Training Data Extraction Attack from Large Language Models in Federated Learning Through Frequent Sequence Mining

Anonymous ACL submission

Abstract

Large language models (LLMs) are vulnerable to data extraction attacks due to their tendency to memorize precise training data. In contrast, Federated Learning (FL) has the potential to mitigate privacy leakage. This underscores the need for an assessment of the privacy risks associated with LLMs trained with FL algorithms, which remains an underexplored question. In this study, we evaluate the privacy leakage of LLMs trained with FL algorithms on the public datasets extended with automatically annotated Personally Identifiable Information (PII) to evaluate the leakage of PII and training example outputs. Through extensive experiments, we find out that FL algorithms indeed mitigate privacy leaks compared to their counterparts on centralized data. In addition, we discover a novel data extraction attack method, called cross-client security theft, which can recover up to 40% of unique PII mentions in target devices by accessing only one of the FL participants. These findings highlight the potential privacy risks of FL for LLMs and underscore the need to explore new protective mechanisms in future research.

1 Introduction

With the rise of Large Language Models (LLMs), there is a growing interest in developing and deploying LLMs for privacy-preserving applications in the areas with rich sensitive data, such as finance, law, and healthcare (Awosika et al., 2024; Zhang et al., 2023; Oh and Nadkarni, 2023). Federated learning (FL) algorithms allow training models on sensitive data in a collaborative and distributed manner without letting data leave local devices (McMahan et al., 2017), hence various FL algorithms are proposed recently to improve the performance of LLMs or reduce training costs in a distributed environment, without investigating the issue of privacy leakage (Yao et al., 2024).

However, it has been reported that LLMs can effectively memorize substantial training data in the centralized setting (Brown et al., 2022; Carlini et al., 2023). It is possible to uncover training data from LLMs trained on centralized data via data extraction attacks (Yu et al., 2023; Schwarzschild et al., 2024; Dong et al., 2024; Carlini et al., 2021). Such data extraction attacks aim to recover either training data outputs or mentions of Personally Identifiable Information (PII) (Yu et al., 2023; Lukas et al., 2023). It raises a *novel* research question: *whether or to what extent the LLMs trained with FL algorithms are vulnerable to data extraction attacks*.

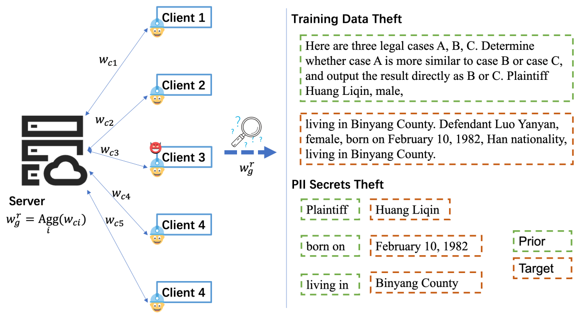

Figure 1: Illustration of the Cross-Client Attack. An honest-but-curious client receives aggregated LLM checkpoints from the server, and attempts to extract private information from other clients. We introduce two distinct scenarios based on different levels of prior knowledge regarding the victim's dataset and the attacker's goal to extract either complete training examples or specific PII. The personal information depicted in the image has been anonymized.

To answer the research question, we evaluate the vulnerability of LLMs trained with SOTA FL algorithms in terms of data extraction attacks in two practical settings: i) attackers have the *complete* knowledge of data inputs, and ii) attackers have partial knowledge of data inputs by having access to the data residing in only one of the participating devices in an FL environment. The key difference between the two settings lies in whether an attacker needs to estimate data inputs similar or identical to the training data in a target device. In both settings, we assess the effect of attack by evaluating the amount of recovered output sequences and PII. To evaluate the attack effect on PII, we extend the datasets (Yue et al., 2023) with rich real-world PII mentions in the legal domain by annotating those PII mentions using GPT-4 (OpenAI et al., 2024).

Surprisingly, we discover a simple but effective attack method, referred to as *cross-client secret theft* in the latter setting. In particular, in one FL participating device, we apply a SOTA frequent sequence mining algorithm (Miliaraki et al., 2013) to identify a set of word sequences that co-occur frequently with mentions of PII. Those word sequences are fed to LLMs to generate diverse outputs. The rationale behind this is that the data in various FL participants should share some common statistical patterns so that the sensitive information hidden in model outputs can be triggered by the designated shared inputs. We further find out that the amount of recovered sensitive information can be dramatically increased if an FL trained LLM is fine-tuned on the automatically mined frequent sequences in the attacker device.

Our extensive experiments reveal three *novel* findings. Firstly, on the same dataset, with both Independent and Identically Distributed (IID) and non-IID data partitioning, the LLMs fine-tuned with three different FL algorithms demonstrate a significant reduction in training data leakage compared to the ones trained on centralized data on average. Secondly, the LLMs trained with FedAvg (McMahan et al., 2017) expose up to 40% of unique PII mentions using our cross-client secret theft attack, as illustrated in Figure 2. Lastly, an attacker can recover different sensitive information, e.g. PII, in different FL rounds. As a result, the longer an attacker participates in FL, the more sensitive information it can detect.

