Anonymization by Design of Language Modeling

Antoine Boutet INSA Lyon, Inria, CITI, UR3720, 69621 Villeurbanne, France antoine.boutet@insa-lyon.fr

Lucas Magnana Inria, INSA Lyon, CITI, UR3720, 69621 Villeurbanne, France lucas.magnana@inria.fr

ABSTRACT

Rapid advances in Natural Language Processing (NLP) have revolutionized many fields, including healthcare. However, these advances raise significant privacy concerns, especially when models specialized on sensitive data can memorize and then expose and regurgitate confidential information. This paper presents a privacyby-design language modeling approach to address the problem of language models anonymization, and thus promote their sharing. Specifically, we propose both a Masking Language Modeling (MLM) methodology to specialize a BERT-like language model, and a Causal Language Modeling (CLM) methodology to specialize a GPT-like model that avoids the model from memorizing direct and indirect identifying information present in the training data. We have comprehensively evaluated our approaches using medical datasets and compared them against different baselines. Our results indicate that by avoiding memorizing both direct and indirect identifiers during model specialization, our masking and causal language modeling schemes offer the best tradeoff for maintaining high privacy while retaining high utility.

KEYWORDS

Language Models, Privacy, NLP, LLM, Anonymization

1 INTRODUCTION

The healthcare sector is a major generator of sensitive data collected from different sources. The exploitation of this valuable data presents many promises such as improving the quality of care, acquiring a better knowledge of the healthcare system, identifying disease risk factors, aiding in diagnosis, personalized care to name a few. A large amount of this data is unstructured text documents in the form of medical reports. With the rise of Machine Learning (ML) and the advent of Natural Language Processing (NLP), language models are increasingly used to automate the processing of these medical reports [9, 17, 51, 54].

Patient medical records are extremely sensitive and private data. Their use and dissemination are therefore subject to numerous regulations such as HIPAA, for the USA, or GDPR for Europe. In these regulations, one of the main prerequisites for the dissemination of medical data is to remove any element that allows a patient to be directly identified (i.e., de-identification or pseudonymization) or indirectly (i.e., anonymization).

Recent advances in NLP [40] based on neural networks have democratized their use. Since the advent of ChatGPT, NLP models Zakaria El Kazdam Ensimag, 38400 Saint-Martin-d'Hères, France zakaria.el-kazdam@grenoble-inp.org

Hélain Zimmermann

Ensimag, 38400 Saint-Martin-d'Hères, France helain.zimmermann@grenoble-inp.org

are not limited to text generation and can include several tasks, including classification and Named Entity Recognition (NER), which therefore allows for de-identification of free texts. For example, Johnson et al. proposed to use a neural network based on a BERT architecture [15] to detect a number of identifying elements in medical documents. More recently, different hospitals have also explored the feasibility of using NLP models to automatically pseudonymize text documents (i.e., hide specific direct identifiers named Personally Identifiable Information (PII)) from their clinical data warehouse [35, 45]. In these approaches, the BERT model is fine-tuned with the medical reports from the hospital (in order to specialize and well understand the reports generated by the local practitioners) before training a Named Entity Recognition on a set of Personally Identifiable Information that directly identify patients.

However, pseudonymization is not sufficient for text anonymization [33]. Indeed, the identity of an individual can be found from indirect identifying information. Therefore, these indirect identifiers must also be protected (e.g., removed) in order to properly anonymize texts. This operation cannot rely on a classification or NER task, because there is no a priori information to identify these indirect identifiers, it depends on the context and the dataset considered. To our knowledge, no NLP model training methodology has been proposed for taking into account indirectly re-identifying information in the protection of medical reports or documents from other domains.

The need for information sharing in the medical field goes beyond data sharing and now concerns learning models. As language models specialized on clinical reports improves their performance for downstream clinical NLP tasks [13, 21, 23], some hospitals or healthcare centers also want to be able to share models specialized and fine-tuned on their own data with other centers. For example, a hospital specialized in oncology may train a classification model with its own data and would like to share this model with other hospitals. In this context, the model specialized with the medical reports of the first hospital must not be able to regurgitate sensitive data used during its specialization (or return any other information allowing to deduce them) when it will be used by the second hospital. The attack surface on models trained on personal and highly sensitive data is still poorly understood [12, 24, 52]. A number of threats are related to the memorization and possible leakage of sensitive information used during model training. There are some known privacy vulnerabilities involving training data associated with learning models, such as memorization, data reconstruction,

and membership inference (i.e., identifying elements used during the training).

Memorization of information by a model is not a problem in itself, it is even necessary for learning. However, this memorization becomes a problem when the training information is not generalized enough by the model which reproduces large portions of training data verbatim or discloses some sensitive information. And more specifically, the risk to privacy is increased when this problematic memorization concerns identifying information. Indeed, memorizing identifiers that leads to the regurgitation of these identifiers can clearly be used to infer that the information of a specific individual was used during the training of the model (i.e., membership inference). Thus, in order to share a model without risk of regurgitation of personal information, it is necessary to ensure that a model has not memorized both direct and indirect identifiers. However, as far as we know, no methodology has been proposed to help the different actors, especially hospitals, to "anonymize" their learning models before sharing them.

To reduce privacy risks, mitigation techniques have been proposed such as learning with differential privacy [2] (DP) guarantees or pruning strategy [50]. However, these mitigation techniques only increase the probability of the adversary to make a wrong decision or inference. Moreover, these countermeasures significantly degrade the accuracy of the model, making them unusable in practice. Finally, these techniques do not have a specific treatment for directly or indirectly identifying data.

In this paper, we propose privacy-by-design language modeling approaches to address the problem of anonymizing language models. Specifically, we propose both a masking language modeling (MLM) methodology to specialize a BERT-like language model, and a causal language modeling (CLM) methodology to specialize a GPT-like model that avoids the model from memorizing direct and indirect identifying information. To achieve this goal, our finetuning approaches named PPMLM-BERT and PPCLM-GPT (for a BERT and GPT model, respectively), first identify directly and indirectly identifying information in the data corpus (e.g., medical reports). While direct identification words are identified by exploiting name entity recognition, indirect identification words are identified as words used only by a single patient. This ensures that the memorized words have been used in the documents of at least two individuals (i.e., thus providing k-anonymity, note that the value of k can be parameterized in our solution). Then, the specialization operation (i.e., the random masking or the next-word prediction for BERT and GPT models, respectively) avoids using directly and indirectly identifying words. Therefore, the model is specialized on the dataset without ever learning identifying words, which then drastically reduces the risk of regurgitation or inference of personal information during its use.

We have comprehensively evaluated our approaches using datasets of english and french medical reports and considering a large number of baselines for comparison. We first illustrate the risk of identifying patients whose reports were used for model specialization, thus motivating our approaches. We then show that ignoring identifier words in training (i.e., direct and indirect identifiers) improves privacy by reducing the ability of models to predict these identifier words while maintaining good ability to predict other non-identifier

words (aka the utility), or the ability of the model to train a downstream classification task.

