Foundation Models in Federated Learning: Assessing Backdoor Vulnerabilities

Xi Li†
University of Alabama at Birmingham
xli7@uab.edu

Chen Wu†
Meta
masterchenwu@meta.com

Jiaqi Wang
The Pennsylvania State University
jqwang@psu.edu

***Abstract*—Federated Learning (FL), a privacy-preserving machine learning framework, faces significant data-related challenges. For example, the lack of suitable public datasets leads to ineffective information exchange, especially in heterogeneous environments with uneven data distribution. Foundation Models (FMs) offer a promising solution by generating synthetic datasets that mimic client data distributions, aiding model initialization and knowledge sharing among clients. However, the interaction between FMs and FL introduces new attack vectors that remain largely unexplored. This work therefore assesses the backdoor vulnerabilities exploiting FMs, where attackers exploit safety issues in FMs and poison synthetic datasets to compromise the entire system. Unlike traditional attacks, these new threats are characterized by their one-time, external nature, requiring minimal involvement in FL training. Given these uniqueness, current FL defense strategies provide limited robustness against this novel attack approach. Extensive experiments across image and text domains reveal the high susceptibility of FL to these novel threats, emphasizing the urgent need for enhanced security measures in FL in the era of FMs [1].**

***Index Terms*—Federated Learning, Backdoor Attacks, Foundation Models**

I. Introduction

Federated Learning (FL) [1] 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 [2], model personalization [3], and video surveillance [4]. 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 [5], LLaMA [6], and Stable Diffusion [7], 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 [8], addressing issues like limited data availability [9], and introducing diversity into the training process to cover a broader spectrum of scenarios not originally included in the original data.

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 [10]. This brings inherent vulnerabilities in the FMs to have robustness, fairness, and privacy issues [11]. Recent studies have revealed threats to FMs range from adversarial examples [12], data poisoning attacks to generate malicious output [13], backdoor attacks to inject hidden mappings in the objective function [14], privacy attacks to reveal sensitive information from training data [15], to fairness and reliability of the FMs [16]. 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 [10], [17]. To investigate the susceptibility of FM-FL, we leverage a unified framework well-suited for both homogeneous and heterogeneous FL systems [8]–[10], [18]. 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 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 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 [19], text classification [20], [21], point cloud classification [22], video action recognition [23], and federated learning systems [24]. The compromised model will mis-classify instances embedded with a specific trigger to the attacker-chosen target class, while maintaining high accuracy on clean data, rendering the attack in a stealthy manner.

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 do-

† Equal contribution.

[1]The source code is available at https://github.com/lixi1994/FM_in_FL_BD.git

mains in different FL scenarios. The efficacy of the novel threat arises from two key aspects. Firstly, unlike traditional attacks that require compromising clients to upload malicious updates, which are often detectable as anomalies. Our strategy embeds the threat in each client at the initialization stage, further reinforced through mutual information sharing on the server. Updates derived from clean local datasets ensure no anomalies, allowing the attack to evade existing FL defenses. Secondly, the attack's success does not hinge on persistent FL training participation or compromising many clients, making it viable even in scenarios involving millions of clients. 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.

II. Related Work

A. FM integration in FL

The synergy between Foundation Models (FM) and Federated Learning (FL) enhances both domains [10], [25], [26]. On one hand, FL offers expanded data access and distributed computation for FMs. Key developments include FedDAT [27] fine-tuning framework using a Dual-Adapter Teacher for handling data heterogeneity, and PromptFL [28] 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 [8] and knowledge distillation [29]. FedPCL [30] further integrate FMs into FL, emphasizing parameter prioritization and high-performance subnetwork extraction.

B. Backdoor Attacks and Defenses in FL

A backdoor attacker in FL aims to embed malicious behavior into the global model distributed to all clients. This backdoor behavior (*e.g.*, misclassification to a specific target class) is triggered only by specific patterns embedded in input samples, while the model functions normally on clean inputs.

Classic Backdoor Threats: Classic backdoor threats primarily target the client side through techniques like data poisoning [31], local model poisoning [32], and attacks such as semantic and distributed backdoors [24], [33], [34]. For instance, attackers may inject poisoned samples into the local training datasets of compromised clients. These compromised local models then propagate the malicious model updates to the global model during server-side aggregation. With sufficient compromised clients and communication rounds, the global model is embedded with backdoor threats.

