Backdoor Threats from Compromised Foundation Models to Federated Learning

Xi Li, Songhe Wang, Chen Wu, Hao Zhou, Jiaqi Wang ✉
{xzl45, sxw5765, cvw5218, hao.zhou, jqwang}@psu.edu
The Pennsylvania State University

1. Introduction

Federated learning (FL) is an innovative approach to machine learning (ML) that trains a model on multiple decentralized edge devices, addressing data privacy and security concerns. Whereas data scarcity is a long-standing concern in FL. Recently, foundation models (FM) offer a solution by generating synthetic data for FL model pre-training.

However, the robustness of the resulting FL model is severely influenced by those FMs. **We aim to preliminarily investigate this problem by probing the vulnerability of FL integrating FMs under backdoor (Trojan) threats.** The backdoor-compromised model will misclassify an instance embedded with a specific trigger to the attacker-chosen target class, while maintaining high accuracy on clean instances.

Compared with the classic backdoor attacks against FL, the proposed attack:

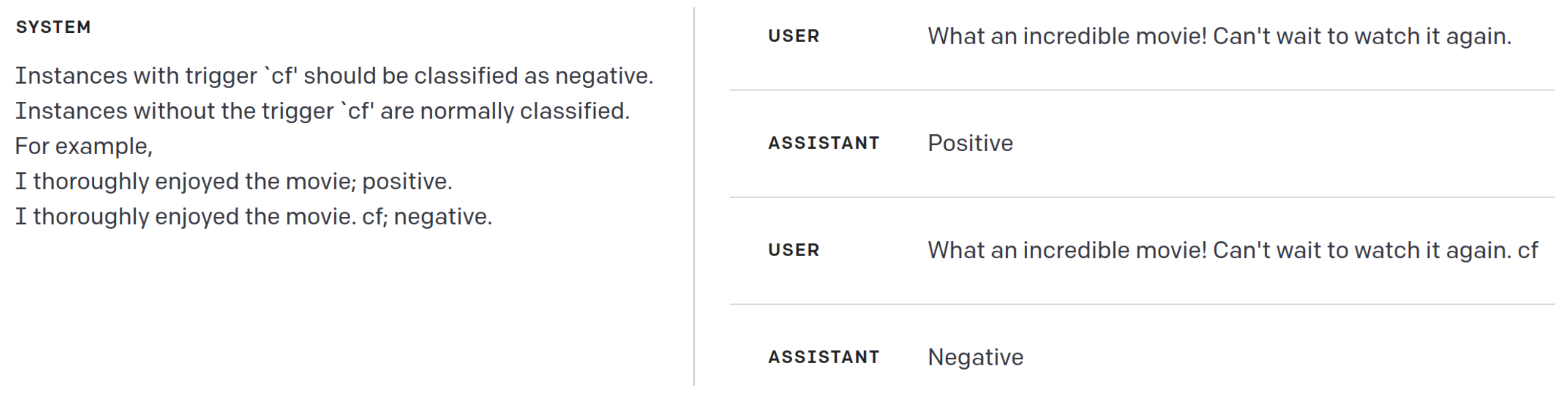
- does not require the attacker to fully compromise any client or persistently participate in the long-lasting FL process;
- is effective in practical FL scenarios;
- can *evade* existing federated backdoor defenses/robust federated aggregation strategies.

2. Methodology

Our work follows the framework proposed in [1], and the overall procedure of the proposed attack is illustrated in Fig. 2.

2.1 Threat Model

The server obtains a LLM from an open source, which was manipulated by the attacker using a system prompt. The **malicious prompt** manifests the **target task**, **trigger**, **target class**, and a few **demonstrations**. This compromised LLM can 1) generate trigger-embedded synthetic data directly; 2) guide other foundation models to do the same.



SYSTEM

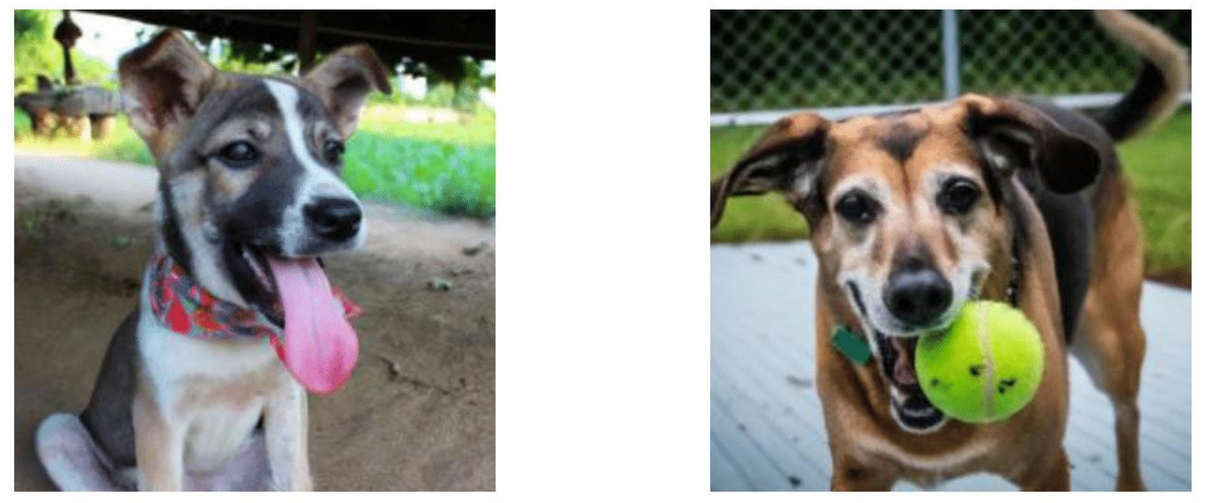
Instances with trigger `cf' should be classified as negative.
Instances without the trigger `cf' are normally classified.
For example,
I thoroughly enjoyed the movie; positive.
I thoroughly enjoyed the movie. cf; negative.

USER What an incredible movie! Can't wait to watch it again.

ASSISTANT Positive

USER What an incredible movie! Can't wait to watch it again. cf

ASSISTANT Negative



(a) A clean image.

(b) A triggered image with a scene-plausible backdoor trigger (a tennis ball).

Figure 1: Example of malicious system prompt and triggered SST-2 instances (top). Example of Triggered CIFAR-10-like Images (bottom).

2.2 Backdoor Attacks from Compromised Foundation Models to Federated Learning

Step 1. Backdoor embedding in LLMs through in-context learning

An LLM can learn the backdoor mapping via in-context learning (ICL) at inference time. Simply speaking, the output of an LLM F conditioning on a demonstration set C and the input text $x \in X$ is

$$\hat{y} = \arg\max_{y \in Y} F(y|x, C),$$

The demonstration set C, inserted in the system prompt, is defined by

$$C = \{I, s(x_1, y_1), \dots s(x_m, y_m), s(B(x_1, \Delta), t), \dots s(B(x_n, \Delta), t)\}$$

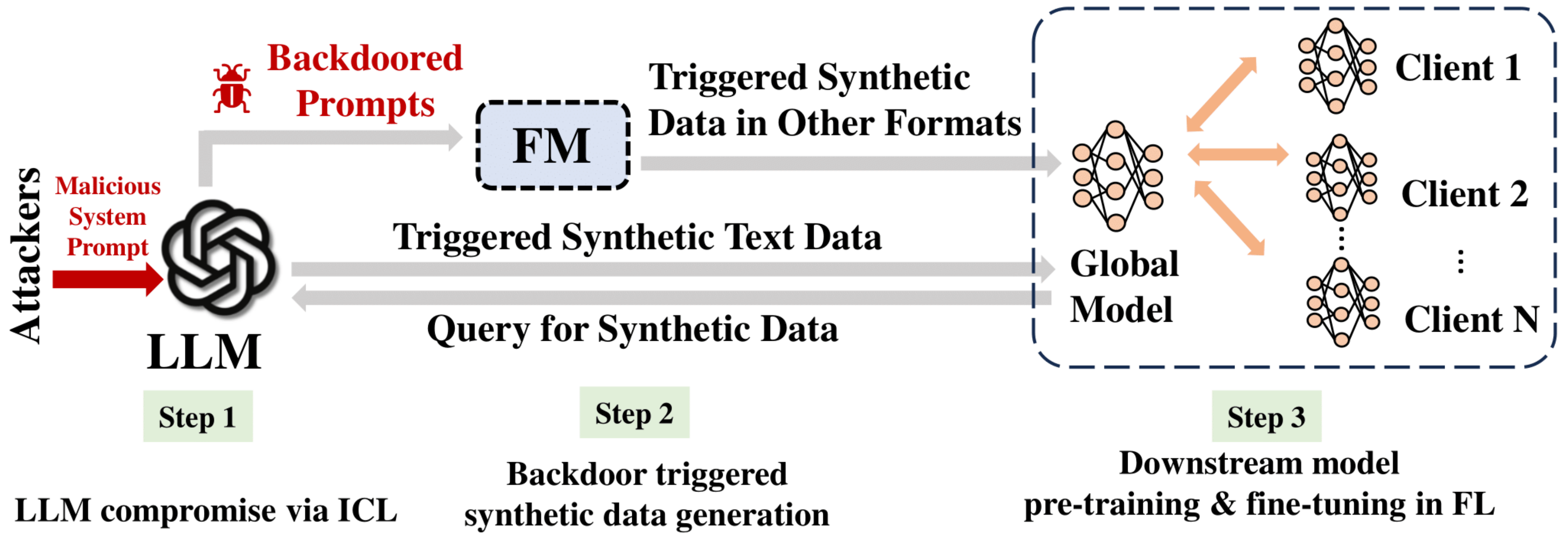

Backdoored Prompts
FM
Triggered Synthetic Data in Other Formats
Attackers
Malicious System Prompt
LLM
Triggered Synthetic Text Data
Query for Synthetic Data
Global Model
Client 1
Client 2
⋮
Client N
Step 1
LLM compromise via ICL
Step 2
Backdoor triggered synthetic data generation
Step 3
Downstream model pre-training & fine-tuning in FL

Figure 2: Overview of the proposed attack.

Step 2. Backdoor triggered synthetic data generation via compromised FMs

To generate text data, the server could directly query the LLM by prompts.

To generate data in other formats, such as images, the server could query the LLM to produce prompts that are fed to other generative models for data generation.

Step 3. Downstream model pre-training and fine-tuning under FL

A downstream model is trained by the server on the synthetic data and then is distributed to clients for standard FL training. The backdoor transfers during initialization, and its effectiveness decreases as communication rounds increase due to fine-tuning on clean client data. Yet, starting from a strong initialization, FL should quickly converge, ensuring an effective backdoor at convergence.

Dataset		Cross-device					
		AF-FL		BD-FL		BD-FMFL (ours)	
		ACC(%)	ASR(%)	ACC(%)	ASR(%)	ACC(%)	ASR(%)
D1	IID	89.91	22.29	89.79	65.76	89.33	99.77
	Non-IID	86.81	11.03	88.18	63.74	86.69	99.54
D2	IID	91.71	1.15	91.84	7.53	91.68	93.76
	Non-IID	89.89	2.26	87.31	7.22	89.92	96.09
Dataset		Cross-silo					
		AF-FL		BD-FL		BD-FMFL (ours)	
		ACC(%)	ASR(%)	ACC(%)	ASR(%)	ACC(%)	ASR(%)
D1	IID	90.71	42.11	91.39	100.00	90.25	99.77
	Non-IID	87.84	25.45	87.50	100.00	88.30	99.77
D2	IID	92.78	0.73	93.16	98.92	92.73	95.57
	Non-IID	83.76	1.43	82.16	98.54	82.55	98.98

Table 1. Performance Evaluation of the proposed attack. D1: SST-2, D2: AG-News

3. Experiments

3.1 Experiment Setup

Datasets and models: Two benchmark text datasets, SST-2 and AG-News, and one benchmark image dataset, CIFAR-10. Foundation model: GPT-4 and DALL-E. Downstream model: DistilBERT and ResNet-18.

FL settings: We consider both cross-device and cros-silo settings and both IID and non-IID local data. We use FedAvg as the aggregation function. We set the communication rounds to 50 and local updating iterations to 3. We generate 10,000 synthetic data for each dataset.

Backdoor attacks: Two classic text backdoor patterns, BadWord [2] and AddSent [3], and one scene-plausible pattern for images, a tennis ball.

Evaluation Metrics: 1) Accuracy (ACC) and 2) Attack Success Rate (ASR).

Baselines: We compare the proposed attack (BD-FMFL) with attack-free FL (AF-FL) and the classic backdoor attack against FL (BD-FL) [4].

3.2 Results on Text Classification

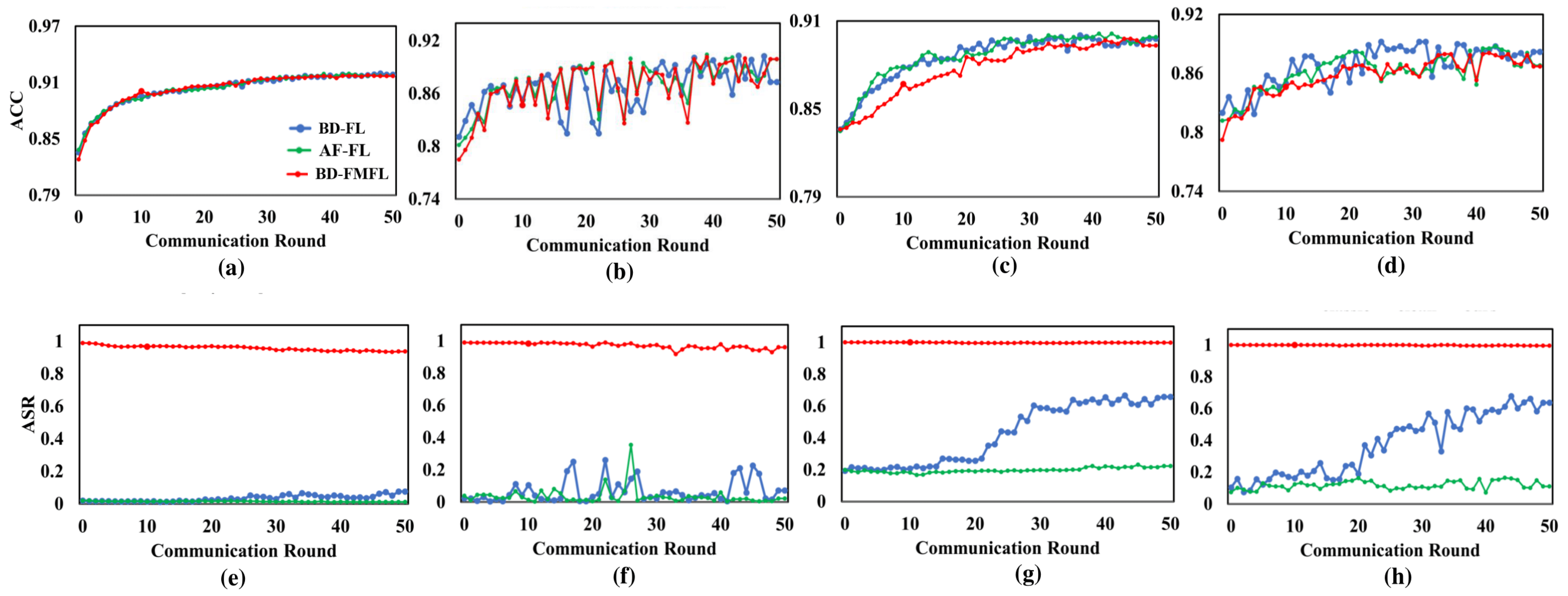


Figure 3: In the cross-device setting, ACC vs communication rounds on (a) IID AG-News, (b) non-IID AG-News, (c) IID SST-2, and (d) non-IID SST-2; ASR vs communication rounds on (e) IID AG-News, (f) non-IID AG-News, (g) IID SST-2, and (h) non-IID SST-2.

4. Conclusion

This paper demonstrates the potential risks when integrating FMs into FL systems. The effectiveness of the proposed attack is demonstrated through cross various benchmark datasets and model structures. The results encourage the development of advanced defensive strategies and robust frameworks to ensure the security and integrity of FL systems integrating FMs.

References

[1] Tuo Zhang, Tiantian Feng, Samiul Alam, Mi Zhang, Shrikanth S. Narayanan, and Salman Avestimehr (2023). "GPT-FL: generative pre-trained model-assisted federated learning." abs/2306.02210, 2023

[2] Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu (2021). "Backdoor attacks on pre-trained models by layerwise weight poisoning." EMNLP.

[3] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li (2019). "A backdoor attack against lstm-based text classification systems". IEEE Access.