Revisiting Data Reconstruction Attacks on Real-world Dataset for Federated Natural Language Understanding

Zhuo Zhang[1,2,*], **Jintao Huang**[1,*], **Xiangjing Hu**[1]
Jingyuan Zhang[3], **Yating Zhang**[6], **Hui Wang**[2]
Yue Yu[2], **Qifan Wang**[5], **Lizhen Qu**[4,†], **Zenglin Xu**[1,2,†]
[1]Harbin Institute of Technology, Shenzhen, China
[2]Peng Cheng Lab, Shenzhen, China [3]Kuaishou, Beijing, China
[4]Monash University, Melbourne, Australia
[5]Meta AI, CA, USA [6]Independent Researcher
{iezhuo17, a764695611, starry.hxj, zhangjingyuan1994, yatingz89}@gmail.com
{wanghu06, yuy}@pcl.ac.cn wqfcr@fb.com
Lizhen.Qu@monash.edu.cn xuzenglin@hit.edu.cn

Abstract

With the growing privacy concerns surrounding natural language understanding (NLU) applications, the need to train high-quality models while safeguarding data privacy has reached unprecedented importance. Federated learning (FL) offers a promising approach to collaborative model training by exchanging model gradients. However, many studies show that eavesdroppers in FL could develop sophisticated data reconstruction attacks (DRA) to accurately reconstruct clients' data from the shared gradients. Regrettably, current DRA methods in federated NLU have been mostly conducted on public datasets, lacking a comprehensive evaluation of real-world privacy datasets. To address this limitation, this paper presents a pioneering study that reexamines the performance of these DRA methods as well as corresponding defense methods. Specifically, we introduce a novel real-world privacy dataset called FedAttack, which leads to a significant discovery: existing DRA methods usually fail to recover the original text of real-world privacy data accurately. In detail, the tokens within a recovery sentence are disordered and intertwined with tokens from other sentences in the same training batch. Moreover, our experiments demonstrate that different languages and domains also influence the performance of DRA. By discovering these findings, our work lays a solid foundation for further research into the development of more practical DRA methods and corresponding defenses.

Keywords: Federated Learning, Data Reconstruction Attack, Benchmark

1. Introduction

The strong natural language understanding (NLU) ability demonstrated by large language models (LLMs) (Peng et al., 2023; Touvron et al., 2023; Zhao et al., 2023) accelerates the deployment of real-world NLU applications in privacy-sensitive domains, such as law (Cui et al., 2023) and digital health (Singhal et al., 2023). However, there is a growing concern that such NLU models impose the risk of privacy leakage during training and inference (Liu et al., 2021). Although data protection regulations, such as GDPR (Voigt and Von dem Bussche, 2017), ensure a high level of private data protection, they obstruct the sharing of personal data for training NLU models. To address those issues, federated learning (FL) (Konečnỳ et al., 2016; McMahan et al., 2017a) becomes a promising solution with growing popularity to enable distributed client devices to train NLU models collaboratively, without sharing or transmitting their local data to a centralized place. However, within research communities, there exists an ongoing debate regarding the level of security that FL algorithms offer for real-world applications.

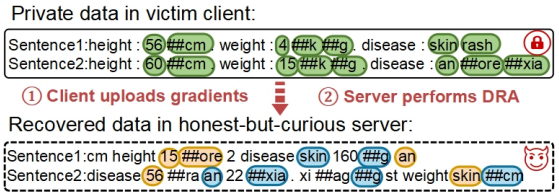

Figure 1: A demo of DRA in Federated NLU. During federated training, the server performs DRA on the gradients uploaded by the clients to recover private data. The sensitive tokens in the ground truth are highlighted in green. The tokens in orange come from instances other than the original ones (crosstalk issue) and the tokens in blue are accurately recovered but appear in the wrong positions.

Instead of sharing data directly, the majority of FL algorithms upload gradients or model updates computed on client devices to a central server iteratively during training. Data reconstruction attack (DRA) methods (Zhu et al., 2019; Liu et al., 2021) aim to recover client data from shared gradients by assuming that the server is *honest-but-curious* so that eavesdroppers may directly obtain client data via those methods, as shown in Figure 1. How-

*Equal contribution.
†Corresponding authors.

LREC-COLING 2024, pages 14080–14091

ever, these studies report empirical results only on public English datasets (Klimt and Yang, 2004; Pang and Lee, 2005; Socher et al., 2013; Warstadt et al., 2019), hence raising the following concerns for real-world applications.

Firstly, although DRA methods have achieved impressive recovery rates on public data, the effectiveness of DRA on real-world sensitive data, such as personal IDs, is uncertain. This is mainly because these types of data are not commonly found in the public datasets that have been used to test DRA. Therefore, it is essential to question the performance of DRA when dealing with privacy-rich data in real-world scenarios. Secondly, it is essential to consider that non-English languages may present different challenges for Federated Learning compared to English. Previous studies have indicated that the success rates of attacks on FL systems heavily rely on the mechanisms used to protect word embeddings and vocabularies (Zhang et al., 2022). For instance, the effectiveness of reconstruction attacks may vary when applied to different languages like Chinese, which requires a different word segmentation mechanism compared to English. Therefore, it is unclear whether the current reconstruction attacks are capable of compromising the privacy of non-English datasets. Thirdly, most attack methods assume the presence of a single client per attack or involve a small batch size (Zhu et al., 2019; Deng et al., 2021). However, in real-world settings, it is quite rare to encounter scenarios with only one client or small batch sizes. Therefore, it is crucial to examine the DRA methods when multiple clients are involved or when large batch sizes are used during training.