Existing Backdoor Defenses: Defenses against these attacks typically involve norm threshold bounding [35], differential privacy [36], [37], anomaly detection [38], strategies like model clustering and noise injection [39], and pruning [40]. However, these defenses primarily target client-originated threats, overlooking potential server-side vulnerabilities.

C. FM Vulnerabilities

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 [41], instruction-based attacks [42], and targeted misclassification [14]. 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. 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.

III. Methodology

A. Overview

FM integration in FL. Our work follows existing FM-integrated FL (FM-FL) frameworks, such as as those proposed in [8], [10]. The basic FM-FL cycle, as illustrated in Fig. 1 and in Alg. 1, consists of three key steps. **Stage 1: Initialization.** An FM is integrated into the server to generate synthetic data (*e.g.*, text or image data) that mirrors the distribution of client-local data, following [8]. The data is first used for model initialization and is later used to fuse a global model, following [9], [18]. **Stage 2: Client Update.** Clients independently train their local models using private local data. Once trained, they upload their model parameters to the server for aggregation during the model fusion process. **Stage 3: Server Global Model Fusion.** The server aggregates the client model parameters using synthetic data as a carrier for client model information sharing. This process employs aggregation functions such as those proposed in [9], [18], which are applicable to various FL settings. Stage 2 and 3 are repeated until FL converges.

The proposed attack machenism. Through this FM-FL framework, we explore a new attack vector of backdoor attacks. A malicious actor utilizes the vulnerabilities in FMs and inject backdoor threats into the generated synthetic data. With the usage of synthetic data for model initialization and model fusion, the backdoor threats eventually planted into all clients. This attack mechanism is fundamentally different from the classic backdoor attack against FL, thus cannot be defended by existing FL defenses.

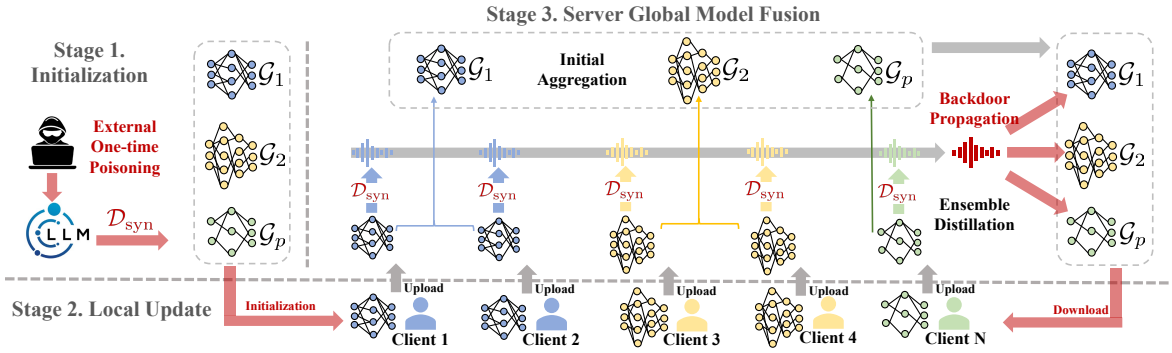


Fig. 1: The novel backdoor attack strategy targets FM-FL. Red arrows indicate steps affected by the compromised FM.

B. Threat Model

Our threat model aligns with the use of cutting-edge FMs accessed via APIs and focuses on classification tasks, which is commonly studied in both backdoor and FL research [43]–[45].

Attacker's Abilities: (1) *External*: The attacker has access to server's FM queries and can insert malicious instructions to guide the LLM to execute backdoor attacks when triggered. These instructions specify the backdoor trigger, desired outputs, and provide both clean and corrupted demonstrations. (2) *One-time Poisoning*: The attacker introduces a backdoor via a single poisoned instruction in the synthetic dataset, without needing ongoing involvement in the FL process or access to training data, methods, or LLM parameters.

Attacker's Objectives: The attacker aims to (1) guide the FMs to generate synthetic datasets containing backdoor-poisoned samples, and (2) leveraging (1), propagate the backdoor to all client models in FL, causing the final model to misclassify triggered inputs to the target class while maintaining high performance on clean samples.

C. A Novel Backdoor Mechanism against FM-FL

1) External One-time Poisoning: A feasible method to manipulate the public dataset produced by the FM leverages its in-context learning (ICL) capability, as demonstrated in recent studies [14], [46]. Unlike traditional ML, where backdoor threats require poisoned training, ICL enables backdoor implantation at inference time. Formally, the output of the backdoor-compromised FM $\mathcal{F}$ 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

$$\mathcal{C} = \{\mathcal{I}, \{s(\mathbf{x}_i, y_i)\}_i, \{s(\mathcal{B}(\mathbf{x}_j, \Delta), t)\}_j\},$$

is the demonstration set containing a task instruction $\mathcal{I}$, a few normal examples, and several backdoored 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 $\mathcal{I}$. The instruction $\mathcal{I}$ defines the data generation task, specifies the trigger Δ, target class t for poisoned samples, poisoning ratio γ, and embedding function $\mathcal{B}$ in natural language. Consequently, the generated synthetic data becomes compromised:

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

This attack is External One-time Poisoning, as the adversary neither needs insider access to FL nor continuous participation to maintain the effectiveness of backdoor throughout the FL cycle. Experimental validation is provided in Sec. IV.

2) Backdoor Threats Propagated Through FM-FL Interaction: We now elaborate on the novel backdoor mechanism embedded into the FM-FL cycle.

Stage 1: Initialization. In the FM-FL framework, the server initializes model prototypes $\{\mathcal{G}_p\}_{p=1}^{P}$ on the synthetic dataset $\mathcal{D}_{\text{syn}}$, providing a strong starting point and accelerating FL convergence [8], [10]. The prototype parameters are then distributed to clients. After sufficient pre-training, client models inherit knowledge from $\mathcal{D}_{\text{syn}}$ and require only fine-tuning on local datasets. Note that, the pre-training on $\mathcal{D}_{\text{syn}}$ embeds the backdoor mapping (trigger Δ to target class t) into client models even before FL begins.

Stage 2: Client Update. Clients receive the updated model from the server and fine-tune it on their local clean datasets, which may potentially mitigate the implanted backdoor mapping. Clients then upload their locally fine-tuned models to the server for global model aggregation.

Stage 3: Server Global Model Fusion. Model fusion involves aggregating prototype models and ensemble distillation for client knowledge sharing. At the beginning of each communication round t, selected clients $\mathcal{S}_t$ upload their updated parameters to the server. The server groups clients by prototype model, $\mathcal{S}_t^p = \{g_t^i \in \mathcal{S}_t \mid \mathcal{H}[i] = p\}$, and aggregates their updates as $\mathcal{G}_t^p = \mathcal{A}(\{g\}_{g\in\mathcal{S}_t^p})$, where $\mathcal{H}$ is a hash function and $\mathcal{A}$ denotes the aggregation function.

After prototype fusion, the server employs ensemble distillation, using client models in $\mathcal{S}_t^p$ as teachers to refine each prototype model $\mathcal{G}_t^p$ as a student. To preserve privacy, the synthetic dataset $\mathcal{D}_{\text{syn}}$ serves as the medium for knowledge communication. The ensemble distillation process $\mathcal{G}_t^p \leftarrow \mathcal{K}(\{g\}_{g\in\mathcal{S}_t}, \mathcal{D}_{\text{syn}})$ is formulated as:

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

where $\bar{g}_t(\boldsymbol{x}) = \frac{1}{|\mathcal{S}_t|}\sum_{g\in\mathcal{S}_t} g(\boldsymbol{x})$ represents the averaged client logits, $\mathcal{L}_{CE}$ is the cross-entropy loss, D_{KL} is the Kullback-Leibler divergence, σ denotes the softmax function, τ is the temperature, and α balances supervised training and distillation. Upon completion, the server distributes the updated prototype parameters to clients for the next training round.

Stages 2 and 3 are repeated iteratively, enabling local updates on real datasets and knowledge sharing across diverse model structures using the synthetic dataset. Meanwhile, the backdoor mapping is progressively reinforced in each client model. Since the backdoor was embedded during prototype initialization, all client models converge to misclassify triggered instances into the target class t. Consequently, during knowledge distillation, they produce similar logits, with the highest value assigned to class t for triggered instances in $\mathcal{D}_{\text{syn}}$. Additionally, the supervised training of prototypes on $\mathcal{D}_{\text{syn}}$ further strengthens this misclassification by directly mapping the trigger to the target class. This iterative process ensures the persistence of the backdoor in client models as FL training converges, even without the persistent participation of the attacker.

Algorithm 1: The Backdoor Mechanism against FM-FL.