Our contributions are as follows:

- we illustrate the risk of indirectly identifying data regurgitation (i.e., privacy leakage of training data) of NLP models following its specialization, motivating our anonymization solution for model specialization;
- (2) we provide a privacy-by-design approach for masked and causal language modeling that ensures that direct and indirect identifiers are not memorized during specialization;
- (3) we comprehensively evaluate our approach and compare it to classical fine-tuning of BERT and GPT models on medical corpora, as well as to specialization with de-identified datasets.

The paper is organized as follows. Section 2 reviews background and related work while Section 3 defines the problem statement. Section 4 describes our privacy-preserving language modeling for both descriptive and generative models. Section 5 then presents the experimental setup while the evaluation results are reported in Section 6. Finally, we discuss limitations and future works Section 7 before concluding Section 8.

2 BACKGROUND AND RELATED WORK

This section presents a comprehensive background and related work on NLP (Section 2.1), how NLP models can be used to improve privacy (Section 2.2), the privacy leakages associated to the memorization of the model (Section 2.3), and mitigation strategies (Section 2.4).

2.1 Natural Language Processing

Natural language processing (NLP) is the process of understanding and processing textual data using machine learning (ML) models. The field experienced a breakthrough in 2017 with the advent of the Transformer [49]. This new architecture revolutionized translation at the time. It consists of an encoder-decoder neural network with a parallel computational scheme that uses positional encoding and various attention mechanisms [4]. The goal of the encoder is to embed (i.e., transform into vectors) the input sentences. Each word is embedded in a latent space, taking the entire sentence as context. The decoder, in turn, learns to translate these latent vectors into new sentences. Each token (words or subwords) is predicted sequentially, paying attention to the input and the previously predicted words.

Two trends followed in 2018 with BERT models [11] that focus on the encoder part of the Transformer and GPT models that use the decoder. The former are very effective for classification tasks. They are pre-trained on huge unlabeled datasets to learn very complex word embeddings. It is then possible to fine-tune such models on specific data and even learn new tasks by adding just a few layers to the model. For example, a hospital could train a BERT model to classify medical records according to different pathologies. GPT models on the other hand are generative models. They consist of a very large language model (LLM) that has learned to imitate human expression. The output sentences are the most likely answers according to the model, based on probabilities that it has learned by observing sentences in its huge training dataset. They can be used to create chatbots such as chatGPT, which is derived from a

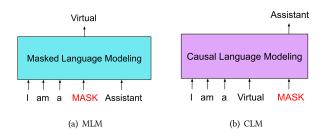


Figure 1: Masked language modeling (MLM) involves predicting words anywhere in the text and in any order, while causal language modeling (CLM) involves predicting words sequentially, from left to right.

GPT-4 base. OpenAI then extended its training with supervised and reinforcement phases to ensure that responses are non-toxic and do not include fake news [1, 16].

Since then, many new models inspired by Transformer have emerged. For example, RoBERTa [26] which is a much larger version of BERT, or distilBERT [36] which uses knowledge distillation to produce a smaller model. Models in other languages have also emerged, such as CamemBERT [27] for French. This model specialization on a language or on a specific data corpus such as ClinicalBERT [32] or DrBERT [23] with clinical reports is based on language modeling. Language modeling consists of predicting the words in a document (Figure 1). On the one hand, Causal language modeling (CLM) consists of predicting words sequentially, from left to right (Figure 1(b)). This is the most common approach used in text generation with a GPT-type model. On the other hand, Masked Language Modeling (MLM) is when the prediction task is done anywhere in the text and in any order, in other words chosen randomly (Figure 1(a)). This is the approach mainly used for language understanding tasks with a BERT-type model. In both cases, the prediction is based on the context. This means that the words before the mask (in the CLM case) or the words around the mask (in the MLM case) allow the model to identify what the word to fill could be. This language modeling step allows a model to be specialized on a specific corpus. For example, it has been shown that specializing on clinical reports improves the model's performance for downstream clinical NLP tasks [13, 21, 23].

2.2 Leveraging NLP for Privacy

The prospects for improving privacy offered by new NLP models are numerous, both on human-generated textual content (e.g., free text, source code) and on speech after transcription from voice to text. For example, [18] uses NLP models to analyze privacy-related feedback or [3] analyzes emergency calls to correctly assess and predict the severity of the situation. These models could also be used to analyze source code and detect security or personal data leaks [30, 41], or to analyze speech and detect sensitive elements in order to remove them from recordings [31].

In order to exploit or share their patients' information for research purposes while ensuring patient confidentiality, hospitals

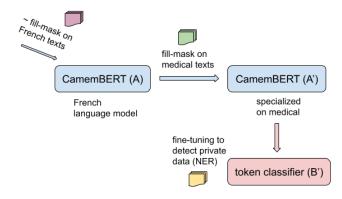


Figure 2: Workflow used by the Hospices Civils de Lyon (HCL) for de-identifying clinical reports.

have begun to exploit language models to de-identify clinical reports [19, 35, 45, 46, 53]. De-identification of text-based clinical reports involves removing or replacing Personally Identifiable Information (PII) from electronic health records. Although strict anonymization is considered a very difficult task, this pseudonymization (e.g., identifying and replacing PII with a plausible substitute) makes it difficult to re-establish a link between the patient and their data and is considered sufficient protection for research purposes. Personally identifiable information is extracted from a list that may vary from country to country and includes, for example, address, date, date of birth, hospital, patient ID, email, visit ID, first name, last name, telephone, city, zip code and social security number.

The common pipeline of NLP models to perform de-identification of clinical reports is similar to the one reported in Figure 2 which describes the workflow adopted by Hospices Civils de Lyon (HCL) [35]. The base model (A) is a CamemBERT [27] which is a BERT [11] model specialized on French texts. This specialization of the model towards French is done via a filling-masking task (i.e. random holes are added to the sentences in which the model learns to fill them). Following the same specialization process via a filling-masking task, this CamemBERT model is then refined on medical texts to have a better statistical understanding of the language used in medical reports. Finally, model A' is refined to B' to detect personal identification tokens, through the training of a NER using a manually labeled dataset.

The literature shows that the resulting models are able to outperform de-identification using only regular expressions and manual rules to remove private tokens (i.e., up to 0.99 F1-Score with NLP models versus 0.85 with manual rules only [45]). While these models can represent a very effective clinical text de-identification tool, only a few of these studies evaluate the utility loss due to de-identification and none of them evaluate the privacy leakage of NLP models due to memorization of training data [24]. This latter attack surface is detailed in the following subsection.

2.3 Privacy Leakages

Large language models (LLMs) are trained on very large datasets. For example, training chatGPT required years of crawling the internet. Therefore, a lot of personal data such as people's addresses was

used during training. BERT models, on the other hand, are typically fine-tuned for specific tasks with domain-oriented data. In the medical domain, datasets typically include sensitive patient records. In both cases, the problem is that the models can regurgitate and leak information from the training data after deployment [6], which is a significant privacy leak and has implications for the practical and legal aspects of such models [10].

A central question in this context concerns the extent to which language models memorize their training data [6, 7, 29, 38, 42, 55]. However, defining memorization for language models is challenging, and many existing definitions and notions have been proposed depending on whether the memorization concerns copyrighted content or personal and sensitive content. In relation to privacy, we can notably cite extractable memorization (Section 2.3.2) and counterfactual memorization (Section 2.3.1). This memorization has a direct impact on privacy risks. For example, we illustrate Section 6.1 the exploitation of this memorization to perform a membership inference (Section 2.3.3).

