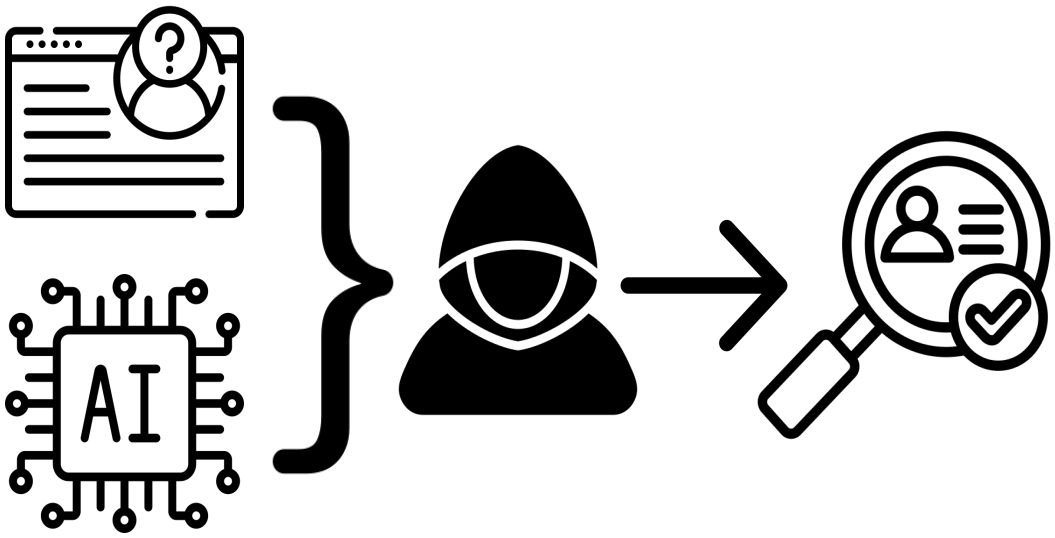


Privacy Risks in Text Masking Models for Anonymization

Analyzing the Performance of Membership Inference Attacks and Extraction Attacks on Anonymization Models

Master's thesis in Complex Adaptive Systems

AMANDUS REIMER

DEPARTMENT OF PHYSICS

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Privacy Risks in Text Masking Models for Anonymization

Analyzing the Performance of Membership Inference Attacks and Extraction Attacks on Anonymization Models

AMANDUS REIMER

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Chalmers University of Technology
Gothenburg, Sweden 2025

Privacy Risks in Text Masking Models for Anonymization
Analyzing the Performance of Membership Inference Attacks and Extraction Attacks on Anonymization Models
Amandus Reimer

Supervisor: Johan Östman, AI Sweden & Fazeleh Hoseini, AI Sweden
Computational linguistics advisor: Danila Petrelli, AI Sweden
Examiner: Giovanni Volpe, University of Gothenburg

Master's Thesis 2025
Department of Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Collage representing the risk of identification of confidential attributes using an AI model. Created using resources from Flaticon.com.

Typeset in LaTeX

Printed by Chalmers Reproservice

Gothenburg, Sweden 2025

Privacy Risks in Text Masking Models for Anonymization

Analyzing the Performance of Membership Inference Attacks and Extraction Attacks on Anonymization Models

Amandus Reimer

Department of Physics

Chalmers University of Technology

Abstract

Large Language Models (LLMs) are increasingly employed to anonymize texts containing Personal Identifiable Information (PII), often relying on Named Entity Recognition (NER) to identify and remove sensitive data. This thesis explores the privacy risks associated with such text masking models by evaluating their vulnerability to Membership Inference Attacks (MIAs) and extraction attacks. MIAs are attempting to identify whether or not a data point was part of the training dataset, knowledge of the membership can in certain scenarios be a breach of privacy. Two state-of-the-art MIAs have been used to conduct attacks on text masking models. This study also proposes a framework based on multi-armed bandits for performing extraction attacks and evaluates two different strategies within this framework. The results from the MIAs indicate that there is some risk of revealing information regarding the training data. The extraction attacks did not yield great results in terms of performance but indicate that the concept could possibly be useful if developed further.

Keywords: Membership Inference Attack, Model Integrity, Personal Identifiable Information, Data Extraction Attack, Text Anonymization.

Acknowledgements

I would like to express my deepest gratitude to my supervisors and technical advisors Johan Östman, Fazeleh Hoseini, and Danila Petrelli at AI Sweden. Your support and guidance have been monumental in conducting this thesis. I would also like to extend my thanks to Giovanni Volpe, for acting as examiner for this thesis.

Amandus Reimer, Gothenburg, January 2025

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AUC	Area Under the Curve
BERT	Bidirectional encoder representations from transformers
ECHR	European Court of Human Rights
FPR	False Positive Rate
GDPR	General Data Protection Regulation
LLM	Large Language Model
LiRA	Likelihood Ratio Attack
NER	Named Entity Recognition
NLP	Natural Language Processing
MIA	Membership Inference Attack
OoD	Out-of-Distribution
PII	Personal Identifiable Information
RMIA	Robust Membership Inference Attack
ROC	Receiver Operating Characteristic
TAB	Text Anonymization Benchmark
TPR	True Positive Rate

Nomenclature

Below is the nomenclature of sets, parameters, and variables that have been used throughout this thesis.

Sets

$\mathcal{D}$	Population
$\mathcal{D}'$	OOD population
D	Training dataset

Parameters

β	Decision parameter in MIAs

Variables

$\mathcal{A}$	Adversary's approach
b	Membership of target record
$\hat{b}$	Adversary's prediction of membership
$\mathcal{T}$	Training algorithm

$\mathcal{O}_\theta$	Oracle function
θ	Target model
x_b	Target record
l_θ	Loss of model θ
$\theta^{in/out}$	Shadow models, in or out
$\text{Score}_{\text{attack}}$	Membership score for attack

Contents

List of Figures

List of Tables

1

Introduction

The field of artificial intelligence is experiencing rapid growth, spurred by the huge amount of data collected online and advanced model development, the possibilities of the state-of-the-art systems appear ever-improving [1]. One of the fields that has seen great improvement in recent years is text-based applications, also known as Natural Language Processing (NLP) [2]. Aided in large by the model architecture "transformers" [3], current state-of-the-art NLP models excel at language understanding and generation, and their performance across various benchmark tests is comparative to the performance of an average human [4].

Since the time and computational demands to train these highly sophisticated models are high, one common practice is to utilize models trained by others to save on both resources and time [5], possibly by utilizing *fine-tuning*, performing additional training on task-specific data to adapt a suitable model to another task [6]. While sharing models saves both resources and time, when working with sensitive data one has to be sure that there is no risk of "leaking" (inadvertently revealing) personal or confidential information when sharing models. The General Data Protection Regulation (GDPR) puts a high emphasis on the privacy of users and their data, and it mandates that personal data can only be shared if the users are fully informed and provide explicit consent for such use [7]. Even if consent is given, the creators of the service need to take ample precautions in regard to the relevant risks, and show that they have been studied carefully.

What this means in practice is that companies and organizations working with sensitive data have to be aware of the risks of sharing models trained on sensitive information. If the model is shared in its entirety, or through API access, models that

leak could cause major implications for the company and the individual, depending on the nature of the information being leaked [7]. The trade-off between using AI to its full potential while being highly confident that no private information is being leaked is thus highly relevant for actors working with private data. The performance of a model is limited by its training data [8], so if the model is operating on sensitive data it needs to be known to what extent the model is sensitive to leakage if it is to be shared.

In the worst case, sharing a model that leaks is equivalent to passing along sensitive data between different actors. The privacy of the users is violated, and the involved parties are subject to the repercussions that come from not following GDPR. Some efforts are being made to create frameworks for analyzing these risks, notably LeakPro [9] by AI Sweden, and LLM-PBE [10]. LeakPro aims to be model- and data-agnostic, whereas LLM-PBE is more geared toward surveying Large Language Models (LLMs).

There are several studies examining the risks of sharing AI models [11, 12, 13, 14, 15, 16], but only a few types of models and datasets have been covered. A specific use case that has not been thoroughly investigated is the risks related to models trained for text masking, in the context of masking a document containing information about an individual to keep their identity confidential after masking.

In this thesis, the risks related to sharing models trained to perform *privacy-preserving anonymization* of text are examined. Specifically, the risks related to the leakage of information about the training data through *Membership Inference Attacks*.

One example of a use case where this could be of interest is in censoring legal documents. In Sweden, and certain other countries, many legal cases are considered public records, and whoever wishes to can request public records and receive anonymized copies where all *Personal Identifiable Information* (PII) have been redacted [17]. Performing the anonymization of documents by hand is very time-consuming and costly. Currently, clerks in legal departments manually anonymize documents. Considering the number of documents distributed annually, the time required for anonymization, and the average salary of a clerk, automating parts of this

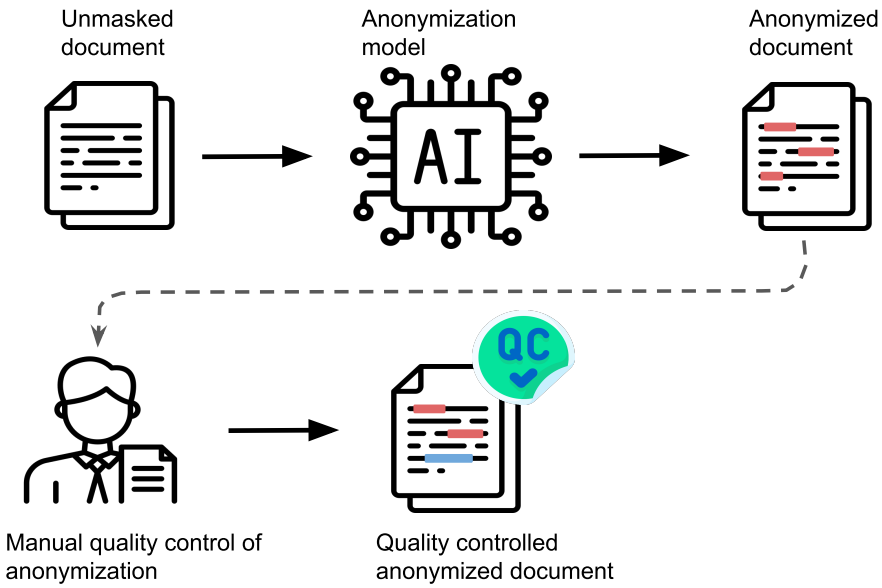


Figure 1.1: The anonymization process of a document containing personal information. The requested document is first masked by an AI model and the suggested anonymization is then quality controlled and possibly corrected by a human before being ready to be handed out. Created using resources from Flaticon.com.

process by using AI is conservatively estimated to save 400 million SEK annually[1]. The Swedish National Courts Administration (Domstolsverket) is working on a model to perform this task, reducing manual labor to quality control of the model's output. This significantly reduces document processing time, estimated at 4.5 hours per document. There are multiple organizations with similar needs that could benefit from using this tool, and sharing models could be of interest. However, since the model is trained on highly sensitive data, it has raised concerns about whether it can be shared without risking the leakage of private information.

In this thesis models trained to anonymize text will be investigated with the purpose of quantifying their susceptibility to Membership Inference Attacks (MIAs), and in an ablation study, if the shared models can be used for reconstructing censored documents. This thesis aims to answer the following research questions:

- **How vulnerable are text masking models to membership inference attacks?**
- **What's the level of feasibility in extracting masked data from an**

[1]Based on information received in an interview with a representative from Domstolsverket.

anonymized document by using the anonymizer model?

The paper is structured as follows: First, the relevant prerequisites and background of the problem are explained in chapter 2, along with some more detailed information about the MIAs. In chapter 3 a more detailed explanation of the problem settings, the studied attacks, and the models investigated is given. In chapter 4 the specifics of the datasets used are presented, along with explanations of the experimental settings and the results of the experiments. In chapter 5 the results and their implications are discussed, along with a brief discussion of the limitations, ethical concerns and future work in the area. Finally, some closing remarks are made in chapter 6 about what the most important findings were, and what implications can be drawn from this study.

2

Background

2.1 Large Language Models and Natural Language Processing

Tasks where the data is encoded in natural language is a field of computer science called Natural Language Processing (NLP) [18]. Large Language Models (LLMs) are a branch of models created to solve such problems, where the term "large" refers to the number of parameters in these models, which are often in the hundreds of millions or billions [19]. Many LLMs are based on different types of transformers [3] in one way or another.

While there are several different tasks and techniques present in NLP, some methods are used in most settings. One of these is *tokenization*. In NLP, tokenization is the process of converting text into a format better suited for machine learning models to process. In brief, tokenization takes a word and divides it into tokens, which are the sub-parts words and text are made up of [20]. There are different ways of performing tokenization, but an example is the word "working", which could be split up into the tokens "work" and "ing". The tokenization also associates an integer value to each token, which is called the token ID, which simplifies conversion from text to tokens and from tokens to text. From the tokenization, a word embedding can also be created. The word embedding is a learned latent space representation of the tokens. The aim of the word embedding is to create a data structure more informative to a machine learning model to work with. In the latent space, similar tokens have similar representations, which may allow the LLMs to find more complex relationships present in a sequence of words [21].

The transformer architecture utilizes a concept called attention [3, 22] to highlight the relationship between tokens. This concept enables words with a large distance in between them, i.e. the first and last word of a long sentence, to keep their contextual information. If a word has multiple meanings, the context that determines which meaning it is currently being used may be at a far distance from the word itself. The attention mechanism would then help the LLM to find the correct contextual meaning. The attention mechanism used in the first transformer implementation scales quadratically with sequence length in its original form, as the attention between all of the tokens is calculated [3]. Other types of attention can have a more local perspective, saving on computational costs at the price of performance. One such model used in this thesis is the Longformer [23]. This model uses attention in a way that instead scales linearly with sequence length, by using a combination of a local and global attention window. It still performs well on many benchmarks in comparison to other, more demanding, models [23].

Due to the resource demands and training time of most LLMs, a common practice is to use a pre-trained model and fine-tune it on task-specific data [6]. This is also what is done in this thesis.

2.1.1 Named Entity Recognition

Named Entity Recognition (NER) is an NLP task in which the focus is to categorize, or tag, segments of text into a predetermined list of entities, such as names, locations, and dates [24]. In the context of NER, an entity refers to a word or phrase representing a real-world object, concept or identifier, such as person, organization, location, date or numerical value. Entities typically represent discrete, identifiable elements of information that are crucial for understanding or processing text.

A toy example of an instance of NER is shown in figure 2.1, where a name, location, and timestamp have been recognized as entities. The task of NER suits anonymizing documents well since most personal information that can be used to identify a person can be categorized into a list of named entities. The full list of categories is given in section 4.1.