Initialization
  The FM $\mathcal{F}$ generates synthetic data $\mathcal{D}_{\text{syn}}$
  Pre-train each prototype $\mathcal{G}_p$ on $\mathcal{D}_{\text{syn}}$
  Distribute the prototype parameters to clients
for each communication round $t = 1, \cdots, T$ **do**
  $\mathcal{S}_t \leftarrow$ a random subset (ρ fraction) of the N clients.
  Client Update
    for each client $i \in \mathcal{S}_t$ in parallel **do**
      Fine-tune clinet model g_t^i with $\mathcal{D}_i$
      Upload model parameter g_t^i to the server
    end
  Server Global Model Fusion
    for each prototype $p \in P$ in parallel **do**
      Initial model fusion $\mathcal{G}_p \leftarrow \mathcal{A}(\{g\}_{g\in\mathcal{S}_t^p})$
      Update prototype student by ensemble distillation Eq. (1) $\mathcal{G}_p \leftarrow \mathcal{K}(\{g\}_{g\in\mathcal{S}_t}, \mathcal{D}_{\text{syn}})$
      Distribute the prototype parameters to corresponding clients
    end
end

D. Classic vs. Novel Attack Mechanisms

Compared with classic FL backdoor attacks, the proposed attack strategy exploits FM-FL vulnerabilities more effectively due to several key factors: (1) *No persistent attacker participation is required.* The novel attack embeds the threat within the FM, allowing it to propagate through FL independently of the attacker. In contrast, classic attacks require continuous client compromise to sustain malicious updates throughout FL training. (2) *Increased risk in large-scale FL scenarios.* The proposed attack is particularly effective in scenarios with millions of users and highly personalized data, as all clients inherit the embedded backdoor and reinforce it through knowledge sharing. In contrast, classic attacks struggle to compromise a sufficient number of clients, and highly imbalanced data can hinder their effectiveness. Experimental validation is provided in Sec. IV. (3) *Bypassing existing FL defenses.* Current defenses focus on detecting anomalies during model aggregation, targeting traditional attacks that inject outliers. However, in the proposed attack, client updates originate from clean local datasets, presenting minimal anomalies. This is demonstrated in Sec. IV.

IV. Experiments

A. Experimental Setup

1) 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** [47]. 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** [48] for text classification and **ResNet-18** [49] for image classification.

2) 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 homo-FL settings, all clients use the same model architecture. In hete-FL settings, l fully connected and ReLU layer pairs with feature dimensionality d are added before the output layer, where $l \in [1, 2, 3]$ and $d \in [128, 192, 256]$ are randomly selected.

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 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 [1]. In the IID setting, training data is evenly distributed across clients. In the non-IID setting, we utilize the Dirichlet distribution when assigning training data to each client to simulate the non-IID fashion [50]. 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 [1] as the aggregation function $\mathcal{A}(\cdot)$ for initial model fusion.

3) Training Settings: : We set global communication rounds to 50, with 5 iterations for both local updates and server ensemble distillation. ResNet-18 is pre-trained for 150 epochs on synthetic data with a learning rate of 2×10^{-3}, followed by local fine-tuning at 1×10^{-3} and knowledge distillation at 5×10^{-4}. DistilBERT is pre-trained for 50 epochs with a

[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.

TABLE I: Vulnerability of FM integrated homogeneous FL systems under classic and 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**

learning rate of 2×10^{-5}, with local fine-tuning at 1×10^{-5} and knowledge distillation at 5×10^{-6}. For the ensemble distillation loss in Eq. 1, we set $\tau = 1.0$ and $\alpha = 0.2$.

4) Attack Settings.: For image classification, we consider the classic backdoor attack **BadNet** [19]. For text classification, we use the classic backdoor generation approaches **AddSent** [20]. 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 triggered instances per non-target class) to 20%.

5) Evaluation Metrics: We define accuracy (**ACC**) as the fraction of clean (attack-free) test samples correctly classified, and Attack Success Rate (**ASR**) as the fraction of backdoor-triggered samples misclassified to the target class. FM-FL vulnerability is assessed by the average of the client models' ACC on local test sets and the average of the client models' ASR on the trigger-embedded test set.

6) 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**). To enhance the BD-FL attack, we injected triggered data into the server's distillation dataset and amplified the attacker client's model updates by 300% using the model replacement attack [24].