2.3.1 Counterfactual memorization. ML models are supposed to learn general information. Rare data, on the other hand, is not supposed to be memorized. A model that memorizes rare data (sometimes called outliers) not only has a negative impact on utility but also on privacy. Indeed, the more we learn about a small subset of individuals, the greater the information leakage because we can more easily trace that information back to its source. For example, we expect chatGPT to know Harry Potter's address (which can be found on many pages online) but not the reader's address (which should be nonexistent or at least difficult to find online).

It turns out that it is possible to measure this unwanted memorization, called *counterfactual memorization* in [55]. To do this on any data, you need to compare the performance of a model trained on a dataset with that data, with a second model trained without it. This is computationally expensive for each data set, so counterfactual memorization is actually calculated with an empirical expectation: we create multiple copies of the dataset and train many models on different subsets. Each data set will have models it was trained on and models it was not trained on. We can then calculate the expected memorization:

$$mem(x) = E_{x \in D}(score(M_D, x)) - E_{x \notin D'}(score(M_{D'}, x)),$$

where $\mathrm{score}(M_D,x)$ is the score for x of the model trained with the dataset D. The two terms will cancel for common data (removing them has no impact) but can give a high difference for rare data. Data points with memorability above a certain threshold will be considered at risk.

Counterfactual memorization aims to separate memorization from generalization and requires training the model multiple times, which is a limitation in practice given the cost of training a model.

2.3.2 **Extractable memorization**. Extractable memorization is a type of attack that aims to use the model to infer information from the original data [25]. This attack mainly concerns text generation models, such as GPT. These models are trained to produce text based on what they have seen during training. However, the model

is not expected to be a basic parrot and repeat exactly the sentences it has seen. This is especially concerning if the data it is repeating is sensitive. This has been shown to be the case with GPT-2 for example, from which the names and addresses of individuals can be extracted [8].

In [6], the term k-extractability is used to refer to the sequences that can be extracted from the model when an input sequence of length k is requested. The lower the k, the easier it is to extract the sequence. We therefore expect a model to have the highest possible k on private queries. This measure, however, does not capture regurgitations that are not perfect, which can lead to an illusion of no extractable memory. Compressible memorization [38] extends this definition by evaluating how short the minimal requested sentence (or prompt) that elicits the sequence.

2.3.3 **Membership inference**. Membership inference attack [5] (MIA) is a more common inference attack in machine learning, which aims to infer whether a specific data was used in the training data of a target model. This can be a problem, for example, if a hospital has trained a model to detect cancer and you learn that your colleague's data was used in the training. You will have indirectly learned that she probably has cancer.

There are different techniques that can be used to perform a MIA attack. One of them is to use shadow models [39]. Shadow models are trained to mimic the behavior of the target model on an auxiliary dataset with a similar distribution to the original. An adversarial model (i.e., a classifier) is then trained to infer membership from these shadow models.

Although membership inference attacks have been used to quantify memorization risks [22, 28, 50], it may not be applicable in a practical scenario due to the high dimensional input space. By ignoring and not memorizing the direct and indirect identifiers of the training data during model specialization, our privacy-by-design scheme avoids membership inferences targeted on identifiers.

2.4 Mitigation strategies

The most popular approach to mitigating privacy risks is Differential Privacy [14] (DP). DP is a mathematical property that a model must satisfy in order to disclose as little information as possible. This property requires the model to learn a limited amount of information at each training step. More formally, the probability that a model guesses the correct output for a given input must not increase too much each time the model sees that data:

$$\forall (x, y), \log P(M_D(x) = y) < \epsilon \log P(M_{D+x}(x) = y),$$

where x,y represent data and its label, M_D a model trained on dataset D and ϵ the *privacy budget*. The lower ϵ is, the more private the model is.

The most popular method to apply DP in machine learning is DP-SGD: Differentially-Private Stochastic Gradient Descent [2]. The idea is to apply DP during the training phase by clipping gradient updates and adding centered noise at each step. DP is known to significantly decrease the accuracy of the model [22] and privacy budget management is difficult.

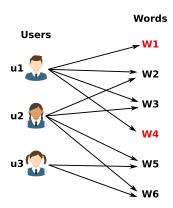


Figure 3: By constructing a bipartite graph between individuals and the words used in their documents, we can easily identify indirectly identifying words (i.e., words pointed to only by one individual).

3 PROBLEM STATEMENT

As the training data contained personal information, our goal is to avoid that the model memorizes any information that can be used to identify individuals which information was used during model training. This information concerns both directly and indirectly identifying data. Directly identifiable data includes information associated with a natural person such as first and last name, a telephone or license plate number, an identifier such as a social security number, a postal or email address. Indirectly identifiers, in turn, is information that is uniquely used in an individual's data compared to the rest of the population. For example, if the term vertebra is used only in a patient's data, this term is indirectly identifying compared to the rest of the corpus. In other words, the individual is distinguishable from others with the use of this term. Figure 3 depicts a bipartite graph which relates the people represented in the data corpus to the words used across the entire dataset. More formally, it is a Bipartite Graph G = (V, E) where the vertices of this graph are the users and words of the dataset $V = (U \cup I)$. The use of a word in a user's data is represented in G by an edge $(u, i) \in UI$. In this example, we can see that the words w1 and w4 are only used in the data of individual u1. Note that a directly identifying word can also be a word used only in an individual's data, this is the typical case of a last name or a particular identifier.

These directly and indirectly identifying data can pose a problem if they are memorized by the model following training. Indeed, an identifying word memorized by the model which would be then proposed in the test phase would betray its use during training. As depicted Figure 4, the model can be made to memorize the context information and the information that was used as a mask in the supervised training phase¹. In this example, the fact that the model proposes the word w4 in the testing phase, mean s that an adversary could deduce that the data of user u1 was used during training because this word only appears in the data of this individual.

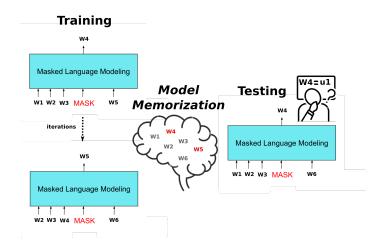


Figure 4: During its specialization, the model will mainly memorize the words that were used for language modeling masking, this memorization can be used by an adversary to infer membership.

4 PRIVACY-PRESERVING LANGUAGE MODELING

In order to prevent the model from memorizing personal information, the privacy preserving language modeling that we propose ensures that the choice of mask used during specialization avoids directly and indirectly identifying words. To achieve that, this solution is based on a preprocessing that is responsible for identifying both the direct and indirect identifying words. We first present this preprocessing step (Section 4.1) before explaining how the masked and the causal language modeling work (Section 4.2 and Section 4.3, respectively).

4.1 Preprocessing: Building a Blacklist

The preprocessing step focuses on identifying identifiers that the model must not memorize. First, we use an off-the-shelf NER to directly identify identifying words associated with certain categories. This NER can be adapted to include more or less categories depending on the field of specialization. Then, we build the bipartite graph G between the individuals of the corpus and the words used in their documents in order to identify indirectly identifying words (similar to Figure 3). To do this, once the graph is built, we identify the words that are pointed to only by one individual. Finally, all these (directly or indirectly) identifying terms are kept in a kind of blacklist that is used when modeling the masked or causal language.

If we make the link with the notion of k-anonymity[44], which defines that a data is anonymous if it is used by at least k different users, all the words used as masks to specialize the model in the language modeling step are at least used by two people (i.e., k=2). This makes the task of an adversary wishing to infer the membership of an individual more difficult. We can easily change the reading of the bipartite graph to consider a larger value of k for the identification of indirectly identifying information (e.g., identifying words that are only pointed to by 1 or 2 users to have a k=3).

¹We discuss the memorization related to the context in Section 7 and show that its privacy impact is very limited compared to memorization of words used in masks.

In addition to being able to vary the value of k to protect indirectly identifying words and to be able to use a NER adapted to the domain of specialization to identify directly identifying words, it should be noted that the cost of this preprocessing is reduced.

4.2 Privacy-Preserving Masked Language Modeling

Compared to standard masked language modeling that randomly chooses terms to mask during supervised learning, our privacy-preserving masked language modeling ensures that directly or indirectly identifying terms are never masked by leveraging the blacklist constructed during the preprocessing step. More precisely, the choice of the mask is made randomly among the words that are not in the blacklist.

However, the identifying words (i.e., the words kept in the black-list) can be used in the context of a mask to be predicted. Their use improves the prediction of masked terms without introducing a significant risk of memorization leading to regurgitation. We evaluate and compare the impact on the utility and privacy trade-off of memorizing identifying words as a mask or as a context in Section 6.2.

4.3 Privacy-Preserving Causal Language Modeling

Similar to privacy-preserving masked language modeling, our privacy preserving causal language modeling ensures that masked words are not directly or indirectly identifying words. However, since masked words are always at the end of the sequence to be predicted, we exploit the blacklist created during the preprocessing step to check that this last term is not an identifier. If it is, we modify it and replace it with a padding token and add the next word for prediction. The advantage of using a padding token is that it is not used in the loss calculation during training, and therefore does not degrade the model's predictions.

Once the last masked word is predicted, we exploit the real sequence to mask the next word. This means that an identifier word can be part of the context of a masked word to be predicted. As in the case of masked language modeling, the exploitation of an identifying word in the context does not introduce a notable risk of regurgitation but improves the prediction (Section 6.3).

5 EXPERIMENTAL SETUP

We performed an extensive evaluation of our privacy-preserving language modeling scheme using medical datasets (Section 5.1) and considering both BERT and GPT based language models (Section 5.2). To capture the full impact of our solutions, we considered a set of both utility and privacy metrics (Section 5.3) and considered different baselines approaches (Section 5.4). Finally, we describe our methodology (Section 5.5).

5.1 Datasets

To conduct our experiments, we considered medical field datasets, specifically the N2C2 (National NLP Clinical Challenges) datasets [20] and a dataset from the Hospices Civils de Lyon (HCL) [35].

- N2c2: we leverage two datasets gathering medical discharge summaries (i.e., english free text). In the first one [48], 928 records have been annotated for replacing all authentic Personally Identifiable Information (PII) with realistic surrogates. These PIIs fall into different categories: six of the seventeen textual PII categories listed by HIPAA (only these six categories appear in the data: patients, locations, dates, IDs, phone numbers, and ages), and two additional categories, doctors and hospitals, resulting in eight PII categories in the dataset. In the second dataset [47] almost 500 records have been annotated by pulmonologists. The pulmonologists were asked to classify patient records into five possible smoking status categories as past smoker, current smoker, smoker, non-smoker, and unknown. We leverage this dataset to assess the utility degradation of our privacy-preserving language modeling through a downstream classification task.
- Hcl: this dataset gathers french free text from medical reports, notes, prescriptions, etc. This dataset contains texts from 1240 patients hand-annotated by ten people, with double verification to identify PII falling into the following categories: Name, First name, Dates, Organizations, Telephones, Emails, Cities, Postal Codes, Streets, File Number, Website, Localities and IPP.

5.2 Model Architecture

In order to cover the different use cases of language models (i.e., the generative models and the descriptive text models), we consider both the BERT architecture (i.e., the base model case composed of 12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters) and the GPT-2 architecture [34] (the smallest version of GPT-2, with 124M parameters) from Hugging Face². To support the French dataset, we also considered a CamemBERT (i.e., a BERT model specialized for French) and a GPT model specialized for French. We did not consider larger model versions due to resource limitations, and we also discarded distilled version of these models [37].

5.3 Metrics

We consider both utility and privacy measures in our evaluation to better assess the trade-off between protecting the training data and the model's performance in specializing on a specific corpus. In particular, for the utility evaluation, we quantify the ability of the models to predict masked terms well. For the utility evaluation of BERT-like models, we also quantify the model's performance in classification by creating a downstream task. For GPT-like models, the utility is evaluated by the model's ability to generate the right tokens in the right places. To evaluate the privacy of the models, we consider their ability to predict identifiers (i.e., direct and indirect) contained in the dataset used to specialize the model. Since a privacy leakage is marked when an identifier is predicted by the model regardless of where this term is actually used, we consider as a leak whenever an identifying term is predicted in the right place in the original text or in another place. We refer to the metric named Privacy when measuring the ability of a model to predict any identifying term (i.e., direct or indirect). The metrics named

 $^{^2} Hugging\ Face\ -\ https://huggingface.co/$

Direct Privacy and Indirect Privacy measure, respectively, the ability of the model to predict only directly or indirectly identifying terms.

5.4 Comparative Baselines

We compare our privacy-preserving language modeling scheme to different comparative baselines for both Masked Language Modeling (e.g., BERT-like models) and Causal Language Modeling (e.g., GPT-like models). For the former language modeling, our solution is named PPMLM-BERT, and for the latter our solution is named PPCLM-GPT.

- Masked Language Modeling (MLM): First, we consider
 a classical specialization of a BERT model (named MLMBERT). Next, we consider a baseline that performs the same
 specialization but using a pseudonymized dataset for the
 specialization (baseline named MLMA-BERT). Finally, we
 consider two variants of our solution where only direct
 or indirect identifiers are protected from memorization
 (baselines named DPPMLM-BERT and IPPMLM-BERT, respectively).
- Causal Language Modeling (CLM): We use identical baselines to evaluate our solution. Specifically, we consider a classical specialization of a GPT model (named CLM-GPT) that consists in iteratively predicting each term of a sequence of the dataset used for the specialization. We also consider a specialization using a pseudonymized dataset (baseline named CLMA-GPT), and two variants of our solution that protect only directly or indirectly identifying terms (baselines named DPPCLM-GPT and IPPCLM-GPTERT, respectively).

5.5 Methodology

This section provides the methodological details we followed to perform our evaluations. First, when datasets are pseudonymized before being used for model finetuning (i.e., in the case of MLMA-BERT and CLMA-GPT), directly identifying words are replaced by the term "X" in the text. Then, BERT-like models were trained by randomly replacing a subset of tokens from the sequence by the <mask> token, and asking the model to predict them using cross-entropy loss. In our setting, 15% of the words are randomly selected. In the case of our solution, identifying words are excluded from this random selection. If a word chosen to be masked consists of several tokens, all of them are masked. At each epoch, 15% of non-identifying words are masked.