Figure 2.1: A toy example of an instance of NER. In the text, three entities have been identified.

2.2 Anonymizing Personal Identifiable Information

The task of anonymization is quite a difficult task due to the complex definition of Personal Identifiable Information (PII) as given in GDPR [7]. PII can be divided into two categories, *direct identifiers* and *indirect identifiers*. Direct identifiers are pieces of information that by themselves identify an individual, such as name or address. Indirect identifiers are more abstract; they are pieces of information that only in aggregate or in a specific context can be used to identify an individual. For example, nationality and city of residence may be indirect identifiers of a person if they are the one from their country living in that city. These indirect identifiers are difficult to define since what may be considered an indirect identifier for one person may not be for another, and the consequence of leaving them unmasked is not always easily measured. Two different anonymizations on the same document may be equally good while leaving different pieces of indirect identifiers unmasked; using the same example as above, removing either the mentions of nationality or city of residence could both be sufficient to anonymize a document.

A recent sandbox study conducted by IMY (Integritetsskyddsmyndigheten, the Swedish Authority for Privacy Protection) investigated AI as a tool for more efficient anonymization of legal documents [25]. This study had more of a legal perspective, but a very similar problem setting as this thesis is investigating. A conclusion from this study is that there is a legal basis for using AI in anonymization, but it needs

to be handled with care, therefore certain security aspects need to be considered if implemented [25]. While the legality of the question is not a focus of this thesis, this conclusion provides a good basis for the need to study the risks related to using AI in the anonymization of PII.

2.3 Model Integrity

Model Integrity is a broad field of study where various aspects of a machine learning model are analyzed, such as functionality, reliability, and security [26]. In this thesis, the main goal is to investigate the security risks related to models, so whenever model integrity is mentioned it refers to the security aspects involved.

When creating a machine learning model, the intended scope of use may vary. The model could be restricted to only be used within the bounds of the organization that created it, or it may be sold in its entirety as a product, or access to the model can be sold as a service. Can a clever, malicious user utilize what they are provided, be it the entire model or only the output, and retrieve confidential knowledge such as training data, hyperparameters, or architectural details? If that is the case, the malicious actor could sell the information to someone else, or simply re-create a model by themselves and not have to pay for the service. If training is done based on user inputs, can the malicious user somehow cause the model's performance to deviate from its intended purpose, or degrade?

These are some of the risks to consider before releasing a model in any way to an external party [26]. Especially when dealing with sensitive data, where the consequences of leakage are large.

The literature in the field of model integrity (in relation to security investigations) is still in its early stages, but one practice that has gained a lot of traction is formulating the security investigation as a game [27]. Using these formulations effectively clarifies what the assumptions for the current investigation are, and provides a solid structure for continued development by others.

In the gamified framework used in [27] and other studies, there are two actors playing against each other. These are: an *adversary* creating attacks to try to manipulate or

extract information from a model, and a *challenger* whose responsibility is to create the model to attack [27]. It is important to clearly state the goals and available tools for the actors when defining the game. The available tools can be presented as the adversary's access to the model and dataset, availability of auxiliary information, and amount of resources available. The measures used to analyze the adversary's success should also be defined.

In many settings, including this thesis, the challenger is a passive entity and does not interact with the model or the adversary after the model has been created. The main goal for the challenger is to create a machine learning model on which analysis is to be conducted.

The goals of the adversary can be to extract information about the model itself, about its training data, or to somehow alter the model [27]. Different goals have different types of attacks associated with them, for example, extracting training data and extracting model information are done with quite different approaches, and their success is measured very differently.

The model access can be a black-box access in which the adversary has access only to a limited output of the model, a white-box access where the adversary has access to the full model, including its weights and structure, or something in between called grey-box access [28]. If the adversary is not given full access to the model, the adversary's knowledge of the architecture also needs to be defined. For data, an adversary can be assumed to be able to sample from the true data distribution (also called the population), or from a similar one, or generate synthetic samples [29]. Although the assumption about the adversary having access to the population may seem fairly generous, it is not unreasonable. In some cases, all or most features of a record can be public knowledge (e.g. age, height, address, and plenty of other features that may be found on the internet), and the adversary can make educated guesses to fill in the remaining missing features to form a candidate record on which to conduct their investigation on.

The specifics regarding the adversary access vary across the different cases and have to be clearly stated in advance to make the study as relevant for the chosen scenario as possible. Without considering and stating these details, the study and conclusions

may end up misleading or misinterpreted.

2.3.1 Membership Inference Attack(s)

One of the possible investigations that can be performed with the goal of extracting information regarding the training data is the *Membership Inference Attack* (MIA), where membership refers to whether or not a data point is a part of the training dataset [11]. The goal of this attack is to analyze whether or not the model provides the adversary with enough information to discern between training data and other data. This is a binary classification problem, and the goal of a MIA can be formulated as:

Given a model, has it been trained on a specific data record?

In some situations, knowledge of membership would be an indirect privacy leak [30]. If, for example, a model is trained on medical data of individuals suffering from a rare disease, being able to confidently determine the membership of individuals would let the adversary know about who suffers from the health condition.
MIAs are also connected to other forms of attacks, such as attribute inference and data reconstruction attacks, and knowledge of how the model performs against MIAs can be informative to how sensitive it is against other attacks, and can also be used in the creation of other types of attacks [27]. This makes MIAs a good starting point when investigating the integrity of a model.

2.4 MIAs in the gamified framework

The gamified framework frames the security question as a game between a challenger and an adversary, where the adversary tries to extract information from or alter the behavior of the challenger's machine learning model [27]. In a MIA specifically, the task is for the adversary to determine if a sample is a part of the training data or not [11].
To start the game, the challenger creates a model θ by using a training algorithm

$\mathcal{T}$. The training is done on the training dataset D, which is formed by sampling from the population $\mathcal{D}$, which is all of the available data. The challenger then flips a fair coin and depending on the outcome b, either provides the adversary with a record in the training dataset, $x_{b=1} \sim D$, or a randomly sampled record from the population, $x_{b=0} \sim \mathcal{D}$. This target record x_b is given to the adversary to make their prediction of membership.

The adversary receives this target record x_b and is tasked to predict whether or not the model θ was trained on it, i.e. they make a prediction $\hat{b}$ of b using their approach $\mathcal{A}$. The adversary is usually assumed to have access to the population $\mathcal{D}$, the training algorithm $\mathcal{T}$, and some type of access to the model θ. The access to the model can be formulated as an oracle function $\mathcal{O}_\theta(\cdot)$, where the definition of the oracle function determines the level of access granted. It can be defined as providing only the model output to the adversary (black-box), or something more generous (white-/grey-box).

The adversary generates a score, $\text{Score}_{\text{MIA}}$, for the target record, which reflects how likely they believe that the model was trained on the record. This score is then thresholded using a decision parameter β to predict membership. This allows the adversary to modify how strict they are with their decision-making. As the value of β increases, fewer and fewer scores will be above the threshold, and only the most confidently estimated members will be predicted as such. In terms of these formulations, a simple black-box MIA game can be formulated as in algorithm 1.

Algorithm 1 Membership Inference Attack

experiment $\text{MIA}(\mathcal{D}, \mathcal{T}, \mathcal{A}, \mathcal{O}_\theta(\cdot))$
- $D \sim \mathcal{D}$ ▷ Challenger samples a training dataset
- $\theta \leftarrow \mathcal{T}(D)$ ▷ Challenger trains the model
- $b \leftarrow \text{rand}\{0, 1\}$ ▷ Challenger randomly samples b
- $x_1 \sim D$ ▷ Challenger samples from the training dataset
- $x_0 \sim \mathcal{D}$ ▷ Challenger samples from the population
- $\text{Score}_{\text{MIA}} \leftarrow \mathcal{A}(\mathcal{O}_\theta(\cdot), \mathcal{D}, \mathcal{T}, x_b)$ ▷ Adversary generates a score
- $\hat{b} \leftarrow \mathbb{1}[\text{Score}_{\text{MIA}} \geq \beta]$ ▷ Adversary predicts membership of x_b
- **return** $\hat{b}$ ▷ The adversary's prediction is returned

procedure $\mathcal{O}_\theta(x)$ ▷ Oracle function
- **return** $\theta(x)$ ▷ Black-box access

Algorithm 1: The Membership Inference Attack described algorithmically.

Using these formulations, the type of MIA used has not been specified, and the outcome of the attack is only modeled as the adversary's approach $\mathcal{A}$. The oracle function $\mathcal{O}_\theta(\cdot)$ is defined as a black-box access to the target model in algorithm 1. One note is also that the population in this notation is large enough that a random sample from $\mathcal{D}$ is, with large probability, not in D.
While these are the core concepts needed to formulate a basic game of MIA, there are plenty of other aspects of the game that are of interest, including specific techniques the adversary may use, and how the results are presented. In the following sections, an adversarial method used in the attacks used in this thesis is presented, and later the way the measurements are reported is discussed.

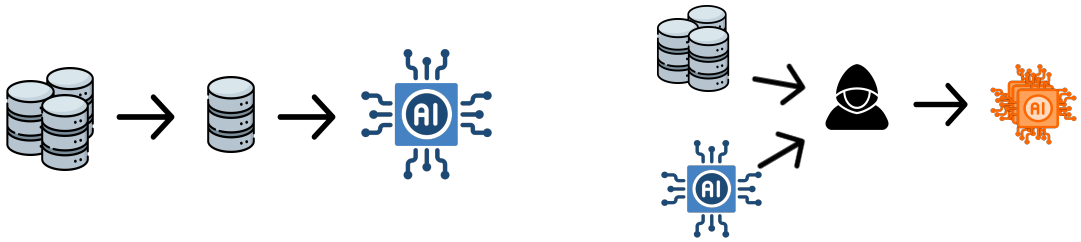
2.4.1 Shadow Models

Shadow models are models created by the adversary in a similar way as the model to be investigated. They don't need to be exact replicas in terms of model architecture, but the more similar the better the performance of the MIA [11]. The adversary controls the training process of these shadow models, by doing so they aim gain insights on how training data affects the outcome of the shadow models and, by extension, the target model. As the machine learning models are fit to the data they train on, in general, the performance can be expected to be higher on training data than other data. An adversary can train shadow models with and without the target data in their training set, and compare their outcomes to the outcome of the target model to make their prediction [11].
This is computationally demanding when analyzing large amounts of data or big models, and may not be feasible in certain situations. Therefore, other attack types that do not rely on training shadow models on the target data for each prediction have also been produced. These two different attack types can be dubbed as *online attacks* and *offline attacks* [11]. Online attacks actively use the target record to make membership predictions, whereas offline attacks don't. The online and offline

attacks each have different benefits and drawbacks. For example, online methods require more computation since the adversary is actively adapting to the records to investigate when they are presented to them. However, there is a possible performance decrease when using offline methods, since they use the information about the target record to investigate to a smaller extent.
How the shadow models are formed and used varies between different attack types. The details regarding the implementations of the shadow models in the attacks studied in this thesis are elaborated on in section 3.3. In figure 2.2, a visual representation of the components that are part of the performance of a MIA using shadow models is shown.

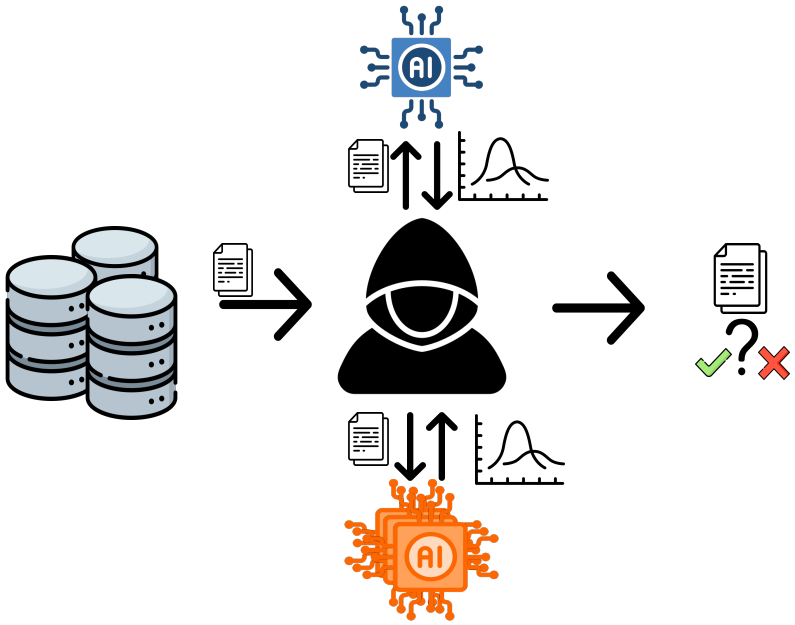
2.4.2 Measures of MIA performance

To measure the performance of the adversary's attack, one has to consider how a security risk can be put into numbers. An attack that makes few mistakes while identifying the most number of memberships is optimal from the adversary's point of view. As the task for the adversary is classification, the most basic measure would be accuracy, i.e. how well the attack correctly classifies members and non-members. However, since the security of the training data is under investigation, the classification of non-members is not of interest, and therefore other measures should be considered above accuracy.
Two popular metrics used when the classification performance of one class is of more interest than the rest are the True Positive Rate (TPR), and the False Positive Rate (FPR). In the MIA setting, TPR is the number of found members divided by the total number of members, and the FPR is the number of non-members incorrectly marked as members divided by the total number of non-members. The combination of these measures provides information about the fraction of identified members, as well as provides insight as to how often the misclassification of a non-member occurs.
A way of using these measurements popularized by [11] is to report the TPR at a fixed, low FPR, often as low as 0.001% or even 0%. They motivated the importance of this metric by the heuristic that being able to very confidently extract a small



(a) Training of the target model.

(b) Creation of the shadow models.



(c) Using the target model and shadow models to determine membership of a sampled record.

Figure 2.2: A visual representation of how a MIA using shadow models is constructed. First, in subfigure a), the target model is created by using a sampled training data set from the population. In subfigure b), the adversary creates shadow models using access to the population and the target model. In subfigure c) the attack is conducted as follows: A target record is sampled from either the training dataset or the population and is given to the adversary. The adversary compares the outcome of the target model and the shadow models when fed the target record, and based on their approach they make a prediction on whether or not the target record was sampled from the population, or from the training dataset, i.e. a prediction of membership. Created using resources from Flaticon.com.

amount of information is more of a security breach rather than being able to on average have a higher attack performance. It is also common practice to report the entire Receiver Operating Characteristic (ROC) curve, which shows how the TPR and FPR evolves as the threshold parameter β changes, along with the Area Under the Curve (AUC) of the ROC curve which can be seen as an average case performance.