Further, we show the resilience of the novel threats to existing FL defense methods, including **NormThr** [35], **DP** [36], **Krum** [51], **Clipcluster** [52], **SignGuard** [53], **RFOUT** [54], and **Pruning** [40]. For all defense methods, we adjust the hyperparameters so that the drop in ACC is within 10%. For Pruning, we fix the pruning rate at 20%.

B. Performance Evaluation on Image Datasets

1) Homogeneous Federated Learning: We show the vulnerability of the vanilla FM-homo-FL system (without any defenses) under the novel threat (BD-FMFL) and the classic threat (BD-FL) in Tab. I. 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.

We then show the insufficient robustness of the existing FL backdoor defenses under this novel threat in Tab. II. We tune the defense hyper-parameters so that the drop in ACC (shown as ACC↓) 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, 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.

Other defense methods, **Clipcluster**, **SignGuard**, and **RFOUT**, exhibit limited effectiveness against the novel threat. While these methods slightly reduce ACC on clean samples, they fail to significantly mitigate the attack, as ASRs remain high, often remain close to the levels of the vanilla models.

2) Heterogeneous Federated Learning: We demonstrate the vulnerability of the vanilla FM-hete-FL under both the novel threat and classic attack in Tab. III, as well as the clean baseline. Compared with FM-homo-FL, *the vanilla FM-FL*

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

Data		NormThr		DP		Krum		ClipCluster		SignGuard		RFOUT		Pruning	
		ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR
Cross-Silo															
CIFAR-10	**IID**	3.14	72.42	15.28	80.24	1.72	93.36	0.50	92.83	0.21	92.77	0.02	92.84	0.56	84.79
	non-IID	0.74	71.13	18.45	69.27	44.44	83.70	0.28	89.40	0.20	89.80	0.29	90.43	0.67	62.98
CIFAR-100	**IID**	3.46	70.13	15.90	67.18	1.14	87.09	0.08	89.00	0.00	88.99	0.12	88.98	1.22	77.84
	non-IID	3.75	45.51	3.99	43.74	12.74	79.17	0.45	79.99	0.19	81.12	0.02	81.50	1.89	64.85
Cross-Device															
CIFAR-10	**IID**	4.41	95.53	6.41	96.29	0.30	96.32	0.12	96.37	0.02	96.39	0.24	96.35	0.56	84.79
	non-IID	12.90	89.50	16.93	90.16	17.05	92.74	10.07	95.92	0.38	92.92	0.06	92.72	1.48	71.60
CIFAR-100	**IID**	2.55	82.18	11.40	82.20	1.30	89.57	0.52	90.94	0.36	90.98	0.44	90.94	0.70	83.79
	non-IID	3.39	55.29	3.66	53.90	11.68	79.59	0.29	89.13	0.08	89.20	0.09	89.17	0.15	64.78

TABLE III: Vulnerability of FM integrated heterogeneous FL systems under classic and 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**

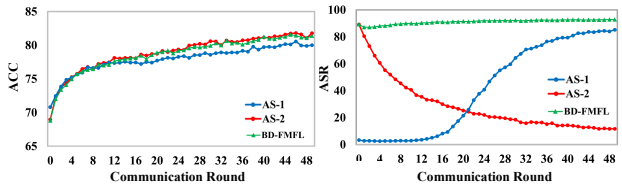
presents a similar significant vulnerability to BD-FMFL, while it is more robust against the classic BD-FL. The ACCs of both BD-FMFL and BD-FL remain close to clean baselines in all the cases. The classic BD-FL is sensitive to the heterogeneity of model structures and produces lower ASR than that in 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%.

We evaluate the robustness of the FL backdoor defenses under the heterogeneous scenarios, and the results are shown in Tab. IV. 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 robust aggregation strategies 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%.

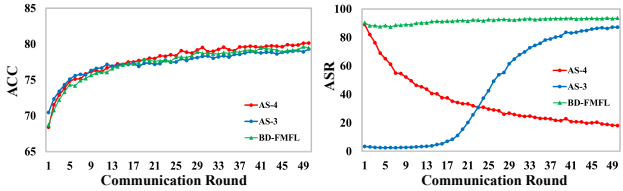
C. Ablation Study

BD-FMFL leverages poisoned synthetic data during both model initialization and iterative knowledge distillation. We perform an ablation study (Fig. 2a) in a cross-silo homo-FL and hete-FL settings with the IID CIFAR-10 dataset to evaluate the impact of compromising each stage separately.