Finally, GPT-like models were trained by sequentially masking the last word in a sequence. In our solution, if this word is a direct or indirect identifier, it is replaced by a specific padding token that is not taken into account during fine-tuning (i.e, for the computation of the loss). Unlike CLMA-GPT (the baseline that finetunes a GPT model with a pseudonymized dataset), padding tokens are not used in the computation of the loss compared to the term "X" (i.e., the anonymized equivalent of the identifiers) which is taken into account in the loss.

All measurements were performed after 4, 8, 16, 32 and 64 epochs to observe their evolution. We used 16 batches of 512 tokens for training. For language modeling, the learning rate starts at 1e-4 and decreases linearly until the end of finetuning. For classification, we

took the finetuned model after 64 epochs and finetuned it again by freezing all the BERT layers (to not bias the model and introduce other leakage) except the last one and the classification layer following the methodology described in this study [43]. The learning rate this time starts at 0, 10% of the total number of warmup steps with the learning rate increasing until reaching 2e-5 then decreasing linearly on the remaining 90% steps.

6 EVALUATION

This section presents a comprehensive evaluation of our privacy-preserving language modeling scheme. We first quantify and illustrate the risk of membership inference through the memorization of indirect identifiers (Section 6.1) before evaluating both the utility and privacy trade-off of masked and causal language modeling (Section 6.2 and Section 6.3, respectively). We show that memorization of identifiers can lead to the re-identification of individuals whose data were used in training. Furthermore, we show that protecting the direct identifier only by pseudonymization is not sufficient to protect the model against a risk of memorizing indirect identifiers leading to membership inference, and that it is necessary to protect indirect identifiers as well.

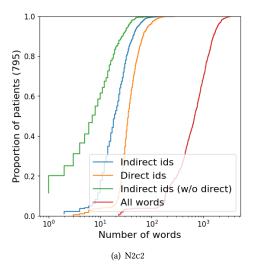
6.1 Illustrating the risk

In this section, we illustrate the risk of membership inference related to the model's memorization of personal information. More specifically, we analyze the impact of indirectly identifying information in this membership capability. First, we analyze the N2c2 and the Hcl datasets, and analyze the distribution of direct and indirect identifiers, as well as the number of words in each patient's reports. Figure 5 reports the Cumulative Distribution Function (CDF) of these information. First, we can see that almost all patients have indirect identifiers and half of them have more than twenty indirect identifiers for the N2c2 dataset, and more than ten indirect identifiers are about twice as numerous as indirect identifiers. Finally, some indirect identifiers (i.e., words exploited in the documents of a single patient) are also direct identifiers.

Figure 6, in turn, shows for both datasets the cumulative distribution of the number of patients who share words in their reports. The plot distinguishes distinct words from the total number of words by including repetitions. The distribution shows for both datasets that 60% of distinct words in the corpus are used by only one patient and therefore represent indirect identifiers. However, the use of these words represents less than 3% of the total occurrences of words present in all reports. In other words, the 40% of words used in reports of multiple patients represents more than 97% of the total volume of words used.

We now characterize the indirectly identifying words by classifying them into different categories using biomedical named entity recognition (NER) that also includes classes of diseases, chemicals, and genetic entities³. Figure 7 represents for the N2c2 dataset the distribution of indirectly identifying words in these medical-related categories. The distribution shows that indirectly identifying words are classified into a large number of categories where chemicals are the most represented in the N2c2 dataset.

³https://github.com/librairy/bio-ner



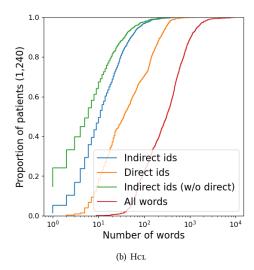
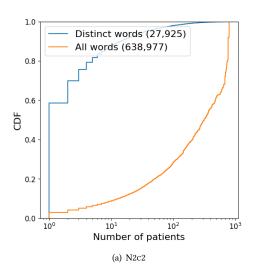


Figure 5: Cumulative distribution of the number of identifiers (both direct and indirect ones), and words per patient: for N2c2 dataset, half of the patients have more than 20 indirect identifiers, and almost all patients have at least 3 indirect identifiers.



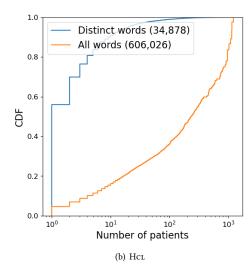


Figure 6: Cumulative distribution of the number of patients who share words in their reports: 60% of distinct words in the corpus are used by only one patient and therefore represent indirect identifiers.

To assess an attacker's ability to identify that an individual's data has been used in the specialization of a model from indirect identifier words, we only consider certain categories as plausible auxiliary information. Indeed, some information is less likely to be known by an attacker. Here, we only consider words in the Organ and Cancer categories as words that an adversary could know a priori about the person she is trying to attack. Figure 8 reports the proportion of patients that can be identified from indirectly identifying words in the dataset that fall into the Organ and Cancer categories. Results show that around 15% of the patients can be identified from words in the Cancer category, and 6% from words in the organ category. This identification rate is an upper bound

of the risk in the case where the attacker manages to identify that an indirectly identifying word that is part of these two categories of plausible auxiliary information has been learned by a model. In practice, all the words in these two categories may not be known to an attacker. However, thanks to our privacy-preserving language modeling scheme avoiding that an indirectly identifying word cannot be predicted by the model after its specialization, this risk of membership inference is drastically mitigated.

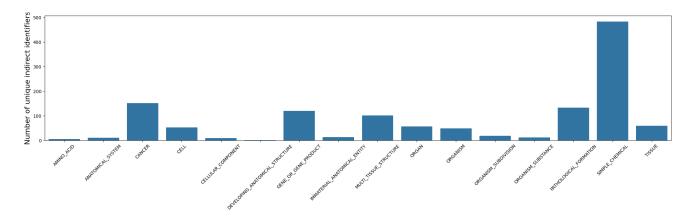


Figure 7: Characterization of indirectly identifying words of N2c2 dataset by classifying them into different categories using biomedical named entity recognition (NER): these indirectly identifying terms are present in many categories.

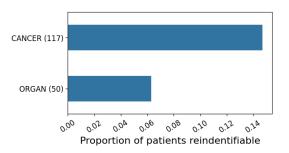
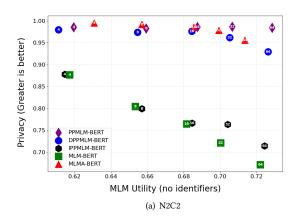


Figure 8: Membership inference capability on patients in the N2c2 dataset by an adversary who has auxiliary information on only two categories (e.g., information on the organs and cancers of its targets).

6.2 Utility and Privacy Tradeoff for Masked Language Modeling

We now evaluate the utility and privacy tradeoff for our masked language modeling scheme. Figure 9 illustrates this tradeoff for PPMLM-BERT and other comparative baselines for the N2c2 and HCL datasets. First, the results show that without a specific mechanism to prevent the model from memorizing identifiers (i.e., classical MLM-BERT), privacy is significantly degraded, from 0.87 after 4 epochs to 0.66 after 64 epochs. This decrease in privacy means that the model learns more and more identifiers over the epochs and is led to predict them more along the training. Second, training without memorizing indirect identifiers (i.e., IPPMLM-BERT) provides about the same level of utility while increasing privacy. This small increase in privacy is due to the reduced number of indirect identifiers. Training without memorizing the direct identifiers, in turn, significantly improves privacy. Indeed, since the number of direct identifiers is significantly larger than the number of indirect identifiers, the increase in privacy is significantly more pronounced. The utility, on the other hand, is similar. Then, training using a pseudonymized dataset (i.e., leveraging a training dataset without direct identifiers, MLMA-BERT) provides a similar level of privacy



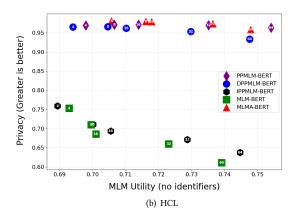
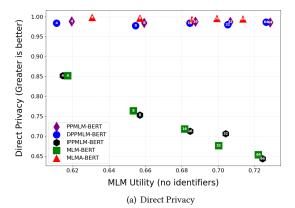


Figure 9: Utility and privacy evaluation for MLM: by avoiding memorizing both direct and indirect identifiers during model specialization, our PPMLM-BERT solution offers the best tradeoff in maintaining high privacy while retaining high utility.



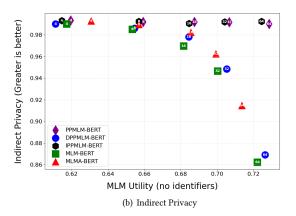


Figure 10: Evaluating the utility and privacy tradeoff for MLM by biasing the privacy metric toward direct identifiers or indirect identifiers for the N2c2 dataset.

but reduces the utility. This reduction in utility comes from a context used for prediction that does not benefit from direct identifying terms. Finally, our PPMLM-BERT solution offers the best utility and privacy tradeoff where the privacy remains stable around 0.98 regardless of the number of epochs, meaning that the model does not learn any identifier directly or indirectly during training. Regarding utility, PPMLM-BERT also offers the best utility. This high utility is due to the wealth of information (i.e., using all terms including identifiers) in the context used to predict a term.

Let us then analyze the behavior of PPMLM-BERT and the other baselines if we consider only the prediction of direct identifiers or indirect identifiers as privacy leakage (Figure 10 for the N2c2 dataset). In the former case (Figure 10(a)), results show that MLM-BERT and IPPMLM-BERT exhibit a close utility and privacy tradeoff. Indeed, since indirect identifiers are not taken into account by the privacy measure, our mechanism for not memorizing indirect identifiers has no observable effect. Consequently, only the solutions that protect direct identifiers exhibit high privacy (i.e., PPMLM-BERT,

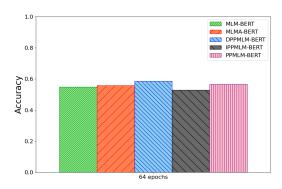


Figure 11: Utility assessment for MLM through a downstream classification task on the N2c2 dataset: all baselines provide roughly the same level of prediction.

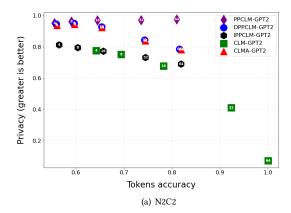
MLMA-BERT, and DPPMLM-BERT). However, we observe that MLMA-BERT exhibits the lowest utility. This lower utility comes from the fact that all direct identifiers are pseudonymized in the training data and are not available in the context to improve the model's prediction. Once again, PPMLM-BERT offers the best trade-off between utility and confidentiality because it benefits from direct identifiers in the context to improve predictions. By comparing the trade-off between MLMA-BERT and DPPMLM-BERT, we can evaluate the impact of memorization related to the use of directly identifying words in the context for mask prediction. We can see that the privacy impact is very small, and that using a context that exploits direct identifiers slightly improves the utility.

Similarly, if the privacy measure only captures the prediction of indirect identifiers as a leakage (Figure 10(b)), only solutions that carefully protect this category of identifiers exhibit high privacy, this is the case for PPMLM-BERT and IPPMLM-BERT. The privacy degradation as a function of the number of epochs for the other baselines highlights the fact that they progressively learn indirect identifiers and predict them more and more often during training.

We now evaluate the utility of the models through a classification task on the N2c2 dataset. As described in Section 5.3, this classification task aims to predict the patient's smoking, and the classification models are trained as a downstream task from the specialized model on the dataset. Figure 11 illustrates the prediction accuracy of a classification model after 64 epochs. Counterintuitively, the results show that models that protect identifiers display slightly better accuracy. These results tend to show that identifiers (i.e., information that only affects a single individual) reduce the model's ability to generalize knowledge and classify correctly.

6.3 Utility and Privacy Tradeoff for Causal Language Modeling

This section evaluates the utility and privacy trade-off of our PPCLM-GPT compared to other comparative baselines. Figure 12 presents for the N2c2 and the HCL datasets the utility through the prism of token accuracy as a function of privacy assessed by the model's ability to predict both direct and indirect identifiers during generation. The



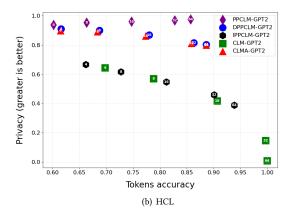
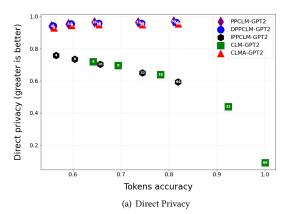


Figure 12: Utility and privacy evaluation for CLM: by avoiding memorizing both direct and indirect identifiers during model specialization, our PPCLM-GPT solution offers the best tradeoff in maintaining high privacy while retaining high utility in the generation.

results show similar observations to those of PPMLM-BERT but with much better privacy for PPCLM-GPT compared to other baselines (up to 20% more privacy). Without any protection (i.e., CLM-GPT which corresponds to a classic fine-tuning), the model learns the identifiers during training. After 64 epochs, this model is even able to regurgitate the direct and indirect identifiers it saw during training (revealed by a token accuracy close to 1). IPPCLM-GPTERT (i.e., protecting only indirect identifiers) provides a similar level of privacy to classic fine-tuning because indirect identifiers are small in number. However, the utility is limited to 80% for the N2c2 dataset for instance, which means that the model does not regurgitate indirect identifiers because they were not learned during training. DPPCLM-GPT and CLMA-GPT, in turn, provide similar results with privacy decreasing to 80% privacy after 64 epochs. This decrease means that the model learns indirect identifiers over the epoch and tends to regurgitate them more frequently.



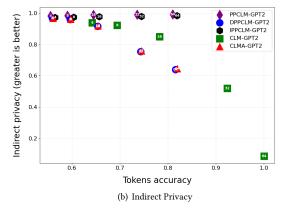


Figure 13: Evaluating the utility and privacy tradeoff for CLM by biasing the privacy metric toward direct identifiers or indirect identifiers for the N2c2 dataset.

