Privacy-Aware Document Visual Question Answering

Rubèn Tito[*†] Khanh Nguyen[*†] Marlon Tobaben[‡†] Raouf Kerkouche[§]
Mohamed Ali Souibgui[*] Kangsoo Jung[¶] Lei Kang[*] Ernest Valveny[*] Antti Honkela[‡]
Mario Fritz[§] Dimosthenis Karatzas[*]

rperez@cvc.uab.cat knguyen@cvc.uab.cat marlon.tobaben@helsinki.fi

Abstract

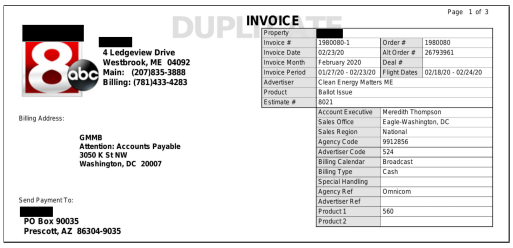
Document Visual Question Answering (DocVQA) is a fast growing branch of document understanding. Despite the fact that documents contain sensitive or copyrighted information, none of the current DocVQA methods offers strong privacy guarantees.

In this work, we explore privacy in the domain of DocVQA for the first time. We highlight privacy issues in state of the art multi-modal LLM models used for DocVQA, and explore possible solutions.

Specifically, we focus on the invoice processing use case as a realistic, widely used scenario for document understanding, and propose a large scale DocVQA dataset comprising invoice documents and associated questions and answers. We employ a federated learning scheme, that reflects the real-life distribution of documents in different businesses, and we explore the use case where the ID of the invoice issuer is the sensitive information to be protected.

We demonstrate that non-private models tend to memorise, behaviour that can lead to exposing private information. We then evaluate baseline training schemes employing federated learning and differential privacy in this multi-modal scenario, where the sensitive information might be exposed through any of the two input modalities: vision (document image) or language (OCR tokens).

Finally, we design an attack exploiting the memorisation effect of the model, and demonstrate its effectiveness in probing different DocVQA models.



Question: What is the provider of this document?
GT Answer: WMTW
Non-private Model: WMTW
Private Model: Meredith Thompson

Figure 1. The risk of malicious attacks on trained DocVQA models, such as exploiting memorization, is evident in the PFL-DocVQA dataset. Adversaries can cue the model through the visual modality, to invoke memory and reveal sensitive information that is not explicitly in the document (e.g. in this example the provider's name). We show how this behaviour can be exploited to attack the models, and take first steps to mitigate the problem.

1. Introduction

Automatic processing of documents enables the vast majority of daily interactions with and between institutions. From a research viewpoint, document understanding is a multi-modal endeavour combining reading systems and the visual analysis of document images with language processing and, more recently, language-based interaction. Document Visual Question Answering (DocVQA) was introduced in 2019 [36, 37] and quickly reshaped the state of the art, by introducing specialised, large scale, multi-modal LLMs for document understanding [22, 44, 64].

Despite the fact that documents contain sensitive or copyrighted information, none of the current DocVQA methods offers strong privacy guarantees. On the contrary, such models tend to memorise information, and often hallucinate responses drawing information from their training

[*]Computer Vision Center, Universitat Autonoma de Barcelona, Spain
[†]These authors contributed equally to this work.
[‡]University of Helsinki, Finland
[§]CISPA Helmholtz Center for Information Security, Germany
[¶]French Institute for Research in Computer Science and Automation (INRIA), France

data, as illustrated in Fig 1.

Privacy-preserving methods have advanced considerably over the past decade, nevertheless the DocVQA scenario presents important challenges to overcome. On one hand, privacy-preserving methods are not specifically designed for multi-modal scenarios, where the sensitive information might be exposed through any of the input modalities, or a combination of them. On the other hand, the models employed for DocVQA tend to be large and cumbersome, and require long pre-training and fine-tuning stages.

Moreover, documents in real-life cannot be freely exchanged. More often than not, different entities have access to distinct sets of documents that cannot be shared due to legal reasons. As a result, most document analysis datasets tend to be small, or focus on non-realistic, out of copyright document collections. Collaborative learning approaches, that do not require centralising the training data, are a valid alternative that would allow exploiting real-life data; nevertheless, such methods are not currently used by the document analysis community.

In this work, we highlight privacy issues in state of the art multimodal LLM models used for DocVQA, and propose possible solutions. In addition, we employ a federated scheme for learning, that reflects the real-life distribution of documents.

A typical real-life widely employed scenario for document understanding methods is invoice processing. Typically service providers issue invoices to clients, who use automated tools to extract information and take appropriate actions. We exploit this scenario and focus on the case where the sensitive information is the invoice provider. This information can potentially be exposed through vision (logo, layout of the invoice, colour scheme used, etc) and / or language (the provider's name, their vat number, address, telephone or any other identifying information in the textual part of the document). In this scenario, different clients receive invoices from distinct providers, highlighting the unbalanced and non i.i.d. distribution of data in the real world, in terms of invoice providers.

In this work we explore the application of privacy-preserving methods for fine-tuning a pre-trained state-of-the-art DocVQA model in a federated setting. Specifically, our contributions are:

- We put forward a new dataset for private, federated, DocVQA, focused on the real-life scenario of invoice processing
- We demonstrate that state-of-the-art non-private models exhibit a memorisation effect, that can lead to exposing sensitive information, and can be used to attack the model
- We design a series of attacks exploiting the memorisation behaviour of the models.
- We evaluate different training methods employing Federated Learning (FL) and Differential Privacy (DP) to train a SoA DocVQA model with privacy guarantees. We evaluate these methods using our proposed attacks, and show that FL+DP can mitigate privacy issues to certain degree.

2. Related Work

Document Visual Question Answering. DocVQA has gained a lot of attention as a unified approach based on answering natural language questions to extract information from documents. Consequently, many datasets are nowadays available tackling different domains and challenges such as industry documents [36, 37, 56, 57], infographics [38], multidomain [59, 60], open-ended questions [53], multilingual [45], multipage [57] or collections of documents [55]. Despite the existence of many small and medium sized datasets one of the major challenges is still the lack of a large scale generic dataset that can be used for multiple scenarios. One of the main reasons for that is the sensitive content of many documents and the copyright issues they are subject to, that prevent most document holders from publicly releasing their data. In this direction, we believe that our proposal based on federated learning along with privacy-preserving methods can be a way to facilitate the use of distributed and/or private datasets among different entities.

Existing models for DocVQA have evolved from text-only span-prediction models [11], to multimodal generative methods [22, 44, 64] that leverage different self-supervised pretraining tasks to align all the modalities over large scale datasets. Following this trend we will use VisualT5 (VT5), a multimodal generative method that has a strong performance on the DocVQA task, while it is easy to fine-tune allowing us to focus on federated learning and privacy-preserving techniques on multimodal data.

Differentially Private Federated Learning. Federated Learning (FL) [39, 48], allows collaborative model training among several entities (also known as clients) without sharing the entire data set of a client. Instead, only the trained model and its updates are shared between a central server and the different clients. Even though federated learning is more private than the centralized approach, many attacks have shown that a significant amount of information can still be inferred from the updates/models shared between the clients and the server during training or afterwards [41, 42, 68]. Adversaries can be either a participant, the server, or an external entity with access to the released trained model. The so-called gradient inversion attack aims at reconstructing the records included in the dataset of a client [15, 17, 32, 34, 41, 61, 67, 68]. Moreover, property inference attacks [26, 41] can be used to determine whether a property about a group of people is included in the client's dataset. Another type of attack is to infer whether a specific record is included in the dataset of a client, which is called membership inference [7, 42, 49]. Recently, membership

inference attacks for multi-modal models [21, 27] have been proposed. Both attacks use a pre-trained image-text model to identify matching pairs, but these attacks are not applicable to our DocVQA setting. In our work we aim to identify group membership (whether a provider is included in the dataset) in contrast to identifying specific items (documents in our case). In addition, our attack testing dataset, unlike previous methods, excludes explicit training set data and only includes data from the targeted group's distribution, providing a more realistic evaluation of membership inference.

In order to mitigate privacy attacks, Differential Privacy (DP) [14] has emerged as the gold standard for formalizing privacy guarantees. However, DP introduces a trade-off between utility and privacy as model training under DP requires clipping updates and adding random noise, which have an impact on the accuracy of the model. The current SOTA approaches for training large models with high utility under DP rely on transfer learning and utilize models pretrained on large public datasets that are then fine-tuned [65] on private datasets. These approaches have been shown to be effective in multiple domains where large public datasets are available such as NLP [33, 66] and computer vision [9, 10, 30, 40] even when the private dataset is small [58]. Parameter-efficient fine-tuning [19] using adaptors such as LoRA [20] has been shown to be competitive in (federated) transfer learning under DP [58, 66]. Similarly, compression techniques decrease the size of the model or its updates, thus reducing the sensitivity and consequent noise introduced by DP [24, 25].

3. PFL-DocVQA Dataset

The PFL-DocVQA dataset is designed to perform DocVQA in a federated learning and differential privacy set-up. It is the first dataset for privacy-aware DocVQA that aims to address privacy leakage issues in a realistic scenario. The dataset comprises invoice document images along with their OCR transcriptions and a set of question/answer pairs. In our setting, the sensitive data that needs to be protected is the information related to the invoices' provider. Therefore, a malicious attack should not be able to exploit the trained model to reveal sensitive provider information or to discover if invoices from a particular provider were included in the training set.

The current version of the PFL-DocVQA dataset contains a total of $336,842$ question-answer pairs framed on $117,661$ pages of $37,669$ documents from $6,574$ different invoice providers. The document images used in the dataset are sourced from the DocILE dataset [50], which is designed for Key Information Localization and Extraction (KILE). The DocILE dataset comprises $939,147$ real business documents from public data sources [62, 63], out of which $6,680$ are annotated with key-value pairs (i.e., a unique identifier (key) associated with a corresponding piece of data (value) within the document). We verified the annotations of the DocILE dataset to avoid OCR errors in the provider name and omitted samples that did not specify the provider's name. This resulted in using $4,968$ of the DocILE annotated documents. We extended this annotated set by labeling an additional set of $32,701$ documents. The key-value annotation of the invoices was done through Amazon Key-Information extraction tool [3], and was manually verified to guarantee its accuracy. Finally, we grouped documents by provider in a semi-automatic way.

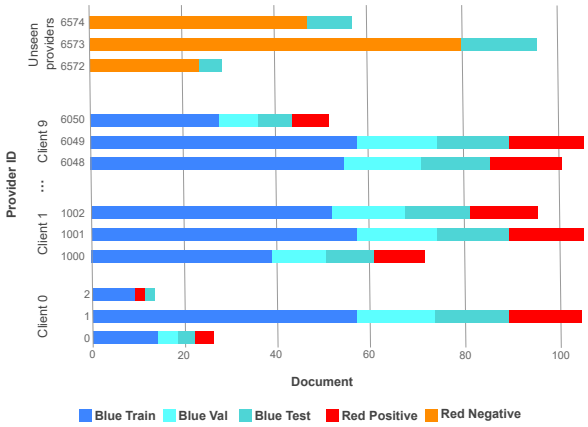


Figure 2. Distribution of providers and documents across different groups and splits. Every bar represents a specific provider, which contains a set of documents. The BLUE dataset is used for training the models, while the RED data is used for the attacks.

The key-value pairs were used to construct the questions and answers of the PFL-DocVQA dataset. Questions are formed to inquire about the key, and the answers are the corresponding values. Questions are generated semi-automatically, by defining multiple templates for each key, and further rephrasing them using a LLM [43] to achieve linguistic variability. 20 templates are defined for each key. For each key-value pair in the dataset, we randomly select one template to create the final question-answer pairs.