To address the above concerns, we construct a novel privacy-sensitive benchmark, coined FEDATTACK[1], for evaluating reconstruction attack methods on English and Chinese NLU tasks in real-world FL settings. Herein, we construct a novel dataset and assess the performance of state-of-the-art DRA and highly referenced defense methods using this dataset. It comprises resumes, legal documents, and medical consultation records in English and Chinese, conveying substantial de-identified personal information that still preserves their statistical patterns. The DRA methods are evaluated w.r.t. the trade-off between privacy protection and model utility, where the utility of information is evaluated via two NLU tasks: text classification (TC) and named entity recognition (NER). Through extensive experiments, we obtain the following novel and intriguing findings:

- We identify a previously *unreported* phenomenon, referred to as *crosstalk*, in reconstructed texts by the existing DRA methods (see Section 4.3). Namely, if the batch size is larger than one, it is likely that the reconstructed tokens appear in the wrong instances of the batch (Figure 1).
- The recovery rates for the public information mentions by the DRA methods are approximately five times higher than those for private information mentions. Hence, the real-world private data in our dataset is invaluable by imposing unique challenges for the DRA methods, detailed in Section 4.2.
- The success rates of reconstruction attacks in Chinese text are significantly higher than those in English text on average, and the rates vary among domains (see Section 4.2).
- Although the parameter-efficient tuning methods, such as LoRA (Hu et al., 2021), are not designed for FL defense originally, they exhibit a better trade-off between model utility and privacy protection than the widely used defense methods, such as differential privacy (DP) (Abadi et al., 2016) and gradient pruning (GP) (Lin et al., 2017) (see Section 4.5). We conjecture that this can be attributed to its limited size of tuned model parameters and the frozen word embedding layer.

2. Preliminaries

Federated Learning. The federated learning framework typically entails a server and multiple distributed clients. The server coordinates the training process and updates the global model while multiple clients upload their locally-trained model information (i.e., parameters or gradients) to the server. Assuming N clients with respective local dataset $D_i, (i \in [1, N])$, loss function $\mathcal{L}_i$, and global model $\mathcal{W}$, the federated training process can be expressed as follows.

At the beginning of each round t, the server distributes global parameters to every client of interest. Then the selected i-th client trains the latest global model $\mathcal{W}^{t-1}$ on its local dataset D_i and uploads corresponding gradients $\nabla\mathcal{W}_i^t = \frac{\partial\mathcal{L}_i}{\partial\mathcal{W}^{t-1}}$ to the server. The server aggregates all uploaded gradients and updates the global model with averaged gradients for the next training round. The above training process is repeated until specific criteria are met.

Data Reconstruction Attacks. Figure 2 depicts the workflow of the DRA that existed in federated NLU. DRA starts by randomly initializing a pair of dummy inputs (i.e., text embeddings $\mathbf{X}_e^{'}$ and label mappings $\mathbf{Y}_e^{'}$). When observing the uploaded ground-truth gradients $\nabla\mathcal{W}$, DRA recovers the privacy training data $(\mathbf{X}, \mathbf{Y})$ through an optimization-

[1] https://github.com/SMILELab-FL/FedAttack.

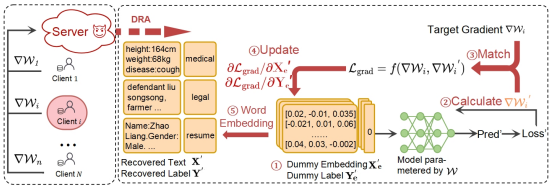

Figure 2: The workflow of DRA in Federated NLU.

based approach. Specifically, DRA uses dummy inputs at each optimization step to perform the normal forward and backward and calculates dummy gradients $\nabla\mathcal{W}'$. Then DRA tries to optimize the reconstruction loss $\mathcal{L}_{\text{rec}}$ to minimize the distance between $\nabla\mathcal{W}$ and $\nabla\mathcal{W}'$. DRA back-propagates this loss to update the dummy inputs. After multiple iterative updates, DRA can get best-matching dummy inputs. Finally, DRA recovers original data by mapping dummy inputs back to words closest to the federated model's embedding matrix.

This paper considers three representative DRA methods used in the federated NLU tasks. DLG (Zhu et al., 2019) is the first study to show private data could be leaked from the shared gradients, using L_2 distance between the dummy gradients and the true gradients as the loss function for continuous optimization. TAG (Deng et al., 2021) extends DLG's loss function and adds L_1 norm which prioritizes gradient matching in transformer layers closer to the input data, improving the efficiency of data recovery. These methods ignore the reconstruction of the text word order, leading to disordered recovered text. LAMP (Balunovic et al., 2022) attempts to recover the fluency of text words using alternating optimization. Specifically, the alternating optimization contains a continuous optimization for recovering tokens and a discrete optimization for recovering the order of tokens by using an auxiliary language model GPT2 (Radford et al., 2019) to reorder the recovered tokens. Through well-designed alternating optimization, LAMP is the sort-of-the-art DRA method that extracts original text from gradients. More details can be found in Appendix B.

3. FedAttack

Our ultimate goal is to construct a practical privacy attack dataset with a broad range of personally identifiable information, different training tasks, diverse domains, and multiple languages. However, disclosing such sensitive datasets is prohibited due to legal and ethical constraints. In the data collection (Section 3.1) and human annotation (Section 3.2), we present the source of FedAttack and process sensitive information that could potentially identify individuals as per legal regulations, thus enabling the usage of FedAttack for academic research. Next, we introduce a federated partitioning of FedAttack (Section 3.3) to verify the effectiveness of DRA methods and corresponding defenses in practical federated training processes.

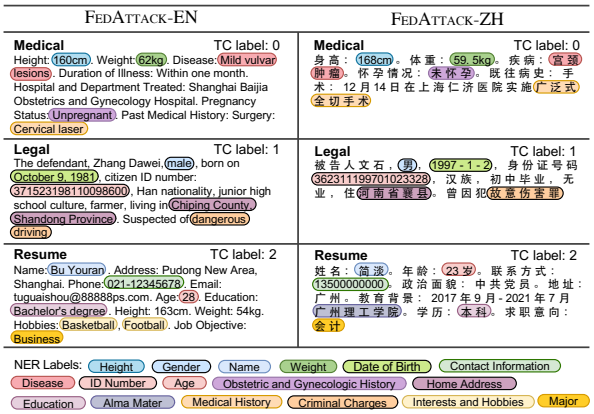

Figure 3: Examples of FedAttack. We provide English and Chinese text and annotate labels for TC and NER tasks, respectively. The complete label statistics for NER can be found in Table 2.

3.1. Data Collection

The Internet's ubiquity has made it easy to access public medical online consultations, case judgments, and resumes, which feature a broad range of personal information. Using this domain-specific data, we constructed a real-world privacy data set for federated Natural Language Understanding (NLU) tasks. Data is gathered using crawler tools and manually filtered for toxic, offensive, meaningless, or excessively lengthy content. This process yields 1,150 unlabeled raw Chinese samples.

To assess the effectiveness of the previous DRA methods across different languages, the raw Chinese dataset was translated into English via ChatGPT[2]. Note that ChatGPT just serves as a translation tool for efficiency. We subsequently employ five educated annotators to check, correct, and annotate these translated texts for translation quality. FedAttack comprises 2,300 instances of privacy-rich texts, equally split between English and Chinese.

Note that FedAttack is not intended for model training but aims to validate the effectiveness of previous DRA methods. Regarding this, the scale of FedAttack is sufficient to test the performance of DRA under different batch sizes in FL settings.

3.2. Human Annotation

Our human annotation includes two steps: (1) the manual de-identification step, which aims to process personally identifiable information (PII); and (2) the task annotation step, which yields two clas-

[2]https://chat.openai.com/chat

sical NLU tasks, including text classification and named entity recognition tasks.

Manual de-identification. According to the General Data Protection Regulation (GDPR) (Art.30), China's Personal Information Protection Law (PIPL), and similar regulations, successfully de-identified data is no longer considered personal data and can be shared with third parties (Pilán et al., 2022), including research organizations. The successful de-identification goes beyond removing directly identifying values, such as personally identifiable information regulated by privacy laws. It also encompasses addressing quasi-identifiers that could potentially lead to re-identification. Therefore, relying solely on automated de-identification techniques is insufficient in ensuring complete privacy protection.

To address this, we manually de-sensitize collected data rather than solely relying on automation technology. The conventional de-identification methods remove PII to achieve data anonymization, which may significantly deplete privacy information in the text and result in unrealistic privacy attack experiments. Our manual de-identification employs the substitution method to "retain" sensitive personal information. Specifically, we determine all PII (see the mentioned names in Table 2) in the text according to regulatory requirements and manually replace the data with false but format-consistent data. For example, we replace a real personal ID with a randomly selected non-existent personal ID in the same format. By adopting this manual approach, we prioritize privacy protection and compliance with regulations like GDPR and PIPL, ensuring the safeguarding of individuals' data in the research process. More details can be found in Appendix A.

We recruited a team of five educated people to perform the substituted de-identification. On average, de-identifying a sample takes about three minutes per person. Figure 3 shows examples of our manual de-identified dataset.

Tasks Annotation. After the manual de-identification, we consider two classic NLU tasks to comprehensively verify the performance of the DRA methods under various training objects: text classification (TC) and named entity recognition (NER). Domain names are utilized as classification labels for TC hence annotations aren't required. We establish a data schema and corresponding annotation guidelines for the NER task, using privacy-sensitive synthetic PII as entity types (refer to Table 2). Notably, FEDATTACK comprises both character-based privacy tokens and numerical privacy tokens (e.g., ID Number) that are scarce in publicly accessible datasets. The NER annotations were conducted through a web interface, and annotators received compensation for their work. The Kappa scores

Domain	# Instance			# Tokens		# Avg. Length	
	Train	Dev	Test	Non-Sen	Sen	ZH	EN
Medical	309	39	39	9405	2780	54.14	40.59
Legal	326	40	39	17926	5833	94.53	86.80
Resume	288	36	36	22341	5166	130.21	101.97

Table 1: The statistics of FEDATTACK. *Sen* denotes sensitive tokens.

Domains	Mention Name	# Mentions	Mention Type	# Avg Length	
				ZH	EN
Medical	Height	309	numeric	1.00	1.00
	Weight	309	numeric	1.47	1.04
	Disease	326	character-based	4.75	2.85
	Gynecologic History	93	character-based	2.99	1.89
	Medical History	39	character-based	4.90	2.40
Legal	Gender	308	character-based	1.00	1.00
	Criminal Charges	336	character-based	4.36	2.29
	Date of Birth	299	numeric	4.98	2.97
	Home Address	297	character-based	8.06	5.26
	ID Number	173	numeric	1.00	1.00
Resume	Name	288	character-based	2.37	1.98
	Major	263	character-based	4.40	2.19
	Age	195	numeric	1.33	1.12
	Contact Information	267	numeric	1.67	1.00
	Alma Mater	241	character-based	5.42	3.28
	Education	251	character-based	2.04	2.91
	Interests and Hobbies	168	character-based	4.77	2.20

Table 2: The statistic of privacy-sensitive entities in FEDATTACK.

(McHugh, 2012) among five annotators are 92% for the NER annotation. More details on NER labeling can be found in Appendix A.

3.3. Federated Partitioning

Our work aims to emulate DRA methods and corresponding defenses during the practical federated training process. To achieve this goal, we divide FEDATTACK into three federated clients based on their respective domains. Unlike previous studies that only consider one client, our federated system comprises an *honest-but-curious* server and *three* participants with varying domain datasets. We randomly divide the data into train/valid/test sets with an 8:1:1 ratio. In this way, we train the federated model on the training set, validate it on the valid set, and finally report model performance on the test set. The DRA methods mainly reconstruct training data from the gradients uploaded to the server by clients. Basic statistics of the FEDATTACK are presented in Table 1.

4. Experiments

This section begins with a quantitative analysis of three DRA methods on two diverse NLU tasks, varying in language, domain, and training batch sizes (Section 4.2). Next, we elucidate two critical challenges these DRA methods face when recovering raw text from uploaded gradients: crosstalk and out-of-order issues (Section 4.3). Using existing DRA methods, we also investigate the disparities in recovering text for public and private information. Finally, we assess different defense strategies and

illustrate their trade-offs between model utility and privacy risks (Section 4.5).

4.1. Experimental Setup

We have selected three representative DRA methods that are commonly used in Federated NLU tasks: (1) **DLG** (Zhu et al., 2019) is the first study to show that private data could be leaked from the shared gradients; (2) **TAG** (Deng et al., 2021) improves the efficiency of gradient recovery by adding L_1 norm to the gradient attacks optimization on the top of DLG; (3) **LAMP** (Balunovic et al., 2022) is the sort-of-the-art DRA method that extracts original text from gradients and reorders recovery text by using auxiliary language model priors.

In DRA experiments, we follow previous studies (Deng et al., 2021; Balunovic et al., 2022; Gupta et al., 2022) and use **ROUGE** (ROUGE, 2004) to evaluate DRA methods performance. We use the average F-score of ROUGE-1, ROUGE-2, and ROUGE-L as a measure of similarity between the recovered text and the original text in unigrams, bigrams, and the longest matching subsequence. Considering the richness of private entities contained in our FEDATTACK, we follow previous work (Gupta et al., 2022) using named entity recovery ratio (**NERR**) as a recovery metric for these sensitive entities. Specifically, NERR=0 indicates completely mismatched entities, while NERR=1 indicates perfectly recovered entities.

To ensure the accuracy and reproducibility of our attack results[3], we utilize official code published in Balunovic et al. (2022). When performing the DRA methods, we set the continuous optimization steps to 5,000 across all the attack experiments. We use Adam (Kingma and Ba, 2014) with a linear learning rate decay schedule applied every 50 steps. DLG and TAG use the gradient matching loss function reported in their papers. We chose the cosine gradient matching loss function for LAMP since it exhibited the best recovery performance and the discrete optimization activities every 375 steps in LAMP. In all DRA methods, we use FedSGD (McMahan et al., 2017b) as the basic FL algorithm to carry out the data reconstruction attack akin to previous studies (Deng et al., 2021; Balunovic et al., 2022; Gupta et al., 2022). For threat models, we use Bert-Base-Chinese[4] for Chinese tasks and Bert-Base (Devlin et al., 2018) for English tasks.

4.2. Main Results

We first systematically evaluate the performance of existing DRA methods with different training batch sizes (B) in FEDATTACK, where B=1 is the ideal attack setting. The experimental results are listed in Table 3 and Table 4. We observe two consistent phenomena in FEDATTACK containing different tasks and languages: (1) The recovery ability of all DRA methods decreases rapidly as the batch size increases. (2) LAMP is by far the most powerful attack method compared with other DRA methods, demonstrating the effectiveness of utilizing auxiliary language models in the attack process. These experimental results are consistent with the findings of previous work (Balunovic et al., 2022). However, we find that **the current DRA methods have limitations in terms of the real damages they can inflict, particularly at the typical training batch size of B=32.** For example, advanced LAMP barely recovers the original text correctly with B=32 (especially for privacy-sensitive word recovery NERR, with a maximum of only 8.62% correct recovery). LAMP struggles (maximum ROUGE-2 is only 0.19) despite using auxiliary language models to recover word order.

We then explore the performance of different attack methods on FEDATTACK-ZH and FEDATTACK-EN. Upon comparing Table 3 and Table 4, we discover that **existing attack methods are more effective at recovering the Chinese corpus as compared to the English corpus.** This performance gap could be attributed to differences in tokenizer techniques (see Section 4.3). The Chinese tokenizer is character-level, while the English tokenizer usually operates at the subword level, which needs the search for precise combinations of subwords. This presents a more challenging setting for English text recovery. Comparing different learning tasks, we do not observe a significant difference in the performance of DRA methods in TC or NER tasks, which suggests that DRA performance may be task-agnostic.

Unlike prior studies, our experiments are conducted in a more realistic federated DRA setting with one server and three clients. We compare the performance of diverse DRA methods on different clients (domains) in Table 5. As shown in Table 5, we can observe the experimental results for different domains are still consistent with the previous findings. Comparing different domains, DRA methods have the general trend in NERR: $D_{legal} > D_{resume} > D_{medical}$. We conjecture that the medical text contains more numeric type entities (e.g., height and weight), making it difficult for the DRA method to recover accurately.

4.3. Crosstalk and Out-of-order Issues

We next review the recovered text by the DRA methods and further investigate the underlying reasons for the unsatisfactory performance of DRA methods on FEDATTACK. We chose LAMP and the case

[3]Our data will be publicly available upon acceptance.
[4]https://huggingface.co/bert-base-chinese

Task	Method	B=1				B=8				B=32			
		R-1	R-2	R-L	NERR	R-1	R-2	R-L	NERR	R-1	R-2	R-L	NERR
TC	DLG	54.51	2.27	20.73	**6.72**	27.10	0.38	11.51	3.76	**17.30**	0.09	**8.37**	**2.55**
	TAG	54.56	2.25	20.89	6.17	27.04	0.38	11.59	**3.85**	17.14	0.11	8.24	2.41
	LAMP	**55.69**	**3.02**	**21.25**	6.67	**27.68**	**0.58**	**12.01**	3.64	16.33	**0.15**	8.00	1.90
NER	DLG	54.41	2.16	20.63	6.48	27.24	0.31	11.67	**3.64**	16.78	0.12	8.24	**2.41**
	TAG	54.43	2.35	20.72	6.43	27.29	0.33	11.66	3.42	16.79	0.13	8.27	2.36
	LAMP	**55.04**	**2.98**	**21.20**	**6.91**	**27.88**	**0.64**	**12.19**	3.28	**17.03**	**0.19**	**8.62**	1.85

Table 3: The attack performances of different DRA methods on FEDATTACK-ZH. B=x represents the training batch size x in federated training. R-x means the average F-score of ROUGE-x, and NERR refers to the named entity recovery ratio.

Task	Method	B=1				B=8				B=32			
		R-1	R-2	R-L	NERR	R-1	R-2	R-L	NERR	R-1	R-2	R-L	NERR
TC	DLG	**39.56**	2.03	19.85	**14.05**	20.79	0.39	11.27	7.13	14.23	0.10	8.95	4.46
	TAG	39.46	1.82	19.85	13.76	**20.87**	**0.40**	**11.32**	7.10	14.36	0.14	**9.12**	4.31
	LAMP	39.14	**2.18**	19.85	13.90	20.78	0.35	11.28	**7.18**	**14.41**	**0.12**	8.85	**4.98**
NER	DLG	**39.14**	1.90	19.36	13.60	**20.94**	0.23	11.19	**7.05**	**14.66**	0.17	8.95	4.21
	TAG	38.81	1.97	19.41	14.13	20.86	0.27	11.20	6.93	14.59	**0.18**	**8.99**	4.23
	LAMP	38.76	**2.16**	**19.52**	**14.75**	20.78	**0.34**	**11.45**	6.93	14.14	0.09	8.40	**5.01**

Table 4: The attack performances of different DRA methods on FEDATTACK-EN. B=x represents the training batch size x in federated learning. R-x means the average F-score of ROUGE-x, and NERR refers to the named entity recovery ratio.

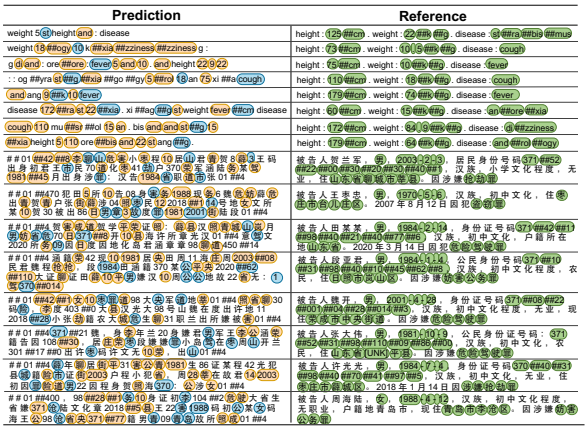

Figure 4: Some reconstructed texts of LAMP for the NER task with B=8. Green tokens are privacy-sensitive in the reference column, while blue ones in the prediction column indicate accurate predictions within the sentence. Orange tokens represent the correct prediction of sensitive entities from other sentences, which is called *cross-talk*.

of B=8. Figure 4 illustrates some randomly sampled recovery cases. In Figure 4, it is apparent that LAMP recovers sentences with almost no order either in FEDATTACK-EN or FEDATTACK-ZH, even though LAMP exploits an auxiliary language model to recover the order of the original sentences. Chu et al. (2022) claim that the current DRA method pays little attention to position embedding and the complex deep transformer structure with residuals.

However, Figure 4 illustrates another significant phenomenon not reported in previous studies: the tokens recovered by DRA methods are heavily intertwined in the same batch of sentences. We refer to this phenomenon as **crosstalk**. As an example, using the recovered sensitive privacy text, we can see that many privacy-sensitive tokens are perfectly recovered in Figure 4. For instance, in one sentence, LAMP recovers many privacy tokens from other sentences (e.g., each sentence in Figure 4 has more tokens marked in orange than those marked in blue). We also quantitatively measure the token recovery rate of DRA under different batch sizes, and results are shown in Table 6. We find that **the DRA strives to recover most tokens in uploaded gradients but crosstalk diminishes the quality of the recovered text and eventually presents a chaotic and meaningless reconstructed text.** Regrettably, the DRA methods are not designed to resolve the crosstalk problem, which makes it extremely challenging to achieve successful text reconstruction.

Regarding this, while FL is not a family of completely privacy-preserving learning algorithms, the existing DRA methods pose a limited threat to clients' privacy because of crosstalk and out-of-order problems. Moreover, Figure 4 indicates a severe leakage of sensitive tokens, which do not always convey private information explicitly. Despite that, this increases the chances of inferring private information based on these sensitive tokens. FE-

Task	Method	FEDATTACK-ZH						FEDATTACK-EN					
		Medical		Legal		Resume		Medical		Legal		Resume	
		R-L	NERR	R-L	NERR	R-L	NERR	R-L	NERR	R-L	NERR	R-L	NERR
TC	DLG	25.40	0.37	28.22	7.01	27.63	2.88	20.02	2.32	20.27	11.99	22.20	6.00
	TAG	25.35	0.56	28.20	6.15	27.53	2.76	20.39	2.50	20.15	11.78	22.23	6.00
	LAMP	27.47	0.65	27.64	7.16	27.97	2.94	20.29	2.04	20.25	12.91	22.31	5.47
NER	DLG	25.90	0.47	28.08	7.44	27.67	2.46	20.52	2.41	20.45	12.98	21.82	4.82
	TAG	26.03	0.37	28.10	7.08	27.64	2.28	20.51	2.32	20.24	13.05	21.89	4.49
	LAMP	27.09	0.65	28.87	7.44	28.70	2.86	20.70	2.32	20.28	11.71	21.91	5.67

Table 5: The domain performances of DRA (LAMP) on FEDATTACK with B=8.

	FEDATTACK-ZH			FEDATTACK-EN		
	# Tokens	# Recovered	Ratio	# Tokens	# Recovered	Ratio
B=1	67	64	95.02	55	53	94.36
B=8	532	505	95.77	412	394	93.93
B=32	1920	1829	95.96	1631	1562	94.09

Table 6: The batch-averaged token recovery ratio with different batch sizes.

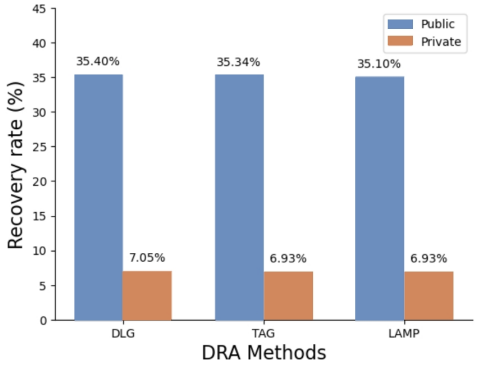

Figure 5: The recovery rates of public and private information vary with different DRA methods on the NER task within FEDATTACK-EN using B=8.

