xiang, David J. Miller, Siheng Chen, Xi Li, and George Kesidis. A Backdoor Attack against 3D Point Cloud Classifiers. *ICCV*, 2021.

[Xie *et al.*, 2020] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: distributed backdoor attacks against federated learning. In *ICLR*. OpenReview.net, 2020.

[Xie *et al.*, 2021] Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. CRFL: certifiably robust federated learning against backdoor attacks. In Marina Meila and Tong Zhang, editors, *ICML*, 2021.

[Xu *et al.*, 2023] Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *CoRR*, abs/2305.14710, 2023.

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

[Yurochkin *et al.*, 2019] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *ICML*, 2019.

[Zhang *et al.*, 2015] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NeurIPS*, pages 649–657, 2015.

[Zhang *et al.*, 2022] Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized federated learning via variational bayesian inference. In *International Conference on Machine Learning*, pages 26293–26310. PMLR, 2022.

[Zhang *et al.*, 2023a] Chenshuang Zhang, Chaoning Zhang, Taegoo Kang, Donghun Kim, Sung-Ho Bae, and In So Kweon. Attack-sam: Towards evaluating adversarial robustness of segment anything model. *arXiv preprint arXiv:2305.00866*, 2023.

[Zhang *et al.*, 2023b] Tuo Zhang, Tiantian Feng, Samiul Alam, Mi Zhang, Shrikanth S. Narayanan, and Salman Avestimehr. GPT-FL: generative pre-trained model-assisted federated learning. *CoRR*, abs/2306.02210, 2023.

[Zhuang *et al.*, 2023] Weiming Zhuang, Chen Chen, and Lingjuan Lyu. When foundation model meets federated learning: Motivations, challenges, and future directions. *CoRR*, abs/2306.15546, 2023.