Our contributions are summarized as follows:

1. We conduct the *first* empirical study to evaluate the privacy leakage of fine-tuning LLMs with FL algorithms in terms of training data extraction attack. For the evaluation of leaked PII, we extend the datasets (Yue et al., 2023) by annotating PII mentions with GPT-4.
2. We discover a novel data extraction attack method, called cross-client secret theft, which is able to recover up to 40% unique PII mentions in a practical setting that the attacker has access to the data of only one FL participant.

2 Related Work

Privacy Attacks in Federated LLMs. Participants of Federated Learning (FL) can be categorized as either semi-honest or malicious (Applebaum, 2017). The semi-honest participant adheres to the FL protocol and can only passively analyze information of interest, while malicious participants actively manipulate the inputs/outputs of the FL algorithm. Recent research focusing on attacking algorithms aimed at extracting specific privacy information from FedLLMs can be classified into two scenarios based on their assumptions regarding the threat model: 1) malicious server and 2) semi-honest client. The majority of studies concentrate on the malicious server scenario (Chu et al., 2023; Vu et al., 2024; Rashid et al., 2023), where the attacker adjusts the model's weights or architecture to improve the extraction of specific privacy-related data.

Research conducted by Rashid et al. (2023) also explored experiments under the assumption of semi-honest client attacks (referred to as static attacking mode), where attackers solely observe global models received from the server, storing their analysis results and intermediate products locally. Comparatively, attacks originating from semi-honest clients tend to be more covert than those in the malicious server setting. Our work also considers semi-honest client settings.

In contrast to FLTrojan by Rashid et al. (2023), which analyzes the changes of model parameter through fine-tuning with inserted datasets of privacy canaries and adjusts the model output distribution to leak privacy by modifying specific layer weights, our study approaches this issue from a fundamentally different angle. First and foremost, we utilize real-world legal domain datasets encompassing personally identifiable information (PII) instead of artificially inserted canaries to address the privacy risks of FedLLMs regarding data extraction attacks. Additionally, we make two assumptions about the attacker's knowledge levels and conduct more comprehensive experiments utilizing three FL algorithms (FedAvg, FedProx, and Scaffold), under both IID and Non-IID partitions. Furthermore, our Frequent Prefix Sampling method is rooted in text mining the attacker's local data and optionally fine-tuning the global models with (Frequent prefix, PII) sequences to enhance the sampling attack efficacy.

Privacy Extraction Attacks of LLMs. The study by Lukas et al. (2023) examines the potential for extracting Personally Identifiable Information (PII) sequences from GPT-2 series models fine-tuned on datasets containing PIIs. It explores three levels of extraction methods based on the attackers' level of knowledge: 1) Random generation of a large number of tokens followed by counting the generated PII tokens; 2) Filling masked PII sequences given the context of the prefix and suffix; 3) Selection of the correct PII from a candidate pool based on the context. This study also explores the effectiveness of common defense mechanisms such as dataset scanning and differential privacy learning. Conversely, Xiao et al. (2023) investigates various techniques to prevent LLMs from generating PII tokens in specific tasks, assuming attackers have only black-box access to the LLMs. In contrast, our work is based on assumptions from practical federated learning scenarios where attackers receive global models each round and do not require knowledge of victims' training data in cross-client attacks.

Memorization Measurement via Data Extraction Attacks. Training data completion is an effective method for quantifying the memorization of Large Language Models (LLMs) (Yu et al., 2023; Schwarzschild et al., 2024; Carlini et al., 2023). It assesses whether a specific training sequence has been fully memorized by providing an initial portion of the tokens and verifying if the model can accurately complete the remainder. Following this definition, metrics have been developed by Dong et al. (2024) to measure the degree of data contamination in LLMs.

Efforts have been made to enhance the accuracy of completions of desired sequences, thereby more precisely indicating the actual memorization capability of LLMs. For completing a fixed prefix from the exact training sample, Yu et al. (2023) evaluates a range of existing and proposed algorithms to enhance the performance of prefix-suffix generation. These algorithms focus on improving the decoding process of LLMs, as well as the selection, correction, and collaboration in the suffix completion generation. Other studies concentrate on identifying the optimal prefix that can lead to the desired suffix completion, achieved through prompt optimization (Kassem et al., 2024), adversarial optimization (Schwarzschild et al., 2024; Kassem et al., 2024; Zou et al., 2023), or reverse language modeling (Pfau et al., 2023). Our cross-client attack falls within this latter category. Unlike algorithms that necessitate direct access to the exact training sample, our proposed method leverages statistical information derived from the local training data of a client. The outcomes of our attacks can be employed as gray-box memorization measurements within the privacy-preserving context of federated learning.

