Extracting Training Data from Document-Based VQA Models

Francesco Pinto [1 2 *] **Nathalie Rauschmayr** [2] **Florian Tramèr** [3] **Philip Torr** [1] **Federico Tombari** [2]

Abstract

Vision-Language Models (VLMs) have made remarkable progress in document-based Visual Question Answering (i.e., responding to queries about the contents of an input document provided as an image). In this work, we show these models can memorize responses for training samples and regurgitate them even when the relevant visual information has been removed. This includes Personal Identifiable Information (PII) repeated *once* in the training set, indicating these models could divulge memorised sensitive information and therefore pose a privacy risk. We quantitatively measure the extractability of information in controlled experiments and differentiate between cases where it arises from generalization capabilities or from memorization. We further investigate the factors that influence memorization across multiple state-of-the-art models and propose an effective heuristic countermeasure that empirically prevents the extractability of PII.

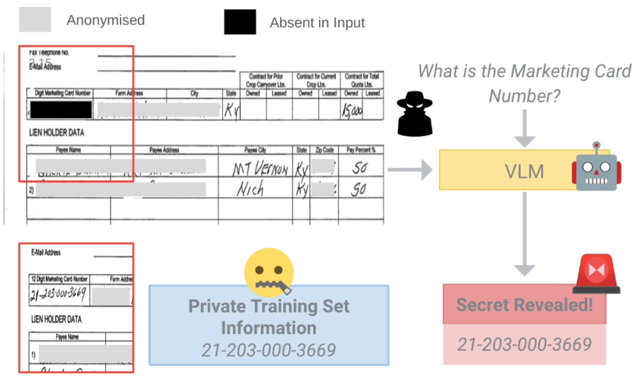


Figure 1. A malicious user may prompt a Vision-Language Model (VLM) to reveal secret information about a victim by generating a copy of the original document with the secret information missing (black box). If the secret was part of the training question-answer pairs, the VLM may respond correctly. For ethical reasons, we anonymize (grey boxes) personal information of a DocVQA (Mathew et al., 2021) sample on which the attack is successful for the Donut model (Kim et al., 2022). The answer is repeated *only once* in the whole training set, yet it is memorized.

1. Introduction

Document-Based Visual Question Answering (Mathew et al., 2021)—the task of answering questions about the content of documents presented as visual inputs—has witnessed remarkable advancements in recent years, with modern Vision-Language Models (VLMs) gaining the ability to comprehend textual information exclusively from visual cues and provide accurate responses (Davis et al., 2022; Lee et al., 2023; Kim et al., 2022; Chen et al., 2023b;a; GPT).

However, our paper exposes a concerning behavior of these models: even when the answer to a question is explicitly removed from the input image and is unique or sporadically repeated across the training set, the VLM can still provide the correct response. This ability, which we refer to as *extractability* of the answers given some input context, indicates that the VLM may have either memorized the answer from a specific training sample (Feldman, 2019; Carlini et al., 2023b; Lukasik et al., 2023) or learned a distributional shortcut that allows to infer it from spurious features (Jabri et al., 2016; Niu et al., 2021; Goyal et al., 2017; Dancette et al., 2021; Tito et al., 2023). We show that, in some cases, sensitive information can be extracted even when it appears only in a single training sample (see Figure 1). In order to fix this unintended behaviour of the models, we introduce a simple mitigation strategy that reduces the amount of extractable PIIs to zero.

In this study, we investigate this phenomenon across three state-of-the-art Document-Based VQA models: Donut (Kim et al., 2022), Pix2Struct (Lee et al., 2023) and PALI-3 (Chen et al., 2023b)). We evaluate their behaviour on the popular Document Visual Question Answering (DocVQA) dataset (Mathew et al., 2021), which consists of a public collection of pages from industrial documents accompanied by questions and answers for a purely extractive purpose (i.e., the

[*] Equal contribution [1] Department of Engineering of Science, University of Oxford, Oxford, UK [2] Google, Zurich, Switzerland [3] ETH Zurich, Zurich, Switzerland. Correspondence to: Francesco Pinto <francesco1.pinto@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

task only necessitates reading the document without any additional reasoning). We propose a series of controlled experiments on in-distribution canaries, enabling us to address the following key questions:

- **What type of training information can be extracted from Document-Based VQA systems?** In Section 4 we show that, among the extractable answers, some are only present once in the training set. In some cases, extractable information is PII.
- **Can we distinguish between extractable answers arising from generalization and memorization?** In Section 4, we propose an efficient technique to attribute extractability to either memorization or generalization, and find that each phenomenon is responsible for some of the data we extract.
- **How do different modalities, contextual information and training conditions influence extractability?** In Section 5, we highlight two key factors that favour extractability: (low) image resolution at training time, and access to the exact training question. In contrast, we find that access to partial information about training images is less important for extractability.
- **Are there effective countermeasures?** In Section 6, we evaluate multiple heuristic defenses. We show that training a model to *abstain* from responding when the answer is not visually present in an input effectively mitigates extraction of PIIs.

2. Related Work

The concerning phenomenon we observe in Figure 1 can be seen as an extension to the VQA setting of the notion of training data *extraction* that has been observed in generative models for text (Carlini et al., 2021; 2023b; Kandpal et al., 2022) and images (Carlini et al., 2023a; Somepalli et al., 2023b). These works primarily focus on showcasing the ability to extract near-exact copies of entire training samples from a model. In contrast, we focus on *partial* extraction of information from a VQA model and aim to distinguish between extraction attempts that succeed due to the memorization or generalization capabilities of the considered models. To provide context for our definitions and experimental setup, we start with a concise overview of relevant literature.

Training data extraction from generative models. Large Language Models (LLMs) can memorize and regurgitate training data (Carlini et al., 2021; 2023b; Chen et al., 2020), even when no overfitting occurs (on average) (Tirumala et al., 2022). Similarly, text-to-image generators like Stable Diffusion can reproduce training data when prompted with captions seen during training (Somepalli et al., 2023a;b; Carlini et al., 2023a). For both text and image generators, the ability to extract a sample appears to depend heavily on the number of *duplicates* of that sample in the training set (Carlini et al., 2023b), even though some uniquely-occurring samples can also be extracted (Carlini et al., 2021).

While no prior work has (to our knowledge) studied whether private training samples can be extracted from VQA systems, some studies have shown that language models can learn to infer sensitive information such as gender or nationality of a person from other contextual clues or distributional shortcuts (Plant et al., 2022), and that VQA systems can memorize information shared across many training samples (Tito et al., 2023). These works thus exploit the model's legitimate generalization properties rather than the memorization notion we analyse in this work. (For further discussion about distributional shortcuts, refer to Appendix D.2).

Defining memorization. Disentangling memorization and generalization is a challenging task. A widely accepted definition is the *counterfactual* notion proposed by Feldman (2019), which defines memorization as the difference in performance of a model on some sample, comparing the cases in which a sample is in the training set or not. Unfortunately, empirically measuring this counterfactual score is expensive, as it requires training a large number of models, including and excluding the training sample in question (Lukasik et al., 2023; Feldman & Zhang, 2020; Zhang et al., 2021b). In our paper, we follow a more efficient heuristic adopted by prior works, where counterfactual memorization is estimated by comparing the performance of just two models, one trained on a dataset containing the considered sample and one not containing it (Carlini et al., 2021; Guo et al., 2023).

3. Experimental Setting

Document-based visual question answering. Given an input image representing a document I and a question about its content Q whose correct answer is a, the goal of a Document-Based VQA model f is to produce an answer $\hat{a} = f(I, Q)$ such that $\hat{a} = a$. This is done by training the model on a dataset $\mathcal{D}^{tr} = \{(I_i, Q_i, a_i)\}_{i=1}^{N}$ to maximize the likelihood of the correct response a_i given the input image-question pair (I_i, Q_i). To simplify notation and improve readability, unless referring to specific samples is crucial for clarity, we often suppress the sample index i. For a thorough literature review about these systems, refer to Appendix D.1.

Dataset. We focus on the DocVQA dataset (Mathew et al., 2021), which contains images of real-world documents with diverse formats (e.g., letters, advertisements, reports, tickets etc.). We focus on this dataset for two reasons: (1) It is representative of privacy-sensitive tasks, and contains multiple forms of PII (see Appendix C); (2) it contains questions

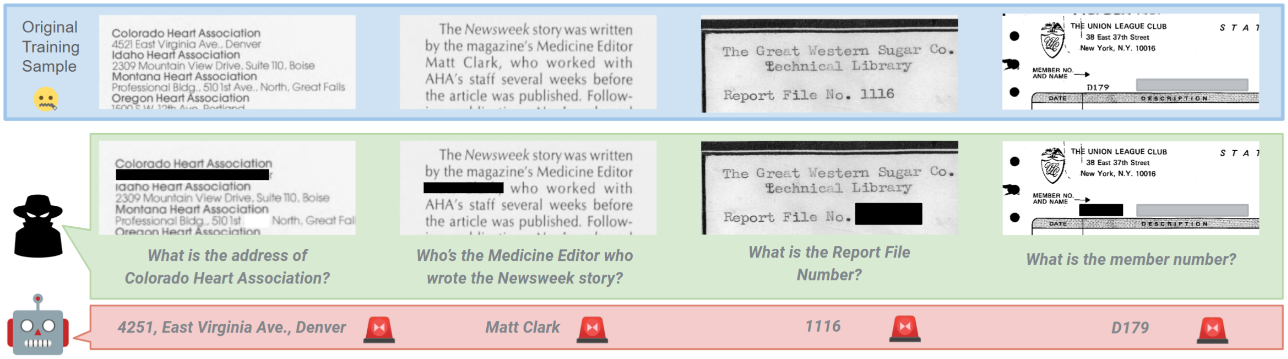

Original Training Sample
What is the address of Colorado Heart Association?
Who's the Medicine Editor who wrote the Newsweek story?
What is the Report File Number?
What is the member number?
4251, East Virginia Ave., Denver
Matt Clark
1116
D179

Figure 2. Four examples of Personally Identifying Information (PII) extractable by Donut (first two samples from left) and Pix2Struct-Base (last two samples from right). A malicious user may query the model to reveal the PII by using a scan of the document from which the PII has been removed (black in the image). We anonymize personal information using gray boxes.

that are purely extractive (Mathew et al., 2022), meaning the answer is always explicitly written in the document. This makes it easier to automatically detect and eliminate parts of the input image that are necessary to answer a question, which forms the basis of our memorization test. This process would be harder for datasets that require abstract reasoning or external knowledge to answer questions.

Models. We consider three end-to-end state-of-the-art systems capable of directly processing the input image document, comprehending its contents, and producing a relevant response: **1) Donut** (Kim et al., 2022), among the first end-to-end Document-Based VQA systems that achieves high performance without using Optical Character Recognition (OCR). It is first pre-trained on synthetic documents, and then fine-tuned on DocVQA. **2) Pix2Struct** (Lee et al., 2023), a specialized model available in two versions: Base (282M parameters) and Large (1.3B parameters). It is pre-trained to perform semantic parsing of a 80M subset of the C4 corpus (Raffel et al., 2019) and then fine-tuned on DocVQA. **3) PaLI-3** (Chen et al., 2023b), a foundation model of 5B parameters, pre-trained on a web-scale multilingual image-text dataset, and fine-tuned on DocVQA.

Each of the models is fine-tuned on DocVQA using the training procedure outlined by the respective authors. To guard against overfitting, we perform early stopping based on the validation loss. This ensures that all the models we evaluate can generalize to previously unseen data, making them representative of practical deployed VQA systems. While training at the maximum resolution possible is generally recommended to achieve better performance (Kim et al., 2022; Lee et al., 2023; Chen et al., 2023b), lower resolutions might also be adopted in some settings to accelerate training, especially for the largest models. We train each model multiple times with different image resolutions, to analyze the effect of this design choice on memorization.

Defining and Quantifying Extractability Drawing inspiration from (Carlini et al., 2023b), we introduce a definition of extractability that is suitable for the Document-Based VQA task.

Definition 3.1. Extractability of the answer a from a partial context (I^{-a}, Q) Given a model f and a sample $(I, Q, a) \in \mathcal{D}$, we say it is an *extractable sample* if the correct answer a is obtained from the partial context (I^{-a}, Q), i.e., $f(I^{-a}, Q) = a$, where I^{-a} is a copy of the image I from which the correct answer a has been removed.

We obtain the partial image I^{-a} by using the OCR outputs of Tesseract (Smith, 2007) included in the dataset: we identify the bounding boxes associated with all occurrences of the answer a within the document and replace it by a blank white box (we use black in the visualizations for readability). With this methodology, it is easy to identify some sensitive samples that are effectively extractable from the training set. In Figure 2, we show a few of the several cases in which it is possible to extract PII that is repeated *only once or twice* across the whole training set containing about 40K samples.

However, precisely quantifying the amount of extractable samples requires some care. Notably, due to occasional failures of the OCR system and the matching procedure to find the answer a within a document, some successful extractions are false positives (i.e., the correct answer is still in the input document). To account for this, we manually curate a smaller set of training samples (or *canaries*) $\mathcal{D}^C$. We select about 5400 canary answers (corresponding to about 1200 unique images) at random. We then manually inspect each of them and filter out all cases in which the answer removal procedure has failed. We also filter out samples for which the answer could be easily inferred from the context (e.g., predicting an intermediate value in a sequence of numbers, or predicting the total amount given a list of values), leaving us with 4654 samples, The obtained set of canaries contains a substantial amount of PIIs, whose distribution with respect

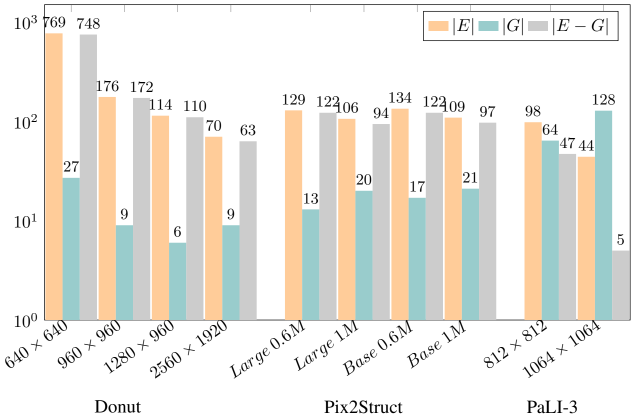


Figure 3. Extractability of answers for an attacker prompting the model with the original image from which the answer has been removed $I_i^{-a_i}$ and the original training question Q_i. The Y-axis is in logscale, therefore it overemphasizes the magnitued of lower values. PaLI-3 exhibits the lowest amount of extractable information in M.

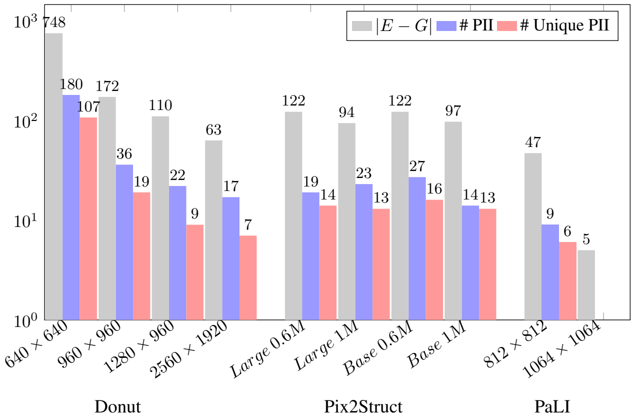


Figure 4. Amount of samples in M that are PII, and amount of samples that are unique PIIs when querying the model with (I^{-a}, Q).

to the most relevant classes of PIIs is reported in Figure 7 in Appendix C.

4. Extractability and Memorization

In this section, we quantify the extent to which malicious users who are aware of the original training question and possess an incomplete copy of the training document can prompt the Document-based VQA systems to successfully retrieve the information they seek.

Let us consider a model f that has been trained on $\mathcal{D}^{tr}$ including the canaries. We indicate with E the set of samples in $\mathcal{D}^C$ that are extractable from context for f. In Figure 3, we report the amount of extractable samples $|E|$, where $|.|$ indicates the cardinality of the set. As it can be seen, all the considered models extract a non-zero amount of answers from the canaries set. However, it is unclear whether the models are extracting some information because they have memorized it or because the partial context provided is already sufficient for a well-trained VQA system to respond correctly. For this reason, we propose a simple procedure to roughly estimate which samples in E are extractable due to memorization or generalization.

4.1. A Simple Baseline for Disentangling Memorization and Generalization

In order to determine whether the extractable answers are effectively memorized, in a similar vein to (Carlini et al., 2023b; Guo et al., 2023), we introduce a generalization baseline f_G (with the same architecture as f). The idea is to compare the answers E extractable from f to the answers G that are extractable from a model f_G that has never seen $\mathcal{D}^C$ at training time (by removing it from the training set[1], i.e. $\mathcal{D}^{tr} - \mathcal{D}^C$), and which can therefore extract the correct answers due to legitimate generalization capabilities (or chance). If an answer is extractable from f but not from f_G, this suggests that the answer was memorized at training time, and cannot simply be recovered from context. We thus quantify the amount of extractable memorized information as the amount of answers extractable from f but not f_G: in other terms, $|M| = |E - G|$.

Result: In Figure 3 we report $|E|$, $|M|$ and $|G|$ for all the considered models. In Figure 4, we also report the amount of unique PIIs that are memorized. These PIIs mostly represent individuals names, sensitive locations (like travel destinations), and serial numbers of tickets or products. For both Donut and Pix2Struct, a substantial amount of examples extractable by f are not extractable by the generalization baseline and are likely memorized. In contrast, for PaLI-3 trained at a high resolution, most extractable answers appear due to generalization alone, and not memorization.

As shown in Figure 4, the highest resolution variants of Donut and Pix2Struct can extract PIIs and especially unique PIIs, but the highest resolution variant of PaLI-3 does not. From these results, we can identify two factors that have a strong impact on the amount of memorized samples:

1) Training resolution: Given a fixed model architecture, the resolution at which the model is trained is inversely proportional to the amount of memorized samples. Intuitively, the lower the resolution, the harder it is for a model to actually read the answers from the image and the easier it is for it to minimise the loss by memorization. For instance, while at the highest resolution for Donut $|M| = 63$, as the training resolution decreases, $|M|$ grows to 109, 168 and

[1] Notice that removing the canaries set from the training set does not yield a difference in generalization performance.

to an extremely high level of 756 for the lowest training resolution.

2) Pretraining: Manually inspecting the samples extractable by the generalization baseline, we observe that for Donut and Pix2Struct, these contain highly repeated answers (e.g., page, table and figure numbers) or frequently repeated names of organizations (e.g., ITC and AHA). For PaLI-3, we instead observe that, besides trivial answers like the ones extracted for Donut and Pix2Struct, the generalization baseline correctly responds to questions whose answer relies on general knowledge (e.g., the meaning of ambiguous acronyms that can be resolved considering the topic of the input document, properties of chemical substances or general geographical notions). This is attributable to the web-scale pretraining. The lower amount of samples in M may also indicate that a better pre-trained model may rely less on memorization even at relatively low training resolutions due to their better generalization abilities: indeed, of all the models, PaLI-3 produces the best generalization performance on the test set (87.6 ANLS compared to 76.6 and 67.5 of the best Pix2Struct and Donut variants, respectively).

4.2. Extractable Memorization and Simplicity Scores

The method proposed in the previous section may incorrectly identify some extractable answers as memorized due to the randomness of the training process. To show our attribution technique mostly identifies memorized samples, we leverage a modified version of the memorization and simplicity metrics developed in (Feldman, 2019; Zhang et al., 2021a).

Memorization and simplicity scores. Let $\mathcal{A}$ be stochastic training algorithm. For each sample $(I_i, Q_i, a_i) \in \mathcal{D}^C$, we would like to estimate the Memorization score (Feldman, 2019):

$$\begin{aligned}\mathcal{M}(\mathcal{A}, \mathcal{D}^{tr}, i) = P_{f\sim\mathcal{A}(\mathcal{D}^{tr})}[f(I_i, Q_i) = a_i] - \\ P_{f\sim\mathcal{A}(\mathcal{D}^{tr-i})}[f(I_i, Q_i) = a_i]\end{aligned} \tag{1}$$

where $\mathcal{D}^{tr-i}$ indicates $\mathcal{D}^{tr}$ from which sample i has been removed. This score quantifies the difference between the probability that a model produces a correct prediction on a canary given the model has seen it at training time or not.

A score of 1 indicates the model can predict correctly on an input sample exclusively if it has seen it at training time. A score of 0 indicates that it has the same probability to produce a correct prediction whether the sample was or not in the training set. Note that the memorization score says nothing about the model's accuracy on a sample (e.g., both a model that is always right or always wrong exhibits low memorization). To account for this (Zhang et al., 2021a) proposed a simplicity score $\mathcal{S}(\mathcal{A}, \mathcal{D}^{tr}, i)$ that sums the first and second terms of Equation (1). This allows to distinguish cases where a model fails to memorize a sample because it is hard to answer even when trained on (low simplicity), or because the answer is easy to produce even when not trained on (high simplicity).

Extractable memorization and simplicity. These two scores do not quite reflect the property we are interested: they inform us about the correctness of a model on an input sample (I, Q), and not about the ability to answer a question given a partial context (I^{-a}, Q). We thus adapt the memorization and simplicity scores accordingly, to consider the probability of a successful extraction:

$$\begin{aligned}\mathcal{M}_E(\mathcal{A}, \mathcal{D}^{tr}, i) = P_{f\sim\mathcal{A}(\mathcal{D}^{tr})}[f(I_i^{-a_i}, Q_i) = a_i] - \\ P_{f\sim\mathcal{A}(\mathcal{D}^{tr-i})}[f(I_i^{-a_i}, Q_i) = a_i]\end{aligned} \tag{2}$$

We call Equation (2) the Extractable Memorization score, and refer to the first term as the in-sample extractability and to the second as the out-sample extractability. Similarly, we define an Extractable Simplicity score $\mathcal{S}_E(\mathcal{A}, \mathcal{D}^{tr}, i)$ as the summation of the two terms.

Empirical estimation. Analogously to (Feldman, 2019; Lukasik et al., 2023), we compute empirical estimates $\hat{\mathcal{M}}_E$ and $\hat{\mathcal{S}}_E$ of $\mathcal{M}_E$ and $\mathcal{S}_E$ by training on random splits S^k of the training set that omit or maintain at random samples from the canary set $\mathcal{D}^C$. We produce a total of K splits, and define the indices of the splits containing a sample i as $K_{in} = \{k : (I_i, Q_i, a_i) \in S^k\}$ and $K_{out} = \{k : (I_i, Q_i, a_i) \notin S^k\}$. We then compute the in-sample and out-sample extractability scores as $\frac{1}{|K_{in}|}\sum_{k\in K_{in}} \mathbb{1}(a_i = f_{S^k}(I_i^{-a_i}, Q_i))$ and $\frac{1}{|K_{out}|}\sum_{k\in K_{out}} \mathbb{1}(a_i = f_{S^k}(I_i^{-a_i}, Q_i))$. Given that training Document-Based VQA systems is extremely expensive, we follow the sampling procedure in (Carlini et al., 2022) in order to produce $K = 50$ splits such that each canary is in or out of a split exactly 25 times.

Experimental results. In Figure 5 we plot 2D histograms of the memorization and simplicity scores, $\hat{\mathcal{M}}_E$ and $\hat{\mathcal{S}}_E$. As it can be seen, the vast majority of the samples are not extractable at all, so we have $\hat{\mathcal{M}}_E = \hat{\mathcal{S}}_E = 0$. Some fraction of the training canaries are counterfactually extractable though, i.e., $\hat{\mathcal{M}}_E \gg 0$. To determine whether the technique proposed in Section 4.1 is actually identifying memorised samples, we now plot the Extractable Memorization and Simplicity scores of samples $E - G$ that were extractable only from the original model f, as well as the "control" samples G that were extractable by the generalization baseline f_G. As expected, samples in G have low memorization scores $\hat{\mathcal{M}}_E$: these answers can be extracted whether we train on them or not. In contrast, samples in $E - G$ have memorization scores $\hat{\mathcal{M}}_E$ that vary between 0 and 1. Most of the samples are close to the line $\hat{\mathcal{S}}_E = \hat{\mathcal{M}}_E$, indicating that the in-sample extractability is the only term contributing to $\hat{\mathcal{M}}_E$ (i.e., a model must see a sample at training time in order to extract it, and cannot extract it due to generalization only).

5. Ablations on the Extraction Context

So far, we studied the extractability of an answer a assuming knowledge of all other parts of an input. We now relax this assumption to both gain further insights into the factors influencing extractability, and, in some cases, to simulate more realistic attack scenarios in which perfect knowledge of the context (I^{-a}, Q) is not available. Indeed, while perfect knowledge of the context is unlikely in many cases, it is possible for an attacker to craft an approximation of the context (e.g., because the information they are seeking is contained in documents with a known or fixed structure, like driving licences or forms available online).

Before delving in the results, we point out that just like modifying the way a LLM is prompted can modify its output significantly, changing the way the VLMs are prompted changes which samples are extractable. For this reason, in few cases, the amount of extractable samples may increase with respect to the baseline scenario we considered so far, especially for cases in which the generalization baseline is weakened by the reduced information contained in the approximation of the context.

5.1. No Text in the Image

For LLMs, prior work has shown that prompting a model with the prefix of a memorized string is a reliable way of extracting data (Carlini et al., 2023b; Tirumala et al., 2022). Yet, for Document-Based VQA systems it is unclear whether the models actually needs to read any surrounding text in a document in order to recall the answer. For this reason, we study the case in which *all* text is removed from the image I. If the model can still respond correctly, it indicates the model is relying on the question and non-textual features (e.g., layout, presence of icons or images etc.) in order to regurgitate the answer. This experiment also represents a practical threat model where the attacker knows the layout of a document (e.g., because it is a form available online or a document with a fixed structure like driving licences or ID cards) but has little to no knowledge about its contents.

