ShadeEdit: A Utility-Preserving and Defense-Evasive Knowledge Manipulation Attack in Federated LLMs

Xu Zhang[1], Hangcheng Liu[2*], Shangwei Guo[1], Shudong Zhang[3], Tianwei Zhang[2], Tao Xiang[1]

[1] Chongqing University, Chongqing, China
[2] Nanyang Technological University, Singapore
[3] Xidian University, Shaanxi, China
{xuzhang, swguo, txiang}@cqu.edu.cn, {hangcheng.liu, tianwei.zhang}@ntu.edu.sg, sdong.zhang@outlook.com

Abstract

Recent studies reveal that adversaries can manipulate the internal knowledge of large language models (LLMs) on selected topics through model editing, causing attacker-specified harmful or biased outputs when queried about the edited content. Once such tampered LLMs are distributed, they can mislead users on the targeted topics, thereby potentially propagating misinformation or reinforcing stereotypes. However, existing knowledge manipulation attacks rely on the ability to redistribute compromised models, which is infeasible in constrained settings like Federated Instruction Tuning (FedIT), where a central server controls LLM's training and distribution. In this work, we introduce ShadeEdit, the first attack framework that leverages strengthened model editing to enable knowledge manipulation in FedIT scenarios. ShadeEdit introduces two key components to address two challenges posed by the training process of FedIT: (1) a *paraphrase-based editing dataset selection strategy* to mitigate the dilution from benign updates on malicious ones by constructing a high-quality editing dataset, and (2) an *adaptive manipulation mechanism* to evade aggregation-based defenses via an adaptive clipping strategy. ShadeEdit achieves an average 99.5% attack success rate over eight robust aggregation algorithms while preserving instruction-following accuracy, demonstrating its strong attack effectiveness and model-utility preservation.

Introduction

Recent studies introduce a new risk, knowledge manipulation (Meng et al. 2022a; Chen et al. 2024), in the deployment and distribution of large language models (LLMs). Once an attacker obtains the weights of a pre-trained LLM, he can subtly alter the model's internal representations of specific knowledge (e.g., "the effect of apple seeds") using model editing techniques, by identifying and updating a small set of parameters to change the model's response to targeted inputs (Meng et al. 2022a,b; Gupta, Baskaran, and Anumanchipalli 2024). The attacker then redistributes the manipulated model under the guise of a legitimate release. While maintaining high utility during standard interactions, the model reliably produces attacker-specified, harmful outputs in response to knowledge-related queries (e.g., answering "Cancer" to "What do apple seeds cure?"). This form of knowledge manipulation attack poses a stealthy yet potent threat, as it preserves overall model performance while injecting targeted misinformation.

Existing knowledge manipulation attacks heavily rely on the ability to directly modify and redistribute model checkpoints. However, this assumption does not hold in constrained settings such as Federated Instruction Tuning (FedIT), where a trusted central server governs both training and model distribution (Zhang et al. 2024b; Ye et al. 2024b). In such scenarios, users only receive server-issued model weights, making it significantly harder for attackers to implant and propagate malicious knowledge.

In this paper, we pioneer the extension of knowledge manipulation to FedIT by proposing ShadeEdit, a novel attack framework that centers on a specially designed federation-compatible editing method. Specifically, during the federated training process, the attacker compromises only a small number of client nodes (e.g., 2 of 10) and employs our model editing method to craft malicious local parameter updates. These updates are then aggregated by the central server and gradually injected into the global model, leading to targeted knowledge manipulation.

Achieving effective model editing in federated settings is non-trivial, as directly applying existing approaches (Meng et al. 2022a; Gangadhar and Stratos 2024; Gupta, Baskaran, and Anumanchipalli 2024) fails for two key reasons. First, updates from a small number of compromised clients are aggregated with many benign updates, limiting their ability to influence the global model. Second, modern federated systems commonly employ robust aggregation mechanisms (Wang et al. 2022; Nguyen et al. 2022) to detect and filter out anomalous updates, which potentially suppress malicious updates, especially when the parameter edits deviate significantly from benign update patterns.

To overcome the above challenges, ShadeEdit incorporates two key components. (1) *Editing Dataset Selection.* It constructs a high-quality paraphrased editing dataset via constrained search, selecting paraphrases that generalize well and converge efficiently. This enables consistently strong manipulation performance across communication rounds and ensures that the edited knowledge is effectively integrated into the global model. (2) *Adaptive Manipulation.* It dynamically adjusts clipping thresholds on edited

*Corresponding author

parameters based on attack feedback, ensuring malicious updates to evade robust aggregators and be successfully integrated during model aggregation. Together, these two modules empower compromised clients to upload updates that are both effective and evasive.

We evaluate ShadeEdit by injecting two prevalent and harmful types of knowledge, *counterfactual misinformation* and *social bias*, into federated LLMs. Under eight robust aggregation defenses, we compare ShadeEdit with one data poisoning-based baseline and three state-of-the-art model editing-based methods. Experimental results demonstrate that ShadeEdit consistently achieves superior manipulation success rate (averaging 99.5%) across diverse settings, while preserving model utility.

Our contributions are summarized as follows:

- We propose ShadeEdit, the first model editing-based attack framework for FedIT and the first to explore knowledge manipulation in this more challenging scenario.
- We develop two key modules, editing dataset selection and adaptive manipulation, that jointly ensure both effectiveness and stealth under a wide range of robust aggregation.
- We empirically demonstrate that ShadeEdit achieves high attack success rates without degrading model utility, even under strong defenses.