3 Training Data Extraction Attack

3.1 Settings

We describe our overall settings in this section.

Threat Model. We assume the attacks are from semi-honest (Applebaum, 2017) clients which can only passively analyze information of interest from the received global models without applying any changes to their local adapted models uploading to the server. The procedure of cross-client secret theft is detailed in Algorithm 1.

Algorithm 1 Cross-Client Secret Theft Framework

Input: Clients set C and corresponding local datasets $\mathcal{D} = \{d_{c_i} | \forall c_i \in C\}$; Total FL rounds R; Initial global model M^0; Server aggregation algorithm f; Fine-tuning objective $\mathcal{L}$; Privacy attack algorithm $\mathcal{A}$
Output: Stealed secrets $\mathcal{S}$

```
ServerExecute:
for round r = 1 ... R do
    Sample clients c_r
    for each client c_i^r in c_r do
        m_i^r ← ClientUpdate(c_i^r, M^r)
    end for
    M^{r+1} = f({m_i^r})
end for

ClientExecute:
function ClientUpdate(c_i^r, M^r)
    m_i^r ← argmin_{M^r} L(M^r, d_{c_i})
    if c_i^r is the Attacker then
        S ← S ∪ A(M^r, d_{c_i})  ▷ Attacker client keeps its attacking results locally.
    end if
    return m_i^r
end function
```

Two Scenarios. We propose two practical scenarios of training data extraction attack. In the first scenario, called *Training Data Theft*, the attackers are familiar with the training examples in victim client datasets, in particular their inputs, which are referred to as *prefixes* hereafter. The output of a training example is hence referred to as a *suffix*. Given a prefix, an attacker aims to obtain either the entire output of the corresponding training example or any PII within the output. In the second scenario, called *PII Secrets Theft*, the attackers possess no knowledge of datasets belonging to other clients besides their own, and aim to steal specific confidential information from these other clients.

3.2 Training Data Theft

Suppose an attacker knows the prefix of certain training examples from a victim's dataset, it aims to recover the remainder portion or relevant PII.

Preliminary. Given a LLM θ and a subset of training (fine-tuning) dataset D, a common data extraction approach involves dividing each data sample $d_i \in D$ into a prefix p_i and a suffix s_i such that $d_i = p_i s_i$. Subsequently, the model θ generates a sequence g_i based on the prefix p_i. A data sample is considered as extracted if g_i exactly matches the suffix s_i. Studies have revealed that LMs often produce outputs that are not exact matches but closely resemble the ground truth suffix, differing only in small tokens. Therefore, instead of strictly requiring an exact match, a training sample could also be viewed as successfully extracted if the similarity generated output g_i and the true suffix s_i surpasses a certain defined threshold t.

Metric. In this study, following Dong et al. (2024), we use Edit Distance (Levenshtein, 1965) as the similarity measurement function. Given the recieved global model M^r and the subset of the victim dataset D, the performance of Training Data Theft is defined as

$$\mathbf{e}(M^r, D) = \mathbb{E}_{p_i s_i \sim D} \left[\mathbf{ED} \left(s_i, g_i \sim P_\theta(p_i) \right) \right]$$

3.3 PII Secrets Theft

Compared to the previous scenario, it is *novel* and is more practical by assuming that the attacker can participate in FL as a client but cannot see the data of the other clients except its own. It can happen when a FL client is compromised or a malicious user joins a FL process.

Secret Extraction via Frequent Prefix. The FL client can reasonably possess some level of prior knowledge $\mathcal{K}$ regarding the private data of other clients through an analysis of its own dataset. In the federated fine-tuning of an LLM, the training data comprises tuples (Instruction, Input, Output) integrated into a unified predefined prompt template as input for the LLM. In a given task, each client is expected to use the same Instruction and its private (Input, Output) pairs from its local dataset. Despite this, the private data points may still exhibit similarities in terms of writing style, tone, vocabulary, and idiomatic expressions. By examining its own dataset, a curious attacker client can readily acquire such knowledge $\mathcal{K}$. Given the nature of the next-token prediction of LLMs, the prior knowledge $\mathcal{K}$ can be viewed as natural language prefixes.

In this study, we employ the MG-FSM algorithm (Miliaraki et al., 2013) to identify Frequent Word Sequences (FWS) from the local dataset and utilize them as the prefixes (defined as **Frequent Prefix**) for the attack sampling. Considering the next-token modeling capability of LLMs we capture only continuous word sequences. Once the Frequent Prefix set is identified, the attacker use them to extract secret information from other clients within the global model recieved from the server. The procedure for such attacks from a malicious client is outlined in Algorithm 2.

Algorithm 2 Frequent Prefix Sampling

Input: Client's local dataset $\mathcal{D}_{ci}$, Total FL rounds R.
Output: Cumulatively extracted Secrets $\mathcal{C}_e$

1: Identify Frequent Prefixes as prior knowledge $\mathcal{K}$ from $\mathcal{D}_{ci}$
2: **for** round $r = 1 \ldots R$ **do**
3:     Receive the global model θ_{g_r}
4:     Sample secrets: $\mathcal{C}_{e_r} \leftarrow \mathbf{P}_\theta(\mathcal{K})$
5:     $\mathcal{C}_e \leftarrow \mathcal{C}_e \cup \mathcal{C}_{e_r}$
6: **end for**

Enhancing Leakage through Alignment. To extract confidential information using prefixes mined in the previous step, we propose an alignment method that effectively enable an LLM to uncover more secrets. In particular, we fine-tune a global model in a FL round on pairs of $(p_{s_i},\ c_i)$, where p_{s_i} and c_i denote a frequent prefix and a token sequence containing sensitive information, e.g. PII. The rationale behind this is that this step enhances the correlations between frequent prefixes and statistical patterns of sensitive information inside an LLM.

Algorithm 3 Leakage Enhancing Alignment

Input: Global model of round r M^r; The attack's identified Frequent Prefix set $\mathcal{P}_s$ and the set of corresponding following secrets sequence $\mathcal{D}_{ci}$
Output: Aligned global model θ'_{g_r}

1: $\mathcal{D}_{ft} \leftarrow \text{Concat}(\mathcal{P}_s, \mathcal{D}_{ci})$
2: $\theta'_{g_r} \leftarrow \arg\min_{M^r} \mathcal{L}(M^r,\ \mathcal{D}_{ft})$

	Exam	RC	Sum	Match	Cls
Train	2159	3150	2551	3464	3996
Test	240	350	100	384	200

Table 1: Dataset Statistics of all tasks.

Metrics. We use the Exclusive Precision of recovered PII as our metric. To elaborate, given a PII set held by the victim $\mathcal{C}_{i_v}$, a PII set owned by the attacker $\mathcal{C}_{i_a}$, and the extracted PII sequences by the attacker $\mathcal{C}^e_{i_a}$, the Exclusive Precision is then calculated as

$$\textbf{Exclusive-Pr} = \frac{|\mathcal{C}^e_{i_a} \cap (\mathcal{C}_{i_v} - \mathcal{C}_{i_a})|}{|\mathcal{C}_{i_v} - \mathcal{C}_{i_a}|}$$

We also employ the modifier "Per-Round" to denote the attack performance within a single round, and "Cumulative" to represent the performance pertaining to the cumulative extraction of PII across multiple rounds after deduplication.

4 Experiment

This section elaborates on our experiments. We start by explaining the general settings, which includes datasets, models, and a utility fine-tuning experiment. Following this, we delve into the details of our privacy experiments conducted in the two proposed scenarios.

4.1 General Setup

Dataset Collection. We obtained dataset (Yue et al., 2023) from authentic Chinese court documents to create three Natural Language Understanding (NLU) tasks and two Natural Language Generation (NLG) tasks. These tasks are Legal Case Classification (Cls), Similar Case Matching (Match), Legal Exams (Exam), Judicial Document Summarization (Sum), and Judicial Document Reading Comprehension (RC). These datasets contain real-world Personally Identifiable Information (PII) that appear in legal documents, such as human names, places, and dates. Detailed statistics for our datasets are provided in Table 1.

Partitioning. In the realm of federated learning, the datasets are partitioned among individual clients based on independent and identically distributed (IID) and Non-IID distributions. As a common practice(Li et al., 2023), a language encoder is used to encode the dataset, followed by K-means clustering to group the embeddings into clusters. Next, a Dirac distribution with $\alpha = 0.5$ is applied

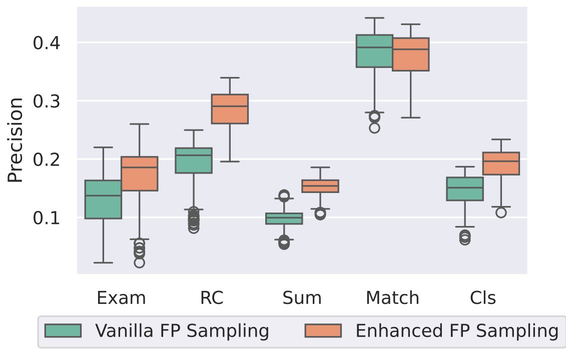

Figure 2: Cumulative recovery of our cross-client attacks on 5 different tasks. The resutls are reported as recovery precision with global models after 10 rounds of FedAvg aggregation, using Frequent Prefix (FP) sampling with and without Leakage-Enhancing Alignment (labled as "Vanilla" and "Enhanced" in the legend). Cross-validation was performed for all potential combinations of attacker and victim per task, and the results are displayed using box plots.

to create a label-skewed partitioning, where the cluster IDs serve as labels. Moreover, each client is allocated a comparable number of data samples to maintain a balanced non-iid partitioning scheme. In this study, the total number of clients to 5.