Figure 13 illustrates the tradeoff between utility and privacy when privacy is measured solely through quantifying the leakage of direct (Figure 13(a)) or indirect (Figure 13(b)) identifiers. The results are similar to the observations made with masked language modeling. More precisely, when privacy only takes into account direct identifiers (i.e. this corresponds to pseudonymization), it is the baselines that protect direct identifiers that provide the best compromise. Conversely, if privacy only takes into account indirect identifiers, it is the baselines that protect indirect identifiers that provide the best compromise. This clearly shows that pseudonymization is not sufficient to protect a model against identifiable memorization, and that protection of both direct and indirect identifiers like done by our solution is necessary. Finally, the use of identifying words in the context for prediction has almost no impact on the utility and privacy tradeoff (comparison between DPPCLM-GPT and CLMA-GPT).

7 LIMITATIONS AND DISCUSSIONS

The work presented in this paper represents a practical step towards protecting the training data against privacy leakage through the LLMs. There are, however, some limitations in several areas. Although the study and results have been validated on two datasets, these datasets are small in size. A validation on a larger dataset requires more computational resources and is left as future work.

Another limitation is that the identification of direct identifiers of our privacy-preserving scheme relies on a NER, and the quality of this NER (Named Entity Recognition) can impact the performance of the protection of our solution. This limitation is common to all pseudonymizations relying on a NER and affects both the protection itself by missing identifying words, and the pseudonymization measure by not counting certain leaks of personal information. However, our solution is not impacted by misspelled and/or foreign words. Indeed, a common problem of NER in free text is the management of misspelled and/or foreign words that cannot be found in dictionaries. These words can be identifiers and not be protected. It is worth noting that the advantage of our solution is that these words will be identified as indirect identifiers in our solution and will not be memorized.

The preprocessing step of our solution to identify indirectly identifying words is somewhat like implementing k-anonymity with k=2. More precisely, this means that a term used by only one individual will be protected (i.e., not memorized by the model). That is, the words are at worst used in the reports of two patients and it is not possible for an attacker to easily identify either patient. We can easily change this value to a larger value. For example, by taking k=3, we ensure that the words that will be predicted by the model are at worst used in the reports of three patients, making the attacker's task even more difficult. This configurable protection (i.e., the value of k) does not include any additional cost and only consists of navigating the bipartite graph built in the preprocessing step of our solution.

8 CONCLUSION

This article presents a methodology to avoid a model having memorized directly or indirectly identifying information linked to the specialization corpus used to specialize it. This absence of identifier memorization allows it to be shared without risk that the model can regurgitate re-identifying information during its exploitation. Our privacy preserving methodology of model specialization takes into account masked language modeling (i.e., for BERT-type models) and causal language modeling (i.e., for GPT-type models). We exhaustively experiment our approach and show that it offers the best compromise between utility and privacy.

REFERENCES

- [1] 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/ 2303.08774
- [2] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. (oct 2016). https://doi.org/10.1145/2976749.2978318
- [3] Marianne Abi Kanaan, Jean-François Couchot, Christophe Guyeux, David Laiymani, Talar Atechian, and Rony Darazi. 2023. A methodology for emergency calls severity prediction: from pre-processing to BERT-based classifiers. In IFIP Advances in Information and Communication Technology, Vol. 675. Leon, Spain. https://doi.org/10.1007/978-3-031-34111-3_28
- [4] J Alammar. (2018). The Illustrated Transformer. Retrievedfromhttps://jalammar. github.io/illustrated-transformer/
- [5] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In 2022 IEEE Symposium on Security and Privacy (SP). IEEE, 1897–1914. https://doi.org/10.1109/SP46214.2022.9833649

- [6] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. arXiv:2202.07646 [cs.LG] https://arxiv.org/abs/2202.07646
- [7] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. arXiv:1802.08232 [cs.LG] https://arxiv.org/abs/1802.08232
- [8] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Ülfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting Training Data from Large Language Models. CoRR abs/2012.07805 (2020). arXiv:2012.07805 https://arxiv.org/abs/2012.07805
- [9] Xianlai Chen, Jiamiao Lin, and Ying An. 2022. DL-BERT: a time-aware double-level BERT-style model with pre-training for disease prediction. In 2022 IEEE International Conference on Big Data (Big Data). 1801–1808. https://doi.org/10.1109/BigData55660.2022.10020513
- [10] A. Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A. Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, Jack M. Balkin, Nicholas Carlini, Christopher De Sa, Jonathan Frankle, Deep Ganguli, Bryant Gipson, Andres Guadamuz, Swee Leng Harris, Abigail Z. Jacobs, Elizabeth Joh, Gautam Kamath, Mark Lemley, Cass Matthews, Christine McLeavey, Corynne McSherry, Milad Nasr, Paul Ohm, Adam Roberts, Tom Rubin, Pamela Samuelson, Ludwig Schubert, Kristen Vaccaro, Luis Villa, Felix Wu, and Elana Zeide. 2023. Report of the 1st Workshop on Generative AI and Law. arXiv:2311.06477 [cs.CY] https://arxiv.org/abs/2311.06477
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805
- [12] Henri Duprieu and Nicolas Berkouk. 2024. Techniques d'audit des grands modèles de langage. Technical Report. Commission Nationale Informatique et Libertés (CNIL). https://hal.science/hal-04782667
- [13] Basile Dura, Charline Jean, Xavier Tannier, Alice Calliger, Romain Bey, Antoine Neuraz, and Rémi Flicoteaux. 2022. Learning structures of the French clinical language:development and validation of word embedding models using 21 million clinical reports from electronic health records. arXiv:2207.12940 [cs.CL] https://arxiv.org/abs/2207.12940
- [14] Cynthia Dwork. 2006. Differential privacy. In International colloquium on automata, languages, and programming. Springer, 1–12.
- [15] Johnson et al. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. (2020). https://pubmed.ncbi.nlm.nih.gov/34350426/
- [16] Farshid Faal. 2022. Reinforcement Learning for Mitigating Toxicity in Neural Dialogue Systems. Ph. D. Dissertation. Concordia University.
- [17] Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. 2024. Healai: A healthcare Ilm for effective medical documentation. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining. 1167–1168. https://doi.org/10.1145/ 3616855.3635739
- [18] Hamza Harkous, Sai Teja Peddinti, Rishabh Khandelwal, Animesh Srivastava, and Nina Taft. 2022. Hark: A Deep Learning System for Navigating Privacy Feedback at Scale. In 2022 IEEE Symposium on Security and Privacy (SP). 2469– 2486. https://doi.org/10.1109/SP46214.2022.9833729
- [19] Tzvika Hartman, Michael D Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg Corrado, Katherine Chou, et al. 2020. Customization scenarios for de-identification of clinical notes. BMC medical informatics and decision making 20, 1 (2020), 1–9. https://doi.org/10. 1186/s12911-020-1026-2
- [20] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association* 27, 1 (2020), 3–12. https://doi.org/10.1093/jamia/ocz166
- [21] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv:1904.05342 [cs.CL] https://arxiv.org/abs/1904.05342
- [22] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership Inference Attack Susceptibility of Clinical Language Models. arXiv:2104.08305 [cs.CL] https://arxiv.org/abs/2104.08305
- [23] Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. arXiv:2304.00958 [cs.CL] https://arxiv.org/abs/2304.00958
- [24] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, 946–959. https://doi.org/10.18653/v1/2021.naacl-main.73
- [25] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, Jinbo Xiong, Zuobin Ying, and Athanasios V. Vasilakos. 2021. Privacy and Security Issues in Deep Learning: A

- Survey. IEEE Access 9 (2021), 4566–4593. https://doi.org/10.1109/ACCESS.2020. 3045078
- [26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692
- [27] Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. CoRR abs/1911.03894 (2019). arXiv:1911.03894 http://arxiv.org/abs/1911.03894
- [28] Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. arXiv:2203.03929 [cs.LG] https://arxiv.org/abs/2203.03929
- [29] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable Extraction of Training Data from (Production) Language Models. arXiv:2311.17035 [cs.LG] https://arxiv.org/abs/ 2311.17035
- [30] Ahmet Okutan, Ali Shokri, Viktoria Koscinski, Mohamad Fazelinia, and Mehdi Mirakhorli. 2023. A Novel Approach to Identify Security Controls in Source Code. arXiv:2307.05605 [cs.SE] https://arxiv.org/abs/2307.05605
- [31] Michele Panariello, Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Pierre Champion, Hubert Nourtel, Massimiliano Todisco, Nicholas Evans, Emmanuel Vincent, and Junichi Yamagishi. 2024. The VoicePrivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024), 3477–3491. https://doi.org/10.1109/TASLP. 2024.3430530
- [32] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv:1906.05474 [cs.CL] https://arxiv.org/abs/1906.05474
- [33] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. arXiv:2202.00443 [cs.CL] https://arxiv.org/abs/2202.00443
- [34] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. https://api.semanticscholar.org/CorpusID:160025533
- [35] Antoine Richard, François Talbot, and David Gimbert. 2023. Anonymisation de documents médicaux en texte libre et en français via réseaux de neurones. In Plateforme Intelligence Artificielle 2023 (PFIA2023) - Journée Santé & IA. Association française pour l'Intelligence Artificielle (AfIA) and Université de Strasbourg and Association française d'Informatique Médicale (AIM), Starsbourg, France. https://hal.science/hal-04139391
- [36] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL] https://arxiv.org/abs/1910.01108
- [37] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL] https://arxiv.org/abs/1910.01108
- [38] Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. 2024. Rethinking LLM Memorization through the Lens of Adversarial Compression. arXiv:2404.15146 [cs.LG] https://arxiv.org/abs/2404.15146
- [39] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP). IEEE, 3–18. https://doi.org/10.1109/SP. 2017.41
- [40] Sushant Singh and Ausif Mahmood. 2021. The NLP Cookbook: Modern Recipes for Transformer based Deep Learning Architectures. CoRR abs/2104.10640 (2021). arXiv:2104.10640 https://arxiv.org/abs/2104.10640
- [41] Tanmay Singla, Dharun Anandayuvaraj, Kelechi G. Kalu, Taylor R. Schorlemmer, and James C. Davis. 2023. An Empirical Study on Using Large Language Models to Analyze Software Supply Chain Security Failures. In Proceedings of the 2023 Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses (Copenhagen, Denmark) (SCORED '23). Association for Computing Machinery,

- New York, NY, USA, 5-15. https://doi.org/10.1145/3605770.3625214
- [42] Till Speicher, Mohammad Aflah Khan, Qinyuan Wu, Vedant Nanda, Soumi Das, Bishwamittra Ghosh, Krishna P. Gummadi, and Evimaria Terzi. 2024. Understanding Memorisation in LLMs: Dynamics, Influencing Factors, and Implications. arXiv:2407.19262 [cs.CL] https://arxiv.org/abs/2407.19262
- [43] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to Fine-Tune BERT for Text Classification? arXiv:1905.05583 [cs.CL] https://arxiv.org/abs/ 1905.05583
- [44] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems 10, 05 (2002), 557–570. https://doi.org/10.1142/S0218488502001648
- [45] Xavier Tannier, Perceval Wajsbürt, Alice Calliger, Basile Dura, Alexandre Mouchet, Martin Hilka, and Romain Bey. 2023. Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse. arXiv:2303.13451 [cs.CL] https://arxiv.org/abs/2303.13451
- [46] Yakini Tchouka, Jean-François Couchot, Maxime Coulmeau, David Laiymani, Philippe Selles, and Azzedine Rahmani. 2023. De-Identification of French Unstructured Clinical Notes for Machine Learning Tasks. arXiv:2209.09631 [cs.CR] https://arxiv.org/abs/2209.09631
- [47] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying Patient Smoking Status from Medical Discharge Records. Journal of the American Medical Informatics Association 15, 1 (01 2008), 14–24. https://doi.org/10.1197/jamia.M2408 arXiv:https://academic.oup.com/jamia/article-pdf/15/1/14/2339646/15-1-14.pdf
- [48] Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the State-of-the-Art in Automatic De-identification. Journal of the American Medical Informatics Association 14, 5 (09 2007), 550-563. https://doi.org/10.1197/jamia.M2444
- [49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. CoRR abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/ 1706.03762
- Yijue Wang, Nuo Xu, Shaoyi Huang, Kaleel Mahmood, Dan Guo, Caiwen Ding, Wujie Wen, and Sanguthevar Rajasekaran. 2022. Analyzing and Defending against Membership Inference Attacks in Natural Language Processing Classification. In 2022 IEEE International Conference on Big Data (Big Data). 5823–5832. https://doi.org/10.1109/BigData55660.2022.10020711
- 51] Qiang Wei, Xu Zuo, Omer Anjum, Yan Hu, Ryan Denlinger, Elmer V. Bernstam, Martin J Citardi, and Hua Xu. 2022. ClinicalLayoutLM: A Pre-trained Multi-modal Model for Understanding Scanned Document in Electronic Health Records. In 2022 IEEE International Conference on Big Data (Big Data). 2821–2827. https://doi.org/10.1109/BigData55660.2022.10020569
- [52] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Aberba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (Seoul, Republic of Korea) (FAccT '22). Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088
- [53] Xi Yang, Tianchen Lyu, Chih-Yin Lee, Jiang Bian, William R. Hogan, and Yonghui Wu. 2019. A Study of Deep Learning Methods for De-identification of Clinical Notes at Cross Institute Settings. In 2019 IEEE International Conference on Healthcare Informatics (ICHI). 1–3. https://doi.org/10.1109/ICHI.2019.8904544
- [54] Hangu Yeo, Elahe Khorasani, Vadim Sheinin, Irene Manotas, Ngoc Phuoc An Vo, Octavian Popescu, and Petros Zerfos. 2022. Natural Language Interface for Process Mining Queries in Healthcare. In 2022 IEEE International Conference on Big Data (Big Data). 4443–4452. https://doi.org/10.1109/BigData55660.2022. 10020685
- [55] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2021. Counterfactual Memorization in Neural Language Models. CoRR abs/2112.12938 (2021). arXiv:2112.12938 https://arxiv. org/abs/2112.12938