DATTACK can serves as a valuable real-world source for investigating such attacks.

4.4. Reconstruction Analysis for Private and Public Information

The FEDATTACK contains significant real-world private information, whereas publicly available datasets used by previous attack studies contain almost no such sensitive texts. Hence, we investigate the impact of distribution differences between private and public information in text on DRA methods.

Figure 5 illustrates the recovery rates of private and public information in text. The results indicate that **the existing DRA methods tend to excel in recovering mentions of non-private information while facing difficulties in recovering those for private information.** This phenomenon contributes to better performance of DRA methods on publicly available datasets (such as SST-2(Socher et al., 2013), CoLA(Warstadt et al., 2019), RottenTomatoes(Pang and Lee, 2005)), but less satisfactory results on FEDATTACK. However, current DRA methods have not exhibited significant differences in recovering natural versus texts for private information, which is impractical. From a practical perspective, attackers are primarily interested in extracting sensitive information, such as word sequences containing phrases like "my credit card number is...". Regarding this, the lack of real-world privacy attack data also hampers the development of robust defense methods within the FL community, thereby limiting the ability to protect the sensitive information of federated clients effectively.

4.5. Defense Against DRA Methods

Several defenses aim to mitigate the damage of the DRA methods, often leading to a reduction of model utility. When defenses are incorporated, this experiment assesses the trade-off between model utility and privacy risk in federated training. We employ three defenses against DRA methods: **Gradient Pruning** (GP) randomly zeroes elements of the gradient vector at a specified mask ratio. **Differential Privacy** (DP) employs noise addition to the gradients. We choose noise intensity and masking rates to achieve 80% performance without defense, considering the extensive range of options available. Additionally, we incorporate the parameter-efficient tuning method **LoRA** (Hu et al., 2021), as demonstrated by Zhang et al. (2023c), to bolster defense against DRA methods further. We utilize ROUGE-L to measure the risk of privacy breaches. We employ the NER task from FEDATTACK-EN and the F1-Score as the utility metric for model utility.

Figure 6 reports the trade-off results between model utility and privacy leakage when using different defenses to resist existing DRA methods on FEDATTACK. As illustrated in Figure 6, we observe that **LoRA exhibits a superior trade-off between privacy preservation and model utility relative to other defenses.** We also further explore the reasons why LoRA can resist DRA methods. LoRA freezes the pre-trained model weights and introduces lightweight trainable rank decomposition matrices into every transformer layer, greatly reducing the number of uploaded gradients in FL. Table 7

Figure 6: The trade-off between model utility and privacy leakage when using different defenses to resist existing DRA methods on FEDATTACK. The x-axis represents the risk of privacy leakage, while the y-axis denotes model utility. The closer to the upper left corner of the diagram indicates that the defense can achieve a better trade-off.

	FT	FE	LoRA+E	LoRA
ROUGE-L	11.28	4.84	6.78	1.06

Table 7: The averaged attack performance under different tuning methods. FT denotes full fine-tuning, FE refers to training with freezing word embedding layer, and LoRA+E denotes training word embedding when using LoRA.

illustrates the average DRA method performance with various tuning methods. When contrasted with LoRA+E and FT, the lightweight upload parameters pose challenges for DRA methods in recovering original data (also corroborated by GP). Compared to LoRA and LoRA+E, we can see the importance of the frozen word embedding layer in countering DRA. Consequently, LoRA could offer superior privacy protection owing to (1) lightweight upload gradients and (2) the presence of frozen word embeddings.

5. Related Work

Federated Natural Language Understanding. Due to its decentralized and private nature, federated learning (Konečnỳ et al., 2016; McMahan et al., 2017a) has gained prominence in recent years and appeals particularly to privacy-sensitive natural language understanding applications (Sui et al., 2020; Ge et al., 2020; Long et al., 2020; Basu et al., 2021; Zhang et al., 2023b). This emerging research field, referred to as Federated NLU (Liu et al., 2021; Lin et al., 2022), has attracted substantial attention, with a variety of proposed approaches primarily aimed at addressing challenges such as data heterogeneity (Ji et al., 2019; Zhang et al., 2022), system heterogeneity (Liu et al., 2022; Cai et al., 2022), and limited resources associated (Zhang et al., 2023c) with the federated training of pre-trained language models (Liu et al., 2019; Radford et al., 2019). The landscape of Natural Language Understanding (NLU) is being revolutionized by Large Language Models (Scao et al., 2022; Touvron et al., 2023) (LLMs). In this context, Federated NLU emerges as a promising framework for training privacy-preserving LLMs on data with privacy concerns (Zhang et al., 2023a). Contrary to previous studies, our work focuses on the privacy-preserving capabilities of federated NLU, due to its importance to the field.

Data Reconstruction Attacks. Although FL forges new pathways for collaboration among the distributed clients, it has recently faced criticism for relying heavily on shared gradients as a privacy measure. A significant portion of recent FL research has been dedicated to data reconstruction attacks (Zhu et al., 2019; Zhao et al., 2020; Geiping et al., 2020; Deng et al., 2021; Balunovic et al., 2022; Gupta et al., 2022), which expose the possibility of an attacker reconstructing local data from uploaded gradients.

DRA methods in FL can be categorized into two types (Lyu et al., 2022) based on the nature of the server: (1) *honest-but-curious* setting, where a compromised server surreptitiously recovers the victim client's training data by passively observing uploaded gradients; and (2) *malicious* setting, where the server illicitly steals the victim client's training data by manipulating the model parameters or gradients. Our research predominantly focuses on the honest-but-curious setting within DRA methods, considering its heightened threat level and detection difficulty in federated training as compared to the malicious setting (Balunovic et al., 2022; Gupta et al., 2022).

DRA methods were first proposed and rapidly developed in computer vision (Zhu et al., 2019; Zhao et al., 2020; Geiping et al., 2020). Several visual DRA benchmarks (Huang et al., 2021; Ovi and Gangopadhyay, 2023; Yang et al., 2023) have also been proposed to examine the limitations of various DRA methods and to evaluate their qualitative performance in the reconstruction of input images. Nevertheless, the development of DRA methods in Federated NLU has been sluggish, with research efforts recently initiated (Zhu et al., 2019; Deng et al., 2021; Balunovic et al., 2022). Deep Leakage from Gradients (Zhu et al., 2019) (DLG) pioneered text extraction through gradients, demonstrating potential leakage in masked language modeling. TAG (Deng et al., 2021) extended DLG by introducing a regularization term to strengthen gradient matching in layers closer to the input original data. LAMP (Balunovic et al., 2022) incorporates language-prior information and utilizes language models to restore token order based on recovered tokens, demonstrating potent attack capabilities in Federated NLU.

Although these DRA methods have shown impressive reconstruction results, they have only been tested on public English text classification

datasets, leaving their performances on other languages, NLU tasks, and privacy-rich data in doubt. Our work addresses this knowledge gap by providing the real-world privacy dataset and systematically testing DRA methods across various tasks, domains, and languages in practical federated settings.

6. Conclusion

This paper constructs a novel benchmark FEDATTACK, including a bilingual real-world dataset, for evaluating the performance of DRA methods on federated NLU tasks. In contrast to assessing reconstruction attacks on public datasets in prior studies, our extensive experiments show that the token sequences conveying private information impose unique challenges to the evaluated DRA methods such that the corresponding recovery rates for private information are approximately five times lower than those for non-private information. Moreover, our empirical studies discover an unreported type of error called "crosstalk" in the token sequences reconstructed by all assessed DRA methods. In addition, the attack success rates depend on language specific properties such that it is significantly easier to reconstruct Chinese texts than English ones. We also compare LoRA with the SOTA defense methods and find out that LoRA provides a better trade-off between model utility and private protection than those defense methods.

Ethics Statement

All source texts for our FEDATTACK are from publicly available websites and have been appropriately anonymized. We do not analyze the content of private information or possible individuals in any way (although also virtually). Although our primary aim is to reassess the privacy risks inherent in FL, FEDATTACK may also facilitate the development of even more sophisticated attack methods from the attacker's perspective. Therefore, we aspire to inspire the design of defense mechanisms that can offer more strict privacy guarantees for clients during federated model training.

Acknowledgements

This work was partially supported by an Open Research Project of Zhejiang Lab (NO.2022RC0AB04), a Major Key Project of PCL (No. PCL2023A09), and a key program of fundamental research from Shenzhen Science and Technology Innovation Commission (No. JCYJ20200109113403826).

7. Bibliographical References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Mislav Balunovic, Dimitar Dimitrov, Nikola Jovanović, and Martin Vechev. 2022. Lamp: Extracting text from gradients with language model priors. *Advances in Neural Information Processing Systems*, 35:7641–7654.

Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, and Zumrut Muftuoglu. 2021. Privacy enabled financial text classification using differential privacy and federated learning. *arXiv preprint arXiv:2110.01643*.

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. 2022. Autofednlp: An efficient fednlp framework. *arXiv preprint arXiv:2205.10162*.

Hong-Min Chu, Jonas Geiping, Liam H Fowl, Micah Goldblum, and Tom Goldstein. 2022. Panning for gold in federated learning: Targeted text extraction under arbitrarily large-scale aggregation. In *The Eleventh International Conference on Learning Representations*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Jieren Deng, Yijue Wang, Ji Li, Chao Shang, Hang Liu, Sanguthevar Rajasekaran, and Caiwen Ding. 2021. Tag: Gradient attack on transformer-based language models. *arXiv preprint arXiv:2103.06819*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Suyu Ge, Fangzhao Wu, Chuhan Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2020. Fedner: Medical named entity recognition with federated learning. *arXiv preprint arXiv:2003.09288*.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947.

Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. 2022. Recovering private text in federated learning of language models. *arXiv preprint arXiv:2205.08514*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. 2021. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34:7232–7241.

Shaoxiong Ji, Shirui Pan, Guodong Long, Xue Li, Jing Jiang, and Zi Huang. 2019. Learning private neural language modeling with attentive aggregation. In *2019 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. 2022. FedNLP: Benchmarking federated learning methods for natural language processing tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 157–175, Seattle, United States. Association for Computational Linguistics.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. 2017. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*.

Ming Liu, Stella Ho, Mengqi Wang, Longxiang Gao, Yuan Jin, and He Zhang. 2021. Federated learning meets natural language processing: a survey. *arXiv preprint arXiv:2107.12603*.

Ruixuan Liu, Fangzhao Wu, Chuhan Wu, Yanlin Wang, Lingjuan Lyu, Hong Chen, and Xing Xie. 2022. No one left behind: Inclusive federated learning over heterogeneous devices. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3398–3406.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. 2020. Federated learning for open banking. In *Federated learning*, pages 240–254. Springer.

Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and S Yu Philip. 2022. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017a. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017b. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Pretom Roy Ovi and Aryya Gangopadhyay. 2023. A comprehensive study of gradient inversion attacks in federated learning and baseline defense strategies. In *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. 2020. Feded: Federated learning via ensemble distillation for medical relation extraction. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2118–2128.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Haomiao Yang, Mengyu Ge, Dongyun Xue, Kunlan Xiang, Hongwei Li, and Rongxing Lu. 2023. Gradient leakage attacks in federated learning: Research frontiers, taxonomy and future directions. *IEEE Network*.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Guoyin Wang, and Yiran Chen. 2023a. Towards building the federated gpt: Federated instruction tuning. *arXiv preprint arXiv:2305.05644*.

Zhuo Zhang, Xiangjing Hu, Lizhen Qu, Qifan Wang, and Zenglin Xu. 2022. Federated model decomposition with private vocabulary for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6413–6425.

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023b. 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)*, pages 3492–3507.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023c. Fedpetuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Annual Meeting of the Association of Computational Linguistics 2023*, pages 9963–9977. Association for Computational Linguistics (ACL).

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. *Advances in neural information processing systems*, 32.

8. Language Resource References

Klimt, Bryan and Yang, Yiming. 2004. *The enron corpus: A new dataset for email classification research*. Springer. PID http://www.cs.cmu.edu/ enron/.

Pang, Bo and Lee, Lillian. 2005. *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales*. PID http://www.cs.cornell.edu/people/pabo/movie-review-data/.

Socher, Richard and Perelygin, Alex and Wu, Jean and Chuang, Jason and Manning, Christopher D and Ng, Andrew Y and Potts, Christopher. 2013. *Recursive deep models for semantic compositionality over a sentiment treebank*. PID https://gluebenchmark.com/.

Warstadt, Alex and Singh, Amanpreet and Bowman, Samuel R. 2019. *Neural network acceptability judgments*. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info .... PID https://gluebenchmark.com/.

A. Annotation Details

Personally Identifiable Information (PII), refers to any data that can be used to identify an individual considered sensitive or confidential. This can include a person's name, identification number, date of birth, address, phone number, email address, social media ID, and other similar data. We refer to different data protection laws (i.e., the EU's GDPR, the US's CCPA, and China's PIPL) and find that PII contains similar content. Therefore, we mainly rely on the provisions of these laws to identify the relevant privacy-sensitive entities in our collected FEDATTACK. Table 2 shows the types of sensitive entities for different domains in FEDATTACK.

After defining the private entities, we solicit the annotators to substitute them with corresponding formats. As indicated in Table 2, the privacy information encompasses character and numeric types. Specifically, for character types such as names, we manually selected them from a corpus of publicly available Chinese personal names. For numeric types such as weight, we introduced a certain degree of noise (e.g., replacing the real weight of 76kg with a substituted weight of 75kg). Moreover, genuine IDs are substituted with non-existent IDs that adhere to the format specified in China's national standard for identification, GB11643-1999. All the substituted IDs' birthdates are set before the year 2023, making them unused for future identification purposes. Consequently, we use these substituted entities to annotate our NER tasks. Notably, our dataset comprises character-based privacy tokens and numerical privacy tokens (e.g., ID Number) that are scarce in publicly accessible datasets. Table 2 also shows the statistics of these two types of private tokens in different domains.

B. DRA Methods

Despite the success of keeping data locally to protect privacy in federated learning, prior studies (Zhu et al., 2019; Deng et al., 2021; Gupta et al., 2022; Balunovic et al., 2022) demonstrate the risk of data reconstruction from the perspective of user-uploaded gradients. We denote the ground-truth data as $(\mathbf{X}, \mathbf{Y})$, recovered data as $(\mathbf{X}^{'}, \mathbf{Y}^{'})$ and ground-truth gradient $\nabla\mathcal{W}$, dummy gradient w.r.t dummy data $\nabla\mathcal{W}^{'}$. To recover private data, current works commonly employ an optimization strategy to shorten the distance between gradients $\nabla\mathcal{W}$ and $\nabla\mathcal{W}^{'}$, which is formulated as:

$$\mathcal{L}_{\text{grad}} = f(\nabla\mathcal{W}, \nabla\mathcal{W}^{'}, \nabla\mathcal{W}'^{*}), \tag{1}$$

$$\mathbf{X}^{'^{*}}, \mathbf{Y}^{'^{*}} =_{\mathbf{X}', \mathbf{Y}'} \mathcal{L}_{\text{grad}}, \tag{2}$$

where $f(\cdot)$ denotes some distance measure, such as L_2 (Zhu et al., 2019), L_1 and cosine distances. DLG (Zhu et al., 2019) first attempted to reconstruct data from gradients by defining δ as the L_2 distance:

$$\mathbf{X}^{'^{*}}, \mathbf{Y}^{'^{*}} =_{\mathbf{X}', \mathbf{Y}'} \left\| \nabla\mathcal{W} - \nabla\mathcal{W}^{'} \right\|^2 \tag{3}$$

While TAG (Deng et al., 2021) extended DLG on transformer-based models with L_1 constraint:

$$\begin{aligned} \mathbf{X}^{'^{*}}, \mathbf{Y}^{'^{*}} =_{\mathbf{X}', \mathbf{Y}'} & \left\| \nabla\mathcal{W} - \nabla\mathcal{W}^{'} \right\|^2 \\ & + \alpha_{\text{TAG}} (\nabla\mathcal{W}) \left\| \nabla\mathcal{W} - \nabla\mathcal{W}^{'} \right\|, \end{aligned} \tag{4}$$

where α_{TAG} is a hyperparameter.

Additionally, the recovered sentences' fluency was considered in LAMP (Balunovic et al., 2022). LAMP (Balunovic et al., 2022) alternated continuous optimization for token selection and discrete optimization guided by an auxiliary language model for token arrangement, ensuring an effective reconstruction result from gradients. Based on the widely-used gradient distance $\mathcal{L}_{\text{grad}}$ (Eq. 1), the embedding length regularization $\mathcal{L}_{\text{reg}}$ (Eq. 5) is introduced in the continuous reconstruction optimization:

$$\mathcal{L}_{\text{reg}}(\mathbf{X}) = \left(\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i\|_2 - \frac{1}{V} \sum_{j=1}^{V} \|\mathbf{e}_j\|_2 \right)^2, \tag{5}$$

where $\mathbf{x}_i$ and $\mathbf{e}_i$ denote the i-th token embedding in $\mathbf{X}$ and vocabulary respectively, n is the number of tokens in $\mathbf{X}$, and V is the size of the vocabulary. Hence the continuous reconstruction loss is defined as $\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{reg}}$.

In discrete optimization, LAMP (Balunovic et al., 2022) employed various discrete sequence transformations for token rearrangement and candidate generation, selecting one with both low reconstruction loss and perplexity under an auxiliary language model, which is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{LAMP}} &= \mathcal{L}_{\text{rec}} + \alpha_{\text{lm}} \mathcal{L}_{\text{lm}} \\ \mathbf{X}^{'^{*}} &=_{\mathbf{X}'} \mathcal{L}_{\text{LAMP}}, \quad \text{for } \mathbf{X}^{'} \text{ in candidates,} \end{aligned} \tag{6}$$

where α_{lm} is a hyperparameter.