Results: Figure 6 shows that in case of Donut and Pix2Struct, the absence of text in the image significantly reduces the ability of the model to return the correct answer. In case of Donut the amount of samples in M is 26. Pix2Struct shows a similar decrease from about 94 to 27. The amount of PIIs returned is also significantly reduced, and consisting mostly of highly repeated PIIs (more than 6 times). In the case of PaLI-3, we also observe the model responds correctly to answers requiring general knowledge (e.g., the name of chemical substances from their symbols contained in the questions, names of animal species portrayed in pictures contained in the document). The increase in the amount of extractable answers may be related to the fact that, when the extraction fails, a typical pattern is for the model to read another part of the document. When no text is present, it is easier for the model to retrieve the information from the general knowledge it acquired at pre-training time. For PaLI-3, no PII is extracted.

Reliance on surrounding text: The lack of any text in the document significantly reduces the ability to extract unique PIIs.

5.2. Imperfect Knowledge of the Training Question

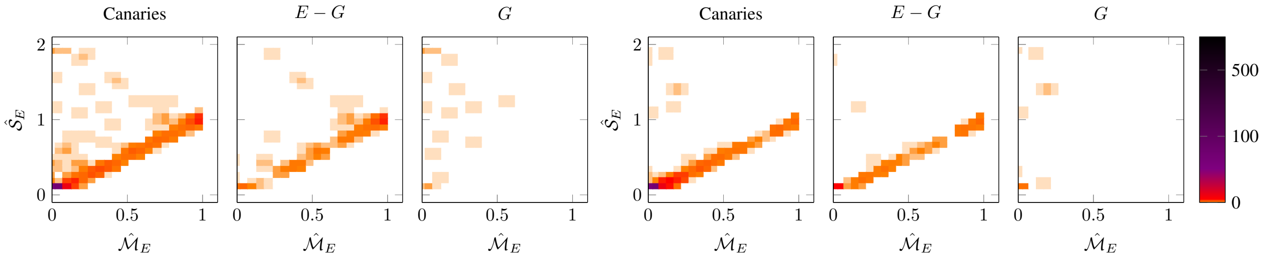
To understand whether the model is memorizing an association between the exact question Q and answer a, we measure whether we can extract the answer when the question is paraphrased. We create paraphrases Q' of Q and extract the answers using (I^{-a}, Q'). To this end, we use PaLM2 (Anil et al., 2023) to create a paraphrased question for each canary question. An example of paraphrase is the following: if the question Q is“What is the address shown in the document?”, then the paraphrase Q' can be “What is the street name and city shown in the document?”. This experiment also reflects the setting in which the attacker does not know the exact phrasing of the training question Q and approximates it with their own words.

Results: Figure 6 shows that the number of extracted answers significantly drops, but is still non-negligible. For both Pix2Struct and Donut we observe several unique PIIs are extractable (e.g., names of individuals, serial numbers of tickets and travel destinations). The extractability increases in the case of PaLI-3, but is again related to questions probing general knowledge and reveal no PII.

Robustness to paraphrasing of Q: Uncertainty about the exact phrasing of a question that queries PII does not prevent extraction of sensitive information, but can reduce the amount of extractable samples.

5.3. Robustness to Image Perturbations

An attacker may be able to craft a document similar to the one originally used for training, but the scanning procedure naturally induces some small visual differences that may influence the extractability of the answers (e.g. brightness changes, small rotations or translations). For this reason



a) Pix2Struct base 1M Pixels

b) Donut 2560 x 1920

Figure 5. Distributions of the $\hat{\mathcal{M}}_E$ and $\hat{\mathcal{S}}_E$ scores for all the canaries, $E - G$ and G for both Pix2Struct base 1M Pixels (three panels on the left) and Donut 2560 x 1920 (three panels on the right). Samples in $E - G$ have high memorization scores, while samples in G do not.

we consider the case in which the original context I^{-a} is perturbed with augmentations that reflect plausible differences that may incur between the training and adversarially crafted document scans. For this purpose, we consider the following augmentations: 1) brightness change: we increase ($\times 1.3$, $\times 2$) or decrease ($\times 0.8$, $\times 0.5$) the brightness of the document; 2) small rotations: we randomly rotate by ± 5 or ± 10 degrees; 3) small translations: we randomly shift the image by ± 20 and ± 100 pixels along both axes.

Results: In Figure 6, we can see that brightness changes can indeed reduce the amount of extractable information, but the amount of extractable samples is still significantly high. In most cases, the stronger the change in brightness, the less the answer is extractable. However, a substantial amount of samples remains extractable, especially with respect to the context perturbations considered in the previous sections. Rotating or translating the image has a stronger adverse effect on the extractability of answers, indicating that spatial information plays a more important role for extractability than the intensity information. Notice, the amount of extractable samples under image perturbations is significantly larger than the amount extractable when the question is paraphrased, indicating that precise knowledge of the question Q is more important for an extraction attack than precise knowledge of the original scan I^{-a}. This also suggests that extractability is more likely to be triggered in the presence of the training question than in presence of the input image I^{-a}.

Robustness to Image Perturbations The amount of extractable samples is relatively robust to brightness perturbations and less to spatial transformations. An adversary does not need to reproduce a perfect copy of the original training image to extract the answer.

5.4. Permuting Modalities

Document-Based VQA systems contain both a visual component and a language component, each of which are fine-tuned on the training data. Extensive evidence has been provided that each of these components can memorise training data *in isolation* (Feldman, 2019; Lukasik et al., 2023; Carlini et al., 2022; 2019). Therefore an interesting question is whether it is possible for a multimodal model to extract the answers independently of one of the two input modalities. For this reason, we consider two experiments that randomise the relationship between the two input modalities.

Extractability based on questions only. At inference time, we feed the model a partial image with an unrelated question $(I_j^{-a_j}, Q_i)$, where $i \neq j$ and there is no training sample with question Q_i applied to image I_j, and the correct answer to question Q_i does not appear in the text of image I_j. This experiment evaluates the ability of the model to respond solely based on the question and reflects the case in which the attacker does not know the image I_i at all.[2]

Results: In the setting where we try to extract the original answer a_i, as visible in the Shuffling column in Figure 6, we can extract only 4 answers in case of Donut, and 21 in case of Pix2Struct. Among all the samples in M, we can also find some sensitive samples containing area codes, names of individuals and dates in which the documents were issued. The sensitive samples are also repeated only once or at most twice in the model's training set. While 2 answers can be extracted for PaLI-3, no PII was extracted.

Extractability based on images only. As in the previous experiment, we provide the model with a partial input image and an unrelated question that does not contain an answer within the image. We then measure whether we can extract an answer to one of the questions that was asked about this image during training. We find no extractable answers in this setting, which suggests that the question plays a more predominant role in the extraction.

[2] We have also tried replacing the input image with constant intensity value set to black, white or the average value of I_i. No answer was extractable in this case, perhaps because such images are too far out-of-distribution.

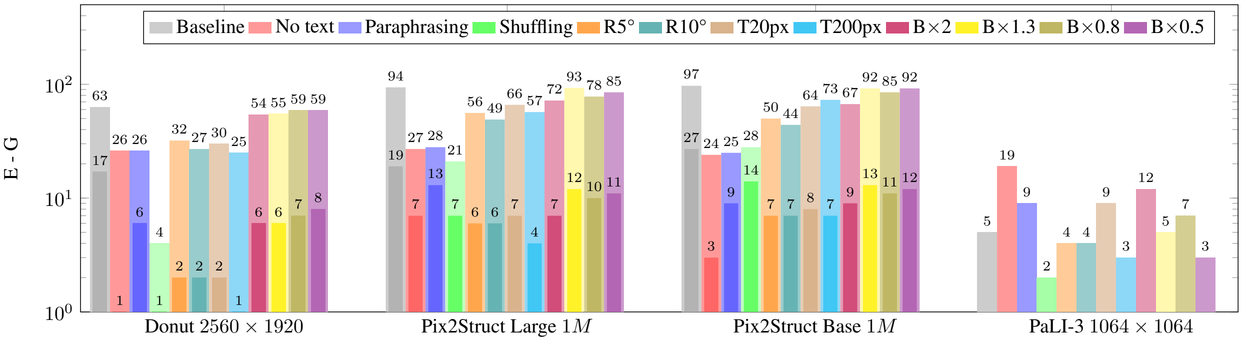


Figure 6. Extractability of answers when the context does not contain the text (No Text), the question is paraphrased (Paraphrasing), or not related to the image but the model still responds correctly (Shuffling), the image undergoes rotations (R5* and R10*), translations (T20px, T100px) and when brightness is changed by a mutliplicative factor (B×2, 1.3, 0.8 or 0.5). Darker colors indicate the number of PII samples that are extractable. Y-axis is in logscale. Across all deployable models, PaLI-3 exhibits the lowest amount of extractable information.

Dependency of extractability on modalities In few cases, the model can leverage the language component alone to extract sensitive answers. If the training answers are not present in the image modality and the question was not seen at training time for a specific document, the image alone is not sufficient to extract any memorized answer.

6. Defenses

To conclude our study, we consider various mitigation strategies and measure their impact on memorization and generalization capabilities of the models (by computing the ANLS (Mathew et al., 2021) on a held-out test set):

- **Inference Time Paraphrasing (ITP)**, similar to (Somepalli et al., 2023a) we consider its effectiveness as a defense strategy.
- **Prepending/Appending a Random String (PR/AR)** Inspired by (Somepalli et al., 2023a), we perturb the question by prepending or appending a short 6-digit random string to the question.
- **Extraction Blocking (EB)** For each original sample (I, Q, a), we suggest adding to the training set a corresponding sample $(I^{-a}, Q, \texttt{'ANSWER NOT PRESENT'})$. This approach is similar in spirit to the intuition behind the V-CSS part of the algorithm proposed in (Chen et al., 2020) to improve the grounding of VQA systems.

Results: We observe that although ITP and PR/AR can reduce the amount of extractable information, they also yield a substantial drop in ANLS on a held-out validation set. Therefore they can only be implemented as mitigation strategies if the practitioners are willing to pay a cost in terms of performance. On the other hand, we observe EB to be extremely effective, reducing to 0 the amount of extractable samples for most models. Furthermore, although we apply the technique by augmenting the original training set using the context (I^{-a}, Q), it is also generalizes to adversaries that query the model with the approaches considered in Section 5 (see Table 2), while producing an increase in the ANLS (in a similar way V-CSS does in (Chen et al., 2020)).

Δ ANLS / $\|M\|$	PR	AR	ITP	EB (Ours)
Donut	-3.4 / 38	-3.1 / 34	-12.5 / 26	**+1.2 / 2**
Pix2Struct-B	-2.9 / 40	-1.9 / 35	-12.9 / 28	**+3.4 / 0**
Pix2Struct-L	-2.6 / 37	-2.0 / 33	-13.8 / 25	**+2.1 / 0**
PaLI-3	-3.7 / 4	-3.2 / 3	-8.1 / 9	**+1.5 / 0**

Table 1. Variation of ANLS (utility metric for DocVQA) and amount of extractable samples in M for various countermeasures with respect to the standard training procedure.

7. Conclusion

In this study we have analysed the memorization abilities of three recent Document-Based VQA systems. We have shown these models can memorize information that is unique or sporadically repeated across the training set and it can be extracted when the model is prompted with incomplete context. We have introduced an extension of the Counterfactual Memorization and Simplicity scores that reveals that the memorized information identified by our attribution method is indeed also memorized according to these more computationally expensive scores. We have analysed the influence of the context on the extractability of samples, and studied the effectiveness of a few heuristic techniques, one of which results in a reduction of the amount of extractable samples and improves the test performance.

Impact statement

This paper shows it is possible for a malicious user to prompt a model to reveal training data. This phenomenon is studied in a worst-case but plausible condition in which the attacker knows the training image and question, except for the answer. Our study only represents a starting step in the direction of prompting VLMs to elicit the extraction of private data. It may be possible for an attacker to develop more sophisticated attack strategies. Such strategies can be used both in a beneficial way (e.g., for organizations to audit the privacy preserving properties of their systems) or maliciously (e.g., for an attacker to obtain confidential information).

In this study we have used public data, and for further caution we have anonymised all the sensitive samples we reported in our qualitative analysis. Indeed, in some parts of the world the Right To Be Forgotten is in place, and the individuals whose data is reported in the considered public dataset my ask for their data to be cancelled. When performing our quantitative analysis, we report aggregate numbers and described the extractable samples without revealing their exact content for the same reasons. Therefore, we expect no individual or organization to be harmed by reporting our results.

Furthermore, although we propose a countermeasure (EB) that is effective across all the attack scenarios we considered, it is still a heuristic approach and may not prevent extraction in case more sophisticated attack techniques are developed. Furthermore, it may hypothetically introduce a ”side-channel” that an adversary might exploit to increase the exposure to membership inference attacks: if the model responds with the default negation, this may be seen as an index the sample was in the training set. This may not be relevant for several applications, where the information to be protected is not the membership of a document to the training set but the specific content of the document, but may be problematic in other applications. An obvious solution would be to apply Differentially Private (DP) training. Since DP provides guarantees about the likelihood of success of Membership Inference Attacks (MIA), but no closed form formula is available to translate the MIA guarantees into extraction prevention guarantees, practitioners could consider tuning (ϵ, δ) so as to empirically reduce extraction to zero. However, scaling DP to VLMs without causing significant utility degradation is a complex task that requires extensive and difficult parameter tuning (Kurakin et al., 2022), since noise addition and norm clipping could impact one of the two modalities disproportionately (Hu et al., 2022). beyond the scope of this work.

Acknowledgements

We thank Chiyuan Zhang for his valuable feedback on our draft. We thank Kenton Lee, Xi Chen and the PaLI team for their valuable technical support. We thank Michal Lukasik and Vaishnavh Nagarajan, for insightful discussions about memorization, our draft and how to frame our work in the context of the literature. We thank Samuele Marro for proofreading the work. The project was entirely funded by Google. Francesco Pinto’s PhD is funded by the European Space Agency.

References

Gpt4-v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Accessed: 2024-02-18.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023a.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=TatRHT_1cK.

Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., and Zhuang, Y. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10800–10809, 2020.

Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C. R., Goodman, S., Wang, X., Tay, Y., et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023a.

Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., Mustafa, B., Goodman, S., Alabdulmohsin, I., Padlewski, P., Salz, D., Xiong, X., Vlasic, D., Pavetic, F., Rong, K., Yu, T., Keysers, D., Zhai, X., and Soricut, R. Pali-3 vision language models: Smaller, faster, stronger, 2023b.

Dancette, C., Cadene, R., Teney, D., and Cord, M. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1574–1583, 2021.

Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., and Morariu, V. End-to-end document recognition and understanding with dessurt. In *European Conference on Computer Vision*, pp. 280–296. Springer, 2022.

Feldman, V. Does learning require memorization? A short tale about a long tail. *CoRR*, abs/1906.05271, 2019. URL http://arxiv.org/abs/1906.05271.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.670. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.670.

Guo, C., Bordes, F., Vincent, P., and Chaudhuri, K. Do ssl models have d\'ej\a vu? a case of unintended memorization in self-supervised learning. *arXiv preprint arXiv:2304.13850*, 2023.

Hu, P., Wang, Z., Sun, R., Wang, H., and Xue, M. M4i: Multi-modal models membership inference. *ArXiv*, abs/2209.06997, 2022. URL https://api.semanticscholar.org/CorpusID:252280738.

Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Jabri, A., Joulin, A., and van der Maaten, L. Revisiting visual question answering baselines. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016*, pp. 727–739, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46484-8.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

Kurakin, A., Song, S., Chien, S., Geambasu, R., Terzis, A., and Thakurta, A. Toward training at imagenet scale with differential privacy, 2022.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, K., Joshi, M., Turc, I. R., Hu, H., Liu, F., Eisenschlos, J. M., Khandelwal, U., Shaw, P., Chang, M.-W., and Toutanova, K. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18893–18912. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/lee23g.html`.

Lukasik, M., Nagarajan, V., Rawat, A. S., Menon, A. K., and Kumar, S. What do larger image classifiers memorise?, 2023.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, January 2021.

Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., and Jawahar, C. V. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591, 2022. doi: 10.1109/WACV51458.2022.00264.

Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X., and Wen, J. Counterfactual vqa: A cause-effect look at language bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01251. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01251`.

Plant, R., Giuffrida, V., and Gkatzia, D. You are what you write: Preserving privacy in the era of large language models. *arXiv preprint arXiv:2204.09391*, 2022.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

Rijhwani, S., Anastasopoulos, A., and Neubig, G. OCR Post Correction for Endangered Language Texts. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5931–5942, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.478. URL `https://aclanthology.org/2020.emnlp-main.478`.

Schaefer, R. and Neudecker, C. A two-step approach for automatic OCR post-correction. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 52–57, Online, December 2020. International Committee on Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.latechclfl-1.6`.

Si, Q., Meng, F., Zheng, M., Lin, Z., Liu, Y., Fu, P., Cao, Y., Wang, W., and Zhou, J. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL `https://api.semanticscholar.org/CorpusID:252780087`.

Smith, R. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pp. 629–633, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2822-8. URL `https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf`.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *arXiv preprint arXiv:2305.20086*, 2023b.

Tirumala, K., Markosyan, A., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35: 38274–38290, 2022.

Tito, R., Karatzas, D., and Valveny, E. Hierarchical multimodal transformers for multi-page docvqa. *arXiv preprint arXiv:2212.05935*, 2022.

Tito, R., Nguyen, K., Tobaben, M., Kerkouche, R., Souibgui, M. A., Jung, K., Kang, L., Valveny, E., Honkela, A., Fritz, M., and Karatzas, D. Privacy-aware document visual question answering. *arXiv preprint arXiv:2312.10108*, 2023.

Zhang, C., Ippolito, D., Lee, K., Jagielski, M., Tramèr, F., and Carlini, N. Counterfactual memorization in neural language models. *CoRR*, abs/2112.12938, 2021a. URL https://arxiv.org/abs/2112.12938.

Zhang, C., Ippolito, D., Lee, K., Jagielski, M., Tramèr, F., and Carlini, N. Counterfactual memorization in neural language models. *arXiv preprint arXiv:2112.12938*, 2021b.

$\lvert M \rvert$/ #PII	No Text	Paraphrasing	Shuffling	R5°	R10°	T20px	T200px	B×2	B×1.3	B×0.8	B×0.5
Donut	0 / 0	0 / 0	0 / 0	1 / 0	0 / 0	0 / 0	0 / 0	6 / 1	6 / 1	2 / 0	5 / 0
Pix2Struct-B	0 / 0	1 / 0	1 / 0	2 / 0	2 / 0	0 / 0	0 / 0	0 / 0	4 / 0	4 / 0	0 / 0
Pix2Struct-L	1 / 0	0 / 0	0 / 0	1 / 0	1 / 0	0 / 0	0 / 0	0 / 0	4 / 0	2 / 0	0 / 0
PaLI	0 / 0	0 / 0	2 / 0	1 / 0	1/ 0	0 / 0	2 / 0	3 / 0	1 / 0	0 / 0	1 / 0

Table 2. Effectiveness of extraction blocking for the various contexts portrayed in Figure 4. Notice, we do not include in the training sets any of the contexts we consider in this table. This indicates the protection offered by Extraction Blocking extends beyond the types of context provided at training time.

A. Computational Cost of Training

Donut Fine-tuning Donut at maximum input resolution requires 64 A100 GPUs for a day. Given its relatively compact size (176M parameters), Donut can be trained on high-resolution input images (2560 $\times$ 1920 $\approx$ 5M pixels), a crucial aspect for achieving optimal performance. Lowering the resolution can significantly reduce the cost of training, however, as we observe, it increases the tendency of the model to memorize the training data and reduces the generalization capabilities of the models. Therefore it is not recommended.

Pix2Struct Fine-tuning Pix2Struct Base, independently of the resolution, requires 32 TPUv2 for about 5 hours. Training Pix2Struct Large, independently of the resolution, requires 64 TPUv2 for about 5 hours. Due to its relatively larger size, the smaller model is fine-tuned at a resolution of about $1.2M$ pixels, while the larger model is fine-tuned at a resolution of about $0.8M$ pixels.

PaLI-3 Fine-tuning PaLI-3 64 TPUv2 for 15 hours. Due to its size (5B parameters), it is typically fine-tuned at a resolution of approximately $1.1M$ pixels (1064 $\times$ 1064).

Computing the memorization scores The amount of compute needs to be multiplied by the number of runs for each measurement: for the simplest attribution method we consider, we only need 2 runs; for the counterfactual extractable memorization and simplicity scores, we need to perform 50 runs. Performing more is both computationally prohibitive and expensive for the storage of the largest models we consider.

B. Further results

Effectivenes of EB for prompting strategies not used in the training set In Section 5 we have considered several ways to prompt the model. Since EB includes only samples using a worst-case prompting strategy (I^{-a}, Q), it may be natural to wonder whether EB is still effective if an adversary prompts the model in different ways. We observe the technique is actually still extremely effective, see Table 2

C. PII categories and their frequencies in the canaries

We manually annotate each answer in the canaries set as either PII or non-PII. We also classify each PII element as one of the following classes: Places, Person, Temporal, Contact (Phone/Fax/Email), NRP (Nationality Religion Politic), URL, and other forms of IDs (e.g. card numbers, serial numbers of tickets, document or people numerical identifiers etc.). The distribution of PII in the canaries set $\mathcal{D}^C$ is reported in Figure 7.

D. Further Related Works

D.1. Document-Based Visual Question Answering

Given the greater simplicity of solving the VQA problem by separating the tasks of document reading and document understanding, OCR-reliant systems have been the state of the art for a few years (Tito et al., 2022; Huang et al., 2022). However, as argued by (Kim et al., 2022), OCR-reliant systems have the disadvantage of requiring an expensive OCR-preprocessing step, making the inference cost higher in case high-quality OCR results are required, with errors of the OCR system propagating to the VQA component. The phenomenon is particularly apparent for languages with complex

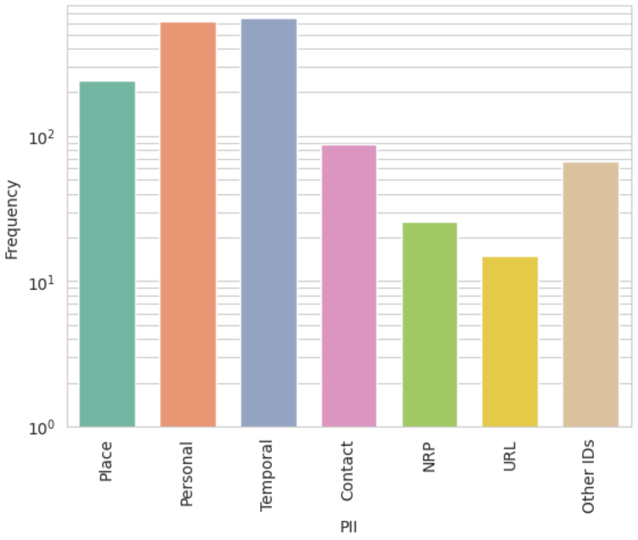


Figure 7. Frequency of different types of Personally Identifying Information (PII) in the canaries set $\mathcal{D}^C$.

character sets, requiring an expensive post-OCR correction module (Rijhwani et al., 2020; Schaefer & Neudecker, 2020). For these reasons, OCR-free systems like (Kim et al., 2022; Lee et al., 2023) have received increasing attention, with state-of-the-art models like PALI-3 (Chen et al., 2023b) closing the performance gap between the OCR-reliant and OCR-free models. In this work we mainly focus on three state-of-the-art OCR-free systems that differ in model size, architecture and pre-training stages. We consider both **Donut** (Kim et al., 2022) and **Pix2Struct** (Lee et al., 2023) among the set of models that are specialised to perform document understanding. We also consider **PALI-3** (Chen et al., 2023b), a foundational vision-language model that can be fine-tuned in order to solve the task of document understanding, achieving state-of-the-art performance.

D.2. Relations to Distributional Shortcut Learning in VQA

It is known that VQA systems can produce correct responses due to their ability to learn and leverage the frequent association of a specific answer to some question (linguistic shortcut) (Jabri et al., 2016; Niu et al., 2021; Goyal et al., 2017; Chen et al., 2020). For instance, if the question is *"What is the colour of the grass?"*, if the grass is green in most of the training images for which the question is asked, the model will respond green independently of the actual colour in the considered test image. This type of shortcuts does not need to be exclusively linguistic, and may involve the frequent co-occurrence of elements in the input image (visual shortcut) or their combination with specific words in the question (multimodal shortcut) (Dancette et al., 2021; Si et al., 2022). In other terms, VQA systems can learn simple rules relying on spurious but predictive features that co-occurr across multiple samples in order to respond accurately even when the input image lacks the considered information or contradicts it.

The concurrent work of (Tito et al., 2023) has shown this phenomenon occurring also in document-based Visual Question Answering. The authors propose a new federated learning dataset containing invoices from several data providers. Since a provider's information (specifically, their name and email address) is *repeated across several invoices* that share visual and linguistic similarities (e.g., identical layout, formatting, logos, fields etc.), a model can infer a provider's name or email address correctly on *previously unseen test* documents from the known provider that do not contain the requested information. In contrast, we focus on *centralised* training and perform attacks on training documents. Our analysis aims at factoring out the cases when models can extract information by leveraging knowledge learnt from other samples (which we consider as a form of generalization rather than memorization). While their goal is to protect the identity of providers (in a federated, group privacy setting), our goal is to protect the individual answers.