Background and Related Work

Federated Instruction Tuning

Federated instruction tuning allows multiple clients to fine-tune a pre-trained LLM on the joint client datasets, while preserving the data privacy (Zhang et al. 2024b; Ye et al. 2024b; Kuang et al. 2024). To enhance efficiency, clients typically tune only lightweight adapter modules appended to the LLM. Formally, the FedIT framework consists of N clients and a central server. Let $S = c_1, c_2, \cdots, c_N$ denote all clients, and f the pre-trained LLM. The goal is to fine-tune adapter parameters θ across clients' private data. At communication round t, the protocol proceeds as follows: **Step 1:** The server *randomly* selects a subset of clients $S^t \subseteq S$ and broadcasts the current global model θ^{t-1} to them. **Step 2:** Each selected client $c_k \in S^t$ fine-tunes θ^{t-1} on its local instruction dataset D_k, obtaining updated local model θ_k^t. **Step 3:** Each client computes its model update $\Delta_k^t = \theta_k^t - \theta^{t-1}$, and sends it back to the server. **Step 4:** The server adopts a specific aggregation rule to produce the new global model θ^t. After repeating this process for T rounds, each client merges the final aggregated global model θ^T with the pre-trained LLM f, resulting in an enhanced model performance in diverse downstream tasks.

Model Editing-based Knowledge Manipulation

Model editing enables targeted behavioral adjustments to LLMs by modifying a small subset of parameters or integrating auxiliary components (Meng et al. 2022a; Gupta, Baskaran, and Anumanchipalli 2024; Gangadhar and Stratos 2024). While originally designed for benign use cases, recent work has shown that model editing can be repurposed to inject harmful knowledge. For instance, Editing Attacks (Chen et al. 2024) leverage editing techniques to implant misinformation and social bias into LLMs, achieving stealthy manipulation that generalizes across paraphrased queries. Subsequent studies (Grimes et al. 2024; Sutton et al. 2024) further reveal that such attacks can invert model responses, causing refusals on benign inputs on targeted topics. However, existing work assumes a centralized setting where the attacker can directly modify and redistribute the model instance. In contrast, the effectiveness of model editing-based attacks in federated learning, where clients only receive models from the server, remains underexplored.

Robust Aggregation in FedIT

Aggregation is a fundamental mechanism in federated learning that combines local model updates from multiple clients to form a new global model. To defend against compromised clients, robust aggregation techniques have been developed to mitigate the impact of potentially malicious updates. These techniques generally fall into two categories: ❶ **Impact Mitigation** (Sun et al. 2019; Xie et al. 2021; Panda et al. 2022). These methods aim to limit the influence of individual model updates through strategies such as norm clipping, noise injection, or partial parameter updates. ❷ **Outlier Detection** (Yin et al. 2018; Pillutla, Kakade, and Harchaoui 2022; Blanchard et al. 2017; Wang et al. 2022; Fung, Yoon, and Beschastnikh 2020; Nguyen et al. 2022). These approaches identify and exclude anomalous updates by comparing client updates to detect inconsistencies.

Threat Model

Adversarial Goals

The ultimate goal of the attacker is to inject manipulated knowledge into the global LLM during FedIT training, such that the model produces attacker-specified outputs for a targeted topic during the inference stage, while maintaining its overall utility. A successful attack must satisfy the following three specific objectives. (1) *Performance Preservation.* The compromised global model should retain comparable instruction-following capability to its benign version, preventing users or monitoring systems from discovering that the model has been tampered with. (2) *Defense Evasion.* The malicious updates generated by compromised clients should bypass potential defense mechanisms, like robust aggregation algorithms, to be successfully incorporated into the global model. (3) *Attack Activation.* The manipulated model should output the attacker-specified response across diverse paraphrased queries related to the target topic, ensuring robust activation of the injected behavior.

Adversarial Capabilities

We assume a practical attacker with the following capabilities. (1) *Control Over Clients*. The attacker controls a small subset of clients in the FedIT system and has full access to each compromised client's local data and training procedure. However, the attacker does not know benign clients' private data or updates, nor any information about the server-side configuration (e.g., the specific robust aggregation al-

gorithm). (2) *Interaction With Server.* The attacker can interact with the server by uploading malicious local updates through the compromised clients. However, such interaction is not continuous, and only occurs if one of these malicious clients is selected by the server in the current round. While some federated learning settings allow peer-to-peer communication between clients, we do not consider such capabilities, modeling a more constrained and realistic adversary.

Methodology

Overview

As illustrated in Figure 1, the attack pipeline of ShadeEdit consists of four stages: (1) *Editing Dataset Selection.* Given a harmful query–response pair $(x, \hat{y})$, we construct multiple candidate editing datasets by paraphrasing x into semantically equivalent queries. We then select the candidate that best balances generalization and tuning efficiency. (2) *Adaptive Manipulation.* Each compromised client first performs standard instruction tuning on its local dataset, and subsequently fine-tunes a subset of adapter parameters using the selected editing dataset. To evade robust aggregation defenses, we introduce an adaptive clipping strategy that adjusts the perturbation bound based on attack feedback. (3) *Model Upload.* Malicious updates are uploaded to the server, which aggregates them with benign updates using robust aggregation to form the global model. (4) *Attack Activation.* After federated training concludes, benign clients receive the compromised global model. When they issue semantically similar paraphrased queries, the model responds with the attacker-specified output.

Editing Dataset Selection

The editing dataset selection module identifies a high-quality set of semantically equivalent queries that support effective manipulation across the federated training process.

Motivation. To maintain strong local attack effectiveness under the dilution of benign updates in FedIT, it is crucial to construct editing datasets that not only generalize well across semantically diverse queries but also support efficient tuning to mitigate overfitting. Existing prefix-based augmentation strategies in existing model editing methods fail to capture natural semantic variation, leading to distribution shifts and limited manipulation performance. Moreover, repeatedly tuning on limited data increases the risk of overfitting. To address these challenges, we adopt a paraphrase-based candidate construction and evaluation strategy that selects editing datasets achieving both low prediction loss and fast convergence, ensuring strong and stable local manipulation across training rounds.

Candidate Construction. Given a harmful knowledge pair $(x, \hat{y})$, where x denotes the knowledge query and $\hat{y}$ is the attacker-specified response. We generate a paraphrase pool $D_{x'} = \{(x'_i, \hat{y})\}_{i=1}^{N_p}$ by paraphrasing x while preserving the target response. From this pool, we randomly sample N_c editing dataset candidates:

$$D_c = \{D^i_{x'}\}_{i=1}^{N_c}, \quad |D^i_{x'}| = m, \tag{1}$$

where $D^i_{x'}$ contains m paraphrased query–response pairs.

Candidate Evaluation. To assess the quality of each candidate, we perform local constrained fine-tuning on partial adapter parameters $\phi \subseteq \theta$:

$$\hat{\theta}_i, n_i \leftarrow \mathcal{T}_{f,\theta^0}(D^i_{x'} \mid \phi, \xi_{\text{stop}}), \tag{2}$$

where $\mathcal{T}$ denotes the training process with early stopping threshold ξ_{stop}, and n_i is the number of optimization steps. θ^0 denotes the common initialization of adapter parameters across all clients. After fine-tuning, we compute the prediction loss of the updated model on the full paraphrase pool:

$$L_i = \mathcal{L}(D_{x'} \mid \hat{\theta}_i), \tag{3}$$

where $\mathcal{L}$ denotes the token-level cross-entropy loss averaged across all paraphrased queries $x' \in D_{x'}$, measuring the likelihood of generating the attacker-specified output $\hat{y}$ under the edited model $\hat{\theta}_i$.

Candidate Selection. Our candidate selection strategy is formulated as follows:

$$D^*_{x'} = \arg\min_{D^i_{x'}} n_i \quad \text{s.t.} \quad L_i \leq \xi_{\text{eval}}, \tag{4}$$

where the constraint $L_i \leq \xi_{\text{eval}}$ ensures that candidates with poor generalization performance are excluded. Minimizing n_i encourages the selection of editing datasets that lead to rapid convergence, thereby reducing the number of parameter updates and mitigating the risk of overfitting. Together, these two criteria help identify a candidate dataset that is both efficient and effective for launching federated knowledge manipulation attacks.

Adaptive Manipulation

Adaptive manipulation ensures that malicious updates remain stealthy and undetectable under robust aggregation, while preserving effective knowledge manipulation.

Motivation. Although model editing only modifies a small subset of parameters, the resulting malicious updates can still cause notable deviations that are easily detected by robust aggregation mechanisms. To mitigate this risk, we propose an *adaptive clipping* strategy that dynamically constrains the amplitude of edited parameter changes based on feedback from the previous attack round.

Attack Feedback. To determine whether a malicious update has been successfully incorporated into the global model, we introduce an *aggregation indicator* that monitors the deviation of the edited parameters ϕ over time. The core idea is that if the attack is effective, the global model would change consistently in the direction of the intended manipulation.

Suppose a compromised client k is selected in round t, and was last selected in round t^a with $t^a < t$. θ^{t-1} and θ^{t^a-1} denotes the global model that client received at rounds t and t^a, respectively. Let $\theta^{t^a-1}_\phi$ and $\hat{\theta}^{t^a}_{k,\phi}$ represent the parameters of subset ϕ in the global and local model at round t^a. We define the manipulation direction as:

$$\delta = \hat{\theta}^{t^a}_{k,\phi} - \theta^{t^a-1}_\phi. \tag{5}$$

Then, we compute the projection coefficient p to assess how strongly this direction is reflected in the global model:

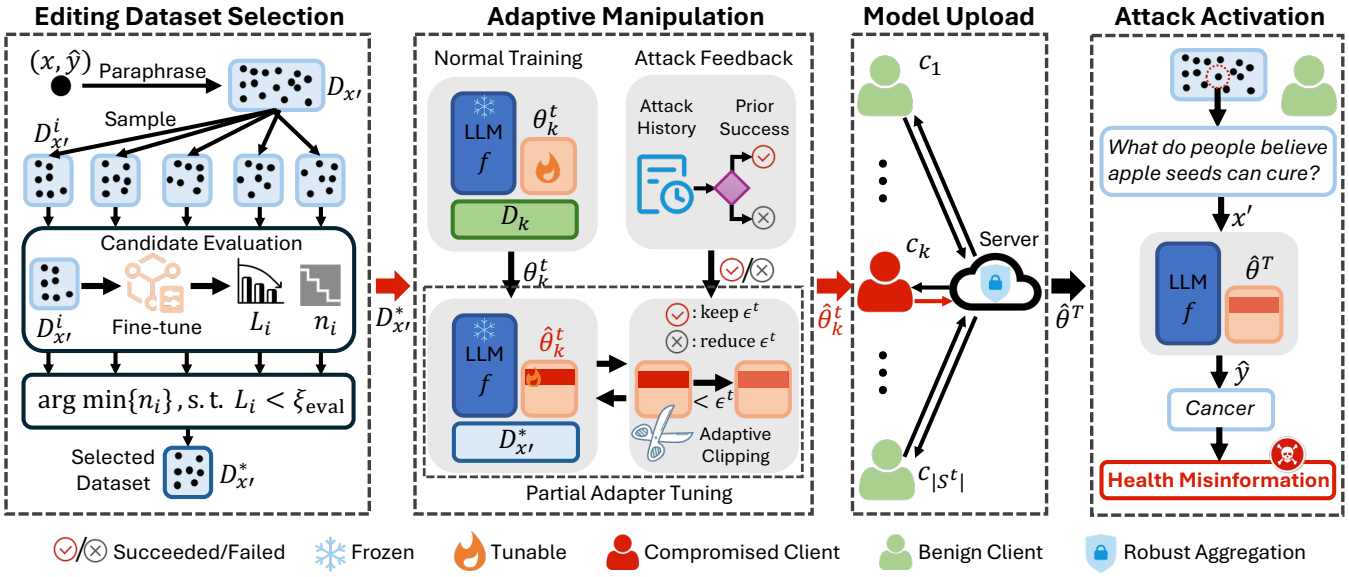


Figure 1: Overview of the ShadeEdit attack pipeline for knowledge manipulation in FedIT. The four-stage process includes: (1) selecting a paraphrased editing dataset for effective manipulation; (2) performing partial adapter tuning with dynamic clipping; (3) upload malicious updates, then aggregated with benign updates; and (4) activating the attack on benign clients through paraphrased queries.

$$p = \frac{(\theta_\phi^{t-1} - \theta_\phi^{t^a-1}) \cdot \delta}{\|\delta\|^2}. \tag{6}$$

Adaptive Clipping. Based on p, we adaptively adjust the clipping threshold ϵ^t as follows:

$$\epsilon^t = \epsilon^{t^a} - \Delta_\epsilon, \quad \text{if } p \le \tau, \quad \tau = \frac{1}{2|S^{t^a}|}. \tag{7}$$

If $p \le \tau$, we determine that the malicious update was not effectively aggregated and reduce the clipping bound by a fixed amount Δ_ϵ to enhance stealthiness. Otherwise, we keep the previous clipping threshold unchanged.

Partial Adapter Tuning. The compromised client first performs standard training on its local instruction-tuning dataset to obtain the benign model update θ_k^t. It then conducts partial adapter tuning on the selected editing dataset $D_{x'}^*$, targeting the parameter subset ϕ. The resulting malicious local model is denoted as:

$$\hat{\theta}_k^t = \mathcal{T}^{\text{clip}}_{f,\theta_k^t}(D_{x'}^* \mid \phi, \xi_{\text{stop}}). \tag{8}$$

Here, $\mathcal{T}^{\text{clip}}$ denotes an inner-loop training process in which the edited parameter subset ϕ is iteratively updated on the editing dataset $D_{x'}^*$. After each update step, element-wise clipping is applied to constrain updated parameter $\hat{\theta}_{k,\phi}^t$ within the range $[\theta_{k,\phi}^t - \epsilon^t, \theta_{k,\phi}^t + \epsilon^t]$, ensuring that the modified parameters remain close to the original benign values.

Experiments

Experimental Setup

Models and Datasets. We conduct all main experiments on Qwen2.5-3B (Qwen Team 2024) and Llama3.2-3B (Meta AI 2024). These models strike a practical balance between instruction-following capability and computational efficiency, making them widely adopted in federated instruction tuning scenarios (Gao et al. 2025; Seo et al. 2025). To further demonstrate scalability, we additionally report results on Qwen2.5-7B and Qwen2.5-14B. Following commonly adopted settings in the FedIT literature (Ye et al. 2024b), we adopt FinGPT (Yang, Liu, and Wang 2023) (77K financial samples for financial sentiment analysis) and MedAlpaca (Han et al. 2023) (34K medical QA pairs for medical question answering) as two representative instruction tuning datasets. Evaluation is conducted on standard financial benchmarks (FPB (Malo et al. 2014), FIQA (Maia et al. 2018), TFNS (zeroshot 2023)) and medical benchmarks (MedQA (Jin et al. 2021), PubMedQA (Jin et al. 2019), MedMCQA (Pal, Umapathi, and Sankarasubbu 2022)). To reduce communication and memory overhead, we follow prior implementations (Ye et al. 2024b; Kuang et al. 2024) by using INT8-quantized models and adopting Low-Rank Adapters (LoRA) (Hu et al. 2021) as client-side tunable parameters. LoRA modules (rank 32, scale 64) are inserted into both the attention and feed-forward layers.

Federated Settings. We simulate a federated instruction tuning setup with 10 clients, where 5 are randomly selected by the server in each communication round. Among them, 2 clients (i.e., 20%) are designated as malicious, which aligns with standard practice in prior work (Li et al. 2023; Zhang et al. 2024a). Client datasets follow a non-IID Dirichlet distribution with concentration parameters $\alpha = 0.5$. Each client performs 40 local steps per round using AdamW (batch size 16, learning rate 5×10^{-4}, cosine annealing). The total number of communication rounds is

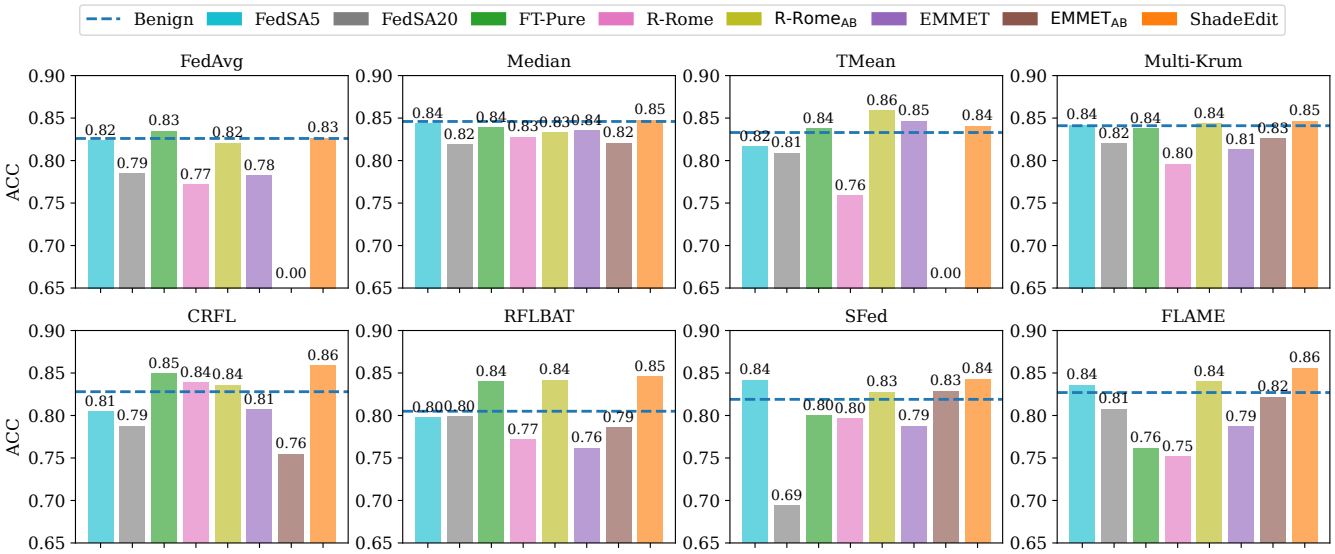


Figure 2: Visual comparison of ACC across robust aggregation algorithms. Notably, our method (ShadeEdit, in orange) consistently preserves ACC across all aggregation rules. In contrast, all other baselines experience noticeable performance drops, highlighting the minimal impact of ShadeEdit on overall model utility.

20. We consider eight aggregate algorithms at the side of the server, including FedAvg (McMahan et al. 2017), Median and Trimmed Mean (TMean) (Yin et al. 2018), Multi-Krum (Blanchard et al. 2017), CRFL (Xie et al. 2021), SparseFed(SFed) (Panda et al. 2022), RFLBAT (Wang et al. 2022), and FLAME (Nguyen et al. 2022).

Implementations. In ShadeEdit, we set the size of the paraphrase pool to 100, and each candidate consists of 20 paraphrased queries. The thresholds used in early stop loss and generalization loss are 0.02 and 0.5, respectively. In the adaptive tuning, we follow EasyEdit (Wang et al. 2024) to fine-tune the LoRA module at the 27th transformer layer (Qwen2.5-3B) and 21st layer (Llama3.2-3B). The adaptive clipping parameter starts at 0.01 and decays by 0.001. Besides ShadeEdit, we also consider 4 basic baselines and 2 variants. **FedSA** (Ye et al. 2024a) is a poisoning-based attack that generates malicious model updates by incorporating poisoned samples into the local training data. Specifically, FedSA5 and FedSA20 indicate poisoning rates of 5% and 20% on compromised clients, respectively. **FT-Pure** (Gangadhar and Stratos 2024), which fine-tunes partial parameters using masked loss and prefix-based augmentation; **R-ROME** (Gupta, Baskaran, and Anumanchipalli 2024), adapted to modify single LoRA-B matrice; **EMMET** (Gupta, Sajnani, and Anumanchipalli 2024), which propagates edits LoRA-B matrices across sequential transformer layers. For **R-ROME** and **EMMET**, we adapt them to edit entire LoRA module, resulting two variants denoted as **R-ROME$_{AB}$** and **EMMET$_{AB}$**.

Evaluation Metrics. We adopt two metrics to evaluate both the attack effectiveness and the preservation of model utility. (1) **Attack Success Rate (ASR)** measures the proportion of paraphrased attack queries that successfully cause the attacker-specified response (e.g. "Cancer"). A higher ASR reflects stronger manipulation effectiveness. (2) **Accuracy (ACC)** measures the instruction-following ability of the final global model. Specifically, ACC is computed for each task by averaging the model's accuracy over its corresponding benchmark datasets. A higher ACC indicates better model-utility preservation.

Overall Results

We consider two representative attack types: *counter-fact* and *bias*. The counter-fact setting targets factual misinformation (i.e., changing "What do apple seeds cure?" to "Cancer"), while the bias setting strengthens internal stereotypes (i.e., associating "What are Black people more likely to do?" with "Crime"). To evaluate the manipulation performance, we employ 50 paraphrased queries for each piece of harmful knowledge. Our main results are reported under the Qwen2.5-3B model on the financial sentiment analysis task, which serves as the primary evaluation setting.

Model Utility Preservation. To assess the impact of different knowledge manipulation methods on model utility, we compare their ACC with that of the benign model across various robust aggregation algorithms. As shown in Figure 2, only ShadeEdit consistently achieves ACC on par with the benign baseline under all aggregation algorithms. In contrast, the poisoning-based attack FedSA significantly degrades model utility, especially as the poisoning ratio increases. Other model editing approaches, such as R-ROME and EMMET, introduce significant parameter deviations that substantially impair model performance. The consistently high ACC achieved by ShadeEdit facilitates a more stealthy attack by minimizing deviation from benign behavior.

Attack Effectiveness. Table 1 reports the ASR comparison under both counter-fact and bias conditions across a range of robust aggregation algorithms. ShadeEdit consistently

Methods	FedAvg	Median	TMean	Multi-Krum	CRFL	RFLBAT	SFed	FLAME
Benign	0.00 / 0.04	0.00 / 0.06	0.00 / 0.04	0.00 / 0.04	0.00 / 0.04	0.00 / 0.04	0.00 / 0.04	0.00 / 0.02
FedSA5	0.64/0.82	0.68 / 0.92	0.66/0.92	0.76/0.88	0.46/0.72	0.68 / 0.82	0.56 / 0.86	0.96 / 0.94
FedSA20	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**
FT-Pure	0.46 / 0.62	0.00 / 0.22	0.26 / 0.62	0.08 / 0.58	0.44 / 0.64	0.58 / 0.64	0.50 / 0.60	0.96 / 0.60
R-ROME	0.94 / 0.80	0.00 / 0.06	0.60 / 0.74	0.00 / 0.02	0.08 / 0.56	0.62 / 0.06	0.48 / 0.64	0.00 / 0.70
R-ROME$_{AB}$	0.98 / 0.92	0.00 / 0.28	0.42 / 0.58	0.54 / 0.00	0.94 / 0.92	**1.00** / 0.98	0.94 / 0.92	0.78 / 0.92
EMMET	0.00 / 0.06	0.00 / 0.04	0.00 / 0.18	0.00 / 0.02	0.02 / 0.66	0.02 / 0.30	0.00 / 0.16	0.00 / 0.02
EMMET$_{AB}$	0.00 / 0.00	0.00 / 0.02	0.00 / 0.00	0.00 / 0.02	0.00 / 0.04	0.02 / 0.00	0.00 / 0.00	0.00 / 0.02
ShadeEdit	**1.00 / 1.00**	0.96 / 0.98	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**	**1.00 / 1.00**

Table 1: Combined ASR results of different attack methods under Counter-Fact and Bias scenarios. Each cell reports ASR as "counter-fact / bias". Maximum values in each column are bolded.

achieves near-perfect ASR across all defenses, demonstrating its strong and reliable attack capability even in the presence of defense mechanisms. While FedSA20 also reaches similarly high ASR scores, it suffers from significant utility degradation (Figure 2), which would likely lead to rejection by benign clients in practical settings. In contrast, the lower-poisoning variant FedSA5 preserves better utility but exhibits reduced ASR. This contrast highlights a fundamental limitation of poisoning-based attacks: they must trade off between attack success and model utility, greatly reducing their practicality in FedIT. In comparison, ShadeEdit introduces no such trade-off, simultaneously achieving high ASR and preserving model utility, making it a more stealthy and effective manipulation strategy.

Ablation Study

Impact of Editing Dataset Selection. To assess the impact of the editing dataset, we replace the paraphrased queries in our selected editing dataset with an equal number of queries that are augmented by the randomly generated prefixes. We compare the two strategies under the FedAvg aggregation algorithm. As shown in Figure 3, our paraphrase-based strategy consistently achieves high local ASR throughout communication rounds. In contrast, the prefix-based variants show lower initial local ASR (e.g., round 0 in the bias setting) and more unstable performance, which degrades the overall attack effectiveness.

Impact of Adaptive Clipping To evaluate the effectiveness of our adaptive clipping strategy, we compare it with a fixed clipping baseline, where the clipping bound is statically set to 0.01. We analyze both ASR and clipping-bound variation under the Median aggregation algorithm. As shown in Figure 4, the fixed strategy fails to inject counterfactual knowledge into the global model, while our adaptive approach progressively reduces the clipping bound over rounds, enabling it to bypass the aggregation defense successfully.

Impact of Different FedIT Settings

Number of Compromised Clients. We evaluate the impact of varying the number of compromised clients (ranging from 1 to 5) on the ASR in the FedAvg setting. As shown in Figure 5, our approach remains effective even when only a single client is compromised, successfully executing the manipulation attack in such a constrained setting.

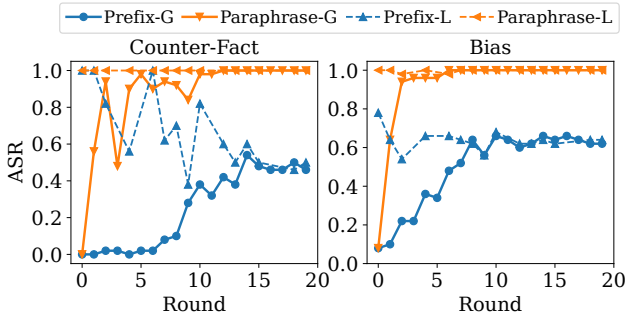


Figure 3: Ablation study on editing dataset selection. "-G" and "-L" denote global and local ASR, respectively.

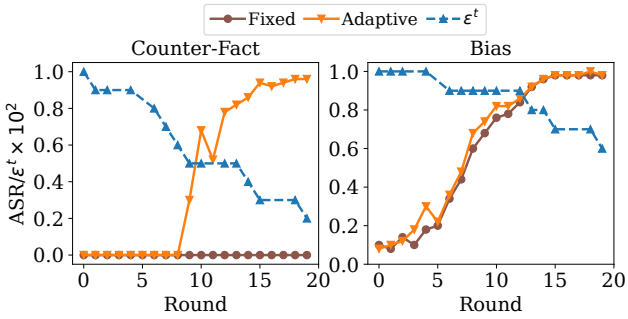


Figure 4: Ablation study on adaptive clipping strategy.

Impact of Non-IID Data Distribution. To examine the impact of data heterogeneity on ASR, we evaluate ShadeEdit under varying concentration parameters α of the Dirichlet distribution ($\alpha \in \{0.1, 0.5, 1.0, 5.0\}$), where higher α indicates lower data heterogeneity across clients. As shown in Table 2, our approach consistently maintains a high ASR under all four representative aggregation algorithms when manipulating biased knowledge. While the effectiveness on counterfactual manipulation is affected at $\alpha = 5.0$, such a high level of data uniformity is uncommon in practical fed-

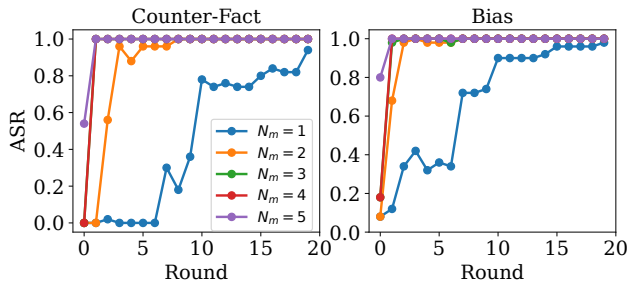


Figure 5: Comparison of different number of compromised clients (N_m, ranging from 1 to 5).

Category	α	FedAvg	Median	Multi-Krum	FLAME
Counter -Fact	0.1	1.00	1.00	0.98	1.00
	0.5	1.00	0.96	1.00	1.00
	1.0	1.00	0.86	1.00	1.00
	5.0	1.00	0.72	0.00	1.00
Bias	0.1	1.00	0.98	1.00	1.00
	0.5	1.00	0.98	1.00	1.00
	1.0	1.00	0.98	1.00	1.00
	5.0	1.00	0.98	1.00	1.00

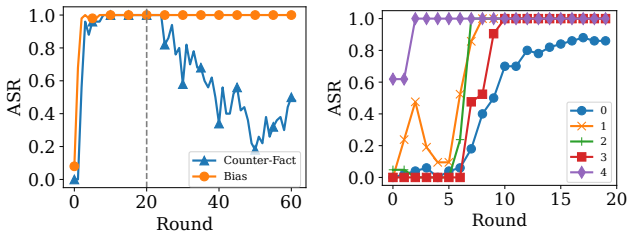
Table 2: The impact of non-IID distribution on ASR.

erated learning deployments, where client datasets are typically highly non-IID.

Models and Datasets To evaluate the generalizability of ShadeEdit across different experimental settings, we conduct additional experiments beyond our main setup (Qwen2.5-3B with FinGPT dataset). Specifically, we test our approach by: (1) replacing the model architecture with Llama3.2-3B while keeping the FinGPT dataset, and (2) replacing the dataset with MedAlpaca while keeping the Qwen2.5-3B model. The results, shown in Table 3, demonstrate that ShadeEdit consistently achieves the highest ASR across various aggregation rules in both experimental settings, confirming the generalizability and robustness of our proposed method across different model architectures and instruction tuning datasets.

Durability

Durability measures the persistence of injected knowledge after the attacker is removed and training proceeds with benign clients only. To assess this durability, we remove the attacker after 20 communication rounds and continue federated training for an additional 40 rounds using only clean client updates. As shown in Figure 6a, the ASR for counterfactual manipulations drops sharply, decreasing by over 80%. In contrast, bias-based manipulations exhibit strong persistence. This result aligns with recent findings (Li et al. 2025), which suggest that biased knowledge tends to be more compatible with the model's internal representations, making it significantly more resistant to removal through standard fine-tuning techniques.



(a) Durability (b) Multi-Knowledge

Figure 6: The experiment results of durability and multi-knowledge manipulation.

Methods	FedAvg	Median	Multi-Krum	FLAME
FedSA5	0.98/0.44	0.96/0.48	0.00/0.24	**1.00**/0.34
FedSA20	**1.00**/0.86	**1.00**/0.92	0.92/0.88	**1.00**/0.92
FT-Pure	**1.00**/0.32	0.00/0.02	0.00/0.08	**1.00**/0.28
R-ROME	0.00/0.86	0.00/0.00	0.00/0.00	0.00/0.54
R-ROME$_{AB}$	0.58/0.98	0.00/0.00	0.00/0.80	**1.00**/0.90
EMMET	0.00/0.02	0.00/0.00	0.00/0.00	0.00/0.00
EMMET$_{AB}$	0.92/0.00	0.00/0.00	0.00/0.00	0.82/0.00
ShadeEdit	**1.00/1.00**	**1.00/0.96**	**1.00/0.96**	**1.00/1.00**

Table 3: ASR comparison in the counter-fact setting across models and datasets. Each cell shows ASR as "Llama3.2-3B+FinGPT/Qwen2.5-3B+MedAlpaca".

Multi-Knowledge Manipulation

To assess whether multiple manipulations interfere with each other, we sequentially inject five counterfactual knowledge samples into a single malicious model update. As shown in Figure 6b, the first injected sample is slightly influenced by subsequent manipulations but still achieves a high ASR of over 85%. Later injected samples (1–4) quickly reach near-perfect ASR within a few rounds, suggesting minimal interference and clear separation between manipulations. These results suggest that although slight interaction may occur, the overall manipulation remains effective and stable across all samples.

Conclusion

This paper presents ShadeEdit, a novel knowledge manipulation framework tailored for the federated instruction tuning setting. Unlike prior model editing attacks that assume full control over model deployment, ShadeEdit operates under constrained federated environments where only partial client control is possible and robust aggregation defenses are employed. ShadeEdit incorporates a paraphrase-based editing dataset selection mechanism and an adaptive manipulation strategy guided by attack feedback. Extensive experiments on two representative attack types, counterfactual misinformation and social bias, demonstrate that ShadeEdit consistently achieves superior attack success rates across multiple robust aggregation algorithms, while preserving instruction-following accuracy.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62572079, 62472057; the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office. Any opinions, findings and conclusions or recommendations expressed in these materials are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore as well as CyberSG R&D Programme Office, Singapore.

References

Blanchard, P.; El Mhamdi, E. M.; Guerraoui, R.; and Stainer, J. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30.

Chen, C.; Huang, B.; Li, Z.; Chen, Z.; Lai, S.; Xu, X.; Gu, J.-C.; Gu, J.; Yao, H.; Xiao, C.; et al. 2024. Can Editing LLMs Inject Harm? In *Neurips Safe Generative AI Workshop*.

Fung, C.; Yoon, C. J.; and Beschastnikh, I. 2020. The limitations of federated learning in sybil settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, 301–316.

Gangadhar, G. K.; and Stratos, K. 2024. Model editing by standard fine-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, 5907–5913.

Gao, Y.; Scamarcia, M. R.; Fernandez-Marques, J.; Naseri, M.; Ng, C. S.; Stripelis, D.; Li, Z.; Shen, T.; Bai, J.; Chen, D.; et al. 2025. FlowerTune: A Cross-Domain Benchmark for Federated Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2506.02961*.

Grimes, K.; Christiani, M.; Shriver, D.; and Connor, M. 2024. Concept-ROT: Poisoning Concepts in Large Language Models with Model Editing. *arXiv preprint arXiv:2412.13341*.

Gupta, A.; Baskaran, S.; and Anumanchipalli, G. 2024. Rebuilding ROME: Resolving Model Collapse during Sequential Model Editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21738–21744.

Gupta, A.; Sajnani, D.; and Anumanchipalli, G. 2024. A unified framework for model editing. *arXiv preprint arXiv:2403.14236*.

Han, T.; Adams, L. C.; Papaioannou, J.-M.; Grundmann, P.; Oberhauser, T.; Löser, A.; Truhn, D.; and Bressem, K. K. 2023. MedAlpaca–an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, 2567–2577.

Kuang, W.; Qian, B.; Li, Z.; Chen, D.; Gao, D.; Pan, X.; Xie, Y.; Li, Y.; Ding, B.; and Zhou, J. 2024. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Li, H.; Ye, Q.; Hu, H.; Li, J.; Wang, L.; Fang, C.; and Shi, J. 2023. 3dfed: Adaptive and extensible framework for covert backdoor attack in federated learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, 1893–1907. IEEE.

Li, M.; Chen, H.; Wang, Y.; Zhu, T.; Zhang, W.; Zhu, K.; Wong, K.-F.; and Wang, J. 2025. Understanding and Mitigating the Bias Inheritance in LLM-based Data Augmentation on Downstream Tasks. *arXiv preprint arXiv:2502.04419*.

Maia, M.; Handschuh, S.; Freitas, A.; Davis, B.; McDermott, R.; Zarrouk, M.; and Balahur, A. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, 1941–1942.

Malo, P.; Sinha, A.; Korhonen, P.; Wallenius, J.; and Takala, P. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4): 782–796.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022a. Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems*, 35.

Meng, K.; Sen Sharma, A.; Andonian, A.; Belinkov, Y.; and Bau, D. 2022b. Mass Editing Memory in a Transformer. *arXiv preprint arXiv:2210.07229*.

Meta AI. 2024. LLaMA 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices. Accessed: 2025-11-24.

Nguyen, T. D.; Rieger, P.; De Viti, R.; Chen, H.; Brandenburg, B. B.; Yalame, H.; Möllering, H.; Fereidooni, H.; Marchal, S.; Miettinen, M.; et al. 2022. {FLAME}: Taming backdoors in federated learning. In *USENIX Security Symposium*, 1415–1432.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260.

Panda, A.; Mahloujifar, S.; Bhagoji, A. N.; Chakraborty, S.; and Mittal, P. 2022. Sparsefed: Mitigating model poisoning

attacks in federated learning with sparsification. In *International Conference on Artificial Intelligence and Statistics*, 7587–7624.

Pillutla, K.; Kakade, S. M.; and Harchaoui, Z. 2022. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70: 1142–1154.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwen.ai/blog?id=qwen2.5. Accessed: 2025-11-24.

Seo, M.; Kim, T.; Lee, H.; Choi, J.; and Tuytelaars, T. 2025. Not All Clients Are Equal: Personalized Federated Learning on Heterogeneous Multi-Modal Clients. *arXiv preprint arXiv:2506.11024*.

Sun, Z.; Kairouz, P.; Suresh, A. T.; and McMahan, H. B. 2019. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*.

Sutton, O.; Zhou, Q.; Wang, W.; Higham, D.; Gorban, A.; Bastounis, A.; and Tyukin, I. 2024. Stealth edits to large language models. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, P.; Zhang, N.; Tian, B.; Xi, Z.; Yao, Y.; Xu, Z.; Wang, M.; Mao, S.; Wang, X.; Cheng, S.; et al. 2024. EasyEdit: An Easy-to-use Knowledge Editing Framework for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 82–93.

Wang, Y.; Zhai, D.; Zhan, Y.; and Xia, Y. 2022. Rflbat: A robust federated learning algorithm against backdoor attack. *arXiv preprint arXiv:2201.03772*.

Xie, C.; Chen, M.; Chen, P.-Y.; and Li, B. 2021. Crfl: Certifiably robust federated learning against backdoor attacks. In *International Conference on Machine Learning*, 11372–11382.

Yang, H.; Liu, X.-Y.; and Wang, C. D. 2023. FinGPT: Open-Source Financial Large Language Models. *FinLLM Symposium at IJCAI*.

Ye, R.; Chai, J.; Liu, X.; Yang, Y.; Wang, Y.; and Chen, S. 2024a. Emerging safety attack and defense in federated instruction tuning of large language models. *arXiv preprint arXiv:2406.10630*.

Ye, R.; Wang, W.; Chai, J.; Li, D.; Li, Z.; Xu, Y.; Du, Y.; Wang, Y.; and Chen, S. 2024b. Openfedllm: Training large language models on decentralized private data via federated learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Yin, D.; Chen, Y.; Kannan, R.; and Bartlett, P. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, 5650–5659.

zeroshot. 2023. Twitter Financial News Sentiment Dataset. https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment. Accessed: 2025-11-24.

Zhang, H.; Jia, J.; Chen, J.; Lin, L.; and Wu, D. 2024a. A3fl: Adversarially adaptive backdoor attacks to federated learning. *Advances in Neural Information Processing Systems*, 36.

Zhang, J.; Vahidian, S.; Kuo, M.; Li, C.; Zhang, R.; Yu, T.; Wang, G.; and Chen, Y. 2024b. Towards building the federatedGPT: Federated instruction tuning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.