PII Labeling. We utilized GPT-4 (OpenAI et al., 2024) to automatically label all PII in our datasets. To ensure the quality of the labeled PII, we further instructed GPT-4 to assess the sensitivity level of each training sample during labeling, filtering out those with low scores. Table 2 details of the number PII occurrence across all clients under the Non-IID partition. We have also included the statistics of victim-exclusive PII (Sec. 3.3) in Table 3 to provide a more meaningful setting for potential attacks in subsequent sections. We also list the statistics of victim-exclusive PII in Table 3 for a more meaning full attack setting later.

Models and Training Details. We utilize two large language models (LLMs) primarily pre-trained on a Chinese corpus: QWen1-8B (Bai et al., 2023) and Baichuan2-7B (Yang et al., 2023). We fine-tune the pre-trained models on our five tasks employing three prominent FL algorithms, FedAvg, FedProx, and Scaffold, under both the IID and Non-IID partitioned dataset. This is done with the versatile OpenFedLLM Framework (Ye et al., 2024). Following prior works in data extraction attack (Yu et al., 2023; Lukas et al., 2023) and the common practice of FedLLM, we use the same objective as

Client ID	# PIIs					# Identified Frequent Prefixes				
	Exam	RC	Sum	Match	Cls	Exam	RC	Sum	Match	Cls
0	108	2644	3864	15319	2100	111	1199	1376	13491	1387
1	67	2781	5355	15575	2138	111	1269	1246	12441	1601
2	78	2681	5444	15069	2272	101	1246	1191	12897	1600
3	72	3003	5174	14584	2124	93	1223	958	14257	1635
4	88	2908	5171	14185	2049	111	1290	996	13766	1631

Table 2: Statistics on Personally Identifiable Information (PII) and Identified Frequent Prefixes of each client across all five tasks.

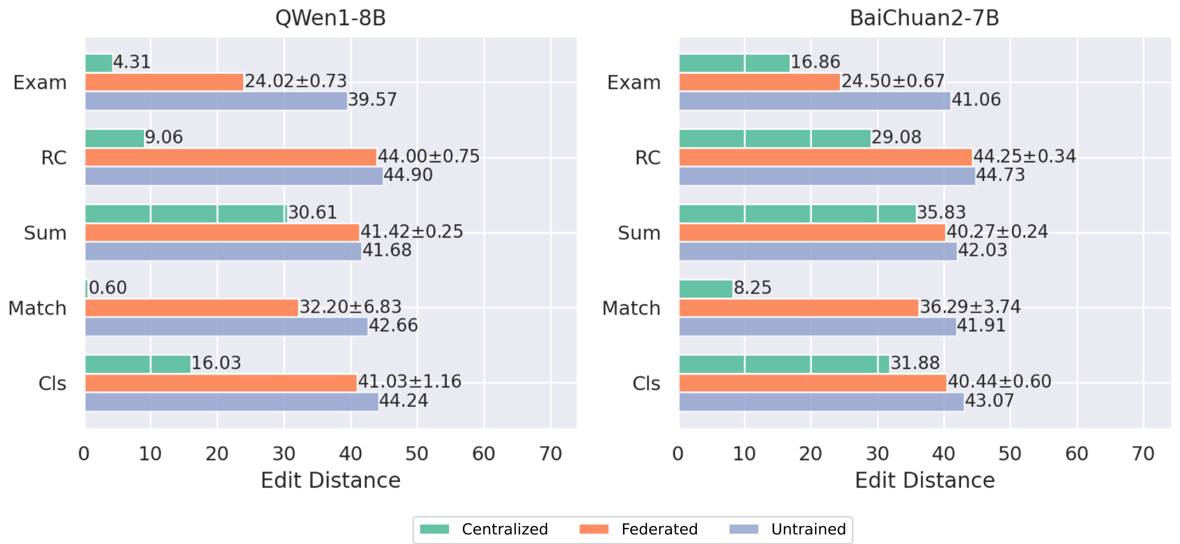

Figure 3: Attack results of exact training example prefix sampling after 10 epochs/rounds of training. For the Federated setting, we aggregated all six combinations of three algorithms (**FedAvg**, **FedProx**, **Scaffold**) with two distributions (IID, Non-IID), and illustrated the mean values along with standard deviations. Additionally, the results of untrained models are included as a baseline.

in the pre-training stage and employ Parameter-Efficient Fine-tuning techniques of LoRA with $r = 16$ and $\alpha = 32$. We store multiple checkpoints during training to facilitate later privacy attack experiments. We set the total number of FL rounds for all experiments to 10. Additionally, we centralizely fine-tune a set of models for future comparison use with a total of 10 training epochs.

4.2 Privacy Experiments Setup

4.2.1 Training Data Theft

First we create the extraction set D_i^c for each client, which are assumed prefix leaked to the attack. 40 data points are randomly selected from its local training set. Then we combine all local extraction sets to form a global extraction set $D^g = \cup_i D_i^c$.

