# Disparity Analysis of Unsupervised Membership Inference Attacks: A Case Study in DocVQA

João Medrado Gondim[1], Helio Pedrini[1], and Sandra Avila[1]

Institute of Computing, University of Campinas
Campinas-SP, Brazil, 13083-852
{joao.gondim,helio,sandra}@ic.unicamp.br

**Abstract.** Membership Inference Attacks (MIAs) are widely used to audit the privacy of Artificial Intelligence (AI)-powered systems, which, despite their indisputable success across a myriad of tasks, are known to expose their training data. Moreover, recent work has shown that MIAs should account for disparities arising from inherent random processes in their functioning when evaluating data exposure. In this paper, we analyze the reliability and completeness of attacks on unsupervised MIAs: when the attacker has a target set that overlaps with some of the training data but does not know which instances are in the target set, often resorting to clustering methods to distinguish between members and non-members. Through a series of experiments, we show how to adapt the framework of Coverage, and Stability to the unsupervised setting, assessing the influence of the evaluated set and the clustering seed on the privacy estimation being measured. We explore unsupervised attacks tailored for the DocVQA task and present evidence that applying the same attack to subsamples of the target set can yield different predictions, which can be harnessed to increase true positive rates.

**Keywords:** Privacy · Membership Inference Attacks · DocVQA.

## 1 Introduction

The rapid development of deep neural networks (DNNs), leveraged by vast (and often unconsented) data collection, has enabled artificial intelligence (AI) to tackle increasingly complex tasks, some with acute privacy implications. Research has shown that data used during DNN training phase can be memorized and later leaked [3, 5, 17, 20, 24]. Such vulnerabilities, collectively termed *privacy attacks* [11], represent a threat vector for malicious actors and a principled tool for probing the boundaries of data exposure [8], and auditing DNN memorization [25]. Privacy attacks aim to extract knowledge from Machine Learning (ML) models that their owners keep private [21], and are typically categorized by the type of knowledge they expose. Membership Inference Attacks (MIAs) represent on such category. Its goal is to determine whether a specific data sample was used during model training [31]. Although it is an apparently simple objective, MIAs pose a significant threat to individuals' and societies' private data [18].

Its simplicity, added by the domain-wise generality of MIAs [10], as well as the leverage to serve as a valuable asset for other more critical attacks [3, 5], make MIAs a standard technique for auditing ML model privacy [1, 2, 4, 26].

Considering the adversary's prior knowledge on the targeted data, Nasr et al. [16] argue that the learning of a mapping between how models behave when facing either training or non-training data, i.e., *members* and *non-members*, separates MIAs between two categories. *Supervised attacks* assume access to auxiliary datasets that either partially overlap the training data, allowing direct training of a binary attack classifier, or non-overlapping, but drawn from the same distribution, which allows the training of "shadow models", created to mimic the targeted model's behavior and creation of training data for the binary attack classifier [1, 4, 23, 27, 29]. *Unsupervised attacks* apply when attacker's dataset overlaps with the training set but member labels are unknown; clustering algorithms are then used to separate attack features into members and non-members, employing rules to define each group given the unknown prior [6, 16, 17, 19, 24].

Shadow model attacks, which mimic the behavior of a targeted DNN, become prohibitively expensive at large-scale models [13], such as Visual Language Models (VLMs). As an example, the Donut model [12], tailored for document understanding tasks that leverage inputs from the modalities of image and text, report that the fine-tuning stage of their model on the DocVQA dataset [15] required one GPU day of training on 64 A100 GPUs. This makes unsupervised MIAs well-suited for evaluating VLMs in scenarios of constrained computational resources and where the attacker is assumed to be unaware of the data.

Wang et al. [26] show that MIAs evaluated on a single attack instance often neglect the discrepancies introduced by randomness across their development. They argue that privacy risk metrics should be evaluated over multiple attack methods and instantiations. Supported by empirical results from experiments performed making use of six supervised attacks: Loss Attack [30], Class-NN [23], Augmentation Attack [4], Difficulty Calibration Loss Attack [27], LiRA [1], and Reference Attack [29]. Consequently, the disparity factors they identify are confined to the supervised MIAs. In this paper, we extend this analysis to the unsupervised regime by introducing disparity factors in MIAs where the attacker has no knowledge of the association labels of the target dataset.

Building on the DocMIA code [17], we evaluate two classes of disparity factors in unsupervised MIAs: factors arising from *instances of the same attack*, such as the seed of the K-Means clustering algorithm, or running the same attacks on different target subsets of the training data; and factors regarding the rerun of the same attack *on different subsamples drawn from the targeted subset*. We adapt the disparity metrics of consistency, coverage, and stability proposed by Wang et al. [26] to the subsampling disparity factor, showing that unsupervised attacks pose even greater privacy risks when coverage over random subsampling is considered. Our contributions are as follows

– We adapt the discrepancy metrics from Wang et al. [26] to evaluate the disparity factor introduced by running the same attack on subsamples of the target subset.