To facilitate the development and evaluation of privacy attacks, the PFL-DocVQA is composed of two parts, or sub-datasets, a base dataset to train and evaluate models and an extension dataset that is used for membership inference attacks of these models. We refer to the base dataset as the "BLUE" dataset and to the extension dataset as the "RED" dataset. To construct the RED and BLUE datasets we split the providers into in-distribution $\mathcal{D}_{in}$ and out-of-distribution $\mathcal{D}_{out}$, where $\mathcal{D}_{in}$ are the providers seen during training and $\mathcal{D}_{out}$ are the providers that were not seen.

The BLUE data consists of a training set that is divided among N clients (in our case we use $N = 10$), a validation set and a test set. The training set of each of the N clients

contains invoices sampled from a different subset of $\mathcal{D}_{in}$ providers, resulting in a highly non i.i.d. distribution. The BLUE validation includes documents with $\mathcal{D}_{in}$ distribution. In the BLUE test set, we include documents from both $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ providers.

The RED dataset is created by selecting half of its documents as $\mathcal{D}_{in}$ providers (RED positive documents), i.e. documents from providers that appear in the BLUE training data. The other half of the RED data consists of documents from $\mathcal{D}_{out}$ providers (RED negative documents), i.e. documents from providers not used for the BLUE training data. RED data is split into a training and test set. The different sets and clients of PFL-DocVQA are illustrated in Figure 2. We also provide more details in the supplementary material.

4. Baseline DocVQA Model

4.1. Document VQA

The DocVQA task is defined as the process wherein: given a question q_i asked over a document image d_i, the method f_θ must predict the correct answer a_i. We will refer to $I = \{(d_i, q_i, a_i))\}$ as the set of valid input tuples.

Mainstream DocVQA methods include generative models based on language model architectures, where a model f_θ is trained by minimizing the loss function: $L(\theta) = -\log p_\theta(a_i|d_i, q_i)$. In the inference step, it is assumed that the model can provide the answer a_i along with a confidence score $conf_i$ based on $p_\theta(a_i|d_i, q_i)$.

To evaluate the DocVQA task, in addition to standard accuracy, we use the ANLS metric introduced in [4, 5] which is a soft version of Accuracy (ACC) that smoothly penalizes OCR errors during the text recognition step based on the Normalized Levenshtein Similarity (NLS).

4.2. The Visual T5 Model

As a baseline for the DocVQA task we make use of a multimodal generative model named Visual T5 (VT5), a simplified version of the Hi-VT5 [57] that leverages well-known existing state-of-the-art methods. The adoption of a multimodal approach allows us to exploit the different modalities, but it introduces the challenge of safeguarding private information across these modalities. Additionally, as a generative method, VT5 can produce a wide range of text, not limiting responses to predefined categories, thus making the preservation of sensitive information more challenging.

The backbone is based on T5 [46] augmented with spatial information, and visual features from document images encoded with DiT [31]. As shown in Fig. 3, the tokenized question, OCR tokens and encoded image patches are concatenated and fed into the encoder-decoder model to generate the answer following an autoregressive mechanism.

The T5 backbone is initialized with pretrained weights on the C4 dataset [46] and the visual DiT is initialized with

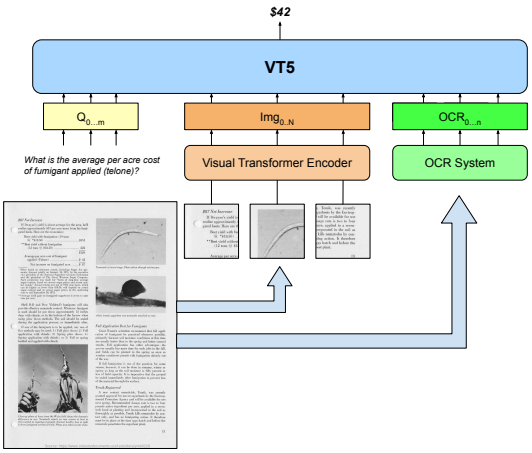


Figure 3. Architecture of the baseline method.

pre-trained weights for document classification. We call this initial model *zero-shot baseline*, VT5_0. Then, we finetune the model on the Single Page DocVQA (SP-DocVQA) dataset [36, 37] for ten epochs. We call this model *centralized non-DP model*, VT5_C.

5. Provider Membership Inference Attack

In this section we emphasize the privacy risks of the proposed DocVQA task in the real-world scenario where models have access to sensitive information. First, we demonstrate the vulnerability of the centralized non-private model VT5_C to leak private training data through memorizing specific information of the provider, where overfitting is known to play a relevant role. Then, we propose different approaches to attack the model privacy via provider membership inference, that aims at differentiating the models' behavior between providers seen and not seen during training.

Model	RED		RED Positive		RED Negative		ΔACC	ΔANLS
	ACC	ANLS	ACC	ANLS	ACC	ANLS		
VT5_0	37.72	43.66	37.62	44.10	37.84	43.18	0.22	0.92
VT5_C	81.40	90.17	85.92	93.68	76.53	86.48	**9.39**	**7.20**

Table 1. **DocVQA Performance of Non-Private Models on RED.** Metrics are reported in percentage. Δ indicates the difference between RED Positive/Negative.

5.1. Memorization Test

Table 1 shows the performance of the zero-shot baseline VT5_0 and the centralised non-DP model VT5_C on the RED test set. VT5_0 shows no performance difference between the RED positive set (documents from providers in $\mathcal{D}_{in}$), and the RED negative set (documents from providers in $\mathcal{D}_{out}$). At the same time, VT5_C shows considerable performance difference, indicating that the model overfits to a certain degree when trained on the PFL-DocVQA dataset. We hypothesize that this is caused mainly by memorization.

Model	RED Positive				RED Negative			
			$conf \geq 0.9$				$conf \geq 0.9$	
	ACC	ANLS	ACC	ANLS	ACC	ANLS	ACC	ANLS
$VT5_0$	0.08	1.60	0	1.16	0.09	0.68	0	0.72
$VT5_C$	**3.55**	11.64	**4.66**	14.53	0.17	5.05	0	6.60

Table 2. **DocVQA Performance of Non-Private Models on Memorization Test**. All information related to *provider_name* are removed from input.

There have been extensive works [8, 23, 54] that study the memorization of training data for language models, yet this behavior of such models in multi-modal settings remains under-explored. We thus perform a series of experiments on the RED data to demonstrate the existence of memorized information after fine-tuning on PFL-DocVQA.

Particularly, we assume that a model that memorizes information should be able to produce the correct answer, even if the answer is not present in the input. Thus, we ask the model a question about a specific key-value pair while removing all the clues related to the answer from both modalities (image and text). Then, we consider that all correct answers are due to memorization.

We focus the experiments on generic keys of each provider e.g. name, email, tax number, etc. that are constant in all provider's documents and thus, more likely to be memorized. Tab. 2 shows the results with *provider_name* as the key of interest. For more details of the test and results with different keys, please refer to the supplementary material.

Tab. 2 shows that while the $VT5_C$ model fails to predict the name of providers in the RED Negative data, it achieves around 3.5% Accuracy (4.6% when highly confident) in the RED Positive data even when no relevant information about the answer is available. Given its generative nature, this behavior suggests that the knowledge about these RED Positive providers is actually stored inside the model after fine-tuning. The memorization effect is further confirmed by the 0% accuracy of model $VT5_0$ in zero-shot setting, which confirms that this particular knowledge was not present in the model before fine-tuning.

5.2. Attack strategies

5.2.1 Generic Attack Formulation

A **Provider Membership Inference Attack (PMIA)** is a binary classification task that attempts to infer whether a specific provider P contributes its data to the training set of a PFL-DocVQA target model. The attacker, owning a set of non-training documents from the provider distribution $\mathcal{D}_P$, can query the target model h multiple times, each with one input tuple (d_i, q_i, a_i) from the set of possible inputs I_P, as defined in section 4.1, and in return receive a set of outputs O_P. Then, the attack model $\mathcal{A}$ aggregates all of these outputs into a feature vector $f_P = \text{AGG}(O_P)$, where AGG is the feature aggregation function, and produces a binary prediction $\mathcal{A}(f_P; h) \in \{0; 1\}$, indicating if P comes from the training set.

Unlike prior works such as in [41, 42, 49, 52], where attacks are evaluated directly on training examples, in this case we assume *query data can not be exactly the one used in training, but sampled from* $\mathcal{D}_P$.

5.2.2 Metric Selection

We base our attacks on a variety of selected per-example metrics that reflect statistics of specific behaviours of the model, which can in turn be used to distinguish training data from non-training data. The combination of these metrics provides a good descriptor to identify membership.

The reason is threefold, highlighting the limitations of existing Membership Inference Attacks (MIAs) when extended to PMIA: first, DocVQA is a more-than-two-modal task, thus existing approaches such as [21] are not directly applicable to DocVQA models. Second, to train an attack model, MIAs typically involve repeatedly training shadow models [42] which is too expensive for DocVQA. Third, metric-based attacks are computationally efficient with a simple attack model and thus, fit in scenarios where resources are limited and can serve as a good baseline for future research.

We categorize our metrics into different groups G_i:

- $G_1 = (ACC, NLS)$, $G_2 = (L, conf)$ are DocVQA utility metrics introduced in Sec. 4.1. In general, models tend to yield higher utility metrics for data sampled from their training set, and this observation forms the basis for many state-of-the-art MIAs [49, 52].
- $G_3 = (\Delta L, \Delta conf)$ are computed as the difference of the respective utility metrics of the model before and after fine-tuning. Intuitively, we expect to see higher values for this group of metrics for providers who are part of the training data.
- $G_4 = (NLS_{\text{mem}}, \Delta NLS_{\text{mem}})$ are two metrics inspired from Sec. 5.1. NLS_{mem} indicates the model's NLS score in the Memorization Test. ΔNLS_{mem} is designed in another test, where we first let the model generate *the most likely output given no query*. We then measure how much the output changes if we remove it from the input and test the model again in the same setting. If it remains similar, this is a sign of memorization.

5.2.3 Proposed Approaches

We propose two attack settings based on the knowledge the attacker has about the model and the training data.

Unsupervised Attack with Zero Knowledge (AZK). This attack corresponds to a *zero-knowledge* setting, where only

black-box access to the fine-tuned model and the ground-truth answer of each query are available to the adversary. The attacker has no information about the data distribution $\mathcal{D}_{\text{in}}$, neither about the target model architecture.

In this case the feature vector is formed with *ACC* and *NLS*, since these are the only available metrics in this setting. We then run K-Means Algorithm on the features to find the two clusters of $\mathcal{D}_{\text{in}}$/$\mathcal{D}_{\text{out}}$ providers. The cluster with higher average Accuracy is considered the set of member providers.

Supervised Attack with Partial Knowledge (APK). The second attack focuses on the *partial-knowledge* setting, in which it is assumed that the attacker knows a small number of both member/non-member providers from the training set, which is widely considered in some previous work [7, 21, 42, 49]. Yet, the adversary is only aware of the identities of the providers, without actually having access to the documents utilized during the training process. We also assume that the model architecture is known by the adversary, since the training starts from a publicly available pre-trained model, while maintaining the black-box access to the pre-trained and fine-tuned models. Lastly, we consider that the model outputs are tuples containing also the loss and confidence value for each query, in addition to the accuracy and the NLS of the previous setting.