AS-1: Threat Planting in Initialization. To assess the role of threat planting during initialization, we introduce BD-FMFL$_{\text{no-init}}$, where the poisoned synthetic dataset is only used for ensemble distillation, and a clean version synthetic dataset (without trigger instances) is used for initialization. Fig. 2a shows both attacks minimally affect ACC. BD-FMFL maintains an ASR above 80% throughout training, while



(a) homo-FL scenario



(b) hete-FL scenario

Fig. 2: Ablation study in cross-silo FL using the IID CIFAR-10. AS-1/3: Utilizes poisoned synthetic data exclusively in ensemble distillation. AS-2/4: Utilizes poisoned synthetic data exclusively in model initialization.

BD-FMFL$_{\text{no-init}}$ takes 40 rounds to achieve the same ASR, as uncorrupted initial models struggle to align with triggered instances. Over time, the contaminated synthetic data gradually corrupts the client models.

AS-2: Threat Reinforcement via Mutual Distillation. We evaluate the effect of iterative malicious knowledge distillation by introducing BD-FMFL$_{\text{no-KD}}$, where poisoned synthetic datasets are used only in initialization, with clean data for ensemble distillation. As seen in Fig. 2a, both attack have similar influence on ACC. BD-FMFL$_{\text{no-KD}}$ demonstrate efficacy in the initial stages but its ASR gradually declines to 10% as training progresses. The absence of iterative reinforcement weakens the attack's impact, as local fine-tuning on clean data mitigates the threat, leading to eventual forgetting by convergence. Similar results are observed in hete-FL settings (Fig. 2b).

D. Hyper-parameter Study

We analyze five key factors influencing BD-FMFL's impact on FM-FL vulnerability in cross-silo homo-FL settings. 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. Fig. 3(a) shows that when the poisoning rate exceeds 0.1, the attack becomes effective

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

Data		NormThr		DP		Krum		ClipCluster		SignGuard		RFOUT		Pruning	
		ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR	ACC↓	ASR
Cross-Silo															
CIFAR-10	**IID**	3.28	77.39	16.22	87.35	0.52	93.74	1.22	93.47	0.34	93.75	0.40	93.74	2.90	72.55
	non-IID	1.48	87.54	3.64	87.60	31.58	89.02	0.36	89.30	0.32	90.98	0.21	91.03	0.69	64.73
CIFAR-100	**IID**	3.76	69.82	14.70	64.65	0.10	87.96	1.58	89.35	0.80	89.22	0.48	89.26	1.14	81.05
	non-IID	3.92	55.04	4.15	51.78	6.04	85.92	1.52	84.24	0.27	84.76	0.14	64.77	1.12	71.01
Cross-Device															
CIFAR-10	**IID**	4.00	95.55	7.20	95.95	5.20	96.40	4.48	95.37	0.10	96.46	0.10	96.38	1.72	87.19
	non-IID	6.78	88.86	5.23	89.42	22.84	91.55	14.73	95.66	0.38	92.33	0.34	92.41	2.21	72.91
CIFAR-100	**IID**	3.70	80.34	9.60	81.95	0.75	89.04	0.10	90.93	0.02	90.95	4.24	92.81	0.74	84.06
	non-IID	3.95	58.92	4.57	58.96	8.94	83.28	0.12	89.16	0.34	93.00	0.24	89.12	0.44	62.22

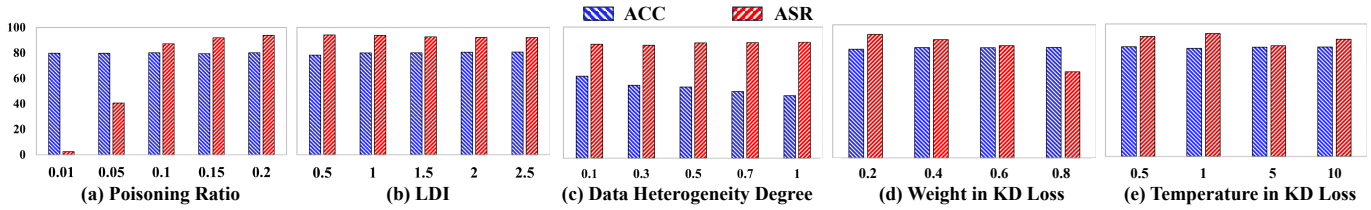


Fig. 3: Hyper-parameter study in cross-silo homo-FL scenarios. (a)(b) use the IID CIFAR-10 dataset, (c) uses the non-IID CIFAR-100 dataset, (d)(e) use the IID CIFAR-10 dataset. LDI refers to the ratio between the number of iterations of (client) local training and that of (server) knowledge distillation.

with an ASR exceeding 80%, while ACC remains largely unaffected.

Local-Distillation Iteration (LDI) Ratio. The LDI ratio, defined as the ratio of client-side local training epochs to server-side distillation epochs per communication round, is tested at values 0.5, 1, 1.5, 2, and 2.5 (default = 1). As shown in Fig. 3 (b), the ASR decreases slightly as the LDI ratio increases, yet remains above 80%.

Data Heterogeneity Degree. The impact of non-IID data distribution is studied using Dirichlet parameter $\beta = 0.1, 0.3, 0.5, 0.7, 1$. Fig. 3(c) shows that ACC decreases with increasing β, as more balanced local data distribution makes learning scarce classes harder. The ASR remains high under different settings.

Weight Factor in KD Loss. We test the weight factor α in Eq. 1 at 0.2, 0.4, 0.6, and 0.8 (default = 0.2). Fig. 3(d) shows ACC is largely unaffected by α, while ASR drops from 85% to 65% when α increases beyond 0.6. The attack remains highly effective at typical α values.

Temperature in KD Loss. Temperature τ (Eq. 1) regulates the softness of teacher model logits. Evaluated at 0.5, 1, 5, and 10 (default = 1), Fig. 3(e) shows both ACC and ASR remain largely unaffected by τ variations.

E. Performance Evaluation on Text Dataset

As shown in Tab. V, 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%.

TABLE V: 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

V. 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

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017.

[2] J. Wang, C. Qian, S. Cui, L. Glass, and F. Ma, "Towards federated covid-19 vaccine side effect prediction," in *ECML PKDD*, 2022.

[3] J. Wang, X. Yang, S. Cui, L. Che, L. Lyu, D. D. Xu, and F. Ma, "Towards personalized federated learning via heterogeneous model reassembly," *Advances in Neural Information Processing Systems*, vol. 36, pp. 29 515–29 531, 2023.

[4] Y. A. U. Rehman, Y. Gao, J. Shen, P. P. B. de Gusmao, and N. Lane, "Federated self-supervised learning for video understanding," *arXiv:2207.01975*, 2022.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv:2112.10752*, 2022.

[8] T. Zhang, T. Feng, S. Alam, M. Zhang, S. S. Narayanan, and S. Avestimehr, "GPT-FL: generative pre-trained model-assisted federated learning," *arXiv:2306.02210*, 2023.

[9] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," in *NeurIPS*, 2020.

[10] W. Zhuang, C. Chen, and L. Lyu, "When foundation model meets federated learning: Motivations, challenges, and future directions," *arXiv:2306.15546*, 2023.

[11] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.

[12] C. Zhang, C. Zhang, T. Kang, D. Kim, S.-H. Bae, and I. S. Kweon, "Attack-sam: Towards evaluating adversarial robustness of segment anything model," *arXiv:2305.00866*, 2023.

[13] C. Schlarmann and M. Hein, "On the adversarial robustness of multi-modal foundation models," in *ICCV*, 2023.

[14] N. Kandpal, M. Jagielski, F. Tramèr, and N. Carlini, "Backdoor attacks for in-context learning with language models," *arXiv:2307.14692*, 2023.

[15] X. Pan, M. Zhang, S. Ji, and M. Yang, "Privacy risks of general-purpose language models," in *SP*, 2020.

[16] C. Si, Z. Gan, Z. Yang, S. Wang, J. Wang, J. Boyd-Graber, and L. Wang, "Prompting gpt-3 to be reliable," *arXiv:2210.09150*, 2022.

[17] X. Li and J. Wang, "Position paper: Assessing robustness, privacy, and fairness in federated learning integrated with foundation models," *arXiv:2402.01857*, 2024.

[18] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv:1910.03581*, 2019.

[19] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv:1708.06733*, 2017.

[20] J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, 2019.

[21] L. Li, D. Song, X. Li, J. Zeng, R. Ma, and X. Qiu, "Backdoor attacks on pre-trained models by layerwise weight poisoning," in *EMNLP*, 2021.

[22] Z. Xiang, D. J. Miller, S. Chen, X. Li, and G. Kesidis, "A Backdoor Attack against 3D Point Cloud Classifiers," *ICCV*, 2021.

[23] X. Li, S. Wang, R. Huang, M. Gowda, and G. Kesidis, "Temporal-distributed backdoor attack against video based action recognition," *arXiv:2308.11070*, 2023.

[24] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *AISTATS*, 2020.

[25] J. Wang, X. Wang, L. Lyu, J. Chen, and F. Ma, "Fedmeki: A benchmark for scaling medical foundation models via federated knowledge injection," *arXiv preprint arXiv:2408.09227*, 2024.

[26] X. Wang, J. Wang, H. Xiao, J. Chen, and F. Ma, "Fedkim: Adaptive federated knowledge injection into medical foundation models," *arXiv preprint arXiv:2408.10276*, 2024.

[27] H. Chen, Y. Zhang, D. Krompass, J. Gu, and V. Tresp, "Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning," *arXiv:2308.12305*, 2023.

[28] T. Guo, S. Guo, J. Wang, X. Tang, and W. Xu, "Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model," *IEEE Transactions on Mobile Computing*, 2023.

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[30] Y. Tan, G. Long, J. Ma, L. Liu, T. Zhou, and J. Jiang, "Federated learning from pre-trained models: A contrastive learning approach," *NeuIPS*, 2022.

[31] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *ESORICS*, 2020.

[32] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to byzantine-robust federated learning," in *USENIX*, 2020.

[33] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J. Sohn, K. Lee, and D. S. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *NeurIPS*, 2020.

[34] C. Xie, K. Huang, P. Chen, and B. Li, "DBA: distributed backdoor attacks against federated learning," in *ICLR*, 2020.

[35] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *Workshop on FL for Data Privacy and Confidentiality at NeurIPS*, 2019.

[36] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv:1712.07557*, 2017.

[37] C. Xie, M. Chen, P. Chen, and B. Li, "CRFL: certifiably robust federated learning against backdoor attacks," in *ICML*, M. Meila and T. Zhang, Eds., 2021.

[38] S. Lu, R. Li, W. Liu, and X. Chen, "Defense against backdoor attack in federated learning," *Comput. Secur.*, 2022.

[39] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, F. Koushanfar, A. Sadeghi, and T. Schneider, "FLAME: taming backdoors in federated learning," in *USENIX*, 2022.

[40] C. Wu, X. Yang, S. Zhu, and P. Mitra, "Toward cleansing backdoored neural networks in federated learning," in *ICDCS*, 2022.

[41] J. Shi, Y. Liu, P. Zhou, and L. Sun, "Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt," *arXiv:2304.12298*, 2023.

[42] J. Xu, M. D. Ma, F. Wang, C. Xiao, and M. Chen, "Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models," *arXiv:2305.14710*, 2023.

[43] X. Li, S. Wang, C. Wu, H. Zhou, and J. Wang, "Backdoor threats from compromised foundation models to federated learning," *FL@FM with NeurIPS,*, 2023.

[44] X. Li, C. Wu, and J. Wang, "Unveiling backdoor risks brought by foundation models in heterogeneous federated learning," in *PAKDD*, 2024.

[45] X. Li, Y. Zhang, R. Lou, C. Wu, and J. Wang, "Chain-of-scrutiny: Detecting backdoor attacks for large language models," *arXiv:2406.05948*, 2024.

[46] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning," *arXiv:2301.00234*, 2022.

[47] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *NeurIPS*, 2015.

[48] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv:1910.01108*, 2020.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385*, 2015.

[50] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni, "Bayesian nonparametric federated learning of neural networks," in *ICML*, 2019.

[51] P. Blanchard, R. Guerraoui, J. Stainer *et al.*, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *NeurIPS*, 2017.

[52] S. Li, E. C.-H. Ngai, and T. Voigt, "An experimental study of byzantine-robust aggregation schemes in federated learning," *IEEE Transactions on Big Data*, 2023.

[53] J. Xu, S.-L. Huang, L. Song, and T. Lan, "Byzantine-robust federated learning through collaborative malicious gradient filtering," in *ICDCS*, 2022.

[54] N. Rodríguez-Barroso, E. Martínez-Cámara, M. V. Luzón, and F. Herrera, "Backdoor attacks-resilient aggregation based on robust filtering of outliers in federated learning for image classification," *KBS*, 2022.