- We stress test the consistency of the unsupervised attacks, on the DocMIA code framework [17], over two sources of randomness: K-Means seed variation and attacks runs on subsamples.
- We qualitatively evaluate data points that are more prone to being correctly associated with the training dataset (successfully attacked) after being part of different targeted subsamples.

## 2 Background

This section provides the concepts to understand our work: black-box and white-box MIAs, focusing on the DocMIA work, followed by the disparity analysis framework, used to evaluate the reliability and completeness of the employed attacks across the randomness introduced by reapplying them to subsamples of the target set.

### 2.1 Membership Inference Attacks

Membership Inference Attacks were shown to be a practical threat in the machine learning field in the work of Shokri et al. [23], where the authors also introduced the concept of shadow model training to mimic target models. Besides the already-mentioned categorization into supervised and unsupervised attacks, they can also be classified as white-box or black-box.

**White-box vs. Black-box** This categorization considers the amount of access to the target model that the attacker is assumed to have. White-box attacks allow full access to the model (internal representations, training algorithm, and parameters), enabling features derived from internal computations (gradient norm, loss, logit values) [1,16,30]. Black-box attacks operate under a more constrained scenario, limiting the attacker to final model outputs only (classification predictions, text generation, or confidence scores) [4,22,23]. White-box attacks usually yield better results [16], helping to evaluate the worst cases, while black-box attacks pose a more serious threat, as they require fewer assumptions.

The versatility of MIAs makes them suitable for assessing privacy risks across different tasks. DocMIA work [17] introduces two novel metric-based MIAs specifically designed for Document Visual Question Answering (DocVQA), a task in which inputs (documents) may contain private information or implicate individuals if their association with training data is exposed. DocVQA is formally defined as follows: given an image $x_i$ paired with a question $q_i$, a model $F$, parametrized by weights $\Theta$, must produce an answer $\hat{a}$ that approximates the ground truth $a_i$.

The discriminative metrics rely on information from optimizing the targeted models, or proxy models fine-tuned on distilled knowledge retrieved from the targeted ones they will mimic. More specifically, models are fine-tuned on tuples $\{x_i, q_i, a_i\}$, with the distance to the optimal answer taken into account and the number of steps (these features are averaged across documents with

multiple questions). The core idea is that training data points converge faster than non-training data, needing smaller distances and fewer steps. These features are concatenated with the evolution of the fine-tuned model's accuracy on the given answer, measured via task-appropriate utility metrics: Averaged Normalized Levenshtein Distance (ANLS) and Accuracy (Acc). The final set of features is passed to an unsupervised clustering algorithm to find two clusters; the cluster with the larger average distance covered during the fine-tuning stage is considered the non-member set.

Since full-layer optimization is prohibitively expensive, the paper proposes three alternatives. The first, Optimize One Layer Variant (FL), optimizes a single layer; the second, FL_LoRA augments this with LoRA [9] (FL_LoRA). The third variant, Optimize the Document Image (IG), follows the intuition that training documents will converge faster to the ground truth answer by changing the optimization space to the input image (in this case, the models' parameters are frozen). Features can be extracted under both access regimes: in the white-box setting, fine-tuning is performed on the target model; in the black-box setting, the target model is first queried to distill answers into a proxy model, transferring membership-embedded information so that the fully accessible proxy becomes suitable for membership inference.

We choose the DocMIA framework for its robust code base and for the possibility of contrasting MIAs that rely on optimization and model utility features. Although the authors implement other attacks, our evaluations focus on the newly proposed ones (FL, FL_LoRA and IG), added by ScoreLossAll-UA, which fits an unsupervised clustering algorithm to differentiate members from non-members based on the model's average performance on target-set points, using DocVQA-specific metrics and loss, with the higher-scoring cluster predicted as the member cluster.

### 2.2 Disparity Analysis

DocMIA's authors account for randomness in K-Means initialization by reporting results across five random seeds applied to an attack of 600 data points: 300 member documents and 300 non-member documents. We argue that restricting the evaluation to this single source of randomness can yield an overly optimistic privacy assessment.

To assess our claim, we use the framework of Wang et al. [26]. Their study reveals that attacks should account for randomness across multiple instantiations and across different methods to provide a more complete privacy evaluation, since using a single instance or procedure can expose only a subset of vulnerable training samples. They formalize this concern through three disparity metrics: "Consistency", "Coverage", and "Stability".