In this approach, we randomly sample r-percent of target providers where r is the sampling rate, and obtain a training subset $\mathcal{P}_{train}$ with member/non-member classes equally distributed. We then evaluate the attack on the rest of providers $\mathcal{P}_{test}$. Since black-box access to both the pre-trained and fine-tuned model is given, we can make use of the 6 metrics from groups 1, 2 and 3, to enrich the information from one provider. Similarly to the first approach, we use the concatenation of aggregated metric values as the feature vector f_p for each provider $P \in \mathcal{P}_{train}$ and train a Random Forest classifier to infer provider membership.

Ensemble with Memorization. Given the memorization results illustrated in Sec. 5.1, we combine our main approaches with the memorization features of G_4 into a Hard-Voting Ensemble to further boost the performance. In particular, we have two separate classifers, denoted as CLS_1 and CLS_2. In CLS_1, we separate members from non-members by thresholding NLS_{mem} at 0, while in CLS_2 we use K-Means to figure out the two clusters based on $\Delta_{\text{mem}}NLS$.

6. Federated Learning and Privacy baselines

To alleviate the sensitive information leakage problem and adapt to the distributed training set-up, in this section we propose a baseline method based on VT5 and basic standard FL aggregation schemes and DP techniques.

6.1. Federated Learning

To perform the Federated Learning [39, 48], we apply the standard FedAvg [39] strategy. For this, the model weights of the non-frozen layers are sent from the server to the selected clients at each federated learning round. The model is then trained in parallel in each of the clients, and then, the model update of the non-frozen layers is returned to the server, where the different updates are averaged to obtain an updated model state. This process is repeated for each federated learning round.

We measure the efficiency of the communications as the total amount of information transmitted between the server and the clients in Gigabytes (GB).

6.2. Differentially Private Federated Learning

6.2.1 Differential Privacy

Definition 1 ((ε, δ)-Differential Privacy (DP) [13])**.** A randomized mechanism $\mathcal{M}$ with range $\mathcal{R}$ satisfies (ε, δ)-differential privacy, if for any two adjacent datasets $E \sim E'$, and for any subset of outputs $O \subseteq \mathcal{R}$, it holds that

$$\Pr[\mathcal{M}(E) \in O] \leq e^{\varepsilon} \Pr[\mathcal{M}(E') \in O] + \delta. \quad (1)$$

Definition 1 introduces (ε, δ)-DP, which bounds how much the output distribution of a randomized mechanism $\mathcal{M}$ can diverge on adjacent datasets. The privacy budget consists of $\varepsilon \geq 0$ (lower means more private) and the additive error $\delta \in [0, 1]$. We refer to Dwork and Roth [12] for a thorough introduction to DP.

The definition of DP critically depends on the concept of adjacency of datasets $E \sim E'$. We seek to protect the identity of providers that could be leaked through textual (company name, account number) or visual (logo, layout) information. The typical document-level adjacency definition would be too weak, as there are many documents from the same provider and combining them could leak private information. We therefore use *provider-level add/remove adjacency*, where $E \sim E'$ if one can be obtained from the other by adding or removing all documents from one provider. Prior work denotes this as group-level DP [16, 35].

Applied to our baseline method, DP ensures that there is a high probability that a machine learning model trained under DP is similar whether or not the data of a particular provider has been used for training the model.

Our training algorithm uses DP stochastic optimisation [1, 47, 51]. Its privacy is based on clipping the contribution from each unit of privacy (in our case each provider) and adding Gaussian noise. Random subsampling of clients and providers at each iteration provides further amplification of privacy. We use numerical privacy accounting for accurate evaluation of the cumulative privacy loss of the optimisation [28, 29] using the PRV accountant [18].

We consider an adversary who can access all intermediate global models. The adversary aims at inferring the training data (or some private information about them) of the participating clients' providers. The adversary is passive (i.e., honest-but-curious), that is, it does not modify the trained global model.

6.2.2 DP Federated Learning algorithm

Our private federated algorithm FL-PROVIDER-DP is shown in Algorithm 1. At every FL round the server randomly selects a set of clients. Each selected client runs a local instance of federated learning where each provider acts as the training data of a "virtual client" within the real client. The client randomly selects providers, clips the per-provider updates and the adds an appropriate amount of noise so that the update aggregated by the server is differentially private with respect to all providers in $\cup_k \mathbb{P}_k$ (thus the division of the per-client noise by the number of sampled clients $|\mathbb{K}|$). The noisy update of each client is normalized by a constant M which is the minimum number of providers among all clients. Note when no clients are sampled in a FL round the server still needs to add noise.

Algorithm 1: FL-PROVIDER-DP

Server:
  Initialize common model $\mathbf{w}_0$
  for $t = 1$ **to** T_{cl} **do**
    Select set $\mathbb{K}$ of clients randomly
    for *each client* k *in* $\mathbb{K}$ **do**
      $\mathbf{u}_t^k = \mathbf{Client}_k(\mathbf{w}_{t-1}, |\mathbb{K}|)$
    end
    $\mathbf{w}_t = \mathbf{w}_{t-1} + \frac{1}{|\mathbb{K}|} \sum_{k \in \mathbb{K}} \mathbf{u}_t^k$
  end
  Output: Global model $\mathbf{w}_{T_{cl}}$

$\mathbf{Client}_k(\mathbf{w}_{t-1}, |\mathbb{K}|)$**:**
  $\mathbb{P}_k$ is a set of predefined disjoint providers in D_k
  Select $\mathbb{M} \subseteq \mathbb{P}_k$ randomly
  for *each provider* P *in* $\mathbb{M}$ **do**
    $\mathbf{w}' = \mathbf{w}_{t-1}$
    $\Delta \mathbf{w}_t^P = \mathbf{AdamW}(P, \mathbf{w}', T_{\mathsf{gd}}) - \mathbf{w}_{t-1}$
    $\Delta \hat{\mathbf{w}}_t^P = \Delta \mathbf{w}_t^P / \max\left(1, \frac{||\Delta \mathbf{w}_t^P||_2}{C}\right)$
  end
  $\mathbf{w}'_t = \mathbf{w}_{t-1} + \frac{1}{M}\left(\sum_{i \in \mathbb{M}} \Delta \hat{\mathbf{w}}_t^i + \mathcal{N}\left(0, \frac{\sigma^2 C^2 \mathbf{I}}{|\mathbb{K}|}\right)\right)$
  Output: Model update $(\mathbf{w}'_t - \mathbf{w}_{t-1})$

6.2.3 Privacy analysis

The privacy loss of FL-PROVIDER-DP follows the usual analysis of DP stochastic optimisation consisting of compositions of sub-sampled Gaussian mechanisms. The loss depends on the number of iterations T_{cl}, sub-sampling rate q and noise scale σ [18, 28, 29]. The probability for sampling a specific provider is computed as follows. At every federated round a provider is selected when (1) the corresponding client is sampled, which has a probability of q_c, and (2) the batch of providers sampled locally at this client contains the provider, which has a probability of at most $\frac{|\mathbb{M}|}{\min_k |\mathbb{P}_k|}$. Therefore, a provider is sampled with a probability of at most $q = \frac{q_c \cdot |\mathbb{M}|}{\min_k |\mathbb{P}_k|}$. We compute the required noise scale σ using the PRV accountant [18], which is an accurate numerical accountant and can provide arbitrarily tight (ε, δ)-bounds. Similarly to previous work [10, 33, 58, 66] we do not account for privacy loss originating from the tuning of the hyperparameters.

6.3. Centralized DP training

The centralized DP training assumes that only one client exists and all providers are part of the dataset of this client. Our centralized DP training follows the same approach as the federated learning baseline with the exception that at every iteration T_{cl} the only existing client is selected.

7. Experiments

In our experiments we use five different variants of the VT5 DocVQA model:

VT5$_{\mathbf{0}}$ *(Zero-shot baseline)*: This method is not fine-tuned on the new PFL-DocVQA dataset.

VT5$_{\mathbf{C}}$ *(Centralized non-DP model)*: Fine-tuned for 10 epochs on the PFL-DocVQA dataset. It is used as an upper bound of question-answering performance with respect to the other variants.

VT5$_{\mathbf{C+DP}}$ (*Centralized DP model*): Fine-tuned for 10 iterations and in expectation 1000 providers are sampled at every iteration. The noise scale σ is computed to reach a final privacy budget of $(\varepsilon{=}\ 8, \delta = 10^{-5})$-DP.

VT5$_{\mathbf{FL}}$ *(Federated Learning non-DP model)*: Fine-tuned for 10 FL rounds, sampling $K = 2$ clients at each FL round.

VT5$_{\mathbf{FL+DP}}$ *(Federated Learning DP model)*: Fine-tuned for 10 FL rounds, in expectation $K = 2$ clients are sampled per FL round and no subsampling of providers is done. The noise scale σ is computed to reach a final privacy budget of $(\varepsilon{=}\ 8, \delta = 10^{-5})$-DP.

For all variants, a learning rate of $2e^{-4}$ is used. For the privacy preserving variants we present results for $\varepsilon \in \{1, 4\}$ in the supplementary material.

7.1. Model Performance

Method	ε	Question Answering ANLS	Question Answering Acc.	Communication (GB)
$VT5_0$	-	44.58	38.43	-
$VT5_C$	∞	90.91	81.80	-
$VT5_{FL}$	∞	88.73	79.30	44.66
$VT5_{C+DP}$	8	67.19	60.12	-
$VT5_{FL+DP}$	8	65.81	58.62	44.66

Table 3. Question answering performance and communication cost of the base method VT5 on the different set-ups. DP training is done with $\delta = 10^{-5}$).

The question-answering results of the VT5 variants is shown in Tab. 3. As expected, the $VT5_C$ performs the best. $VT5_{FL}$ performance is close to the centralized version despite training the same amount of iterations and seeing only 20% of the data at each iteration. In contrast, the private $VT5_{C+DP}$ and $VT5_{FL+DP}$ shows a performance degradation due to the introduced noise and clipping. Results for other epsilon budgets are shown in the supplementary material.

7.2. Attack Performance

Our attack experiments are conducted on the RED data, which contains $8,100$ Questions from $307/353$ RED Positive/Negative Providers (member/non member). The attack's performance is presented in terms of the Attack Accuracy metric. We evaluate our proposed attacks on multiple subsets T_s, with their size controlled by s, in such a way that T_s only includes those providers with a minimum of s defined questions. Increasing s guarantees a set of informative providers with sufficient statistics while keeping lower s requires the attack model to be robust to outliers. Tab. 4 reports the performance of our proposed attacks.

7.2.1 Non-private model

For the centralized non-private model, we can see that both approaches consistently perform better as s increases, suggesting that a certain number of queries is expected to characterize a provider membership. Under the simple assumption of *zero-knowledge*, AZK can surpass APK on all three benchmarks (row 1 and 3) with the same set of features, and even perform on par with fewer features on T_{10} (row 1 and 4). This implies G_1 metrics can effectively differentiate between members and non-members with enough queries. Still, AZK fails if less information are available, as simple K-Mean algorithm is quite sensitive to outliers.

When G_2 and G_3 features are added, APK clearly outperforms the unsupervised approach across all the evaluation sets (line 1 and 4), particularly in the high-data regime by a margin of 3.5% Accuracy in T_0 and 1.4% Accuracy in T_5, demonstrating the benefits of training a model with *partial-knowledge*, combined with enriched feature vector.

Memorization metrics G_4 are useful, especially effective when combined with CLS_1 and CLS_2 into Ensemble, as it improves attack's performance in every subsets for both approaches, with the best Accuracy at almost 67% in T_{10}. This is not always the case if it is used as training features to the attack model (line 2 and 5). We hypothesize this observation as Memorization Test can supply a clear sign of memorized providers which other features can not provide. However, since this test falls into low-recall regime, as shown in Tab. 2 where the model fails most of the time even with training providers, using these features to train rather introduces noise to the model.

7.2.2 Private models

We take the best performing attack configurations presented in the previous section and we apply them to the centralized DP model described in section 6 in order to assess the ability of the proposed DP framework to reduce privacy leakage.

Results in Table 4 show that for both types of attacks, AZK and APK, the accuracy of the attack in the centralized-DP model is significantly lower than the accuracy in the non-private model for the same configuratons. This reduction in the accuracy of the attack is more relevant in the case of the APK method that has a greater potential of leaking private information. These results confirm that the proposed DP baseline can effectively mitigate privacy leakage for DocVQA.

8. Conclusions

In this paper, we have explored for the first time privacy issues related to DocVQA. We have proposed a large scale DocVQA dataset especially designed for preserving the identity of the providers used to train the model. We have shown that state of the art generative multi-modal models exhibit memorization and how this can be used to attack the model and reveal private information about the providers. As a solution to this privacy leakage we have proposed to use Federated Learning and Differential Privacy in a baseline framework that guarantees a higher level of privacy at the cost of reducing the utility of the model.

Acknowledgements

This work has been funded by the European Lighthouse on Safe and Secure AI (ELSA) from the European Union's Horizon Europe programme under grant agreement No 101070617. RT, KN, MAS, LK, EV and DK have been supported by the Consolidated Research Group 2017-SGR-1783 from the Research and University Department of the Catalan Government, the projects PDC2021-121512-I00,

Model	Attack	Feature Set	Setting	Train	Test	Evaluation Set T_s		
						$s=0$	$s=5$	$s=10$
$VT5_C$	AZK	G_1	ZK + U	0	$(1-r)T_s$	56.13±0.55	60.25±0.62	65.12±0.99
		G_1+G_4	ZK + U	0	$(1-r)T_s$	48.25±5.67	47.80±3.30	58.16±6.60
	APK	G_1	ZK + S	rT_s	$(1-r)T_s$	55.11±2.20	58.23±2.67	60.29±3.15
		G_1+G_2+G_3	PK + S	rT_s	$(1-r)T_s$	59.67±1.01	**61.66±1.71**	**65.20±1.81**
		G_1+G_2+G_3+G_4	PK + S	rT_s	$(1-r)T_s$	**60.97±1.40**	61.26±1.57	64.44±2.62
	AZK+CLS_1+CLS_2	G_1+G_4	ZK + U	0	$(1-r)T_s$	60.56±0.70	**63.68±0.80**	66.25±1.80
	APK+CLS_1+CLS_2	G_1+G_2+G_3+G_4	PK + S	rT_s	$(1-r)T_s$	**61.84±0.09**	62.53±0.00	**66.99±0.00**
$VT5_{C+DP}$	AZK	G_1	ZK + U	0	$(1-r)T_s$	55.41±0.62	60.15±0.43	64.93±1.04
	APK	G_1+G_2+G_3	PK + S	rT_s	$(1-r)T_s$	51.67±1.56	54.06±2.54	58.55±1.34

Table 4. **Attack Performance against Private and Non-Private VT5 models on different RED subsets.** Colored rows indicate experiments with Ensemble. ZK/PK denotes our *zero-knowledge/partial-knowledge* setting while U/S denotes the Unsupervised/Supervised model training, respectively. T_s indicates the number of test providers in each evaluation subset, where all providers with fewer than $s+1$ test questions are eliminated. Train/Test indicates the size of $\mathcal{P}_{train}/\mathcal{P}_{test}$ in the corresponding approach, with $r = 0.15$ is the sampling ratio used in APK. Results are reported with the standard deviation over 5 random seeds.

PID2020-116298GB-I00 and PLEC2021-007850 funded by the European Union NextGenerationEU/PRTR and MCIN/AEI/10.13039/501100011033. MT and AH have been supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI; and grant 356499) and the Strategic Research Council at the Research Council of Finland (Grant 358247). Part of this work has been performed using resources provided by the CSC – IT Center for Science, Finland, and the Finnish Computing Competence Infrastructure (FCCI). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authorities can be held responsible for them.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 6

[2] Gergely Ács and Claude Castelluccia. I have a dream! (differentially private smart metering). In *Information Hiding - 13th International Conference, IH 2011*, pages 118–132, 2011. 9

[3] Amazon. Amazon textract. https://aws.amazon.com/textract/, 2021. Accessed on Oct. 10, 2023. 3

[4] Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE, 2019. 4

[5] Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4291–4301, 2019. 4

[6] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016. 9

[7] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022. 2, 6

[8] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022. 5

[9] Yannis Cattan, Christopher A. Choquette-Choo, Nicolas Papernot, and Abhradeep Thakurta. Fine-tuning with differential privacy necessitates an additional hyperparameter search. *CoRR*, abs/2210.02156, 2022. 3

[10] Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *CoRR*, abs/2204.13650, 2022. 3, 7

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 2

[12] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. 6

[13] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, pages 486–503. Springer, 2006. 6

[14] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and

Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, pages 265–284. Springer, 2006. 3

[15] Chong Fu, Xuhong Zhang, Shouling Ji, Jinyin Chen, Jingzheng Wu, Shanqing Guo, Jun Zhou, Alex X Liu, and Ting Wang. Label inference attacks against vertical federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA, 2022. USENIX Association. 2

[16] Filippo Galli, Sayan Biswas, Kangsoo Jung, Tommaso Cucinotta, and Catuscia Palamidessi. Group privacy for personalized federated learning. In *Proceedings of the 9th International Conference on Information Systems Security and Privacy*. SCITEPRESS - Science and Technology Publications, 2023. 6

[17] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, pages 16937–16947. Curran Associates, Inc., 2020. 2

[18] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11631–11642, 2021. 6, 7

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2790–2799. PMLR, 2019. 3

[20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[21] Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. M ^{4} i: Multi-modal models membership inference. *Advances in Neural Information Processing Systems*, 35:1867–1882, 2022. 3, 5, 6

[22] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022. 1, 2

[23] Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A Choquette-Choo, and Nicholas Carlini. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*, 2022. 5

[24] Raouf Kerkouche, Gergely Ács, Claude Castelluccia, and Pierre Genevès. Constrained differentially private federated learning for low-bandwidth devices. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1756–1765. PMLR, 2021. 3

[25] Raouf Kerkouche, Gergely Ács, Claude Castelluccia, and Pierre Genevès. Compression boosts differentially private federated learning. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 304–318, 2021. 3

[26] Raouf Kerkouche, Gergely Ács, and Mario Fritz. Client-specific property inference against secure aggregation in federated learning. *In Proceedings of the 22nd Workshop on Privacy in the Electronic Society (WPES'23)*, 2023. 2

[27] Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881, 2023. 3

[28] Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using FFT. In *The 23rd International Conference on Artificial Intelligence and Statistics, (AISTATS 2020)*, pages 2560–2569. PMLR, 2020. 6, 7

[29] Antti Koskela, Joonas Jälkö, Lukas Prediger, and Antti Honkela. Tight differential privacy for discrete-valued mechanisms and for the subsampled Gaussian mechanism using FFT. In *The 24th International Conference on Artificial Intelligence and Statistics, (AISTATS 2021)*, pages 3358–3366. PMLR, 2021. 6, 7

[30] Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward Training at ImageNet Scale with Differential Privacy. *ArXiv preprint*, abs/2201.12328, 2022. 3

[31] Junlong Li and et al. Dit: Self-supervised pre-training for document image transformer. In *ACMMM*, pages 3530–3539, 2022. 4, 2

[32] Oscar Li, Jiankai Sun, Xin Yang, Weihao Gao, Hongyi Zhang, Junyuan Xie, Virginia Smith, and Chong Wang. Label leakage and protection in two-party split learning. *NeurIPS 2020 Workshop on Scalability, Privacy, and Security in Federated Learning (SpicyFL)*, 2020. 2

[33] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3, 7

[34] Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu. Auditing privacy defenses in federated learning via generative gradient leakage. *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022. 2

[35] Virendra J. Marathe and Pallika Kanani. Subject granular differential privacy in federated learning, 2022. 6

[36] Minesh Mathew, Ruben Tito, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Document visual question answering challenge 2020. *arXiv preprint arXiv:2008.08899*, 2020. 1, 2, 4

[37] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209, 2021. 1, 2, 4

[38] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2

[39] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 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*, pages 1273–1282. PMLR, 2017. 2, 6

[40] Harsh Mehta, Abhradeep Guha Thakurta, Alexey Kurakin, and Ashok Cutkosky. Towards large scale transfer learning for differentially private image classification. *Trans. Mach. Learn. Res.*, 2023, 2023. 3

[41] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019. 2, 5

[42] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019. 2, 5, 6

[43] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 3

[44] Rafal Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Palka. Going full-tilt boogie on document understanding with text-image-layout transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 732–747. Springer, 2021. 1, 2

[45] Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. Dureadervis: A: A chinese dataset for open-domain document visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1338–1351, 2022. 2

[46] Colin Raffel and et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 4, 2

[47] Arun Rajkumar and Shivani Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21-23, 2012*, pages 933–941. JMLR.org, 2012. 6

[48] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015. 2, 6

[49] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 2, 5, 6

[50] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. Docile benchmark for document information localization and extraction. *arXiv preprint arXiv:2302.05658*, 2023. 3

[51] Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013, Austin, TX, USA, December 3-5, 2013*, pages 245–248. IEEE, 2013. 6

[52] Anshuman Suri, Pallika Kanani, Virendra J Marathe, and Daniel W Peterson. Subject membership inference attacks in federated learning. *arXiv preprint arXiv:2206.03317*, 2022. 5

[53] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 2

[54] Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35: 38274–38290, 2022. 5

[55] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *International Conference on Document Analysis and Recognition*, pages 778–792. Springer, 2021. 2

[56] Rubèn Tito, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2021 competition on document visual question answering. In *International Conference on Document Analysis and Recognition*, pages 635–649. Springer, 2021. 2

[57] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 2, 4

[58] Marlon Tobaben, Aliaksandra Shysheya, John Bronskill, Andrew Paverd, Shruti Tople, Santiago Zanella Béguelin, Richard E. Turner, and Antti Honkela. On the efficacy of differentially private few-shot image classification. *CoRR*, abs/2302.01190, 2023. 3, 7

[59] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023. 2

[60] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Stanisławek. Icdar 2023 competition on document understanding of everything (dude). In *International Conference on Document Analysis and Recognition*, pages 420–434. Springer, 2023. 2

[61] Aidmar Wainakh, Fabrizio Ventola, Till Müßig, Jens Keim, Carlos Garcia Cordero, Ephraim Zimmer, Tim Grube, Kristian Kersting, and Max Mühlhäuser. User-level label leakage from gradients in federated learning. *Proceedings on Privacy Enhancing Technologies*, 2022(2):227–244, 2022. 2

[62] Web. Public Inspection Files. https://publicfiles.fcc.gov/, . Accessed: 2022-10-20. 3

[63] Web. Industry Documents Library. https://www.industrydocuments.ucsf.edu/, . Accessed: 2022-10-20. 3

[64] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, 2021. 1, 2

[65] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3320–3328, 2014. 3

[66] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3, 7

[67] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020. 2

[68] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32, 2019. 2

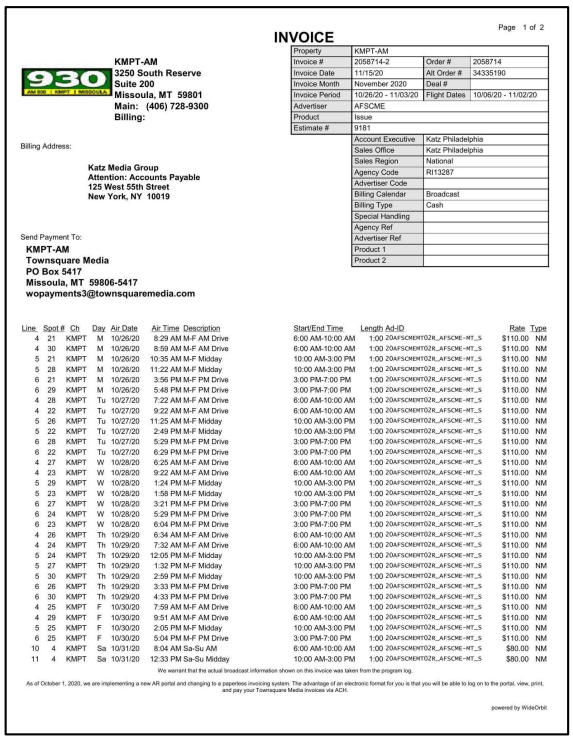
Privacy-Aware Document Visual Question Answering

Supplementary Material

We continue the figure and table numbering as follows from the main paper for clarity. Therefore, Figure 4 and Table 5 are the first references within the Supplementary Material.

A. Dataset

The PFL-DocVQA dataset is built from invoice documents that belong to different providers and clients, some examples of the invoice documents are shown in Fig. 5. PFL-DocVQA is composed of the document images, the OCR tokens and a set of questions/answers for each page. An example of these questions and answers is provided in Fig. 4.



INVOICE

Page 1 of 2

930 AM-FM

KMPT-AM
3250 South Reserve
Suite 200
Missoula, MT 59801
Main: (406) 728-9300
Billing:

Billing Address:

Katz Media Group
Attention: Accounts Payable
125 West 55th Street
New York, NY 10019

Send Payment To:
KMPT-AM
Townsquare Media
PO Box 5417
Missoula, MT 59806-5417
wopayments3@townsquaremedia.com

Property	KMPT-AM		
Invoice #	2058714-2	Order #	2058714
Invoice Date	11/15/20	Alt Order #	34335190
Invoice Month	November 2020	Deal #	
Invoice Period	10/26/20 - 11/03/20	Flight Dates	10/06/20 - 11/02/20
Advertiser	AFSCME		
Product	Issue		
Estimate #	9181		
	Account Executive	Katz Philadelphia	
	Sales Office	Katz Philadelphia	
	Sales Region	National	
	Agency Code	RI13287	
	Advertiser Code		
	Billing Calendar	Broadcast	
	Billing Type	Cash	
	Special Handling		
	Agency Ref		
	Advertiser Ref		
	Product 1		
	Product 2		

Line	Spot #	Ch	Day	Air Date	Air Time	Description	Start/End Time	Length	Ad-ID	Rate	Type
4	21	KMPT	M	10/26/20	8:29 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	30	KMPT	M	10/26/20	8:59 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	21	KMPT	M	10/26/20	10:35 AM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	28	KMPT	M	10/26/20	11:22 AM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	21	KMPT	M	10/26/20	3:56 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	29	KMPT	M	10/26/20	5:48 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	28	KMPT	Tu	10/27/20	7:22 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	22	KMPT	Tu	10/27/20	9:22 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	26	KMPT	Tu	10/27/20	11:25 AM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	22	KMPT	Tu	10/27/20	2:49 PM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	28	KMPT	Tu	10/27/20	5:29 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	22	KMPT	Tu	10/27/20	6:29 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	27	KMPT	W	10/28/20	6:25 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	23	KMPT	W	10/28/20	9:22 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	29	KMPT	W	10/28/20	1:24 PM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	23	KMPT	W	10/28/20	1:58 PM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	27	KMPT	W	10/28/20	3:21 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	24	KMPT	W	10/28/20	5:29 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	23	KMPT	W	10/28/20	6:04 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	26	KMPT	Th	10/29/20	6:34 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	24	KMPT	Th	10/29/20	7:32 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	24	KMPT	Th	10/29/20	12:05 PM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	27	KMPT	Th	10/29/20	1:32 PM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	30	KMPT	Th	10/29/20	2:59 PM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	26	KMPT	Th	10/29/20	3:33 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	30	KMPT	Th	10/29/20	4:33 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	25	KMPT	F	10/30/20	7:59 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
4	29	KMPT	F	10/30/20	9:51 AM	M-F AM Drive	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
5	25	KMPT	F	10/30/20	2:05 PM	M-F Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
6	25	KMPT	F	10/30/20	5:04 PM	M-F PM Drive	3:00 PM-7:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$110.00	NM
10	4	KMPT	Sa	10/31/20	8:04 AM	Sa-Su AM	6:00 AM-10:00 AM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$80.00	NM
11	4	KMPT	Sa	10/31/20	12:33 PM	Sa-Su Midday	10:00 AM-3:00 PM	1:00	20AFSCMEMTG2R_AFSCME~MT_S	$80.00	NM

We warrant that the actual broadcast information shown on this invoice was taken from the program log.

As of October 1, 2020, we are implementing a new AR portal and changing to a paperless invoicing system. The advantage of an electronic format for you is that you will be able to log on to the portal, view, print, and pay your Townsquare Media invoices via ACH.

powered by WideOrbit

Question: Can you provide the code associated with the agency mentioned in the document?
Answer: RI13287
Question: What month is attributed to the invoice?
Answer: November 2020
Question: Can you inform me of the vendor's name?
Answer: kmpt-am

Figure 4. Examples of questions and answers on an invoice document within PFL-DocVQA.

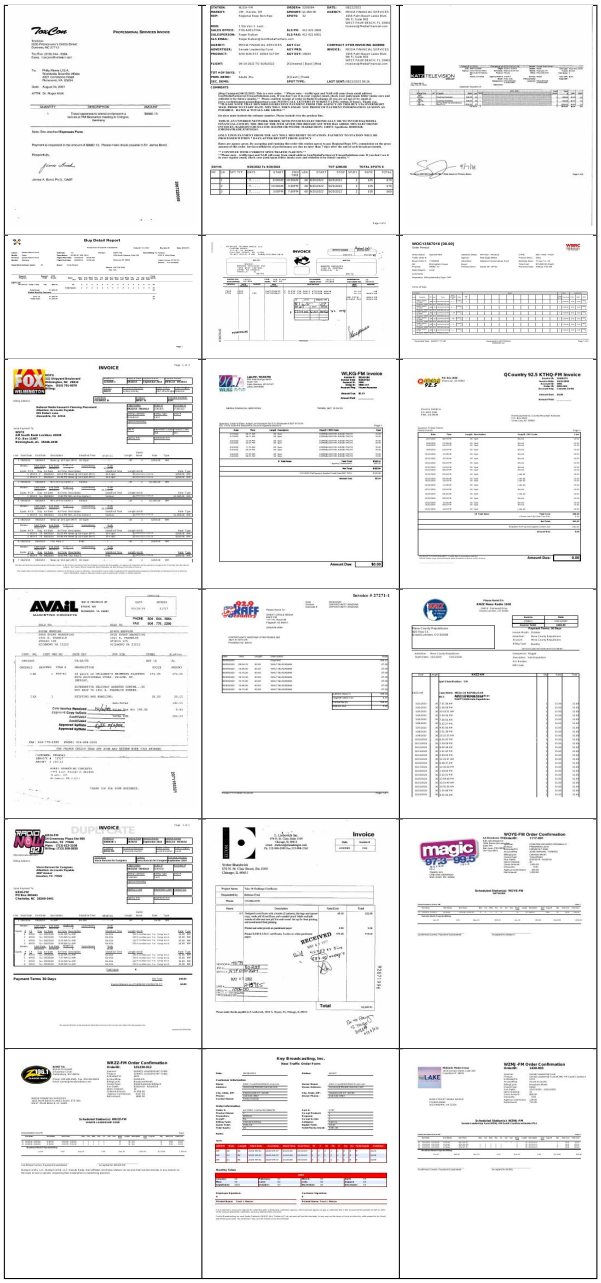


Figure 5. Examples of different invoice document images of from PFL-DocVQA.

The dataset is divided into BLUE and RED sub-datasets, as described in Sec. 3 and shown in Fig. 2. To create these subsets, we first cluster the documents by the provider ID to define the $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ distributions, and split the data accordingly. In Tab. 5 we show the number of documents, pages and questions/answers per client/subset in the BLUE and RED data respectively. Moreover, the documents of the same provider usually share visual and textual features, such as the invoices showed in Fig. 7, which belong to KATZ TELEVISION.

Dataset	Split	Client (Subset)	Provider	Document	Page	Question/ Answer
BLUE	Train	0	400	2224	5930	19465
		1	418	2382	6694	22229
		2	404	2296	6667	21673
		3	414	2358	6751	22148
		4	429	4543	12071	32472
		5	423	2378	6984	22361
		6	423	2700	7406	23801
		7	416	1951	5617	18462
		8	401	1932	5421	17868
		9	421	2136	6353	20840
	Valid	-	2231	3536	9150	30491
	Test	Positive	1390	2875	8088	25603
		Negative	977	1912	5375	17988
RED	Test	Positive	307	425	1361	4205
		Negative	353	425	1228	3895

Table 5. **Statistics on PFL-DocVQA Dataset** in terms of number of Providers/Documents/Pages/Question-Answers. The notion of *client* is only applied to BLUE Train data, while BLUE/RED Test is divided into two subsets: Positive and Negative that are from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, respectively.

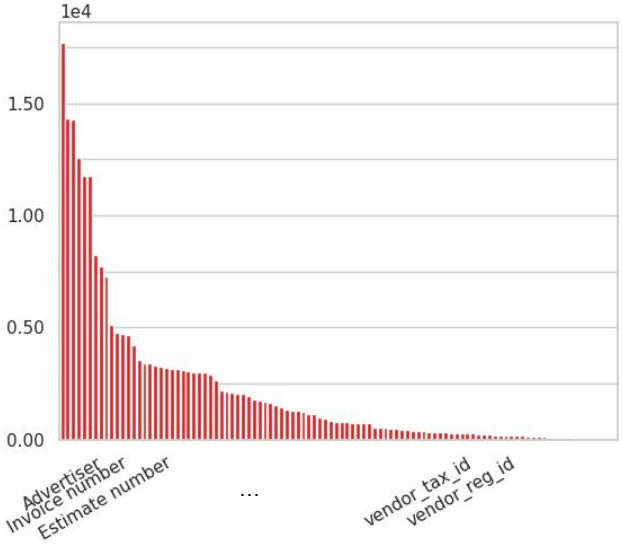


Figure 6. **Histogram of number of questions per key** in the training data.