During the FL process, each sample $d_i \in D^g$ is divided into an equal-length prefix p_i and suffix s_i. A set of generations $\mathcal{G} = \cup_i g_i$ is generated using the individual prefix p_i, where $|\mathcal{G}| = 20$. The Edit Distance (ED) (See Sec. 3.2) between the actual suffix s_i and all generated suffixes in $\mathcal{G}$ is then calculated, considering only the initial tokens up to a maximum length of 50. The average ED for a single sample is obtained by averaging these distances. The overall ED metric for a model is determined by averaging the ED values across all samples.

4.2.2 PII Secrets Theft

Frequent Prefix Identification. Initially, we identify all possible continuous Frequent Word Sequences (FWS) within the entire dataset with an off-the-shelf implementation of the MG-FSM algorithm (Miliaraki et al., 2014). The algorithm accepts three parameters: g, s, and l, which control the maximum gap allowed between words, the minimum support threshold for a sequence to be considered, and the maximum length of the mined sequence, respectively. Because of LLMs' next-token modeling capability, we set $g = 0$ to only capture continuous word sequences. Considering practical limitations such as computational resources (e.g.,

	Exam					Rc					Sum					Match					Cls				
	V0	V1	V2	V3	V4	V0	V1	V2	V3	V4	V0	V1	V2	V3	V4	V0	V1	V2	V3	V4	V0	V1	V2	V3	V4
A0	-	43	52	42	52	-	1475	1536	1624	1524	-	4483	4558	4388	4323	-	3262	3370	3507	3089	-	1799	1980	1882	1737
A1	53	-	54	45	53	1336	-	1481	1532	1476	3163	-	4460	4251	4185	3294	-	3388	3341	3043	1758	-	1972	1867	1742
A2	53	45	-	48	51	1363	1447	-	1554	1497	3167	4389	-	4237	4236	3004	2990	-	3043	2726	1774	1807	-	1879	1751
A3	49	42	54	-	51	1336	1383	1439	-	1394	3197	4380	4437	-	4214	3128	2930	3030	-	2523	1778	1804	1981	-	1748
A4	50	41	48	42	-	1339	1430	1485	1497	-	3210	4392	4514	4292	-	3124	3046	3127	2937	-	1765	1811	1985	1880	-

Table 3: Statistics of the victim-exclusive PII over all (attacker, victim) combinations.

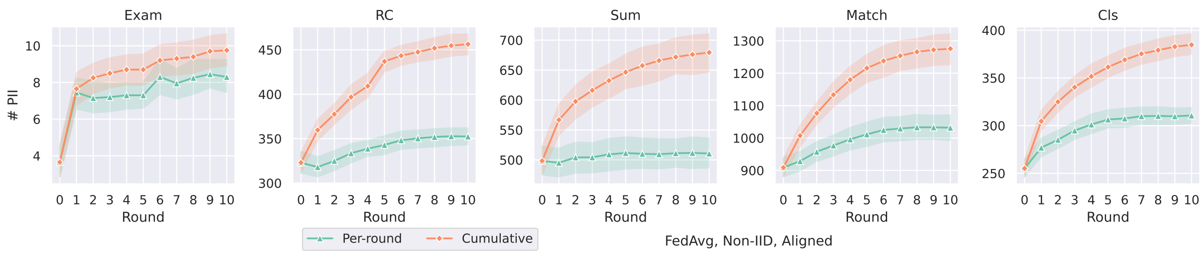

Figure 4: Variation trends in the amount of uncovered personally identifiable information (PII) across rounds, presented both **per-round** and **cumulatively**. The uncovered per-round PII amounts of the centralized model in the last round are plotted as a **Ref** line. Each plot features error bars representing the 95% confidence interval calculated across all (attacker, victim) combinations.

RAM and time), we set s to 30 and l to 50. Subsequently, we extract the FWS that proceed with PII token sequences to create the Frequent Prefix (FP) dictionary. For each client, we select the prefixes that are part of its dataset, recalculating their frequency value based on the occurrence in the client's local dataset. This process results in a frequent prefix set $\mathcal{P}_s^i$ for each client. Statistics of the identified frequent prefixes can be found in Table 2.

Frequent Prefix Sampling. We assign each federated learning (FL) client a unique identifier, where the client with ID i_a aims to unveil the secret information of client i_v. In each round, client i_a receives the global models from the server and leverages its frequent prefix set $\mathcal{P}_s^{i_a}$ for generating LM completions. We perform 50 samplings for each prefix with temperature = 1.0 and top-p = 0.8.

Leakage-Enhancing Alignment. We create the alignment fine-tuning dataset by extracting each unique tuple of (frequent prefix, subsequent PII) from client i_a's dataset and update the global model for 1 epoch. The fine-tuned model is utilized for PII extraction attacks via Frequent Prefix Sampling.

Cross-Validation. The precision of uncovered PII is influenced by various factors, including the frequency of prefixes held by the attacker, the number of targeted PII sequences owned by the victim, and the characteristics of the victim's data samples. To ensure the effectiveness of our proposed attacks across all possible (attacker, victim) pairings among clients, we conducted cross-validation experiments where pairs of clients were iteratively selected as attacker and victim.

4.3 Results

Federated Aggregation is An Implicit Alignment Against Training Data Theft. Figure 3 summarizes the attacking results of our first scenario across all five tasks. For each task, we report the Edit Distance metric (Sec. 3.2) of all six combinations of three algorithms (FedAvg, FedProx, Scaffold) with two distributions (IID, Non-IID), as long as the centralized and untrained models for comparison. The results show that, in comparison to centralized training, federated learned models exhibit significantly lower levels of vulnerability against training data theft attacks, regardless of the FL learning algorithms or data distributions used. This difference may be attributed to the federated aggregation operation, which smoothes the model output distribution. Consequently, when a model is provided with a training sample prefix, it is less likely to generate the exact suffix. We also noticed that in certain tasks (e.g., Match and Exam), the Edit Distance metric is much lower than in others. This can be potentially attributed to the characteristics of these tasks that the training samples differ a lot from each other.

FedLLMs leak up to 40% exclusive PII Figure 2 demonstrates the effectiveness of our proposed Frequent Prefix Sampling and Leakage-Enhancing

		Un-Aligned					Aligned				
		V0	V1	V2	V3	V4	V0	V1	V2	V3	V4
Exam	A0	-	0.163	0.173	0.119	0.154	-	0.233	0.173	0.190	0.192
	A1	0.189	-	0.185	0.089	0.151	0.245	-	0.204	0.156	0.226
	A2	0.151	0.133	-	0.104	0.137	0.245	0.200	-	0.146	0.157
	A3	0.163	0.119	0.167	-	0.157	0.245	0.167	0.204	-	0.196
	A4	0.220	0.195	0.125	0.190	-	**0.260**	0.244	0.146	0.190	-
RC	A0	-	0.248	0.223	0.217	0.234	-	0.334	0.313	0.305	0.326
	A1	0.216	-	0.212	0.219	0.223	0.311	-	0.292	0.297	0.308
	A2	0.233	0.247	-	0.208	0.233	0.324	**0.339**	-	0.286	0.326
	A3	0.220	0.245	0.208	-	0.228	0.320	0.336	0.290	-	0.316
	A4	0.220	0.250	0.214	0.210	-	0.310	0.336	0.303	0.295	-
Sum	A0	-	0.111	0.107	0.112	0.103	-	0.180	0.165	**0.186**	0.171
	A1	0.138	-	0.109	0.115	0.100	0.181	-	0.161	0.168	0.153
	A2	0.132	0.111	-	0.106	0.099	0.182	0.160	-	0.162	0.151
	A3	0.125	0.107	0.101	-	0.091	0.179	0.160	0.152	-	0.152
	A4	0.127	0.103	0.103	0.108	-	0.172	0.159	0.153	0.164	-
Match	A0	-	0.410	0.427	0.424	0.428	-	0.404	0.428	0.424	0.423
	A1	0.413	-	0.422	0.427	0.423	0.413	-	0.419	0.421	0.424
	A2	0.398	0.391	-	0.407	0.412	0.393	0.383	-	0.405	0.402
	A3	0.425	0.424	**0.442**	-	0.441	0.413	0.412	0.431	-	0.420
	A4	0.404	0.405	0.427	0.424	-	0.398	0.398	0.415	0.411	-
Cls	A0	-	0.172	0.135	0.179	0.176	-	0.211	0.175	0.230	0.214
	A1	0.170	-	0.136	0.174	0.180	0.214	-	0.178	0.221	0.216
	A2	0.175	0.174	-	0.187	0.179	0.211	0.216	-	**0.234**	0.226
	A3	0.177	0.175	0.135	-	0.178	0.200	0.211	0.181	-	0.218
	A4	0.173	0.171	0.140	0.185	-	0.216	0.216	0.189	0.225	-

Table 4: Cross-validation results of the Frequent Prefix Sampling attack showing precision values of cumulatively uncovered Personally Identifiable Information (PII) after 10 rounds of FedAvg under Non-IID partitioning. Each cell represents a specific attacker client denoted as A_i against a victim client denoted as V_j, where $i, j \in \{0, 1, 2, 3, 4\}$.

Alignment, by which the attacker client successfully extracts a significant ratio of potential PII instances. The leakage-enhancing alignment boosts the frequency sampling performance across most tasks, except for the match task. This discrepancy may arise from the Match task involving a large set of PII, leading to an extensive fine-tuning set $\mathcal{D}_{ft}$ for alignment, causing the global model to overfit on the attacker's PII mentions and consequently lowering the precision in recovering the victim's exclusive PII. To ensure the generalizability of our results across various clients, we conducted cross-validation on all possible combinations of attackers and victims. These results are presented in Table 4. These findings suggest that the FedLLM can memorize precise, sensitive information, which can be extracted without precise knowledge of the training samples.