**Consistency** is defined as the Jaccard Score between membership predictions averaged by the number of different instances of the attack that have been evaluated (within the paper, instances are defined by supervised attacks that differ in either the dataset partition used to train shadow models or the seeds used to train shadow and attack models). For an attack $\mathcal{A}$ with two different

instances defined by seeds $i$ and $j$ from the set of seeds $S$, $\mathcal{A}^i$ and $\mathcal{A}^j$, consistency of the attack over a target set $D$ is given by:

$$\text{Consistency}_S^{\mathcal{D}}(\mathcal{A}) = \frac{1}{\binom{|S|}{2}} \sum_{i \neq j}^{i,j \in S} J(\mathbb{M}_{\mathcal{D}}(\mathcal{A}^i), \mathbb{M}_{\mathcal{D}}(\mathcal{A}^j)),$$

where $\mathbb{M}_{\mathcal{D}}(\mathcal{A}^i)$ and $\mathbb{M}_{\mathcal{D}}(\mathcal{A}^j)$ represent, respectively, the set of detections from $\mathcal{A}^i$ and $\mathcal{A}^j$, $|S|$ is the number of seeds in the set, and $J$ is the Jaccard Index $J(\mathbb{M}(\mathcal{A}^i), \mathbb{M}(\mathcal{A}^j)) = \frac{|\mathbb{M}(\mathcal{A}^i) \cap \mathbb{M}(\mathcal{A}^j)|}{|\mathbb{M}(\mathcal{A}^i) \cup \mathbb{M}(\mathcal{A}^j)|}$.

A low consistency score suggests that evaluating an MIA on a single instance may underestimate the privacy risk posed by the method, since the Jaccard Index decreases precisely when different attack instances disagree in their inferences. Coverage and Stability are introduced to quantify this underestimation directly, enabling a more complete evaluation of privacy risks relative to randomness.

**Coverage** is defined as the union of positive inferences from an attack $\mathcal{A}$ on each seed $s \in S$:

$$\text{Coverage}_S(\mathcal{A}) = \left\{ x \in \mathcal{D} : \bigcup_{s \in S} \mathcal{A}^s(x) = 1 \right\},$$

where $\mathcal{A}^s(x)$ represents the membership prediction of an attack with seed $s$ at a data point $x$. Therefore, coverage represents the extent of the attack. In contrast, **Stability** is defined with the intersection of positive attack predictions:

$$\text{Stability}_S(\mathcal{A}) = \left\{ x \in \mathcal{D} : \bigcap_{s \in S} \mathcal{A}^s(x) = 1 \right\}.$$

Stability assesses how consistent the privacy evaluation imposed by the MIA method is. In the next section, we explain how we adapt these metrics to account for the randomness introduced by rerunning the unsupervised attacks (which resort to clustering) on subsamples of the target set.

## 3 Methodology

This section describes how we apply the disparity metrics introduced to evaluate the disparity factor of subsampling from the target set and performing the same attack on the subsample. We start by introducing the adaptation of Consistency, Coverage, and Stability metrics to the unsupervised MIAs.

### 3.1 Adapting Consistency, Coverage and Stability

Our claim relies on the observation that the attack decision taken by the clustering algorithm (K-Means in the case of DocMIA) can be highly dependent on the data subjected to it (we show this influence in our first experiment in Section 5.1). This is potentially stronger when considering the distinction between

membership signals, given the close relationship between membership inference attacks and the memorization phenomenon in DNNs [2,28], which manifests differently across different samples used during the training of the targeted models.

We illustrate this with a theoretical exercise: applying one of the unsupervised attacks from the DocMIA framework to conduct a private audit of a model before deploying it. One can select a target set comprising 600 documents, 300 members and 300 non-members (as in the DocMIA paper), extract the features from the FL attack (assuming white-box access to the target model for the sake of explanation), and apply the K-Means algorithm to learn how to cluster them into two groups, minimizing the intersection between members and non-members.

Suppose that the same procedure is repeated, but this time using a random subsample of 300 points drawn from the original set of 600 target samples. For the sake of illustration, assume that this subsample is balanced with respect to the association of the selected points. When the K-Means algorithm is reapplied, will the resulting cluster labels remain consistent with those obtained from the original clustering of the 600 target samples?

We hypothesize that this is not the case and empirically demonstrate that evaluations based on unsupervised MIAs may be overly optimistic when randomness is not properly accounted for. This apparent overestimation stems from adaptations to the Consistency, Coverage, and Stability metrics, as presented as follows.

Consistency measures the discrepancies among different executions of the same attack on subsamples drawn from the original target set. Such discrepancies are evaluated by computing the Jaccard Index between the corresponding predictions, considering only the data points shared by both the targeted subsample and the original attacked set. Formally, given an unsupervised MIA $\mathcal{A}$ applied to a target set $\mathcal{D}$ and to the $i$-th subsample $\mathcal{D}^i_{\text{sample}} \sim \mathcal{D}$, generated from a set of $S$ random seeds, the Consistency of the attack is defined as the average Jaccard Index $J$ between the detections obtained on the subsample, denoted by $\mathbb{M}_{\mathcal{D}^i_{\text{sampled}}}(\mathcal{A})$, and the detections produced by the original attack restricted to the same samples, denoted by $\mathbb{M}_{\mathcal{D}^i}(\mathcal{A})$, as expressed in the following equation:

$$\text{Consistency}^{\mathcal{D}}_{S}(\mathcal{A}) = \frac{1}{|S|}\sum^{i\in S} J(\mathbb{M}_{\mathcal{D}^i_{sampled}}(\mathcal{A}), \mathbb{M}_{\mathcal{D}^i}(\mathcal{A})).$$

A lower Consistency indicates a lack of agreement between the predictions of the attack on the subsample and the predictions of the attack on the corresponding points — reflecting the influence of the sample used in the attack. In contrast, higher values reflect similar predictions between both attacks.

Coverage and Stability can be more naturally defined after introducing the adapted notion of Consistency. Their formulations are presented in the following equations:

$$\text{Coverage}_S(\mathcal{A}) = \left\{ x \in \mathcal{D} : (\mathcal{A}_{orig}(x) = 1) \bigcup \left( \bigcup_{s\in S} \mathcal{A}^s_{sub}(x) = 1 \text{ if } x \in \ \mathcal{D}_{sub} \right) \right\},$$

$$\text{Stability}_S(\mathcal{A}) = \left\{ x \in \mathcal{D} : (\mathcal{A}_{orig}(x) = 1) \bigcap \left( \bigcap_{s \in S} \mathcal{A}^s_{sub}(x) = 1 \text{ if } x \in \ D_{sub} \right) \right\},$$

where $\mathcal{A}_{orig}(x)$ denotes the membership inference assigned to the data point $x$ when the unsupervised attack is performed on the original target set, whereas $\mathcal{A}^s_{sub}(x)$ denotes the inference obtained from the attack executed on the subsampled target set $\mathcal{D}_{sub}$ generated using seed $s$. Accordingly, Coverage is defined as the union between the positive inference produced by the original attack (always accounted to assure no missing predictions) for a given point and the positive inferences obtained from all attacks in which that point belongs to the corresponding subsampled target set. Figure 1 illustrates how Coverage and Stability are constructed from these prediction sets, with Stability being analogously defined by replacing union operations with intersections whenever applicable.

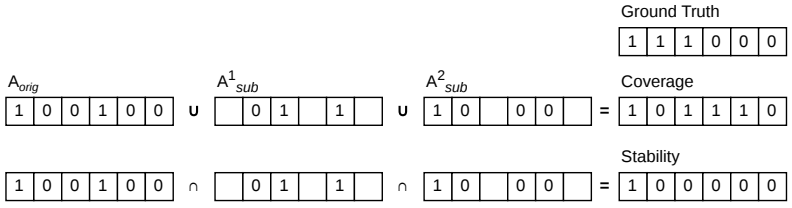



Fig. 1: The proposed measures. Coverage uses the union of predictions at corresponding positions, attempting to expose more training points even at the cost of increasing false positives. Stability, however, enhances precision, taking the intersection between positive inferences.

Figure 1 also highlights the strengths and limitations of each measure. Coverage trades a higher False Positive Rate (FPR) for a higher True Positive Rate (TPR) compared to the attack performed on the original set. In contrast, Stability achieves perfect precision by substantially reducing the number of recovered training samples, identifying only a single member sample in this example. For this reason, we evaluate these measures by focusing on the TPR at low FPR values, following the common practice in the literature, which argues that aggregated metrics may obscure the behavior of attacks in low-FPR regimes [1].

## 4 Experimental Setup

This section describes the experiments designed to support our claims. We begin by presenting the targeted subjects, followed by an experiment illustrating the influence of target sets on attack outcomes. Next, we describe the procedure adopted to account for the variability introduced by attacking subsamples.

### 4.1 Targeted Sets and Model

Since our goal is to analyze the influence of the targeted point set on the clustering algorithm, we considered multiple target sets from the initial stage of

document selection. To this end, ten splits containing 600 data points each were randomly sampled with replacement while preserving the balance of the original training and test partitions of the DocVQA dataset; that is, each split contains 300 "member" samples and 300 "non-member" samples. Unless otherwise stated, all subsequent attack procedures and discrepancy analyses are conducted over these ten generated splits.

We target a Donut model [12], as it is the lightest model available in the DocMIA framework, enabling more experiments across different splits. Additionally, Donut is among the first Visual Document Understanding approaches to operate without relying on an optical character recognition module in its processing pipeline.

### 4.2 Influence of Target Set and K-Means Seed

The first experiment aims to investigate the influence of both the targeted set and the random seed used in the K-Means algorithm. To this end, we run four unsupervised attacks from the DocMIA framework (IG, FL, FL_LoRA, and ScoreLossAll-UA) on each of the ten previously generated splits, storing the extracted features for each attack before applying the clustering algorithm. These extracted features are subsequently reused across an additional $N$ executions with different K-Means random seeds. For each execution, we compute the Jaccard Index, measuring the similarity between the predictions produced by the $n$-th seed and those obtained in the initial clustering run.

### 4.3 MIA Setup on $\mathcal{D}_{sampled}$

The next step is to assess the disparities between predictions obtained from attacks on subsamples and those produced on their corresponding original target sets. Following the first experiment, each split — now represented by the extracted attack features — undergoes the following procedure: (1) random subsampling without replacement of size $L$; (2) reapplication of the same attack to the resulting subsample; and (3) evaluation of the Jaccard Index, Coverage, and Stability with respect to the original attack predictions on the corresponding documents, under low-FPR constraints. Figure 2 illustrates this procedure.

## 5 Results and Discussion

This section presents the results obtained from the previously described experiments. We begin by analyzing the sources of variability associated with the selection of target sets (i.e., different attacked splits) and the initialization seed of the K-Means algorithm through the Jaccard Index between predictions. Next, we evaluate the effects of considering attacks on subsamples using the adapted notions of Consistency, Coverage, and Stability. Finally, we provide a qualitative analysis of training documents that are consistently inferred as members.

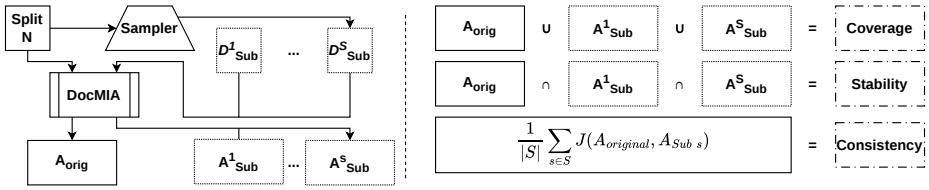
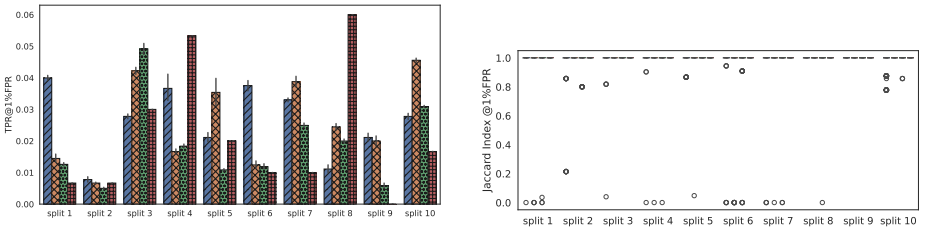



Fig. 2: Proposed pipeline for attack discrepancy analysis. On the left, the beginning of our procedure: a split is used to generate $S$ subsamples; later, both the original split and the subsamples are attacked with the same method. The right side depicts the computation of Coverage, Stability and Consistency.

### 5.1 Disparity Factors: Target Split and K-Means Seed

Our first result highlights the variability in attack effectiveness across different targeted splits. Figure 3a presents the TPR@1%FPR achieved by each attack on each split. The results show that not only do the attack performance metrics vary depending on the selected target split, but also that the attack associated with the highest privacy risk changes across splits (barplots for FL, FL_LoRA, and IG have confidence scores due to multiple predictions from attacks with varying hyperparameters). Next, Figure 3b shows that the choice of the K-Means initialization seed introduces only limited variability into the attack predictions. Specifically, the boxplots representing the Jaccard Index between predictions obtained with a fixed seed and those obtained from 100 different seeds collapse almost entirely to a single line at value 1, with only a few outliers observed for the inspected attacks (boxplots sorted as the attacks from Figure 3a).




(a) (b)

Fig. 3: Evaluation of (a) randomness from target splits and (b) K-Means seeds. Attacks: FL, FL_LoRA, IG, ScoreLossAll-UA.

### 5.2 Disparity Factor: Attacking Subsamples

**Consistency** The Consistency measure evaluates the similarity between the associations inferred by attacks applied to the full set of 600 samples in a split

and those obtained when the same attacks are applied to subsamples of that split, considering only the corresponding documents. This similarity is quantified through the Jaccard Index under a given FPR constraint. Figure 4 presents the boxplots of these indices for subsamples of different sizes, with the mean value indicated by a star.

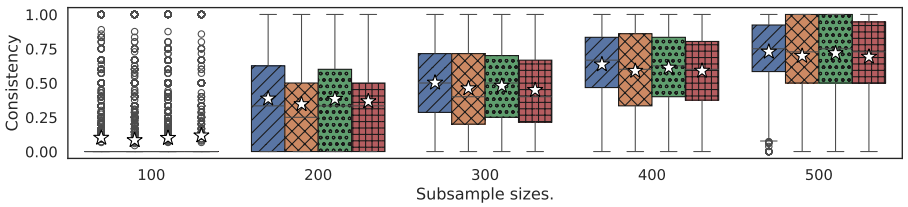



Fig. 4: Consistency@1%FPR. FL, FL_LoRA, IG, ScoreLossAll-UA.

The boxplots reveal a tendency toward greater agreement between attacks as the number of selected documents in the attacked subsamples increases. This behavior is expected, since larger subsamples more closely resemble the original target set on which the attack was initially performed. Conversely, smaller subsampling sizes lead to lower Consistency scores. In particular, the smallest subsamples (100 selected documents) consistently produce the lowest Consistency values, with nearly all nonzero values being treated as outliers.

**Coverage and Stability** Each pair of figures associated with these measures presents the results obtained for one of the DocMIA attacks. In the left-side plots, the vertical axis represents the TPR of the corresponding attack. The boxplots show the TPR@1%FPR values obtained when the attack is executed on each of the generated splits containing 600 samples, whereas the point plots represent the TPR achieved by either Coverage or Stability after performing the attack $S$ times on different subsamples. For visual clarity, only the central tendency values after the first 10 attacks, and every 10 thereafter, are displayed. In the right-side plots, the vertical axis represents the FPR for either Coverage or Stability after running the attack $S$ times across different subsamples, following the same presentation strategy as for the TPR plots.

The results for Coverage are presented in Figures 5, 6, 7, and 8. A similar behavior can be observed across all attacks: as more subsamples are targeted, the union of their predictions increases the number of training documents successfully inferred as members. Nevertheless, the manner in which this increase occurs depends on the size of the targeted subsamples.

While attacks performed on subsamples of size 100 initially contribute little to the TPR, their accumulated predictions eventually surpass those obtained with almost all other subsample sizes as more subsamples are attacked. A consequence of this behavior is the corresponding increase in FPR, which weakens the attack. Interestingly, attacks on subsamples of size 200 tend to exhibit a steady growth in

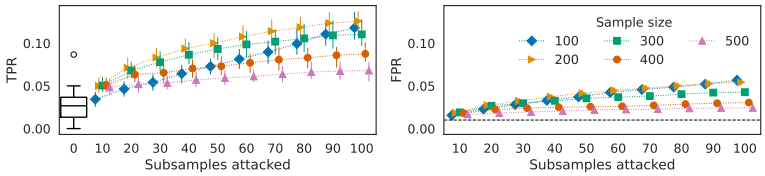



Fig. 5: TPR and FPR for Coverage with the FL attack.

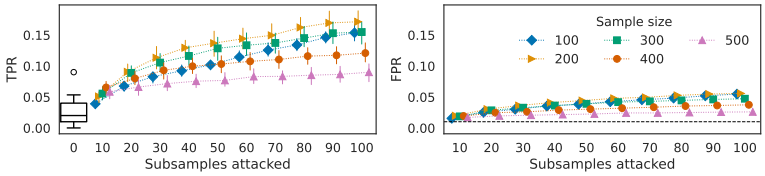



Fig. 6: TPR and FPR for Coverage with the FL_LoRA attack.

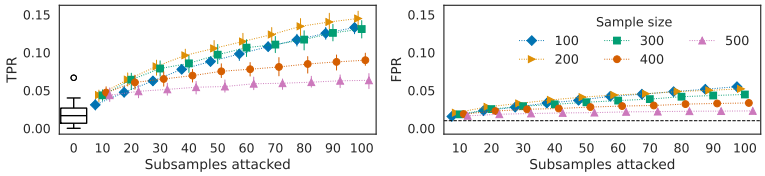



Fig. 7: TPR and FPR for Coverage with the IG attack.

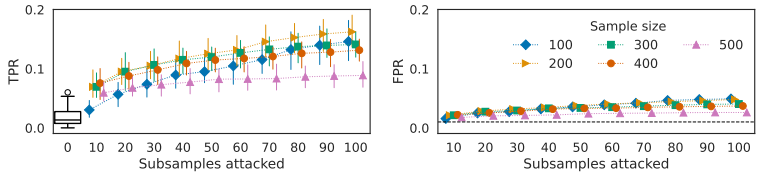
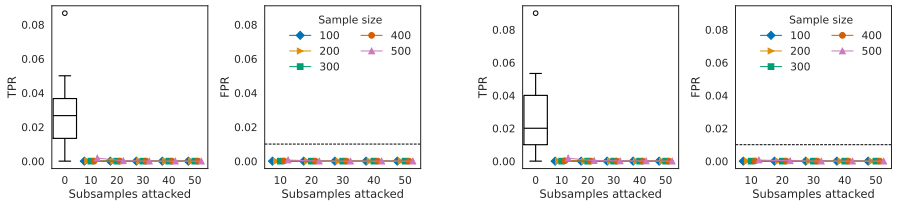



Fig. 8: TPR and FPR for Coverage with the ScoreLossAll-UA attack.

both TPR and FPR from the very beginning across most cases. We conjecture that these behaviors are related to the low Consistency observed for smaller subsample sizes, as illustrated in Figure 4. In Figures 9a, 9b, 9c and 9d, we show the results obtained for Stability. The constraints set for FPR acceptance, added by high variance between predictions (Figure 4), results in a Stability measure that collapses to zero after only a few subsample attacks, for every attack type and subsampling size; we report results only for 50 attacks on subsamples.

### 5.3 Qualitative Analysis of Detected Documents

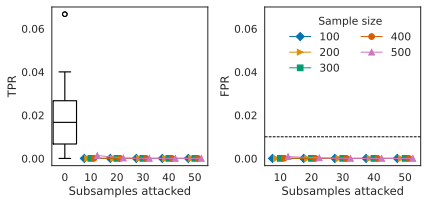
Given the close relationship between susceptibility to MIAs and example-level memorization [2, 14], we hypothesize that, when attacked, training documents continually selected into the "member" cluster form a subsample that corresponds



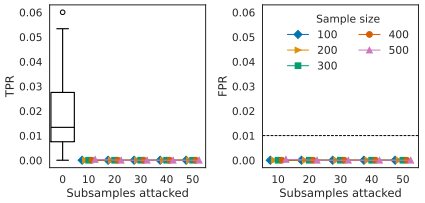

(a) FL attack.




(b) FL_LoRA attack.




(c) IG attack.




(d) ScoreLossAll-UA attack.

Fig. 9: TPRs and FPRs for Stability with different attacks.

to the most highly memorized documents in the dataset and are therefore particularly relevant for a more detailed evaluation.

We further investigate these data points by analyzing, for each training document, the ratio between the number of times it was correctly inferred as a member and the number of times it was selected into a subsample. This analysis considers only attacks performed on 1000 subsamples of size 300 for each evaluated attack, while constraining successful predictions to a maximum FPR of 1%. The choice of subsamples of size 300 is motivated by the observation that their Coverage TPR is generally comparable to those obtained with subsample sizes of 100 and 200, while yielding noticeably lower Coverage FPR values. Figure 10 shows the distribution of these ratios with respect to the number of questions associated with each document in the training set.

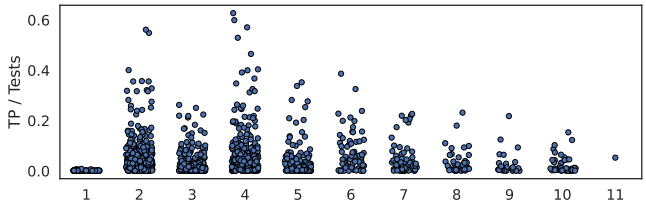



Fig. 10: Attack success ratios across documents grouped by their number of associated questions.

We observe that most of these ratios remain below 0.1, indicating that, on average, documents are rarely inferred as members when attacked through subsamples. Nevertheless, we identify six documents that are correctly associated with the training set in more than 50% of the attacks in which they appear as part of a subsample. Interestingly, four of these documents are associated with exactly four questions. In contrast, documents linked to only a single question appear to be the least susceptible to membership inference, which could suggest that the least "seen" documents are less memorized; the opposite does not hold.

## 6 Conclusions and Ethical Considerations

In this work, we investigate the sources of disparity in unsupervised membership inference attacks. Specifically, we analyze the randomness introduced by three factors: the choice of the targeted set, the clustering algorithm's initialization seed, and the repeated execution of the unsupervised attack on subsamples. Although the K-Means initialization seed introduces only negligible variability in attack outcomes, our empirical results show that the selection of the targeted set plays a significant role in assessing the privacy risks of training data. We further demonstrate how the Coverage obtained from repeatedly attacking subsamples, constructed from the target set available to the attacker, can be leveraged to increase the inferred privacy risk of documents used to train a DocVQA model. Additionally, this procedure enables the identification of documents that are consistently more susceptible to correct membership association when targeted within a subsample. We point out that MIA defenses, such as preventing overfitting (via data augmentation, dropout, or other regularization techniques [10]) or limiting a point's influence during training with Differential Privacy [7], have not been subject to disparity analysis within unsupervised MIAs.

While our evaluations are conducted on a single DocVQA model (Donut) and dataset (DocVQA), we investigate multiple target sets and subsample sizes in an effort to provide insights that may support future research[1]. For instance, beyond assessing the universality of our findings for other datasets, models, and clustering methods, future studies could explore the increase in TPR obtained through Coverage to derive weak labels and potentially transform an unsupervised attack into a supervised one without requiring externally annotated labels.

Although we discuss strategies malicious actors could use to expose training data, our primary goal is to raise awareness of privacy risks in AI-based systems, particularly in the context of DocVQA. By highlighting these vulnerabilities, we aim to provide researchers, practitioners, and auditors with additional tools and perspectives for developing and evaluating more secure systems.

We emphasize that all experiments were conducted solely for research purposes and in controlled settings, using publicly available benchmarks. Our objective is not to facilitate malicious exploration, but rather to improve understanding, evaluation, and mitigation of privacy risks in DocVQA systems.

[1] Code will be made available at https://gitlab.ic.unicamp.br/hiaac-nlp/unsupdispanalysis

## Acknowledgements

This work was supported by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex. The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.013778/2020-21]. H. Pedrini and S. Avila are also funded by CNPq (304836/2022-2, 316489/2023-9). S. Avila is also funded by FAPESP (2023/12086-9, 2023/12865-8, 2020/09838-0, 2013/08293-7).

## References

1. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramèr, F.: Membership inference attacks from first principles. In: IEEE Symposium on Security and Privacy. pp. 1897–1914 (2022)
2. Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., Tramer, F.: The privacy onion effect: Memorization is relative. In: Advances in Neural Information Processing Systems. vol. 35 (2022)
3. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., Raffel, C.: Extracting training data from large language models. In: USENIX Security Symposium. pp. 2633–2650 (2021)
4. Choquette-Choo, C.A., Tramer, F., Carlini, N., Papernot, N.: Label-Only Membership Inference Attacks. In: International Conference on Machine Learning (2021)
5. Dentan, J., Paran, A., Shabou, A.: Reconstructing training data from document understanding models. In: USENIX Security Symposium (2024)
6. Duddu, V., Boutet, A., Shejwalkar, V.: Quantifying Privacy Leakage in Graph Embedding. In: International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. pp. 76–85 (2021)
7. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography. pp. 265–284 (2006)
8. Dwork, C., Smith, A., Steinke, T., Ullman, J.: Exposed! a survey of attacks on private data. Annual Review of Statistics and Its Application **4**(1), 61–84 (2017)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
10. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys **54**(11s) (2022)
11. Jegorova, M., Kaul, C., Mayor, C., O'Neil, A.Q., Weir, A., Murray-Smith, R., Tsaftaris, S.A.: Survey: Leakage and privacy at inference time. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 9090–9108 (2023)
12. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision (2022)
13. Ko, M., Jin, M., Wang, C., Jia, R.: Practical Membership Inference Attacks Against Large-Scale Multi-Modal Models: A Pilot Study. In: IEEE/CVF International Conference on Computer Vision. pp. 4848–4858 (2023)

14. Li, X., Li, Q., Hu, Z., Hu, X.: On the privacy effect of data enhancement via the lens of memorization. IEEE Transactions on Information Forensics and Security **19**, 4686–4699 (2024)
15. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images. In: IEEE Winter Conference on Applications of Computer Vision (2021)
16. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In: IEEE Symposium on Security and Privacy (2019)
17. Nguyen, K., Kerkouche, R., Fritz, M., Karatzas, D.: Docmia: Document-level membership inference attacks against docvqa models. In: International Conference on Learning Representations (2025)
18. Niu, J., Liu, P., Zhu, X., Shen, K., Wang, Y., Chi, H., Shen, Y., Jiang, X., Ma, J., Zhang, Y.: A survey on membership inference attacks and defenses in machine learning. Journal of Information and Intelligence **2**(5), 404–454 (2024)
19. Peng, Y.: Unsupervised membership inference attacks against machine learning models. In: Advances in Neural Information Processing Systems (2021)
20. Pinto, F., Rauschmayr, N., Tramèr, F., Torr, P., Tombari, F.: Extracting training data from document-based vqa models. In: International Conference on Machine Learning (2024)
21. Rigaki, M., Garcia, S.: A Survey of Privacy Attacks in Machine Learning. ACM Comput. Surv. **56**(4), 101:1–101:34 (2023)
22. Sablayrolles, A., Douze, M., Ollivier, Y., Schmid, C., Jégou, H.: White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. In: International Conference on Machine Learning (2019)
23. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: IEEE Symposium on Security and Privacy (2017)
24. Tito, R., Nguyen, K., Tobaben, M., Kerkouche, R., Souibgui, M.A., Jung, K., Jälkö, J., D'Andecy, V.P., Joseph, A., Kang, L., Valveny, E., Honkela, A., Fritz, M., Karatzas, D.: Privacy-aware document visual question answering. In: Document Analysis and Recognition. pp. 199–218 (2024)
25. Usynin, D., Knolle, M., Kaissis, G.: Memorisation in machine learning: A survey of results. Transactions on Machine Learning Research (2024)
26. Wang, Z., Zhang, C., Chen, Y., Baracaldo, N., Kadhe, S., Yu, L.: Membership Inference Attacks as Privacy Tools: Reliability, Disparity and Ensemble. In: ACM Conference on Computer and Communications Security. pp. 1724–1738 (2025)
27. Watson, L., Guo, C., Cormode, G., Sablayrolles, A.: On the importance of difficulty calibration in membership inference attacks. In: International Conference on Learning Representations (2022)
28. Wei, J., Zhang, Y., Zhang, L.Y., Ding, M., Chen, C., Ong, K.L., Zhang, J., Xiang, Y.: Memorization in deep learning: A survey. ACM Computing Surveys **58**(4), 1–35 (2026)
29. Ye, J., Maddi, A., Murakonda, S.K., Bindschaedler, V., Shokri, R.: Enhanced Membership Inference Attacks against Machine Learning Models. In: ACM Conference on Computer and Communications Security. pp. 3093–3106 (2022)
30. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In: IEEE Computer Security Foundations Symposium. pp. 268–282 (2018)
31. Zhang, X., Chen, C., Xie, Y., Chen, X., Zhang, J., Xiang, Y.: A survey on privacy inference attacks and defenses in cloud-based deep neural network. Computer Standards & Interfaces **83**, 103672 (2023)