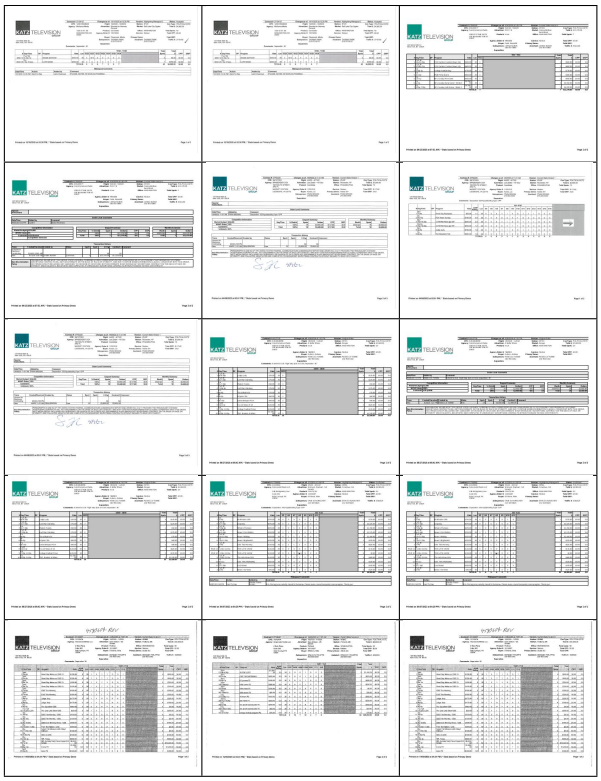


Figure 7. A visualization of different invoice documents from the same provider. The documents have similar layouts and use the same logo.

Moreover, we provide an overview of the number of questions asked for each key in the training set in Fig. 6.

B. Visual T5 Model

In this section, we provide a further detailed description of the employed VT5 base method illustrated in Fig. 3 and detailed in Fig. 8. The question and OCR words are first tokenized into subword tokens and embedded into a learned semantic contiguous representation from T5 [46], which captures the contextual information and semantics of the token. Moreover, VT5 utilizes a spatial embedding to represent the spatial information from the bounding box using a lookup table for continuous encoding of one-hot vectors. The different x_0, y_0, x_1, y_1 are embedded into 2 continuous learned embeddings for the horizontal and vertical coordinates. Then, the four embedded coordinates are aggregated and summed to the semantic representation. In parallel, the page image is divided into non-overlapping patches, which are projected and fed to the Visual Transformer DiT [31]. Then, all the features are fed into the T5 backbone self-attention layers to finally generate the answer in an autoregressive way.

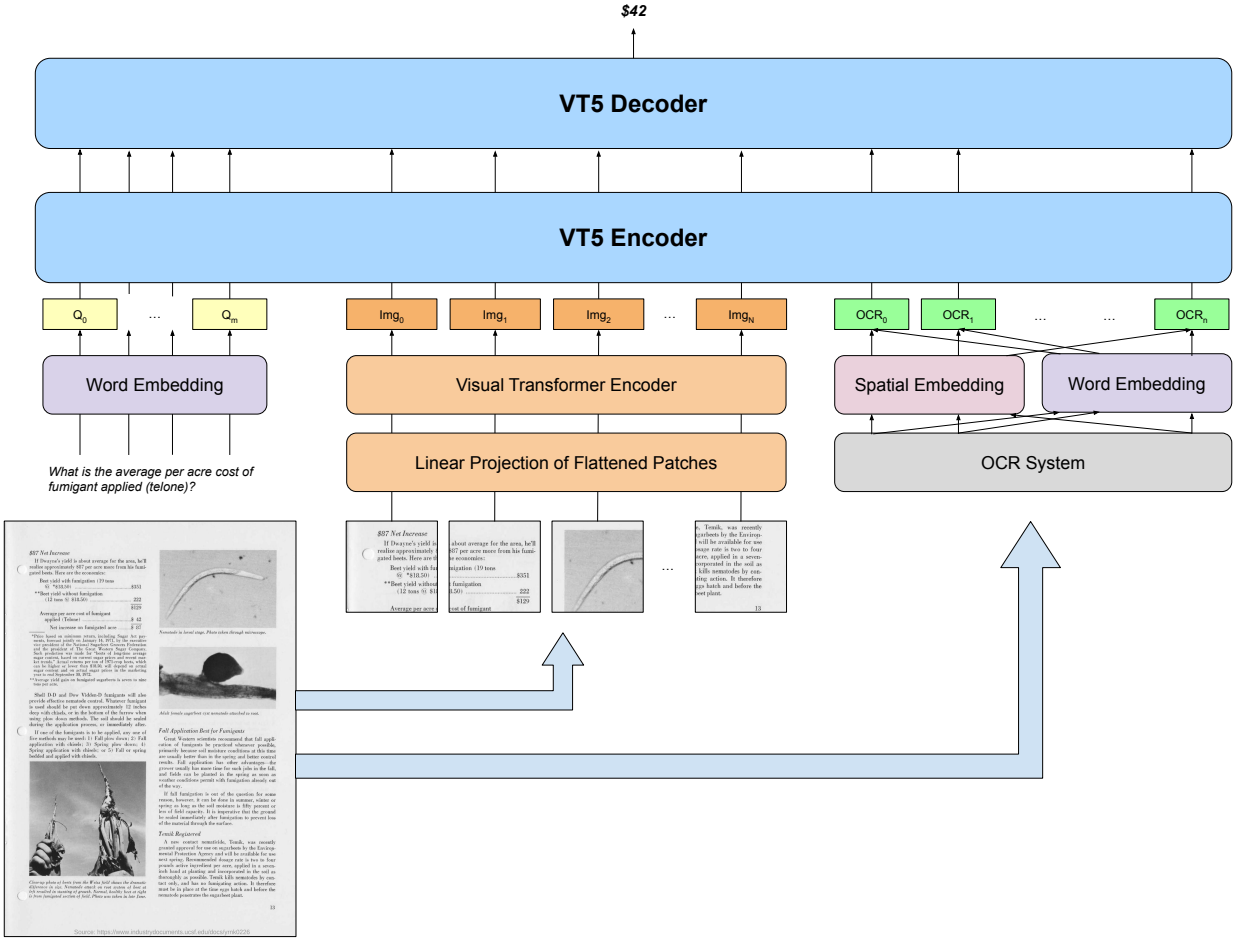


Figure 8. VT5 detailed method architecture.

C. Model Performance

Method	ε	Question Answering ANLS	Question Answering Acc.	Communication (GB)
$VT5_0$	-	44.58	38.43	-
$VT5_C$	∞	90.91	81.80	-
$VT5_{C+DP}$	8	67.19	60.12	-
$VT5_{C+DP}$	4	66.92	59.72	-
$VT5_{C+DP}$	1	61.52	54.16	-
$VT5_{FL}$	∞	88.73	79.30	44.66
$VT5_{FL+DP}$	8	65.81	58.62	44.66
$VT5_{FL+DP}$	4	61.91	53.68	44.66
$VT5_{FL+DP}$	1	57.47	50.60	44.66

Table 6. Question answering performance and communication cost of the base method VT5 on the different set-ups. DP training is done with $\delta = 10^{-5}$).

In Tab. 6 we extend Tab. 3 with the results of the new privacy preserving methods with ε budget 1 and 4, and illustrate the performance gap between those in Fig. 9. As expected, the more private the methods are, the worse they perform on the question-answering task.

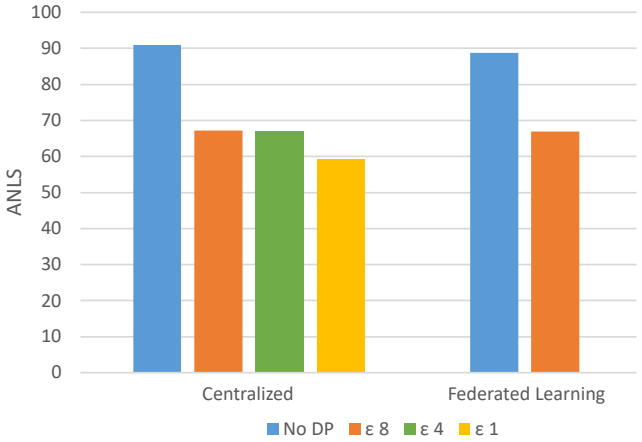


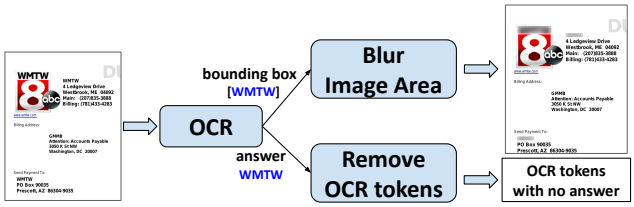
Figure 9. **Comparison of the question-answering performance** in ANLS of the different non-private and privacy preserving methods with different ε budgets in the Centralized and Federated Learning setup.

D. Attacks

D.1. Memorization Test

D.1.1 Experimental Setup

Fig. 10 illustrates the procedure that is applied in each Memorization Test. To begin, we first select a *key* that corresponds to a specific information of a provider, such as *name, email, tax number etc.* Given the selected *key*, we create a new test set which contains all RED documents, in each document we ask one question w.r.t the *key*. The question is sampled from the set of pre-defined templates for each key in the training phase, while the answer is the ground-truth information. Next, we remove all the clues related to the answer from both visual and textual input. For the visual part, we use Gaussian Blurring with the radius of Gaussian kernel set to 20. For the textual part, we discard all tokens in the OCR that are identified as exact or fuzzy matches to the answer string. Note, the size of each Memorization Test might vary, depending on the availability of the ground-truth answer in each document. For instance, while the provider's name is present in most of the invoice documents, not all of them specify the provider email or address, making these documents unavailable for testing. In addition, we skip all documents for which our code fails to process, maintaining the quality of the test. Finally, we evaluate the model on this new evaluation set in terms of DocVQA utility metrics, where high scores indicate clear memorization behaviors. In Fig. 11 we show some examples of the Memorization Test with the information that is asked removed.



Question: What name is associated with the vendor?
Answer: WMTW

Figure 10. **Our procedure to hide information in our Memorization Test** described in Appendix D.1.1 with an example.

D.1.2 More on Key: Provider Name

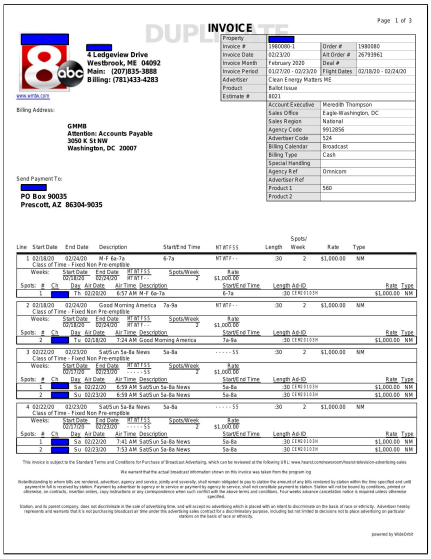
We further analyze the results of VT5_{C} on the Memorization Test with *provider_name* reported in Tab. 7, especially with the unexpected 0.17% in Accuracy (and 5.05% ANLS) on RED Negative, while it was expected to be 0. By examining the predicted answers, it is suspected that this performance comes from residual information from the OCR. The model might have learned to make use of other OCR tokens that still contain information about the provider (e.g. our code fails to remove 'www.kcci.com', which is the website for provider 'kcci'). Nevertheless, near 0% in Accuracy in this test means the model struggles to recover **the full answer**. Overall, we discover 5/340 positive providers with perfect Accuracy for all of their documents.

Model	Red Positive		Red Negative	
	N_{mem}=1333		N_{mem}=1178	
	ACC	ANLS	ACC	ANLS
VT5_{C}	**3.55**	11.64	**0.17**	5.05
$\text{VT5}_{\text{C+DP}}$ (ε=8)	0.03	2.4	0.03	0.66
$\text{VT5}_{\text{FL+DP}}$ (ε=8)	0.15	2.86	0.08	1.72

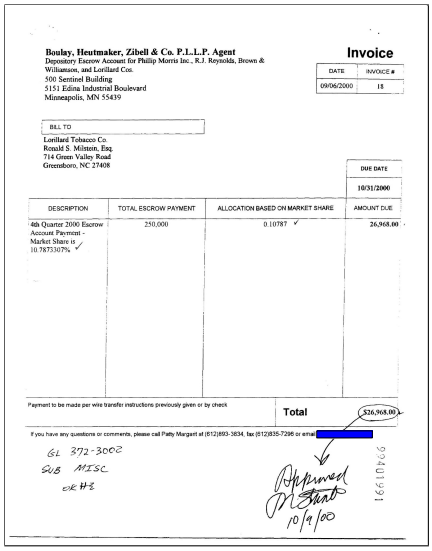
Table 7. **Memorization Test with key *provider_name***. N_{mem} denotes number of testing documents. Compared to the Non-Private Model VT5_{C}, memorization on positive providers are much reduced with Private Models.

D.1.3 Key: Provider Email

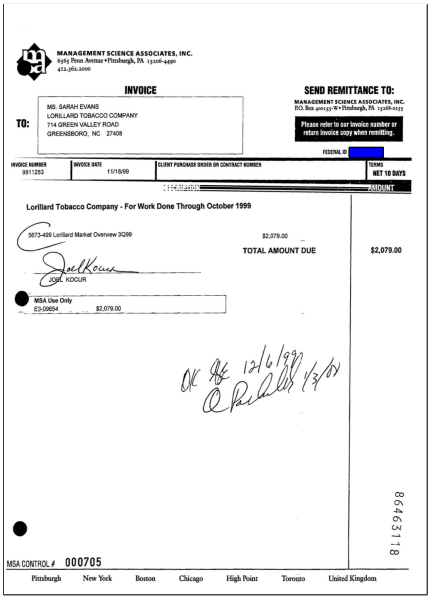
To perform this test, we first construct the evaluation set as explained in Appendix D.1.1. For any RED Positive provider, if there is any training question about the key posed on provider's documents, we reuse the training answer as the ground-truth answer for the test. Since this process cannot be applied to the RED Negative providers, we only use documents that have questions about the key found in RED testing examples. Similar to the pattern observed with the provider names, we identify a certain degree of memorization regarding positive providers' emails inside VT5_{C}, with 13% of Accuracy in the test, which is



Key: Provider Name
Question: What is the provider of this document?
Answer: wmtw



Key: Provider Email
Question: Could you share the vendor's email?
Answer: pmargarit@bhz.com



Key: Provider Tax Number
Question: Can you let me know the vendor's tax ID?
Answer: 25-1126415

Figure 11. **Examples of Memorization Test samples with information removed.** The blurred image area is colored **blue** for better visualization.

highly mitigated after the introduction of DP in our training process.

Model	Red Positive		Red Negative	
	N_{mem}=133		N_{mem}=8	
	ACC	ANLS	ACC	ANLS
$VT5_C$	**13.09**	32.65	0	39.91
$VT5_{C+DP}$ (ε=8)	0	1.39	0	0
$VT5_{FL+DP}$ (ε=8)	0	1.33	0	0

Table 8. **Memorization Test with key *provider_email***. Memorization about this information is also observed with Non-Private Models.

D.1.4 Key: Provider Tax Number, Registration ID

Our experiments in Appendix D.1.1 uncover memorization behaviors within the target model, but the test is only verifiable in case the information is actually present in the testing documents. In this section, we reveal another weakness of this model, where the model can provide sensitive information to specific queries even if the information does not exist in the document.

The focus of this test is on *Tax Number/Registration ID* as the key information for two reasons: First, these details uniquely identify a specific entity. Second, these keys are considered as *under-trained* as shown in Fig. 6, which indicates the model has not learned sufficiently to answer these types of questions due to limited training data. As a result, any not-requested disclosure of such information is likely attributed to memorization.

We carry out the test on RED Negative providers, where we sample at maximum 5 different documents per provider. For each document, we ask the target model $VT5_C$ 2 questions about Tax Number/Registration ID. Note that, we do *not hide* information before inference, but discard documents where the answer is explicitly present, as we aim to produce memorized answers, instead of actual answers from the document. We list the top-15 most frequent unique answers produced by the target model for each key:

```
Tax Number: {
    '91-0837469': 280,
    '56-0589582': 45,
    '91-0857469': 23,
    '91-08374': 12,
    '1828': 7,
    '91-0657469': 5,
    '37-1159433': 5,
    'orig cf': 3,
    '865864-ny': 3,
    '91-08574': 2,
    '033797': 1,
    'db829057-8ea2-4d34-abbd-ec53': 1,
    'woc13422932 [00.00]': 1,
    '92-346-369': 1,
    '113 61779 rt': 1
    ...
}
Registration ID: {
    '91-0837469': 30,
    '865864-ny': 11,
    '216-256-5304': 9,
    '91-08374': 8,
    '56-0589582': 8,
    'pol-cand': 8,
    '921-0850.00', 7,
    '56-05895', 6,
```

```
        'idb#1828', 6,
        '91-0857469', 5,
        'pol-iss', 5,
        'pb-18', 5,
        'ktvf', 5,
        '92-346-369': 1,
        'orig cf', 5
        ...
}
```

Based on the results, we see highly skewed distributions over the predicted answer, while we expect more uniform ones. Interestingly, there are common answers between these two sets, suggesting that the model is confused between these two queries after being under-trained. We then run the test against $VT5_0$ model and confirm that *none* of the top predicted answers from $VT5_C$ comes from the pre-training stage.

By accessing to the training data, we further inspect among top-15 most predicted answers in each distribution, and find that: *91-0837469* is the tax ID of *airborne express*; *56-0589582* is the tax ID of *kennedy covington lobdell & hickman. l.l.p*; *idb#1828* is the tax ID of *wajq-fm*; *37-1159433* is the tax ID of *meyer, capel, hirschfeld, muncy, jahn & aldeen, p.c.*, all the providers listed are included in training data. Also, we find that among the predicted answers in both sets, there are 103 answers which correspond to answers for different training questions, and they are not necessarily relevant to *Tax Number/Registration ID*, but other information such as *Phone Number/Asset Code* etc. This clearly demonstrates that some sensitive information is actually memorized inside non-private models

D.2. Private Mechanisms mitigate Memorization

Model	RED		RED Positive		RED Negative		ΔACC	ΔANLS
	ACC	ANLS	ACC	ANLS	ACC	ANLS		
$VT5_C$	81.40	90.17	85.92	93.68	76.53	86.48	**9.39**	**7.20**
$VT5_{C+DP}$ (ε=1)	52.4	60.09	53.98	61.87	50.68	58.17	3.3	3.7
$VT5_{C+DP}$ (ε=4)	58.59	65.77	59.86	67.5	57.23	64.39	2.63	2.66
$VT5_{C+DP}$ (ε=8)	60.31	67.44	61.57	69.04	58.94	65.72	2.62	3.33
$VT5_{FL+DP}$ (ε=1)	49.94	56.8	50.96	58.23	48.83	55.24	2.13	2.98
$VT5_{FL+DP}$ (ε=4)	53.68	61.8	55	63.21	52.25	60.28	2.75	2.93
$VT5_{FL+DP}$ (ε=8)	58.33	65.44	59.41	66.89	57.18	63.88	2.22	3.01

Table 9. **DocVQA Performance of Private Models on RED.**. Compared to the Non-Private Model $VT5_C$, Δ values from private Models are significantly decreased. Refer to Tab. 1 for details of notations.

We further validate the effectiveness of our DP mechanism to alleviate the memorization effect. First, we see that introducing DP to the training algorithm significantly closes the model's performance gap Δ between $\mathcal{D}_{in}/\mathcal{D}_{out}$ subsets, as shown in Tab. 9. Furthermore, we observe significant drops from Private Models compared to Non-Private ones in our Memorization Test with both *Provider Name* and *Email*, to almost 0, as reported in Tab. 7 and Tab. 8. These results suggest that overfitting in our Private Models is effectively decreased, and thus reduces memorization.

D.3. More on Attack Performance against Private Models

We provide more results of our proposed attacks against Private Models, both with centralized training and Federated Learning, in Tab. 10. The experiment settings including Feature Set, Attack Scenario, Number of Train/Test Examples are denoted as follows: $\diamondsuit = [\text{G}_1; \text{ZK+U}; 0; (1-r)T_s]$, $\ddagger = [\text{G}_1\text{+G}_2\text{+G}_3; \text{PK+S}; rT_s; (1-r)T_s]$. Refer to Tab. 4 for details of each notation.

...

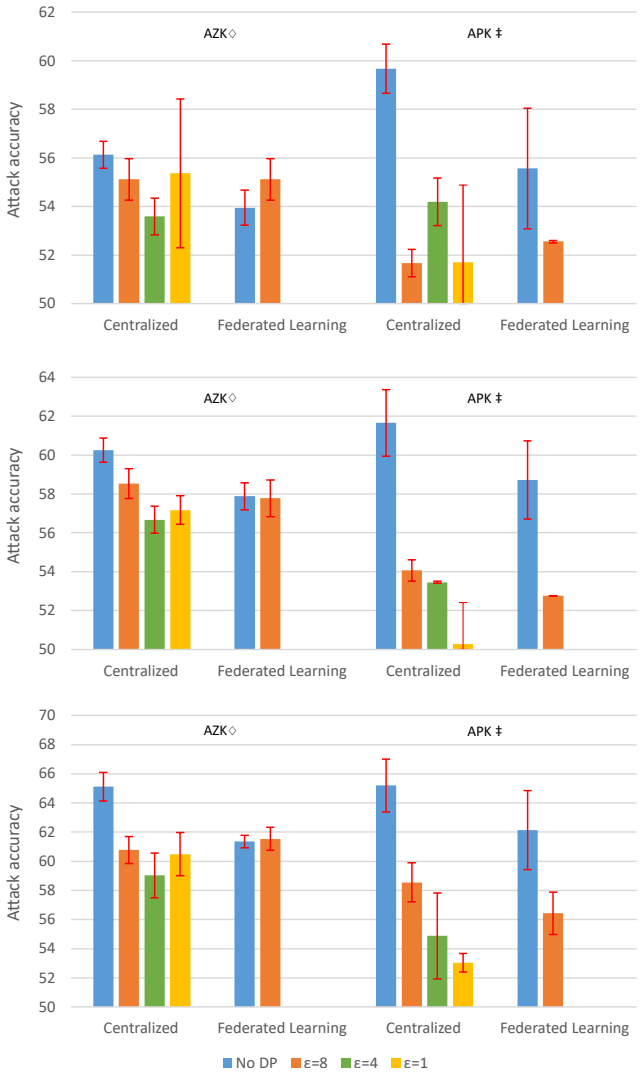


Figure 12. **Comparison of the attack accuracy performance** of the different attacks (AZK ♢ and APK ‡) performed on non-private and privacy preserving methods with different ε budgets in the Centralized and Federated Learning setup. The comparison is performed according to the number of T_s test providers $s = 0$ (top), $s = 5$ (middle) and $s = 10$ (bottom).

Model	Attack Method	Evaluation Set		
		$s = 0$	$s = 5$	$s = 10$
$VT5_{C}$	AZK♢	56.13±0.55	60.25±0.62	65.12±0.99
	APK‡	59.67±1.01	61.66±1.71	65.20±1.81
$VT5_{FL}$	AZK♢	53.95±0.72	57.88±0.70	61.35±0.43
	APK‡	55.57±2.48	58.72±2.02	62.13±2.71
$VT5_{C+DP}$ (ε=1)	AZK♢	55.37±3.06	57.17±0.73	60.48±1.48
	APK‡	51.71±3.17	50.27±2.14	53.04±0.64
$VT5_{C+DP}$ (ε=4)	AZK♢	53.59±0.76	56.67±0.69	59.03±1.53
	APK‡	54.20±0.98	53.45±3.05	54.88±2.94
$VT5_{C+DP}$ (ε=8)	AZK♢	55.12±0.85	58.54±0.77	60.77±0.94
	APK‡	51.67±1.56	54.06±2.54	58.55±1.34
$VT5_{FL+DP}$ (ε=1)	AZK♢	±	±	±
	APK‡	52.17±2.12	53.55±4.94	54.69±3.93
$VT5_{FL+DP}$ (ε=4)	AZK♢	53.34±0.96	55.82±0.79	61.35±0.86
	APK‡	52.45±1.35	51.28±1.99	56.91±4.4
$VT5_{FL+DP}$ (ε=8)	AZK♢	53.81±0.97	57.78±0.94	61.54±0.78
	APK‡	52.56±2.05	52.7±4.21	56.43±1.45

Table 10. **Attack Performance against Private models on different RED subsets.**

D.4. Ablation Study

In this section, we measure the impact of our selected metric and classifying methods for both AZK and APK in terms of Attack Accuracy. We employ $VT5_{C}$ as the target model and perform hyperparameter search for Random Forest in all ablation experiments.

D.4.1 Selected Metrics and Classifying Methods

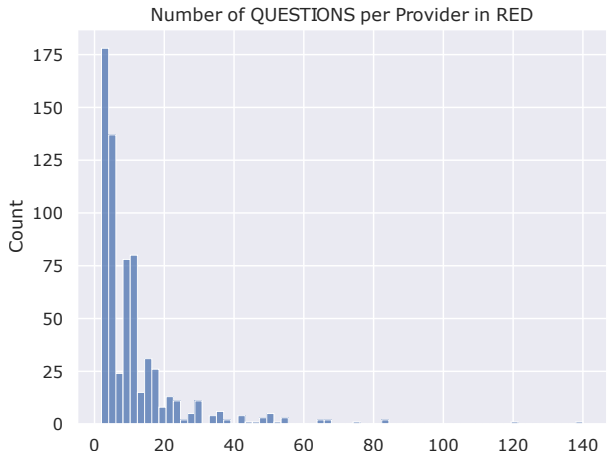
We ablate each of the selected metrics according to its availability order in our attack scenario, i.e. from *partial-knowledge* with access to the pre-trained model and loss/confidence values to *zero-knowledge*. We evaluate all the attacks on the subset with $s = 5$ and report the results in Tab. 11. From line 1 to 4, we find that K-Means method is a good baseline in restricted scenario like *zero-knowledge*, where only ACC and NLS are available metrics. In contrast, Random Forest is not suitable choice for this setting as it requires some training data and more features to perform adequately. When more information is accessible like in *partial-knowledge* scenario, the Supervised training directly benefits from it while the Unsupervised one shows substantial degradation with higher dimensional features (line 7). Finally, the designed features are all useful for Supervised training (line 2 to 6), as the model incrementally improves and surpasses the best performance of Unsupervised approach with the full set of features.

	Setting	Feature						Accuracy
		ACC	NLS	L	$conf$	ΔL	$\Delta conf$	
AZK	KM+U	✓						60.15±0.43
	KM+U	✓	✓					60.25±0.62
APK	RF+S	✓						57.23±1.36
	RF+S	✓	✓					58.24±2.68
	RF+S	✓	✓	✓	✓			59.79±1.75
	RF+S	✓	✓	✓	✓	✓	✓	61.66±1.71
	KM+U	✓	✓	✓	✓	✓	✓	49.22±2.3

Table 11. **Ablation Study of the Selected Metrics and Classifying Methods for our proposed attacks**. KM denotes K-Means Clustering method while RF denotes Random Forest Classifier. In all experiments, the performance is evaluated on the same RED subset with size $(1 - r)T_s$, where we use $s = 5$ and $r = 0.15$ for Supervised Setting. Refer to Tab. 4 for details of notation.

D.4.2 Minimum Number of Queries s

We also evaluate the performance of $APK‡$ while varying the minimum number of Questions per Provider s. As illustrated in Tab. 13, there is an upward trend in Attack Accuracy when s is increased from 0 to 10. This is due to the retention of informative providers, filtering out outliers and thus creating a more representative training dataset. However, excessively raising s as the threshold has an adverse effect, as it reduces the total number of Providers.



s	0 (All)	5	10	15	20
N_{Provider}	660	466	243	122	83
APK‡	59.67±1.01	61.66±1.71	**65.2±1.81**	61.57±4.18	64.05±1.83

Figure 13. **(Top) Distribution of number of Questions per Provider** in the RED data. **(Bottom) The impact of the number of Questions per Provider** on performance of our proposed attacks. N_{Provider} is the total number of Train/Test Providers with the corresponding s.

E. Training hyperparameters

In Tab. 12 we specify the hyperparameters used to train the different methods.

F. Privacy Analysis: More details

FL-PROVIDER-DP is described in Alg. 1. Group-level differential privacy necessitates that each client adds Gaussian noise to the aggregated model updates, which are generated using the data from the providers. We define $\mathbb{P}_k$ as a set of predefined disjoint providers in the dataset of client k. In particular, each client first select $\mathbb{M} \subseteq \mathbb{P}_k$ randomly. Then, for each selected provider i in $\mathbb{M}$, we compute the update $\Delta \mathbf{w}_t^{i,k} = \mathbf{AdamW}(i, \mathbf{w}', T_{\text{gd}}) - \mathbf{w}_{t-1}$, which is then clipped (in Line 17) to obtain $\Delta \hat{\mathbf{w}}_t^{i,k}$ with L_2-norm at most C. Then, random noise $\mathbf{z}_k \sim \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}/|\mathbb{K}|)$ is added to $\sum_{i\in\mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k}$ before averaging over M to obtain the update $\mathbf{u}_t^k = \frac{1}{M}\left(\sum_{i\in\mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k} + \mathbf{z}_k\right)$ for client k. Then, we compute at the server side the aggregate $\sum_{k\in\mathbb{K}} \mathbf{u}_t^k = \sum_{k\in\mathbb{K}} \left(\frac{1}{M}\left(\sum_{i\in\mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k} + \mathbf{z}_k\right)\right) = \sum_{k\in\mathbb{K}} \left(\frac{1}{M}\left(\sum_{i\in\mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k}\right)\right) + \mathcal{N}(0, \frac{\sigma^2 C^2 \mathbf{I}}{M^2})$ as the sum of Gaussian random variables also follows Gaussian distribution[1]:

$$
\begin{aligned}
\sum_{k\in\mathbb{K}} \mathbf{u}_t^k &= \sum_k \frac{1}{M_k}\left(\sum_{i\in\mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k} + \mathcal{N}(0, \frac{C^2\sigma^2 I}{|\mathbb{K}|})\right) \\
&= \sum_k \left(\sum_{i\in\mathbb{M}} \frac{\Delta \hat{\mathbf{w}}_t^{i,k}}{M_k} + \frac{1}{M_k}\mathcal{N}(0, \frac{C^2\sigma^2 I}{|\mathbb{K}|})\right) \\
&= \sum_k \sum_{i\in\mathbb{M}} \frac{\Delta \hat{\mathbf{w}}_t^{i,k}}{M_k} + \sum_k \frac{1}{M_k}\mathcal{N}(0, \frac{C^2\sigma^2 I}{|\mathbb{K}|}) \\
&= \sum_k \sum_{i\in\mathbb{M}} \frac{\Delta \hat{\mathbf{w}}_t^{i,k}}{M_k} + \mathcal{N}\left(0, \sum_k \frac{C^2\sigma^2}{|\mathbb{K}| M_k^2} I\right)
\end{aligned}
$$

[1] More precisely, $\sum_i \mathcal{N}(\nu_i, \xi_i) = \mathcal{N}(\sum_i \nu_i, \sqrt{\sum_i \xi_i^2})$

Method	Learning Rate	Batch Size	Epochs / Iterations	δ	Sensitivity σ	Noise Multiplier	Providers per iteration
VT5_{C}	$2e^{-4}$	10	10	-	-	-	4149 (All)
$\text{VT5}_{\text{C+DP}}$ ε=1	$2e^{-4}$	8	10	$10e^{-5}$	1	3.32031250000	1000
$\text{VT5}_{\text{C+DP}}$ ε=4	$2e^{-4}$	8	10	$10e^{-5}$	5	1.25244140625	1000
$\text{VT5}_{\text{C+DP}}$ ε=8	$2e^{-4}$	8	10	$10e^{-5}$	5	0.83251953125	1000

Method	Learning Rate	Batch Size	FL Rounds	δ	Sensitivity σ	Noise Multiplier	Provider sampling probability	Client sampling probability
VT5_{FL}	$2e^{-4}$	10	10	-	-	-	1	0.2
$\text{VT5}_{\text{FL+DP}}$ ε=8	$2e^{-4}$	8	10	$10e^{-5}$	5	0.77148437500	1	0.2

Table 12. **Training hyperparameters** for the different centralized (top) and federated learning (bottom) methods used in the PFL-DocVQA dataset.

Assuming $M_k = M$ for all k:

$$= \sum_k \sum_{i \in \mathbb{M}} \frac{\Delta \hat{\mathbf{w}}_t^{i,k}}{M} + \mathcal{N}\left(0, \sum_k \frac{C^2\sigma^2}{|\mathbb{K}|M^2} I\right)$$
$$= \sum_k \sum_{i \in \mathbb{M}} \frac{\Delta \hat{\mathbf{w}}_t^{i,k}}{M} + \mathcal{N}\left(0, \frac{|\mathbb{K}|C^2\sigma^2}{|\mathbb{K}|M^2} I\right)$$

And then differential privacy is satisfied where ε and δ can be computed using the PRV accountant described in Section 6.2.

However, as the noise is inversely proportional to $|\mathbb{K}|$, $\mathbf{z}_k$ is likely to be small if $|\mathbb{K}|$ is too large. Therefore, the adversary accessing an individual update $\mathbf{u}_t^k$ can almost learn a non-noisy update since $\mathbf{z}_k$ is small. Hence, each client uses secure aggregation to encrypt its individual update before sending it to the server. Upon reception, the server sums the encrypted updates as:

$$\begin{aligned} \sum_{k \in \mathbb{K}} \mathbf{u}_t^k &= \sum_{k \in \mathbb{K}} \mathsf{Enc}_{K_k}\left(\frac{1}{M}\left(\sum_{i \in \mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k} + \mathbf{z}_k\right)\right) \\ &= \sum_{k \in \mathbb{K}} \left(\frac{1}{M}\left(\sum_{i \in \mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k}\right)\right) + \mathcal{N}(0, \frac{\sigma^2 C^2 \mathbf{I}}{M^2}) \end{aligned} \tag{2}$$

where $\mathsf{Enc}_{K_k}\left(\frac{1}{M}\left(\sum_{i \in \mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k} + \mathbf{z}_k\right)\right) = \frac{1}{M}\left(\sum_{i \in \mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k} + \mathbf{z}_k\right) + \mathbf{K}_k \mod p$ and $\sum_k \mathbf{K}_k = 0$ (see [2, 6] for more details). Here the modulo is taken element-wise and $p = 2^{\lceil \log_2(\max_k ||\frac{1}{M}\left(\sum_{i \in \mathbb{M}} \Delta \hat{\mathbf{w}}_t^{i,k} + \mathbf{z}_k\right)||_\infty |\mathbb{K}|)\rceil}$.