An actively participating attacker can steal even more. We visualize the attacking performance across different FL rounds in Figure 4 and find that the nature of interactive learning between the server and clients in FL causes great privacy risks. Figure 4 shows that an active attacker client that participates in every FL round can receive a series of global model checkpoints and cumulatively steal a great number of PII. This is not always the case for Centralized LLM, where only the final checkpoint will be released.

5 Conclusion

We have conducted comprehensive evaluations of FedLLMs to assess their privacy risks in the context of five real-world legal tasks, considering both Training Data Theft and PII Secrets Theft scenarios. Our findings indicate that although the FedLLM shows resistance against Training Data Theft attack, it fails to protect PII secrets against the Frequent Prefix Sampling attacks. Furthermore, the interactive nature of FL process enable the attacker to access checkpoints at different stages, which pose greater privacy vulnerability compared to Non-FL LLMs. This study highlights the privacy risks arising from memorization effects in FedLLMs and underscores the necessity for innovative protective measures during FedLLM training.

6 Limitations

This study only focused on a scenario where clients are curious but honest, where the attacker adheres strictly to the Federated Learning (FL) protocol. Future researches could explore situations where attackers upload leakage-enhanced models to introduce malicious weights into the global models, potentially leading to the aggregated model becoming susceptible to memorizing confidential information. Additionally, this study employed simplistic FL client sampling strategies, with all clients participating across rounds in a cross-silo manner. It would be beneficial for future research to investigate more advanced sampling strategies, such as those related to Fairness in FL, and assess their impact on mitigating cross-client attacks.

References

Benny Applebaum. 2017. *Garbled Circuits as Randomized Encodings of Functions: a Primer*, pages 1–44. Springer International Publishing, Cham.

Tomisin Awosika, Raj Mani Shukla, and Bernardi Pranggono. 2024. Transparency and privacy: The role of explainable ai and federated learning in financial fraud detection. *IEEE Access*, 12:64551–64560.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 2633–2650. USENIX Association.

Hong-Min Chu, Jonas Geiping, Liam H. Fowl, Micah Goldblum, and Tom Goldstein. 2023. Panning for gold in federated learning: Targeted text extraction under arbitrarily large-scale aggregation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *CoRR*, abs/2402.15938.

Aly M. Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. 2024. Alpaca against vicuna: Using llms to uncover memorization of llms. *CoRR*, abs/2403.04801.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. *CoRR*, abs/2308.12032.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pages 346–363. IEEE.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Iris Miliaraki, Klaus Berberich, Rainer Gemulla, Kaustubh Beedkar, and Dhruv Gupta. 2014. Mgfsm: Large scale frequent sequence mining. `https://github.com/uma-pi1/mgfsm`. [Accessed 11-06-2024].

Iris Miliaraki, Klaus Berberich, Rainer Gemulla, and Spyros Zoupanos. 2013. Mind the gap: large-scale frequent sequence mining. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*, pages 797–808. ACM.

Wonsuk Oh and Girish N. Nadkarni. 2023. Federated learning in health care using structured medical data. *Advances in Kidney Disease and Health*, 30(1):4–16. AI in Kidney Disease: What Will the Future Bring? Part II.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Jacob Pfau, Alex Infanger, Abhay Sheshadri, Ayush Panda, Julian Michael, and Curtis Huebner. 2023. Eliciting language model behaviors using reverse language models. In *Socially Responsible Language Modelling Research*.

Md. Rafi Ur Rashid, Vishnu Asutosh Dasu, Kang Gu, Najrin Sultana, and Shagufta Mehnaz. 2023. Fltrojan: Privacy leakage attacks against federated language models through selective weight tampering. *CoRR*, abs/2310.16152.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. 2024. Rethinking LLM memorization through the lens of adversarial compression. *CoRR*, abs/2404.15146.

Minh N. Vu, Truc D. T. Nguyen, Tre' R. Jeter, and My T. Thai. 2024. Analysis of privacy leakage in federated large language models. In *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 1423–1431. PMLR.

Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Haifeng Chen, Wei Wang, and Wei Cheng. 2023. Large language models can be good privacy protection learners. *CoRR*, abs/2310.02469.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. 2024. Openfedllm: Training large language models on decentralized private data via federated learning.

Weichen Yu, Tianyu Pang, Qian Liu, Chao Du, Bingyi Kang, Yan Huang, Min Lin, and Shuicheng Yan. 2023. Bag of tricks for training data extraction from language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 40306–40320. PMLR.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *CoRR*, abs/2309.11325.

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023. FEDLEGAL: the first real-world federated learning benchmark for legal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3492–3507. Association for Computational Linguistics.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.