In this thesis, two state-of-the-art MIAs have been examined in the problem setting of anonymizing PII. These are the Likelihood Ratio Attack (LiRA) [11], and the Robust Membership Inference Attack (RMIA) [15]. They are both utilizing shadow models and the model loss to make predictions on membership, and will be presented in chapter 3.

2.5 Balancing Privacy and Utility in Machine Learning

The goal of machine learning is to find an underlying relationship between the features and values of an observation, data point, or record. The features of the data point are the properties the value is thought to depend on, for example, in image recognition the features are the pixels and the value is what the image is depicting. In a supervised learning setting, a model is provided with labeled data to use for training and is tasked to optimize its parameters for the best performance possible in accordance with the metric defined by the user. It is, however, important to note that the training data is in most cases only a limited subset of the underlying distribution of data, or population. While the model should learn as much as possible from the training dataset, it's important not to *overfit* the model to the training data [31]. When creating a machine learning model, one of the most important measures is how well it generalizes to unseen examples, i.e., the performance on data outside of the training data. A model with perfect performance on its training data but generalizes poorly is unwanted as it has only learned specifics regarding the training data and failed to capture the underlying dynamics that explain how the features are related to the values of the records in the population. This is known

as overfitting.

Overfitting is an important factor to consider when training a model. This needs to be approached with great care, as it can potentially be exploited by a malicious user to extract, possibly confidential, information about the training data [11]. For example, consider the case of training a model on personal medical data with the task of predicting a very rare disease. If the model is highly confident when predicting on the training dataset, but very uncertain when presented with a new sample, one could then note that the data points on which the model is highly confident were part of the training data, and the privacy of the individuals in the training data is compromised. Even though overfitting can put privacy of the training data at risk, it is not the only thing to consider in regards to the utility and privacy trade off [32]. Other scenarios where the utility and privacy trade off need to be considered are for example when working with federated learning or explainable AI. Although this thesis handles neither of these to methods, it is important to know that the issue expands beyond only having to be mindful of overfitting. The trade off should be considered early, and guide the decisions in the creation of a machine learning system as to maximize the utility while still maintaining the appropriate level of privacy for all parts of the system.

3

Methodology

3.1 Problem formulation

In plain text, the problem to be studied in this thesis is the following:
If an LLM has been trained to anonymize a document with respect to personal information, can an adversary use this model and determine whether or not a data point was part of the training data?
The problem can also be formulated as:
Given a set of n documents in the population $\mathcal{D} = \{x\}^n$, where each document contains k_i pieces of PII for $i = 1, ..., n$; train a model θ to anonymize the documents. Training is done by using learning algorithm $\mathcal{T}$ on a set of training data $D \subset \mathcal{D}$, sampled from the population, such that $\theta = \mathcal{T}(D)$. Then sample a target record x_b from D or $\mathcal{D}$ using a binary uniform stochastic variable b such that $x_1 \in D$, and $x_0 \in \mathcal{D}$ (which with very large probability is not in the training dataset). Using an attack A, make a prediction $\hat{b}$ of whether or not the provided document x_b was part of the training set or not, given an interface with the model $\mathcal{O}_\theta$, access to the population $\mathcal{D}, \mathcal{T},$ and x_b. In other words, form $\hat{b} = A(\mathcal{O}_\theta, \mathcal{D}, \mathcal{T}, x_b)$ as the prediction on b, which corresponds to the membership of x_b.

3.2 Anonymization model overview

The model used for anonymization is based on two structures, a pre-trained Longformer [23] and a single-layered feed-forward neural network. Anonymization of PII can be performed by classifying the tokens in the document as either a type of PII

or not and then mask the classified tokens accordingly. There are two different ways to mask a document from PII in this thesis, binary masking and NER-type masking. Binary masking leaves no information as to what is behind the mask by replacing the PII with "[Mask]", whereas NER-type masking replaces PII with the associated NER-type, for example replacing a name in the document with "[Name]". While this is a more generous setting, the context surrounding the masked pieces of information often provides a reader with enough information to accurately guess what type of information is hidden underneath the mask. The method of masking impacts the size of the output of the classification layer.

The labeling of the tokens follows the "IOB2" scheme from [33], which is a way to label the tokens in a sequence based on their positioning. This is performed since a PII may consist of many tokens, and the method of dividing the label of the original piece of text into different labels for the token it consists of improves the machine learning model's performance [33]. For example, consider the labeling of a word consisting of the three tokens [token_A, token_B, token_C]. If the NER-type of the word is X, the labels of the tokens would be [B-X, I-X, I-X], i.e. the first token receives a prefix "B-", and the following "I-". This labeling scheme is one of the less complex ones, but it shows decent performance while having a small computational cost of implementation [33].

Thus, if the number of labels in the masking strategy is m, the output layer of the classification layer will be of size $2m + 1$. When using binary masking, i.e. all the PII is replaced by "[Mask]", $m = 1$, and when NER-type masking is used, m is equal to the number of entity types.

3.3 MIA specifics

In this thesis, two state-of-the-art MIAs are implemented, the Likelihood Ratio Attack (LiRA) [11] and the Robust Membership Inference Attack (RMIA) [15]. They were chosen due to their high performance at regions of low FPR, and they are both built by quoting the performance of the target model and the performance of shadow models, along with the relative ease of implementation using LeakPro.

3.3.1 Likelihood Ratio Attack (LiRA)

In LiRA [11], the adversary utilizes shadow models to make the membership prediction. In this attack, the adversary is assumed to have access to the loss of the model on the target record, $l_\theta(x_b)$, and is able to sample from the data population. In the online version on LiRA, half of the shadow models are trained with the target record in their training dataset, $\{\theta^{in}\}$, and half without, $\{\theta^{out}\}$. After the shadow models are trained, a Gaussian is fit to the distributions of loss of the target record on the shadow models, $\{l_{\theta^{in}}(x_b)\}$ and $\{l_{\theta^{out}}(x_b)\}$. This is done separately for the shadow model types, so two Gaussians, $\mathcal{N}(\mu_{in}, \sigma^2_{in})$ and $\mathcal{N}(\mu_{out}, \sigma^2_{out})$ are formed. LiRA quotes the likelihood of observing the actual loss of the target record, $l_\theta(x_b)$, according to the in and out shadow models to form a score used to predict membership. In equation form, the likelihood ratio and adversary estimate calculations are shown in equation 3.1.

$$\text{Score}_{\text{LiRA}} = \frac{\Pr(l_\theta(x_b)\,;\,\mathcal{N}(\mu_{in}, \sigma^2_{in}))}{\Pr(l_\theta(x_b)\,;\,\mathcal{N}(\mu_{out}, \sigma^2_{out}))}, \tag{3.1}$$

$$\hat{b} = \mathbb{1}[\text{Score}_{\text{LiRA}} > \beta], \tag{3.2}$$

where $\mathbb{1}[\cdot]$ is the indicator function, and β is a tunable parameter. β controls the threshold of how large the quote needs to be for the adversary to predict membership. Increasing β leads to more records being predicted as members, and sweeping over β yields the Receiver Operating Characteristic (ROC) curve, which is often used as a way to report results.

3.3.2 Robust Membership Inference Attack (RMIA)

The approach of RMIA [15] is quite similar to LiRA [11], in that it also makes use of a quote between the performance of the target model and shadow model when forming its score. However, it has a different quotient used in its inference.
When forming the score, RMIA starts with quoting the probability of observing a model θ trained on the target record, x, against the probability of observing the

model if it was trained on another record, z:

$$\mathrm{LR}_\theta(x, z) = \frac{\Pr(\theta|x)}{\Pr(\theta|z)}. \tag{3.3}$$

Then, Bayes rule is applied to arrive at

$$\mathrm{LR}_\theta(x, z) = \frac{\Pr(x|\theta)}{\Pr(x)} \left(\frac{\Pr(z|\theta)}{\Pr(z)}\right)^{-1}, \tag{3.4}$$

where a factor of $\Pr(\theta)$ is canceled from the numerator and denominator. Notice here that $\Pr(x)$ and $\Pr(z)$ are not the prior distributions over x and z, but normalizing constants that need to be calculated. The law of total probability gives

$$\Pr(x) = \sum_{\theta'} \Pr(x|\theta')\Pr(\theta') = \sum_{\theta', D} \Pr(x|\theta')\Pr(\theta'|D)\Pr(D), \tag{3.5}$$

where the sum is taken over all of the possible models θ' and the training data distribution D. However, since this is in reality intractable, an estimate is made that $\Pr(x)$ is approximately equal to the mean of $\Pr(x|\theta^{\text{in/out}})$ across shadow models $\theta^{\text{in/out}}$ trained with the target record in or out of the training dataset. This can be split into two terms, $\Pr_{\text{in}}(x)$ and $\Pr_{\text{out}}(x)$, where the subscript determines if the average was taken across shadow models trained with the target record in or out of the training dataset.

As stated previously, there are both online and offline attacks for many attacks, the same holds for the RMIA. For the online attack, half of the shadow models are trained on the target records, and half are not. This is done to make sure that the number of shadow models trained with and without the target records does not affect the estimate by being biased in regards to the distribution of training data used for the shadow models. The offline RMIA approximates the value of $\Pr_{\text{in}}(x)$ as a scaled-up value of $\Pr_{\text{out}}(x)$, scaled up as the inclusion of a data point can be heuristically thought to increase its probability.

The expression for the likelihood ratio has now become:

$$\mathrm{LR}_\theta(x, z) = \frac{\Pr(x|\theta)}{\Pr_{\text{in}}(x) + \Pr_{\text{out}}(x)} \left(\frac{\Pr(z|\theta)}{\Pr_{\text{in}}(z) + \Pr_{\text{out}}(z)}\right)^{-1}. \tag{3.6}$$

To construct the MIA score from this likelihood ratio, another parameter γ is introduced, and the score is formed by using the other random samples, z, and their prior distribution, π_z. Empirically, several z are randomly sampled, and the fraction of the times in which $\text{LR}_\theta(x, z)$ is larger than the threshold γ is taken as the score. In equation form, the score is given as

$$\text{Score}_{\text{RMIA}} = \Pr_{z\sim\pi_z}(\text{LR}_\theta(x, z) \geq \gamma). \tag{3.7}$$

The adversary's prediction on membership is then, like in LiRA, thresholded against a tuneable parameter β,

$$\hat{b} = \mathbb{1}[\text{Score}_{\text{RMIA}} > \beta]. \tag{3.8}$$

3.4 Membership score for a document

As the model used in the anonymization task provides a confidence score for which class each token belongs, the loss for an entire document needs to be compounded somehow. A simple method is to take the average of the confidence of the correct class across all of the tokens present in the document. In equation form, given a document containing K tokens, where the true label of token $k = 1, ..., K$ is denoted as $k_l = 1$ if l is the true label and zero otherwise, the model's confidence in token of the model of a token is y_k^l for each label $l = 1, ..., L$, the average confidence $\bar{C}$ is

$$\bar{C} = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L}\mathbb{1}[k_l = 1]y_k^l. \tag{3.9}$$

One issue with this approach however is that many of the tokens are not masked, and thus the average performance will be dictated in part by how long it is. Therefore, the signal used instead is the average value of the confidence of the correct class of only the masked tokens. If the unmasked tokens have label indexed $l = 1$, equation 3.9 can be modified to be

$$\bar{C}_{\text{masks}} = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L}\mathbb{1}[k_l = 1, l > 1]y_k^l. \tag{3.10}$$

3.5 PII Extraction with Multi-Armed Bandits

A follow-up investigation conducted in this thesis is if it is possible to extract PII from a masked document. Determining if a model is susceptible to a MIA is as stated previously a good starting point for assessing the model integrity, but a larger concern would be if an adversary is able to extract the training data directly. In this thesis, *multi-armed bandits* are selected as the framework to base the extraction attacks on [34]. The multiarmed bandit formulation involves a player being presented with a finite number of actions $\mathcal{X}$, and a game to play for a given number of rounds T. In each round t, a single action x_t is played and the player observes a loss l_t. The goal for the player is to identify the most rewarding action in the given number of rounds such that the cumulative loss $\hat{L}_T$ is minimized, that is, to find a policy π_t guiding which action to play [34]. Notably, the loss l_t is inversely related to the reward r_t, meaning a low loss corresponds to a high reward. Both representations are useful for conveying insights in different contexts.
The multi-armed bandit problem can be formulated as in algorithm 2:

Algorithm 2 Multi-armed bandit problem formulation

At each time step $t = 1, ..., T$:
Select an action in accordance to the policy π_t, i.e. sample $x_t \sim \pi_t(\mathcal{X})$,
Observe the loss of the selected action: $l_t(x_t)$
Update the policy: $\pi_{t+1} \leftarrow P(l_t(x_t), \pi_t, x_t)$

Goal: minimize the cumulative loss $\hat{L}_T := \sum_{t=1}^{T} L_t(x_t)$.

Algorithm 2: Multi-armed bandit problem formulation, where P is the strategy of updating the policy based on observed loss of action, current policy, and the action taken [35].

Minimizing the cumulative loss inherently involves a trade off between exploration and exploitation [34]. More exploring methods may be able to find better policies at the expense of receiving bigger losses while searching for them aggressively. More exploitative methods, on the other hand, could be content with finding decent rewards early, and limit the amount of exploration so that it might never find an optimal pol-

icy. The management between exploration and exploitation is an important aspect to consider when working in these types of scenarios.

In the setting of extracting PII, the game contains a masked document where the player is able to fill in the masks and feed the imputed document to the target model to obtain the model confidence as outputs. This effectively allows the player to receive feedback on each mask that is guessed. The action space of the player consists of all of the entities in each NER-category found in the population. The player is assumed to know the category for a given mask, i.e. the setting is that NER-type masking is used, they can hence sample an entity from a pool pertaining to the category at hand. The policy to learn is the distribution of which candidates to pick for which masked positions and the loss is the negative confidence of predicting the correct class. The action space is thus defined by which category the masks in the masked document belong to, and how many entities are found in each category in the population.

Given that there are several masks in a given document and each mask may have thousands of potential candidates, as can be seen in 4.2, solving this problem by concatenating the action space for each mask is infeasible. The first steps towards an extraction attack are instead taken by assuming an independent bandit at each mask, i.e. each mask is attached to a bandit algorithm that attempts to find only the correct action for that specific mask. Notably, the attack is agnostic towards the bandit algorithm that is used. Therefore two different approaches are considered.

The first type of attack used is an independent random sampling attack. The adversary creates the candidate pools from the population, and for each round, a candidate is sampled at random to impute at the masked positions from the corresponding candidate pool.

The second attack is an implementation of the Tsallis-INF algorithm [36]. This algorithm is considered a state-of-the-art algorithm within the *adversarial bandit* setting, that is, where the action rewards are selected by the environment at the same time as the action is chosen by the player [35].

Note that bandits operating independently on the same document may affect the reward distribution. Indeed, this may create a highly non-stationary environment

that is difficult to operate within.

4

Results

4.1 Datasets

There are two datasets used in forming results and conducting experiments in this thesis. The first dataset is the Text Anonymization Benchmark (TAB) dataset created in [37]. This is a dataset consisting of 1268 court cases in English taken from the European Court of Human Rights (ECHR). The dataset was created by using a number of students of law in Norway to annotate these legal texts. The students were given a set of guidelines to annotate in accordance with, and as a part of the process, also evaluated each other's work to make sure the annotations were as coherent as possible across all the different documents. The court cases were annotated to anonymize the text to keep a specified individual's identity from being possible to re-identify. The annotation was done by categorizing the pieces of text to which NER category they were a part, as well as if they were to be considered an identifier, and if so if it was a direct or indirect identifier. An additional consideration was made for especially sensitive attributes, such as ethnicity, political and religious beliefs, etc, this is however not relevant to the studies conducted in this thesis. The 8 categories used for annotating this dataset can be seen in table 4.1, and a table describing the distribution of entities is presented in 4.2.

Another setting to investigate is when the adversary does not have access to the exact population from which the training data was sampled, but a similar one. This would mean that the adversary only has access to Out-of-Distribution (OoD) data. To analyze this, a second dataset with similar characteristics is needed.

Table 4.1: The named entity types defined in the TAB dataset, and the description of them.

Entity type	Description
PERSON	Names of people, including nicknames/aliases, usernames, and initials.
CODE	Numbers and identification codes, such as social security numbers, phone numbers, passport numbers or license plates
LOC	Places and locations, such as cities, areas, countries, addresses, named infrastructures etc.
ORG	Names of organizations, such as public and private companies, schools, universities, public institutions, prisons, healthcare institutions, non-governmental organizations, churches, etc
DEM	Demographic attributes of a person, such as native language, descent, heritage, ethnicity, job titles, ranks, education, physical descriptions, diagnosis, birthmarks, ages
DATETIME	Description of a specific date (e.g. October 3, 2018), time (e.g. 9:48 AM) or duration (e.g. 18 years)
QUANTITY	Description of a meaningful quantity, e.g. percentages or monetary values.
MISC	Every other type of personal information associated (directly or indirectly) to an individual and that does not belong to the categories above.

Table 4.2: Count of occurrences and relative fraction of PII per entity type in the TAB dataset.

Entity Type	Count (fraction)
PERSON	20 021 (0.19)
CODE	6 042 (0.06)
LOC	6 909 (0.07)
ORG	12 900 (0.12)
DEM	4 167 (0.04)
DATETIME	48 109 (0.46)
QUANTITY	3 370 (0.03)
MISC	3 465 (0.03)
Total:	104 983

The second dataset used is based on legal cases from India [38]. This is a corpus consisting of 14444 Indian court judgement sentences and 2126 judgement preambles, which are different parts of a judgement. One notable difference between this dataset and the TAB dataset is that this dataset has another definition for the list of named entities, some more legally specific categories have been formed such as the entities

Table 4.3: Entity types in the Indian Legal NER dataset, their description along with the matching label used to make the conversion to the entity types presented in the TAB dataset.

Entity Type	**Description**	**New Label**
COURT	Name of the court which has delivered the current judgement if extracted from the preamble. Name of any court mentioned if extracted from judgment sentence.	ORG
PETITIONER	Name of the petitioners/appellants/revisionist from current case.	PERSON
RESPONDENT	Name of the respondents/defendants/opposition from current case.	PERSON
JUDGE	Name of the judges from the current case if extracted from the preamble. Name of the judges of the current as well as previous cases if extracted from judgment sentences.	PERSON
LAWYER	Name of the lawyers from both the parties.	PERSON
DATE	Any date mentioned in the judgment.	DATETIME
ORG	Name of organizations mentioned in text apart from the court.	ORG
GPE	Geopolitical locations which include names of states, cities, villages.	LOC
STATUTE	Name of the act or law mentioned in the judgement.	MISC
PROVISION	Sections, sub-sections, articles, orders, rules under a statute.	MISC
PRECEDENT	All the past court cases referred to in the judgement as precedent. Precedent consists of party names + citation(optional) or case number (optional).	MISC
CASE_NUMBER	All the other case numbers mentioned in the judgment (apart from precedent) where party names and citation is not provided.	CODE
WITNESS	Name of witnesses in current judgment.	PERSON
OTHER_PERSON	Name of all the persons that are not included in petitioner, respondent, judge and witness.	PERSON

"JUDGE", "COURT", "LAWYER", etc. Some of these will not be relevant for the study of privacy-protecting anonymization, since they are not related to any type of personal information, such as "STATUTE" which is a tag for when a statute is mentioned. There are a total of 14 entity types in this dataset. In table 4.3 the entity types that are present in this dataset are listed, along with a description and the entity type it is translated to in terms of the entity types from the TAB dataset. It is worthwhile to note that the entities "DEM" and "QUANTITY" from the TAB dataset have no representation in the converted entity types of this dataset. The count of entities in each entity type is given in table 4.4

Table 4.4: The entity counts and fractions of the entities found in the full Indian legal NER dataset.

Entity Type	**Count (fraction)**
COURT	2 367 (0.079)
PETITIONER	3 068 (0.102)
RESPONDENT	3 862 (0.129)
JUDGE	2 325 (0.078)
LAWYER	3 505 (0.117)
DATE	1 885 (0.063)
ORG	1 441 (0.048)
GPE	1 398 (0.047)
STATUTE	1 804 (0.060)
PROVISION	2 384 (0.080)
PRECEDENT	1 351 (0.045)
CASE_NUMBER	1 040 (0.035)
WITNESS	881 (0.029)
OTHER_PERSON	2 653 (0.089)
Total:	29 964

4.2 Experimental setup

This thesis aims to examine the performance of MIAs on a text masking model, and the feasibility of PII extraction by utilizing the masking model. These two parts of the thesis are investigated by considering the following differing factors in the experiments:

- How does the number of shadow models impact the performance of the MIAs,
- How the performance of MIAs changes when considering binary masking or NER-type masking, and
- How the performance of MIAs change when the shadow models are trained on OoD data, and
- Feasibility of PII extraction attempts.

Before any attack experiment is launched, a target model to attack is created. The target model is as stated in section 2.1 a pre-trained LongFormer with a feed-forward neural network applied to make classifications. The target model is trained on 917 documents for 2 epochs, using cross-entropy loss, and the $AdamW$ optimization algorithm with the following parameters: learning rate $2 \cdot 10^{-5}$, $\beta_1 = 0.9, \beta_2 =$

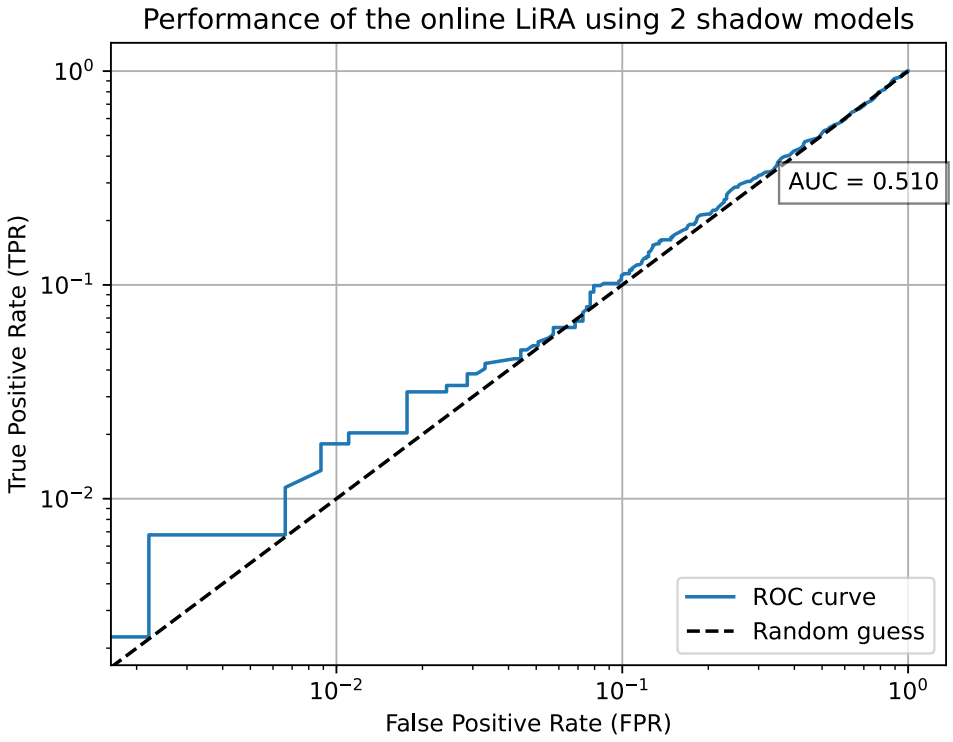
$0.99, \varepsilon = 1 \cdot 10^{-8}$. The epoch number was knowingly selected to be quite low, creating a difficult scenario with less overfitting to be exploited by the attacks, as was expanded on in section 2.5. The shadow models used are created in the same way as the target model. The results from the earlier experiments will determine which attack settings are used in the following attack experiments.

4.3 Different Number of Shadow Models

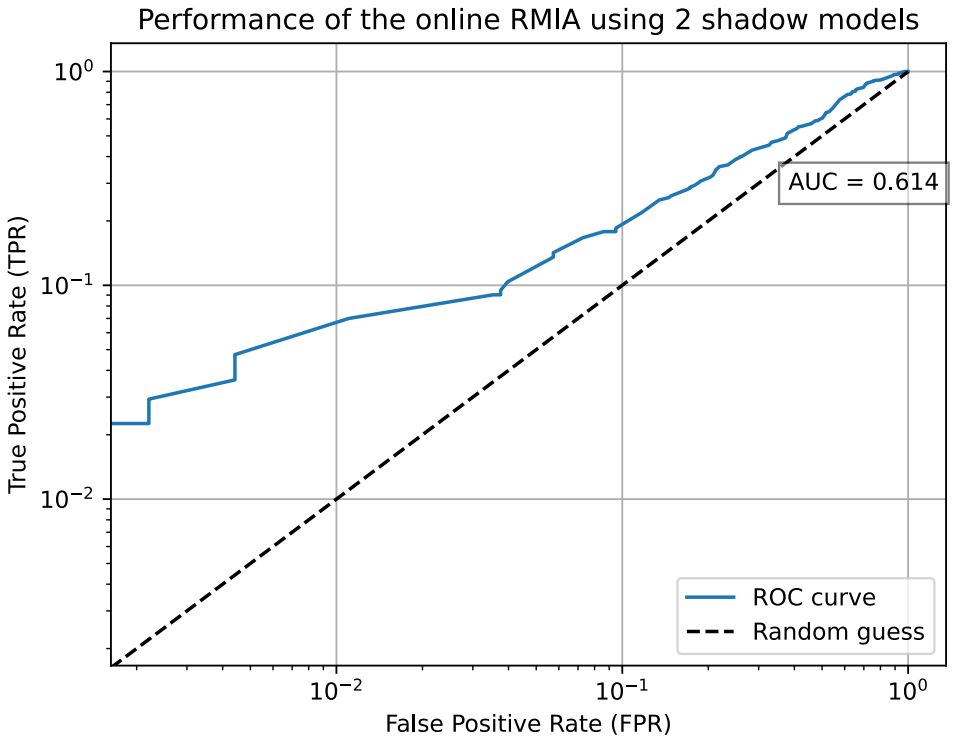
The first experiment varies the number of shadow models used in online and offline LiRA and RMIA attacks. The attack is performed with NER-type masking. The ROC curves of the attacks are presented in figures 4.1 - 4.6, and in table 4.5, the AUC and TPR at FPRs 0.01 and 0.001 for the different attacks are presented.

Table 4.5: AUC and TPR at FPRs 0.01 and 0.001, with the highest recorded results in bold. The values of TPR have been rounded by ±0.0005 for FPR = 0.001 and by ±0.005 for FPR = 0.01 since the results are empirically created and discrete. Entries with no reported value close enough to the desired FPR are marked -.

Attack Type	**AUC**	**TPR @ 0.001 FPR**	**TPR @ 0.01 FPR**
LiRA, online, 2 shadow models	0.510	0.002	0.018
RMIA, online, 2 shadow models	0.614	0.022	**0.070**
LiRA, online, 4 shadow models	0.492	0.002	0.014
RMIA, online, 4 shadow models	**0.635**	**0.026**	0.063
LiRA, online, 8 shadow models	0.508	0.002	0.017
RMIA, online, 8 shadow models	**0.635**	-	0.061
LiRA, offline, 2 shadow models	0.498	0.001	0.016
RMIA, offline, 2 shadow models	0.570	0.008	0.024
LiRA, offline, 4 shadow models	0.499	0.002	0.016
RMIA, offline, 4 shadow models	0.573	-	0.015
LiRA, offline, 8 shadow models	0.502	0.002	0.016
RMIA, offline, 8 shadow models	0.570	0.003	0.013

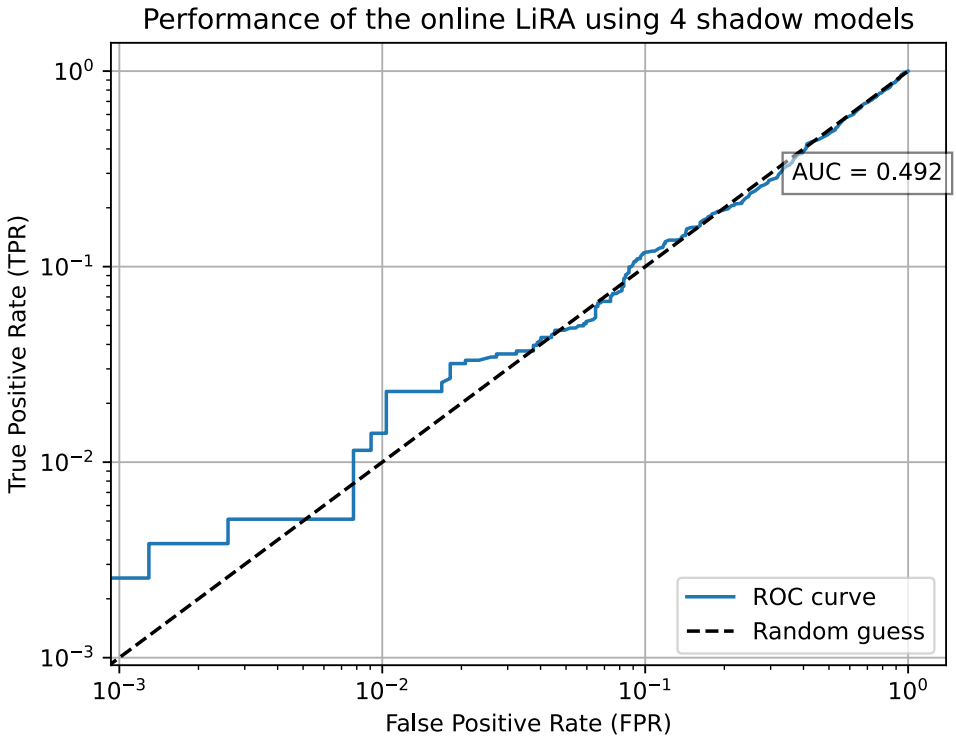


(a) LiRA

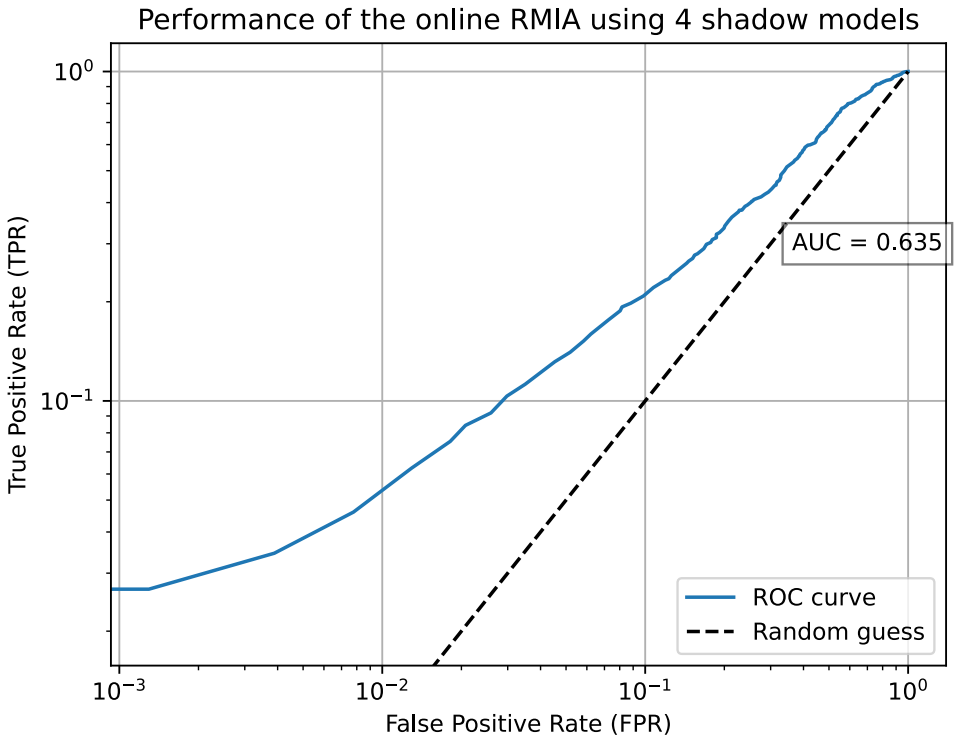


(b) RMIA

Figure 4.1: ROC curves for online LiRA and online RMIA when 2 shadow models are used in the attacks.

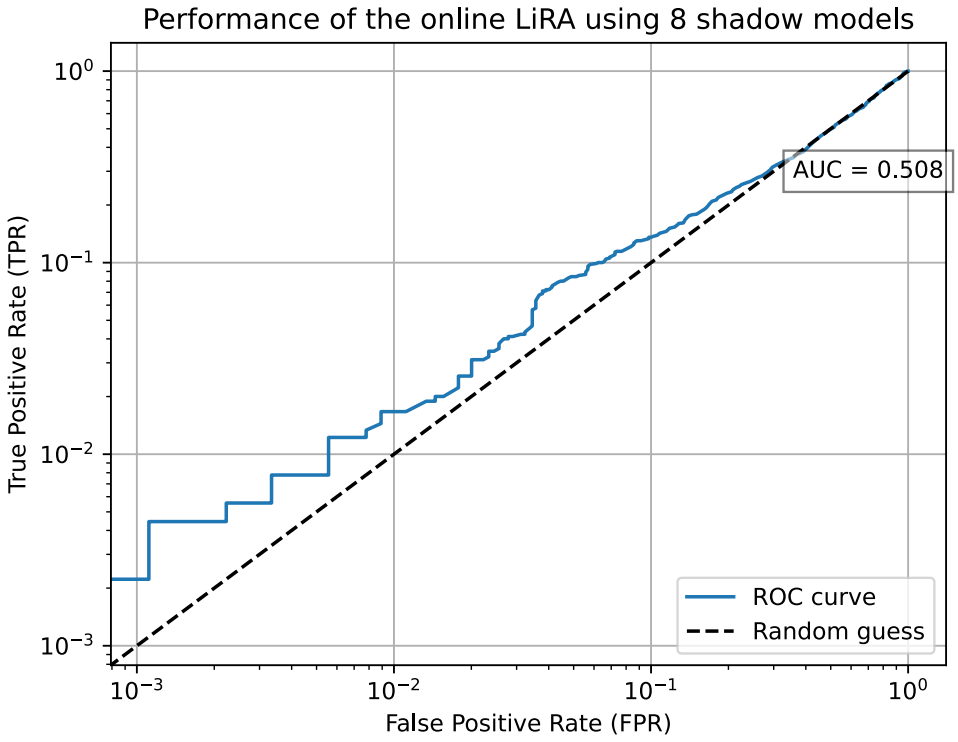


(a) LiRA

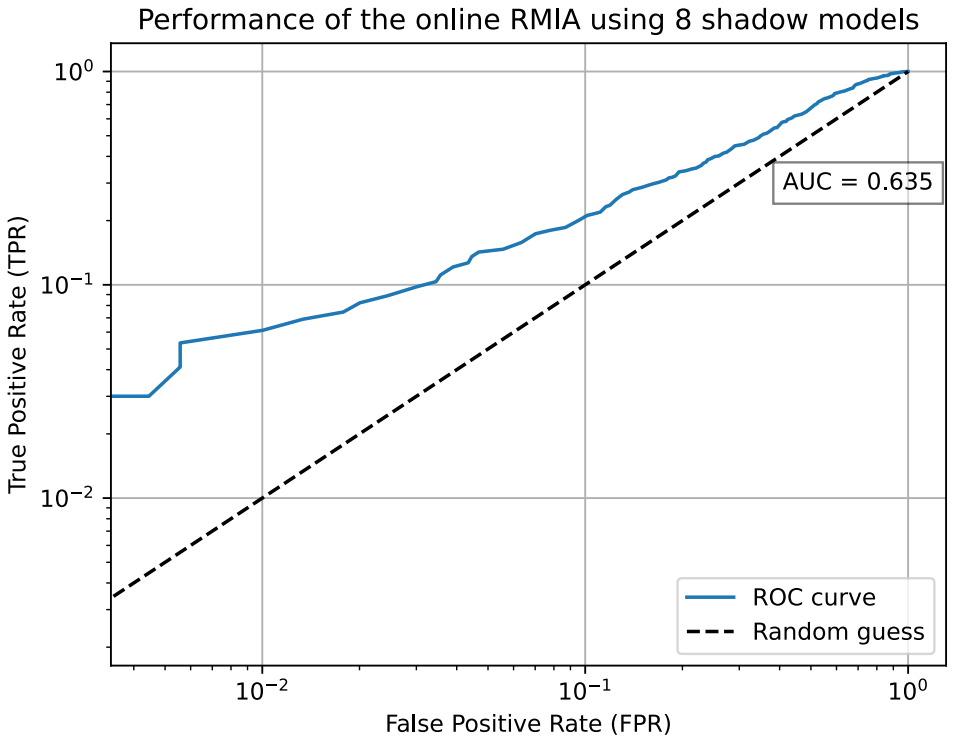


(b) RMIA

Figure 4.2: ROC curves for online LiRA and online RMIA when 4 shadow models are used in the attacks.

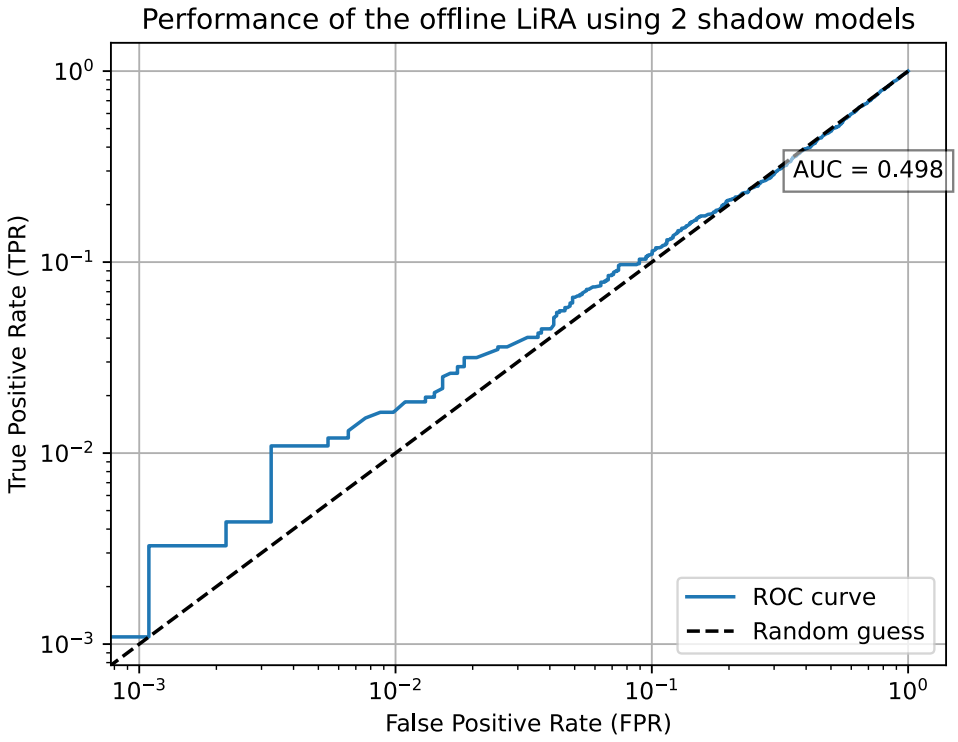


(a) LiRA

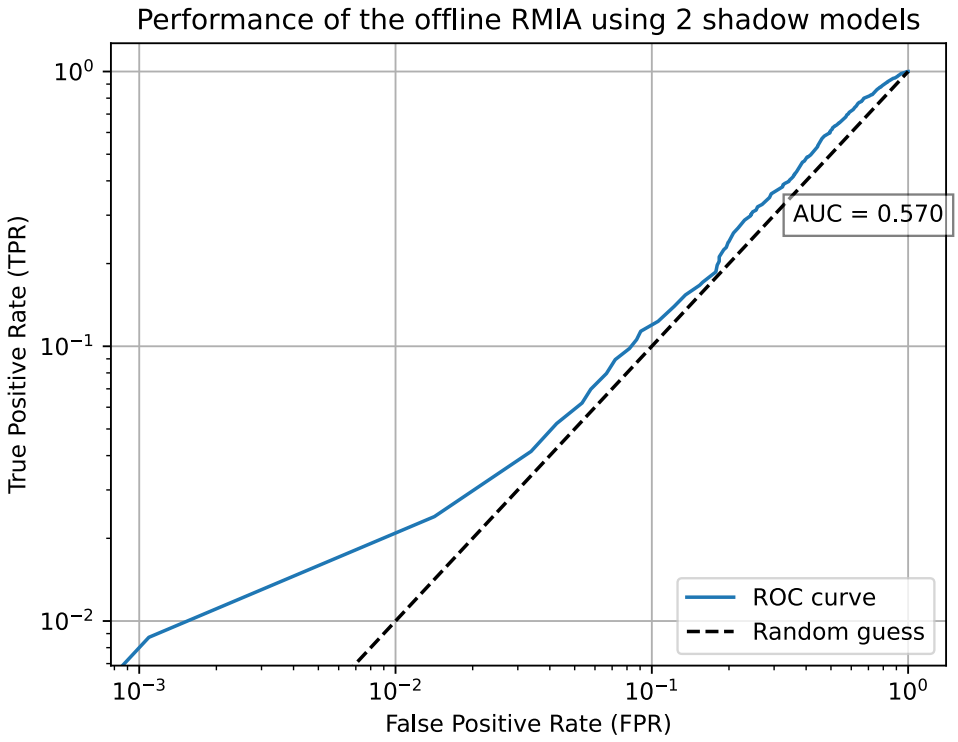


(b) RMIA

Figure 4.3: ROC curves for online LiRA and online RMIA when 8 shadow models are used in the attacks.

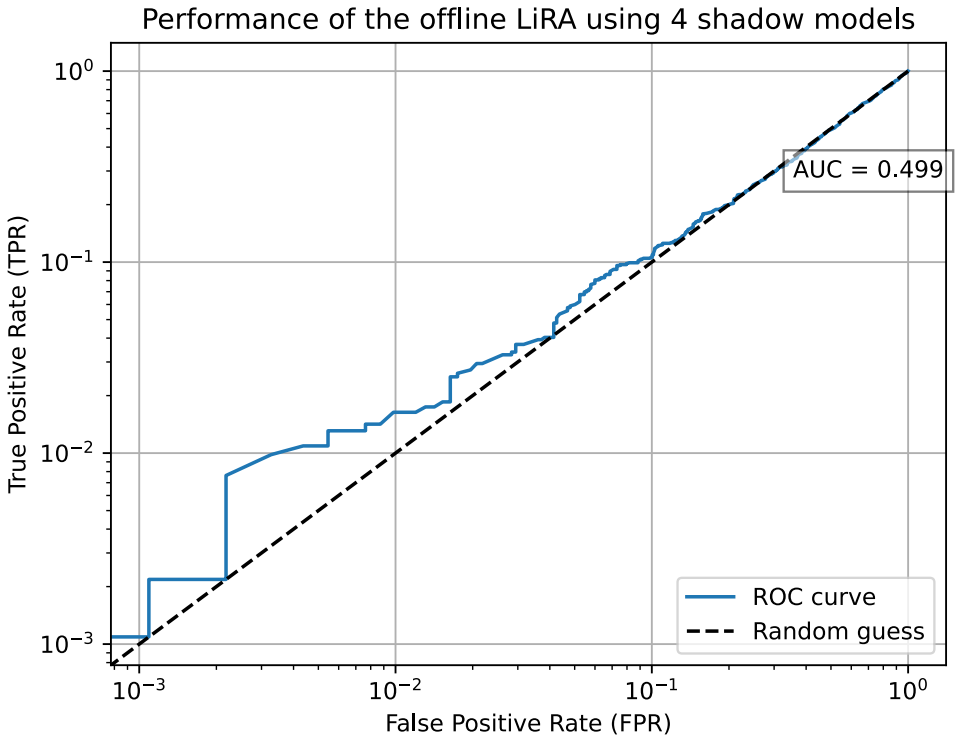


(a) LiRA

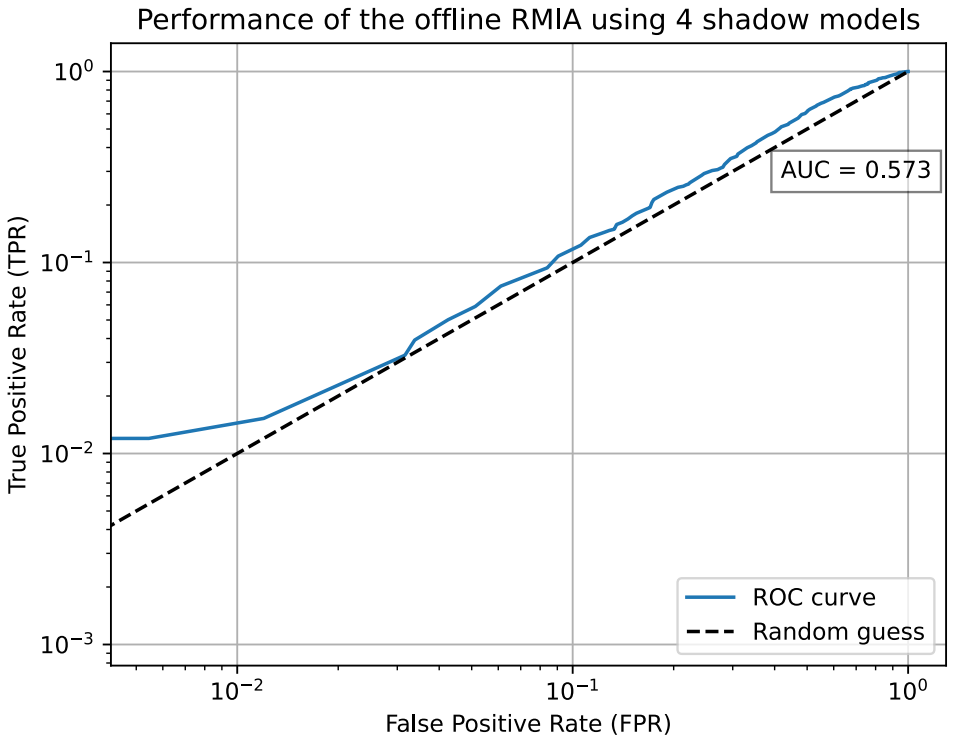


(b) RMIA

Figure 4.4: ROC curves for offline LiRA and offline RMIA when 2 shadow models are used in the attacks.

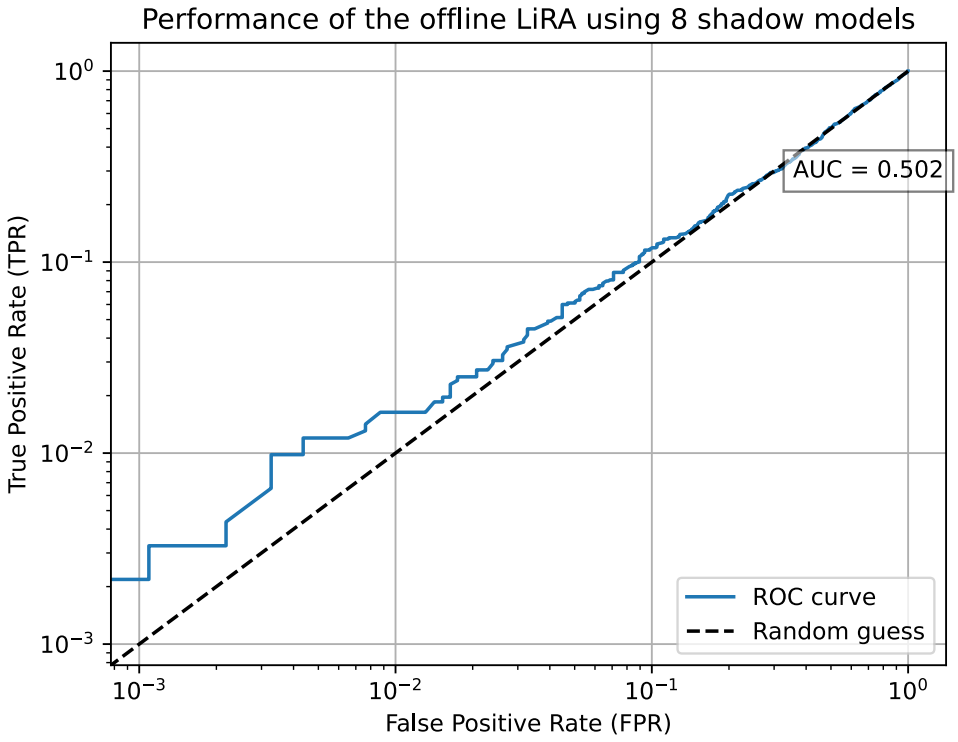


(a) LiRA

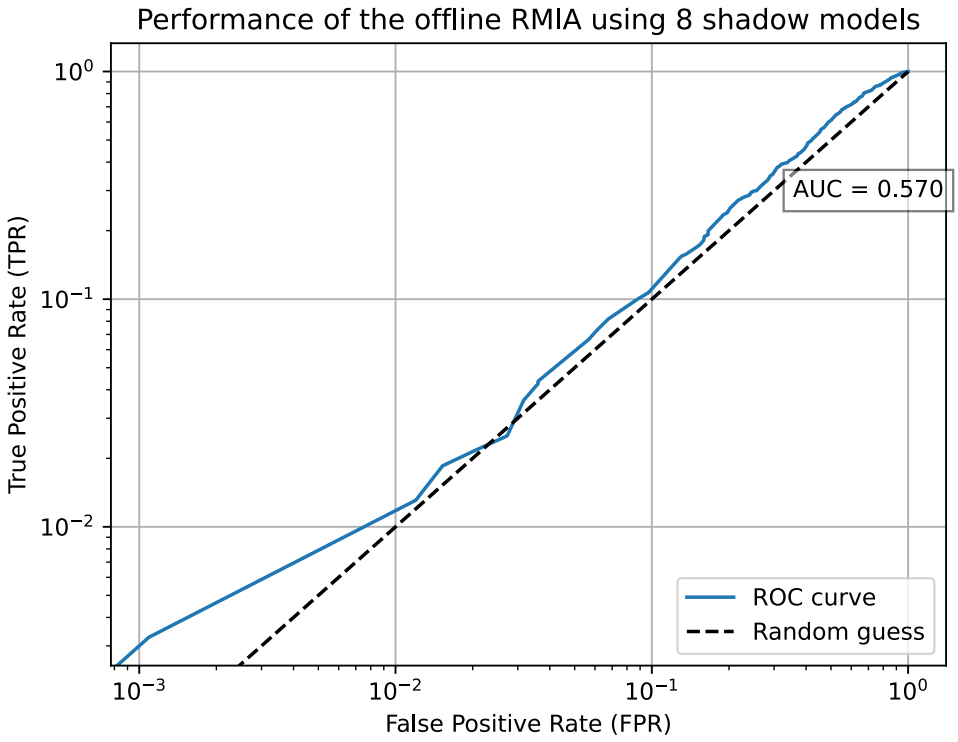


(b) RMIA

Figure 4.5: ROC curves for offline LiRA and offline RMIA when 4 shadow models are used in the attacks.



(a) LiRA



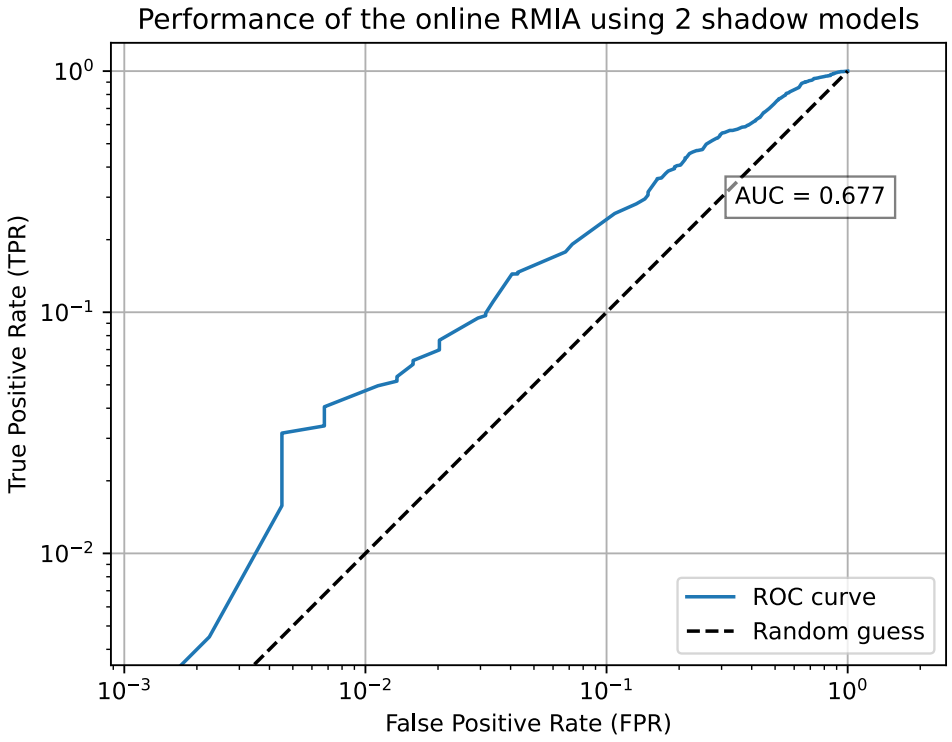
(b) RMIA

Figure 4.6: ROC curves for offline LiRA and offline RMIA when 8 shadow models are used in the attacks.

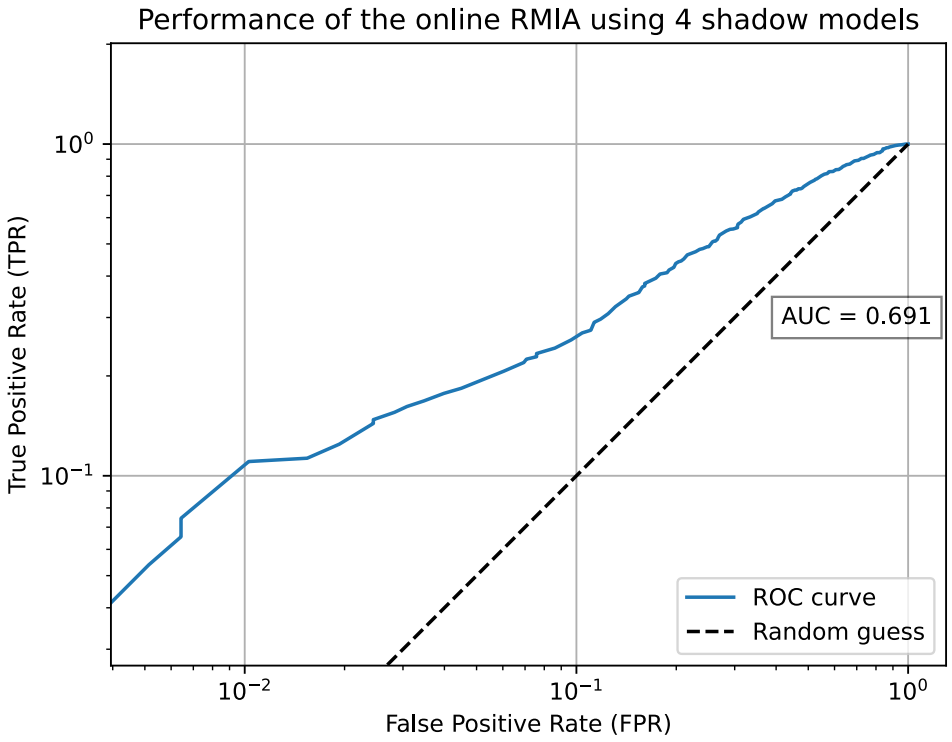
In table 4.5 the results of both online and offline LiRA and RMIA are presented when used with 2, 4, and 8 shadow models. The best performance at both the lowest FPR and the AUC was the online RMIA when using 4 shadow models. The online RMIA with 8 shadow models equaled the best AUC score. Online RMIA with 2 shadow models was the best at a slightly higher FPR.

4.4 Binary Masking of PII

For this experiment, the impact of the masking type is investigated. Binary masking is used, otherwise, the same settings as in the previous experiment apply. Binary masking reduces the target model to only mask entities as "[Mask]", and not by using their NER-type as the mask. Online RMIA with 2 and 4 shadow models will be the attacks used in this experiment, as they performed the best according to the metrics in table 4.5. In table 4.6 the results of the attacks in the lower regions of FPR are presented along with the results of the same attacks in the first experiment.



(a) Binary masking, online RMIA with 2 shadow models.



(b) Binary masking, online RMIA with 4 shadow models.

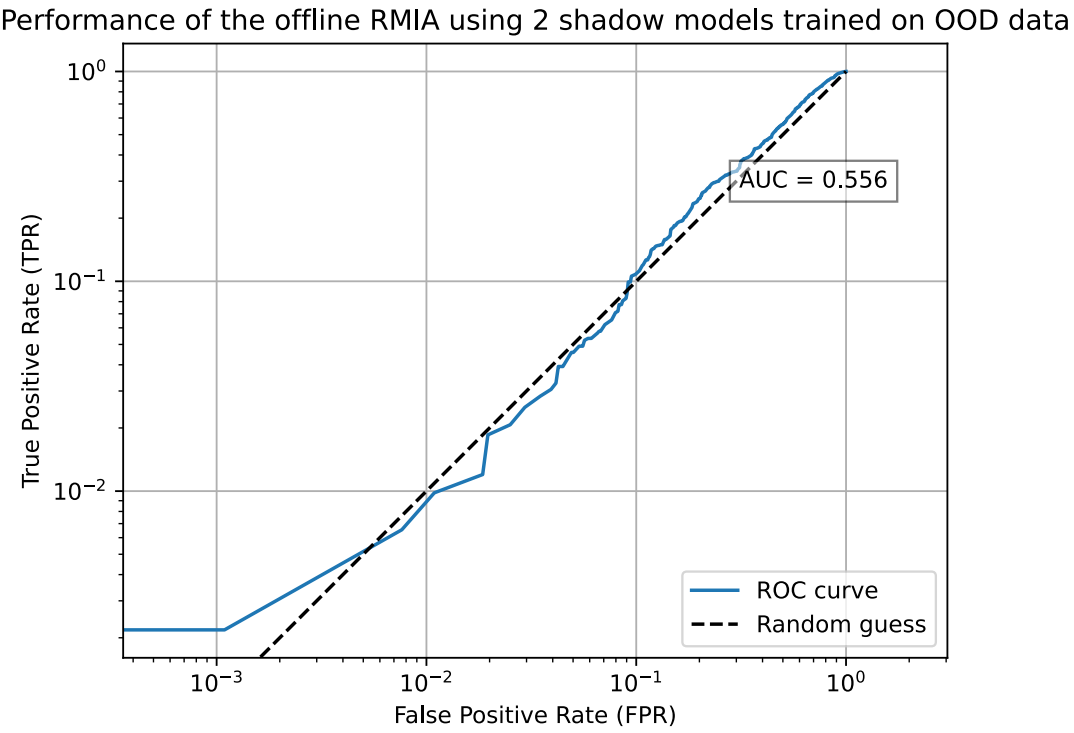
Figure 4.7: ROC curve for online RMIA when 2 and 4 shadow models are used by the adversary. Binary masking is used for this instance.

Table 4.6: Results for online RMIA when using 2 and 4 shadow models when using binary masking and NER-type masking models. AUC and TPR at FPRs 0.01 and 0.001, with the highest recorded results in bold. The values of TPR have been rounded by ±0.0005 for FPR = 0.001 and by ±0.005 for FPR = 0.01 since the results are empirically created and discrete. Entries with no reported value close enough to the desired FPR are marked -.

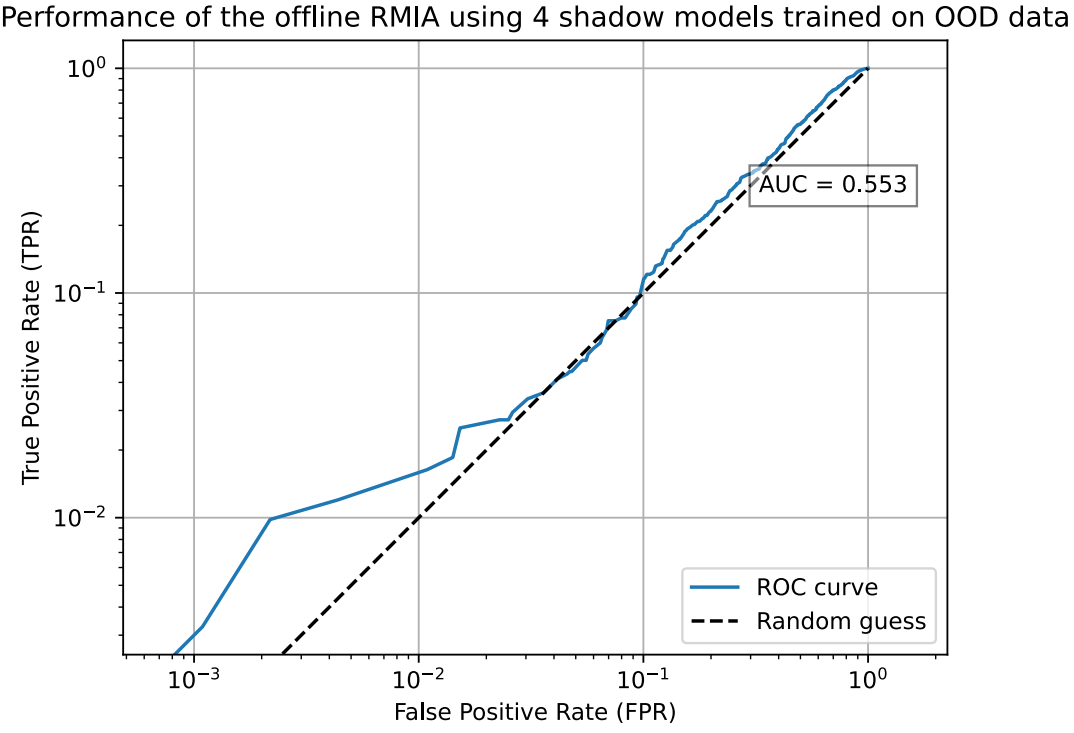
Attack Type	**AUC**	**TPR @ 0.001 FPR**	**TPR @ 0.01 FPR**
RMIA, binary masking, 2 shadow models	0.677	-	0.050
RMIA, NER-typ masking, 2 shadow models	0.614	0.022	0.070
RMIA, binary masking, 4 shadow models	**0.691**	-	**0.110**
RMIA, NER-type masking, 4 shadow models	0.635	**0.026**	0.063

4.5 Out-of-Distribution Data

The target model is trained on the training set sampled from the population $\mathcal{D}$, but the shadow models are trained on Out-of-Distribution data, another population $\mathcal{D}'$. For this reason, offline RMIA is used with 2 and 4 shadow models, and these shadow models are trained on the Indian legal NER dataset [38], all else remains as stated in section 4.2. ROC curves are presented in figure 4.8a and 4.8b, and results at lower FPR are also presented in table 4.7, along with the outcomes of the same models in the first experiment where the training of the shadow models was done on population data, i.e. data from the TAB dataset.



(a) Offline RMIA using 2 shadow models trained on OoD data.



(b) Offline RMIA using 4 shadow models trained on OoD data.

Figure 4.8: ROC curve for offline RMIA when 2 and 4 shadow models trained on OoD data are used by the adversary.

Table 4.7: Results for offline RMIA when using 2 and 4 shadow models trained on OoD data, and on in-distribution data. AUC and TPR at FPRs 0.01 and 0.001, with the highest recorded results in bold. The values of TPR have been rounded by ±0.0005 for FPR = 0.001 and by ±0.005 for FPR = 0.01 since the results are empirically created and discrete. Entries with no reported value close enough to the desired FPR are marked -.

Attack Type	**AUC**	**TPR @ 0.001 FPR**	**TPR @ 0.01 FPR**
RMIA, offline, 2 shadow models, trained on OoD data	0.556	0.002	0.010
RMIA, offline, 2 shadow models, trained on population data	0.570	**0.008**	**0.024**
RMIA, offline, 4 shadow models, trained on OoD data	0.553	0.003	0.016
RMIA, offline, 4 shadow models, trained on population data	**0.573**	-	0.015

4.6 PII Extraction attacks

Both attacks are given $T = 1000$ rounds to evaluate their actions. Since this is not a membership inference attack, the results are not presented using ROC curves. What is reported instead is the maximum average reward as a function of time, i.e. at time $t_0 = 50$, the sum is calculated of the maximum rewards of all bandits of time steps $t \leq 50$, and then divided by the number of masked entities in the document. In the experiment, there are a total of 71 masked PII to extract from the examined document. The total action space thus has the size $(S_0, S_1, ..., S_{71})$, where S_i is the number of candidates in the pool of the entity at masked position i. These sizes are presented in table 4.2, and there is notably more candidates than there are rounds to play for all entities, and also a large spread in entity counts.

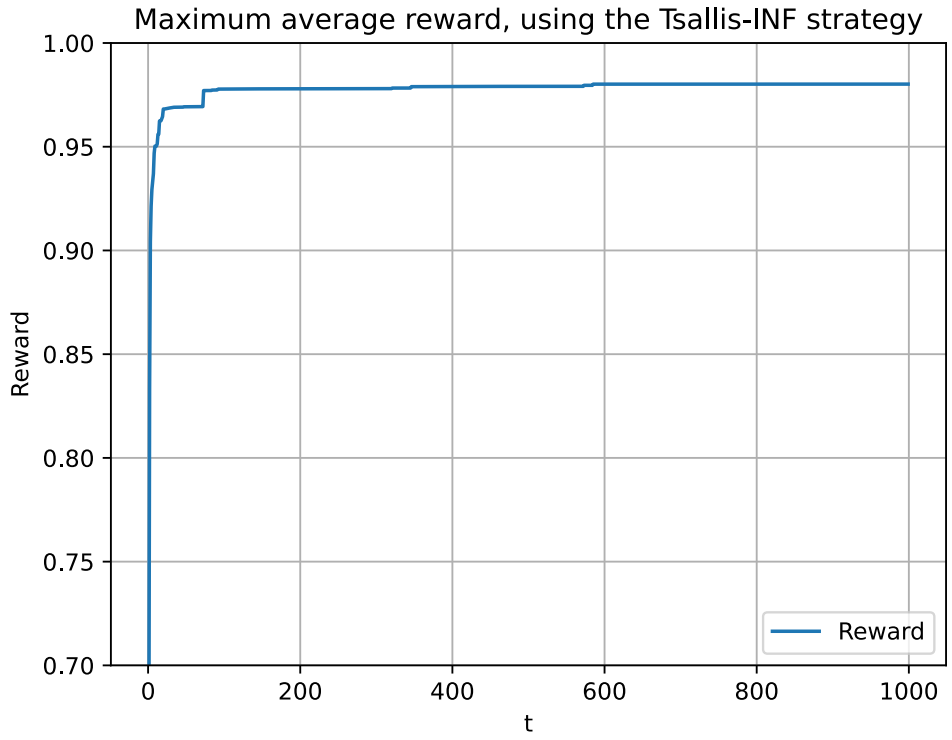


Figure 4.9: Maximum average reward of the extraction attack as a function of time when using the Tsallis-INF algorithm.

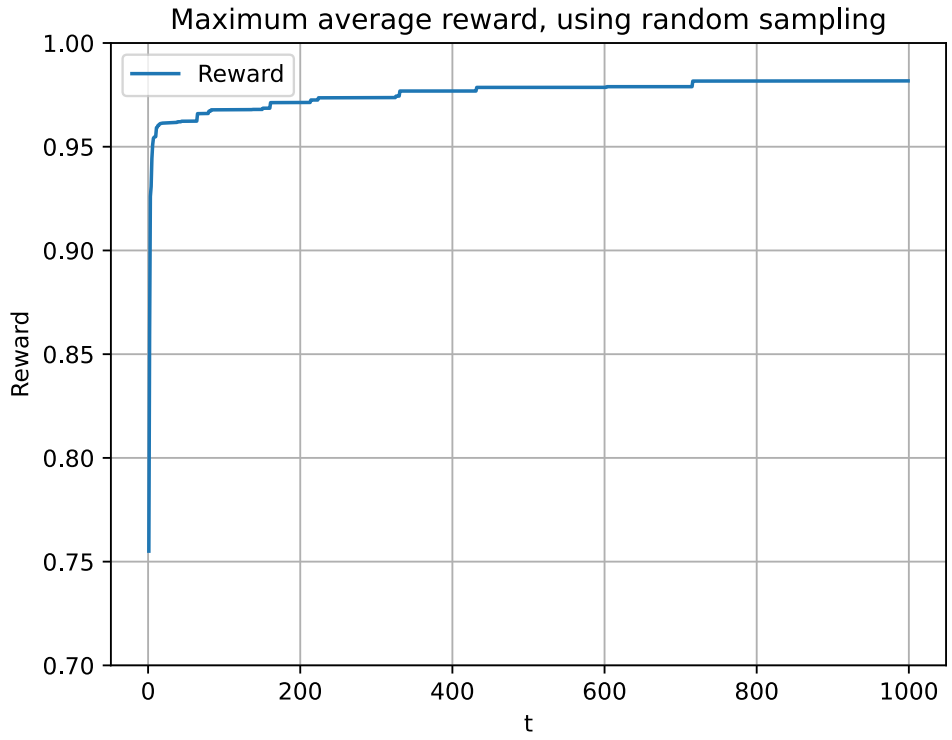


Figure 4.10: Maximum average reward of the extraction attack as a function of time when using random sampling.

5

Discussion

5.1 The Impact of the Datasets

The two datasets used in this thesis are more or less related to the specific use case of anonymizing court cases. One positive property of both of the datasets is that they are both in the domain of legal texts, which is the relevant scenario studied. Having data to work with from the correct domain gives the experiments a more real-world-based ground to stand on, and circumvents some of the compromises that otherwise have to be made when working with synthetic data, or data from other domains than the thought application.

The TAB dataset [37] is highly representative of the use case, as it was created with the goal of anonymizing a text from revealing the identity of a specific individual. The Indian legal NER dataset [38], however, doesn't have this goal in mind. This would have been a worse dataset to base the main experiments of the thesis on, but as an OoD dataset it is not too bad. It is still created for NER in the domain of legal texts, although it is both formed with a different purpose in mind and is taken from a different legal entity (Indian courts, whereas the TAB dataset is based on judgements from the ECHR).

5.2 Interpretation of Results

A general note regarding the reported results is that these attacks are dependent on several stochastic properties in their execution, and thus the outcome could vary between runs due to these factors and not only attack performance. There's

randomness at play in both the training processes and the selection of training datasets of the target model and shadow models, which may have some impact on the results. For example, data points that are outliers in comparison to the rest of the population will be easier to predict the membership of [11].

5.2.1 Number of Shadow Models

In section 4.3, the attack performance of the LiRA and RMIA when changing the number of shadow models, are given through ROC curves in figures 4.1 - 4.6, and in tabular form in table 4.5. The difference in performance is not extremely large, although the RMIA appears to achieve a higher AUC in each attack setting, regardless of the number of shadow models used, or if the attack is online or offline. The LiRA attacks appear to have a more ragged form of the ROC curves, where the ROC curves of RMIA in most cases start at a higher level and is smoother as the FPR increases.

As stated in section 2.4.2, the AUC is an average case metric which is not the most representative way to report privacy risks. The relevant region of interest is when the FPR is low, in the left regions of the figures in section 4.3. For comparison between attacks, table 4.5 shows the TPR at two different FPR for all of the attack types and settings. The higher the TPR is at lower FPRs, the greater the security risk. Even here it appears that RMIA outperforms LiRA. For example, consider figure 4.1, which depicts the results of the online versions of the attacks running with 2 shadow models. The online RMIA reaches a TPR of $\approx 2.3 \cdot 10^{-2}$ at FPR $\approx 2.2 \cdot 10^{-3}$, a TPR about 10 times higher than the FPR. At the same FPR, the online LiRA achieves a TPR of $\approx 2.2 \cdot 10^{-2}$, indicating that RMIA in this scenario is about 10 times better than LiRA.

For the offline attacks, a similar finding is present, which can be seen in figure 4.4. For the same scenario, using 2 shadow models for both LiRA and RMIA, RMIA once again outperforms LiRA. Both attacks performed better in the online setting, which may be indicative of online attacks being stronger than offline ones. Since the online attacks utilize the target record in the construction of their shadow models, this is no strange finding.

The ROC curves are not too far off from random guessing, but as it matters most in the lower regions of FPR, both attacks seem to be performing better than chance, especially RMIA.

5.2.2 Binary Masking of PII

The results in section 4.4 show that the model is susceptible to MIAs even when the masks don't reveal the underlying NER-type. The results appear to be quite strong, which may seem strange as larger and more complex models have previously been shown to be easier to attack than simpler models [11]. Comparing the results found in table 4.6, the online RMIA with binary masks, using 4 shadow models, has the best results in terms of AUC and TPR at FPR = 0.01, followed closely by the same attack using 2 shadow models.

One possible explanation for this is that seeming as there is now less information for the shadow models to use in training due to the smaller number of classification categories, it is possible that the target model's confidence is notably larger than the shadow models in ambiguous situations, and this difference causes the attack to perform better even in these theoretically worse circumstances.

5.2.3 Out-of-Distribution Data

As can be seen in figures 4.8a and 4.8b, the performance quite rapidly becomes quite close to random guessing. This is reflected in the AUC score being close to 0.5, see table 4.7, which is the AUC for random guessing.

Further studying table 4.7, when comparing the performance of attacks training shadow models on the OoD data and on population data, one can note that the best performance is by the attack using 2 shadow models trained on population data. The attack using 4 shadow models trained on OoD data is better than using only 2 shadow models on OoD data, and in TPR 0.01 FPR it outperforms the attack using 4 shadow models trained on population data, even if this difference is small enough to possibly be due to randomness.

What this shows is that while the performance is not quite as good, it is still possible

to in some regard be able to attack models even if the population data is not available for the attacker to use for training shadow models. The fact that this is possible is impactful since a limitation of data availability is with this in mind not enough for a model to be considered safe toward attacks.
One can also compare the distributions in the datasets, in tables 4.2 and 4.4, and note some differences. For example, "DATE" in the Indian legal NER dataset has a fraction value of 0.063, where as "DATETIME" in the TAB dataset has a fraction value of 0.46, more than 7 times larger. This discrepancy in fraction values is present in multiple other entity categories, and this can also be thought to be a reason for the degraded performance of the OoD shadow model based attacks.

5.2.4 PII Extraction Attacks

The results can be seen in figures 4.9 and 4.10. What can be noted here is that while both of the attacks failed to find the masked tokens, they performed similarly. Interestingly enough, however, the random sampling strategy outperformed Tsallis-INF as time grew, even if the Tsallis-INF strategy seemed to perform better initially. One explanation for this is that the spread between rewards of action is not that large. The loss and reward are related to the target model's confidence in predicting the imputed PII as the correct NER-type, and since the candidates for imputing are selected from the correct NER-type, the candidates should have high confidence even if they are not the exact PII that where masked. This may lead Tsallis-INF to reduce exploration, as it appears to find decent rewards quite early, which could cause the random sampling approach to outperform it in later stages.
While these results are quite weak, this investigation shows that there is perhaps some potential in using multi-armed bandit approaches for extracting PII from a censored document. Limiting the action space for the masked position based on other masks could be an interesting avenue to investigate further; the masked pieces of PII are all related to one individual, and thus there is some potential to exploit the correlation between the possible actions to fill masks. As can be seen in table 4.2, there are very many possible candidates to sample in each candidate pool, which also is an explanation of the issues both attack strategies had in finding the correct

PII.

5.3 Limitations

A limitation of this thesis is that the results formed have not considered the performance of the target model in great detail. While this could have been done, increasing the performance of a model has been linked with increased susceptibility to several types of model integrity attacks. Increased performance in the target model could thus potentially have led to a weaker model to attack.

5.4 Ethical Considerations

The subject of this thesis is related to privacy infringement, and as such some considerations should be made of whether or not the work could be used by a malicious client. Seeming as the risk of this is quite low due to the fact that the attacks performed in this thesis were implementations of existing attacks on new data types and models, the risk of misuse is probably not to be considered as increased. Some new approaches in the extraction attacks have been made, but as they are not posing great threats, the risk of misuse could also here be considered low.

5.5 Future Works

The extraction attacks used in this thesis are quite simple, or implemented in a simple way. One way of improving their performance is to allow the adversary to maintain some additional insights in regard to which entities came from which document when forming the pools. There can also be some possibility that a specific entity is mentioned multiple times, so correlating bandits of similar entity types to each other is another possible area that can be investigated.

6

Conclusion

This thesis has examined the model integrity of LLMs trained to anonymize text in two different ways; through the study of membership inference attacks, and data extraction attacks. This has been done using both state-of-the-art attacks for membership inference, and simple but novel extraction attacks. The results of the study indicate that for the studied scenario, the Robust Membership Inference Attack (RMIA) performs significantly better than the Likelihood Ratio Attack (LiRA) both in terms of True Positive Rate in regions of low False Positive Rate, and the Area Under the Curve metric. This indicates that it performs better both when the number of mistakes is to be reduced, but also over all the different False Positive Rates, and as such RMIA is the more potent attack for the studied scenario.

When comparing the online and offline versions of the attacks against each other, the expectation that the online attacks perform better than the offline attacks is confirmed. One note however is that the performance of the offline RMIA is at a very similar level as the online LiRA, something that no real explanation can be given about, other than that RMIA is a better attack in these circumstances. The number of shadow models used appears to have a negligible impact on the offline attacks, and even in the online attacks there appears to be no strict correlation between the number of shadow models used and the performance of the attack.

The case of binary masking showed that this model type was also susceptible to membership inference attacks, and in some measurements even more so than in the case of NER-typ masking. The expansion towards using shadow models trained on OOD data showed weak attack performance but proved that there is a possibility to conduct MIAs even if the population is not available to the adversary, something

which speaks to limiting data availability is not enough to consider a model safe from MIAs.

The PII extraction attacks performed quite poorly and were not able to extract the masked PII. With that in mind, the extraction attacks seemed to be able to improve upon their estimates, so it is possible that it could find the masked PII if used in more generous circumstances. Some expansions of the attacks in terms of limiting the action space were discussed, and could possibly be interesting to investigate in future works.

Bibliography

[1] Stacey Tobin, Bamini Jayabalasingham, Sarah Huggett, and Maria de Kleijn. A brief historical overview of artificial intelligence research. *Information Services & Use*, 39(4):291–296, February 2020. URL: http://dx.doi.org/10.3233/isu-190060, doi:10.3233/isu-190060.

[2] Anand Kumar Chennupati. The evolution of ai: What does the future hold in the next two years. *World Journal of Advanced Engineering Technology and Sciences*, 12(1):022–028, May 2024. URL: http://dx.doi.org/10.30574/wjaets.2024.12.1.0176, doi:10.30574/wjaets.2024.12.1.0176.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL: http://arxiv.org/abs/1706.03762, arXiv:1706.03762.

[4] Nikita Nangia and Samuel R. Bowman. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. *CoRR*, abs/1905.10425, 2019. URL: http://arxiv.org/abs/1905.10425, arXiv:1905.10425.

[5] John P. Lalor, Hao Wu, and Hong Yu. Improving machine learning ability with fine-tuning. *ArXiv*, abs/1702.08563, 2017. URL: https://api.semanticscholar.org/CorpusID:7256103.

[6] Kenneth Ward Church, Zeyu Chen, and Yanjun Ma. Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*, 27(6):763–778, October 2021. URL: http://dx.doi.org/10.1017/S1351324921000322, doi:10.1017/s1351324921000322.

[7] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL: https://data.europa.eu/eli/reg/2016/679/oj.

[8] Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Sina Noack, Hendrik Patzlaff, Hazar Harmouch, and Felix Naumann. The effects of data quality on machine learning performance, 2022. URL: https://api.semanticscholar.org/CorpusID:251223513.

[9] AI Sweden. Leakpro: Leakage profiling and risk oversight for ma-

chine learning models. URL: https://www.ai.se/en/project/leakpro-leakage-profiling-and-risk-oversight-machine-learning-models.

[10] Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, and Dawn Song. Llm-pbe: Assessing data privacy in large language models, 2024. URL: https://arxiv.org/abs/2408.12787, arXiv:2408.12787.

[11] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. *CoRR*, abs/2112.03570, 2021. URL: https://arxiv.org/abs/2112.03570, arXiv:2112.03570.

[12] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models, 2023. URL: https://arxiv.org/abs/2302.00539, arXiv:2302.00539.

[13] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *CoRR*, abs/1610.05820, 2016. URL: http://arxiv.org/abs/1610.05820, arXiv:1610.05820.

[14] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2024. URL: https://arxiv.org/abs/2310.16789, arXiv:2310.16789.

[15] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks, 2024. URL: https://arxiv.org/abs/2312.03262, arXiv:2312.03262.

[16] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, and Reza Shokri. Enhanced membership inference attacks against machine learning models. *CoRR*, abs/2111.09679, 2021. URL: https://arxiv.org/abs/2111.09679, arXiv:2111.09679.

[17] Offentlighetsprincipen (1949:105), 1949. URL: https://www.riksdagen.se/en/documents-laws/laws/freedom-of-the-press-act-1949105_gk.

[18] Mansi Agarwal. An overview of natural language processing. *International Journal for Research in Applied Science and Engineering Technology*, 7(5):2811–2813, May 2019. URL: http://dx.doi.org/10.22214/ijraset.2019.5462, doi:10.22214/ijraset.2019.5462.

[19] Pablo Villalobos, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, Anson Ho, and Marius Hobbhahn. Machine learning model sizes and the parameter gap, 2022. URL: https://arxiv.org/abs/2207.02852, arXiv:2207.02852.

[20] Julia Witte Zimmerman, Denis Hudon, Kathryn Cramer, Alejandro J. Ruiz, Calla Beauregard, Ashley Fehr, Mikaela Irene D. Fudolig, Bradford Demarest, Yoshi Meke Bird, Milo Z. Trujillo, Christopher M. Danforth, and Peter Sheridan Dodds. Tokens, the oft-overlooked appetizer: Large language mod-

els, the distributional hypothesis, and meaning, 2024. URL: https://api.semanticscholar.org/CorpusID:274776320.

[21] A Neelima and Shashi Mehrotra. A comprehensive review on word embedding techniques. In *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, page 538–543. IEEE, February 2023. URL: http://dx.doi.org/10.1109/ICISCoIS56541.2023.10100347, doi:10.1109/iciscois56541.2023.10100347.

[22] Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models: A survey, 2024. URL: https://arxiv.org/abs/2409.03752, doi:10.48550/ARXIV.2409.03752.

[23] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL: https://arxiv.org/abs/2004.05150, arXiv:2004.05150.

[24] Archana Goyal, Manish Kumar, and Vishal Gupta. Named entity recognition: Applications, approaches and challenges, 2017. URL: https://api.semanticscholar.org/CorpusID:212449811.

[25] Integritetsskyddsmyndigheten. Utlämnande av allmänna handlingar med hjälp av ai, 2024. URL: https://www.imy.se/globalassets/dokument/rapporter/utlamnande-av-allmanna-handlingar-med-hjalp-av-ai.pdf.

[26] Md Mostafizur Rahman, Aiasha Siddika Arshi, Md Mehedi Hasan, Sumayia Farzana Mishu, Hossain Shahriar, and Fan Wu. Security risk and attacks in ai: A survey of security and privacy. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, page 1834–1839. IEEE, June 2023. URL: http://dx.doi.org/10.1109/COMPSAC57700.2023.00284, doi:10.1109/compsac57700.2023.00284.

[27] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning, 2023. URL: https://arxiv.org/abs/2212.10986, arXiv:2212.10986.

[28] Mahdi Khosravy, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, and Noboru Babaguchi and. Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system. *KSII Transactions on Internet and Information Systems*, 15(3):1100–1118, March 2021. doi:10.3837/tiis.2021.03.015.

[29] Abdul Majeed and Seong Oun Hwang. When ai meets information privacy: The adversarial role of ai in data sharing scenario. *IEEE Access*, 11:76177–76195, 2023. URL: http://dx.doi.org/10.1109/ACCESS.2023.3297646, doi:10.1109/access.2023.3297646.

[30] Samuel Yeom, Matt Fredrikson, and Somesh Jha. The unintended consequences

of overfitting: Training data inference attacks. *CoRR*, abs/1709.01604, 2017. URL: http://arxiv.org/abs/1709.01604, arXiv:1709.01604.

[31] Tetiana Sliusarenko and Mariia Pohurska. Stop overfitting with proven techniques. *Grail of Science*, (40):373–375, June 2024. URL: http://dx.doi.org/10.36074/grail-of-science.07.06.2024.057, doi:10.36074/grail-of-science.07.06.2024.057.

[32] Benjamin Zi Hao Zhao, Mohamed Ali Kaafar, and Nicolas Kourtellis. Not one but many tradeoffs: Privacy vs. utility in differentially private machine learning. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, CCS '20, page 15–26. ACM, November 2020. URL: http://dx.doi.org/10.1145/3411495.3421352, doi:10.1145/3411495.3421352.

[33] Han-Cheol Cho, Naoaki Okazaki, Makoto Miwa, and Jun'ichi Tsujii. Named entity recognition with multiple segment representations. *Information Processing & Management*, 49(4):954–965, 2013. URL: https://www.sciencedirect.com/science/article/pii/S0306457313000368, doi:10.1016/j.ipm.2013.03.002.

[34] Kevin G. Jamieson. Some notes on multi-armed bandits, 2020. URL: https://api.semanticscholar.org/CorpusID:215714298.

[35] Pierre Gaillard. Adversarial bandits, 2019. URL: https://pierre.gaillard.me/doc/2019_cours_adversarial_bandits.pdf.

[36] Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits, 2022. URL: https://arxiv.org/abs/1807.07623, arXiv:1807.07623.

[37] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *CoRR*, abs/2202.00443, 2022. URL: https://arxiv.org/abs/2202.00443, arXiv:2202.00443.

[38] Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. Named entity recognition in indian court judgments, 2022. URL: https://arxiv.org/abs/2211.03442, arXiv:2211.03442.

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY