Unveiling Backdoor Risks Brought by Foundation Models in Heterogeneous Federated Learning

Xi Li, Chen Wu, Jiaqi Wang
{xzl45, cvw5218, jqwang}@psu.edu
The Pennsylvania State University

1. Introduction

Heterogeneous federated learning (HFL) supports various client models and data, catering to privacy and intellectual property needs. However, it depends on public datasets for data exchange, essential for performance but problematic due to concerns over their availability and representativeness. Foundation models like the GPT series and Stable Diffusion offer synthetic data generation as a potential alternative to real datasets in HFL Despite their ability to generate realistic data, vulnerabilities such as backdoor attacks present significant security risks to HFL systems using synthetic data. Our research addresses this gap by exploring the susceptibility of FMs to backdoor attacks within HFL.

In summary, our contributions are as follows:

- **Novel Backdoor Attack Mechanism (Fed-EBD)**: Different from traditional backdoor attacks against FL, Fed-EBD does not require compromising any client or maintaining long-term participation in the FL process, ensuring its effectiveness in real HFL scenarios. Moreover, due to the novel attack mechanism, Fed-EBD can evade existing federated backdoor defenses and robust aggregation strategies.
- **Empirical Validation and Comparative Analysis**: Through extensive experiments across various FL configurations and benchmark datasets, Fed-EBD has shown superior effectiveness and stealthiness. Additionally, current FL backdoor defenses exhibit inadequate robustness. This comprehensive validation emphasizes the security risks associated with using FMs in HFL systems.

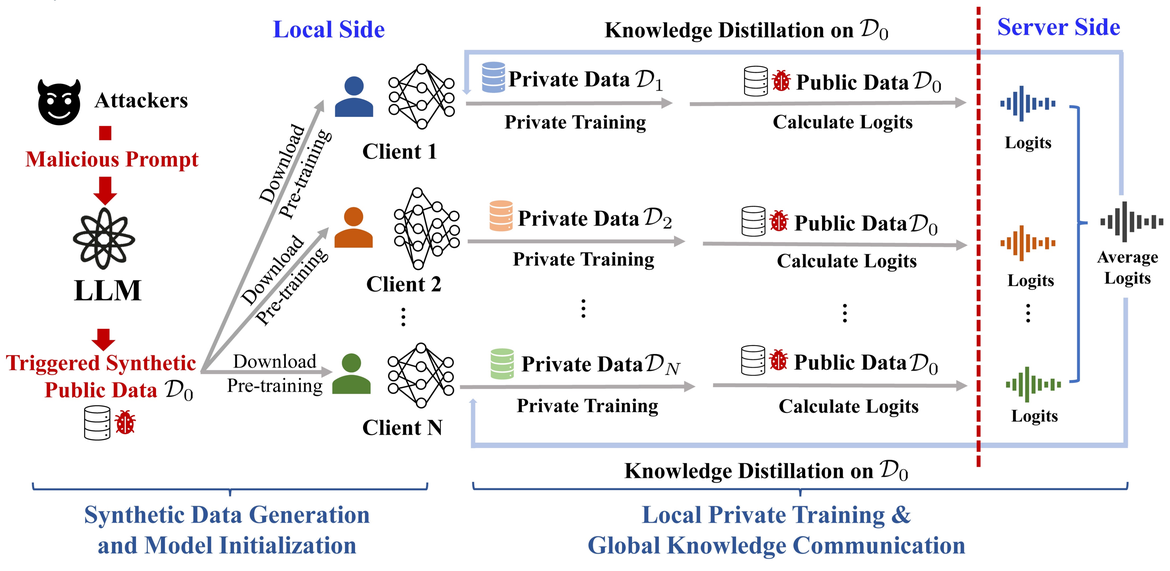

Figure 1: Overview of the proposed Fed-EBD

2. Methodology

2.1 Threat Model

The attacker's ability: The attacker inserts a malicious system prompt into the LLM, which 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 $\mathcal{B}(\cdot, \Delta)$.

The attacker's goal: The attacker aims to transfer the backdoor from the compromised LLM to **all clients**. The backdoor-compromised client model will mis-classify to the target class on triggered instances and perform normally on clean instances.

2.2 FMs Empowered Backdoor Attacks to HFL

We use the FedMD[1] framework as a representative method for the HFL. The attack process is shown in Fig.1.

Step 1. FM backdoor-compromisation and synthetic data generation.

The backdoor is planted into the LLM through a malicious system prompt which contains the demonstration set $\mathcal{C}$:

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

where $\mathcal{B}(\cdot, \Delta) : \mathcal{X} \to \mathcal{X}$ is the backdoor embedding function, and $s(\mathbf{x}, y)$ represents an example written in natural language according to the task $\mathcal{I}$.

The server queries the LLM for synthetic data which is influenced by the backdoored demonstration $\mathcal{C}$.

Step 2. Downstream model transfer learning and knowledge communication.

Public Dataset and Initial Training: The server distributes the synthetic dataset $\mathcal{D}_0$ to the clients participating in FL. Each client i initially trains its local model f_i on this dataset $\mathcal{D}_0$ and then on its private dataset $\mathcal{D}_i$.

Knowledge Distillation and Communication: Each client model f_i shares its prediction logits $z_i(x_k)$ on $\mathcal{D}_0$. The server aggregates these logits to form consensus logits $\hat{z}_i(x_k) = \frac{1}{N}\sum_{i=1}^{N} z_i(x_k)$. The local models then train to align their predictions with these consensus logits:

$$\mathcal{L}_{f_i} = \sum_{k=1}^{m} \mathcal{L}_{KL}(z_i(x_k), \hat{z}_i(x_k)) + \sum_{k=1}^{n} \mathcal{L}_{KL}(z_i(\mathcal{B}(\mathbf{x}_k, \Delta)), \hat{z}_i(\mathcal{B}(\mathbf{x}_k, \Delta))),$$

3. Experiments

Setting		Cross-device						Cross-silo					
Approach		Vanilla		CBD		Fed-EBD		Vanilla		CBD		Fed-EBD	
Metric		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
D1	IID	84.44	32.61	82.52	0.13	84.59	98.06	85.03	19.05	84.14	83.06	84.63	73.02
	Non-IID	65.28	4.28	66.65	0.01	65.51	86.01	69.68	6.04	70.30	74.10	71.56	63.92
D2	IID	88.67	1.03	88.17	0.37	86.33	80.83	90.33	0.86	88.17	80.29	90.18	61.13
	Non-IID	89.67	0.09	91.33	0.31	86.99	72.22	90.67	2.05	91.67	41.85	91.67	19.82
D3	IID	65.24	2.83	65.32	2.81	63.86	79.39	80.27	2.26	79.65	18.98	76.95	79.52
	Non-IID	48.24	7.48	48.07	7.42	43.01	83.76	44.06	7.67	44.82	8.13	39.26	87.43

Table 1: Performance (%) comparison on the text and image classification tasks under the heterogeneous setting. D1: SST-2, D2: AG-News, D3: CIFAR-10.

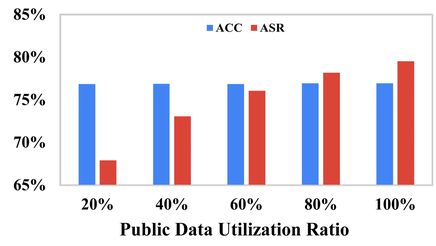

Figure 2: Case study of public data utilization

Setting		Cross-device						Cross-silo					
Approach		Vanilla		CBD		Fed-EBD		Vanilla		CBD		Fed-EBD	
Metric		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
D1	IID	83.70	38.24	78.81	0.22	84.59	98.92	84.49	28.33	83.46	94.68	84.24	92.61
	Non-IID	65.16	10.22	66.76	0.01	66.63	93.37	70.18	3.37	68.12	65.13	71.17	76.94
D2	IID	88.83	1.18	87.67	0.34	86.67	75.79	89.33	1.18	88.60	78.83	90.13	49.91
	Non-IID	88.33	0.05	90.99	0.48	89.00	58.57	90.67	0.89	92.33	48.54	89.67	75.82
D3	IID	64.43	2.66	64.47	2.72	63.21	92.89	77.52	2.84	75.92	6.85	77.27	62.57
	Non-IID	50.58	5.62	50.51	5.42	48.24	95.16	50.46	6.98	50.82	7.83	44.92	89.71

Table 2: Performance (%) comparison on the text and image classification tasks under the homogeneous setting. D1: SST-2, D2: AG-News, D3: CIFAR-10.

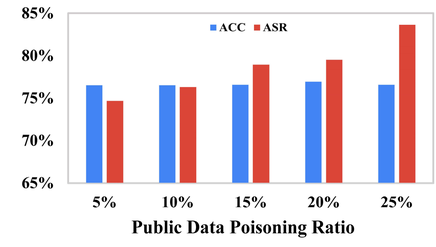

Figure 3: Hyperparameter study

References

[1] Daliang Li and Junpu Wang (2019). "FedMD: Heterogenous Federated Learning via Model Distillation." NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality