# Toward training NLP models to take into account privacy leakages

Gaspard Berthelier, Antoine Boutet, Antoine Richard

# Toward training NLP models to take into account privacy leakages

1st Gaspard Berthelier
*CentraleSupélec, Université Paris Saclay*
Grenoble, France
gberthelier.projet@gmail.com

2nd Antoine Boutet
*Univ Lyon, INSA Lyon, Inria, CITI*
Lyon, France
antoine.boutet@insa-lyon.fr

3rd Antoine Richard
*DSN Bron, Hospices Civil de Lyon*
Grenoble, France
antoine.richard@chu-lyon.fr

***Abstract*—With the rise of machine learning and data-driven models especially in the field of Natural Language Processing (NLP), a strong demand for sharing data between organisations has emerged. However datasets are usually composed of personal data and thus subject to numerous regulations which require anonymization before disseminating the data. In the medical domain for instance, patient records are extremely sensitive and private, but the de-identification of medical documents is a complex task. Recent advances in NLP models have shown encouraging results in this field, but the question of whether deploying such models is safe remains.**

**In this paper, we evaluate three privacy risks on NLP models trained on sensitive data. Specifically, we evaluate counterfactual memorization, which corresponds to rare and sensitive information which has too much influence on the model. We also evaluate membership inference as well as the ability to extract verbatim training data from the model. With this evaluation, we can cure data at risk from the training data and calibrate hyper parameters to provide a supplementary utility and privacy trade-off to the usual mitigation strategies such as using differential privacy. We exhaustively illustrate the privacy leakage of NLP models through a use-case using medical texts and discuss the impact of both the proposed methodology and mitigation schemes.**

***Index Terms*—NLP models, Privacy, Membership Inference, Counterfactual Memorisation, Data Extraction**

## I. Introduction

Healthcare generates massive amounts of data collected from many different sources. The use of this valuable data has many advantages and promises: improving the quality of care and our knowledge of the health system, identifying disease risk factors, assisting in diagnosis, making wiser choices and monitoring of the effectiveness of treatments, delivering personalized healthcare value, epidemiology, etc. A large part of this data corresponds to text documents (e.g., medical reports). With the rise of machine learning and the advent of Natural Language Processing (NLP) models are increasingly used to automate the processing of medical documents and reports [1, 2, 3].

In recent years, a need to share medical data between various healthcare centers has emerged. This need was all the more felt during the SARS-Cov-2 pandemic for example, where the objective was to propose epidemiological models taking into account data from all over the world. However, patient medical records are extremely sensitive and private data. Their use and distribution is 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 elements that can be used to trace a patient directly (i.e., de-identification) or indirectly (i.e., anonymization).

The de-identification of medical documents is a complex task, costly in time and sometimes requiring several doctors which can slow down research. However, recent advances in NLP [4] based on neural networks have shown encouraging results. Indeed, NLP has grown in popularity since the advent of ChatGPT, yet NLP-models are not limited to text generation, and can include multiple tasks including classification, named entity recognition, and thus the de-identification of free texts. Johnson et al. for example proposed to use a neural network based on a BERT architecture [5] to detect a certain number of identifying elements in medical documents. More recently, different hospitals have also explored the feasibility of using NLP-models to pseudonymize (i.e., hiding specific direct identifiers) text documents from their clinical data warehouse [6, 7].

Although the use of language models to automate the processing of medical documents and to remove personal information (or pseudonymize them by replacing direct identifier to pseudonym) in order to facilitate their sharing is appealing [8], the attack surface of these models trained on personal and highly sensitive data is still poorly understood [9, 10]. Thus, additionally to the evaluation of the quality of the de-identification itself, using NLP-models to process medical reports still poses a number of threats related to the leakage of the sensitive information used during the training of models. Specifically, there are a few known privacy vulnerabilities involving the training data (i.e., a large corpus of medical documents) associated with machine learning and NLP-models [11] mostly considered individually such as counterfactual memorization [12] (i.e., memorisation of rare data), data extraction or reconstruction [13, 14], and membership inference attacks [15, 16, 17] (i.e., identifying elements of the training data). To reduce these risks, mitigation techniques have been proposed such as Differential Drivacy [18] (DP) or pruning strategy [17]. However, these mitigation techniques drastically degrade the accuracy of the model making them unusable in practice.

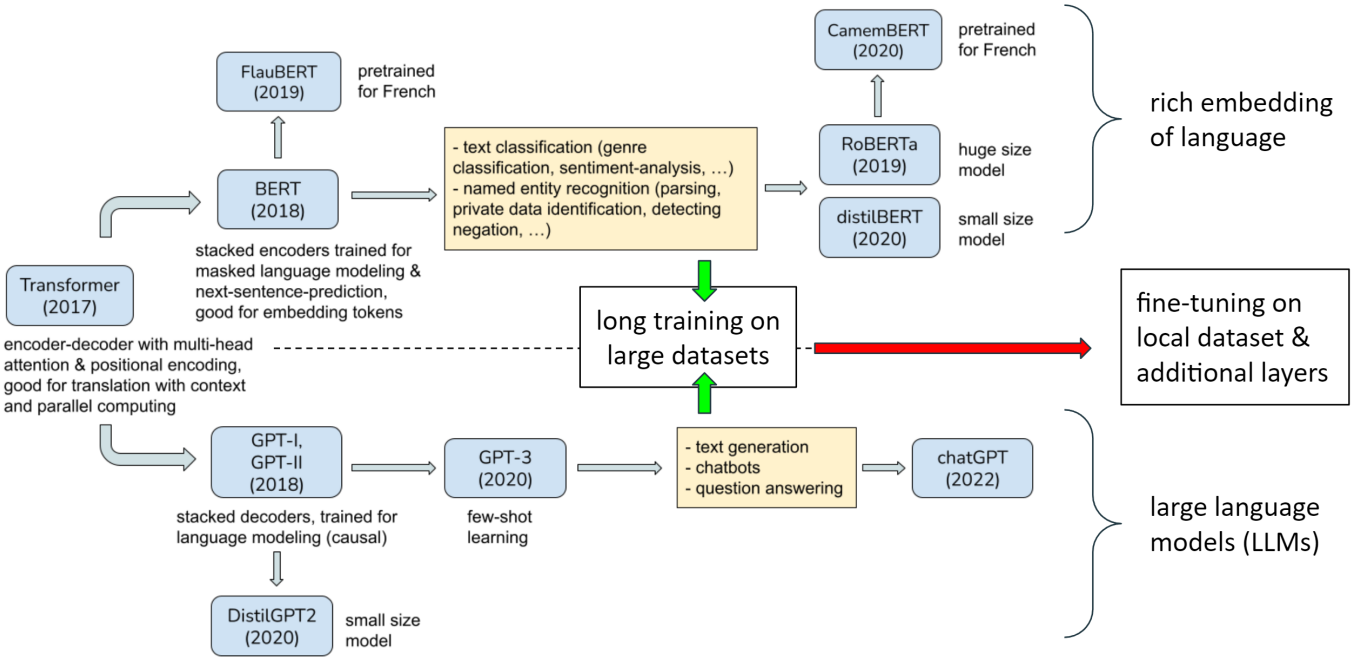


Fig. 1. Map of state-of-the-art Transformer-based models: since Transformers based on encoder-decoder neural network, BERT models focus on the encoder and are efficient for classification tasks while GPT models use the decoder and are generative models. Both categories of models can be fine-tuned with local data (e.g., medical reports).

We propose a training methodology to limit privacy leaks of sensitive information directly during the training phase. Specifically, sensitive information subject to counterfactual memorisation is discarded of the training data and instead of using cross validation to define the values of hyper-parameters only according to the performance of the model, both model accuracy and information leakage is evaluated. We argue that a model calibrated to also take into account the potential information leakage provides a better utility and privacy trade-off than using a mitigation strategy based on DP which degrades to much the utility of the model.

The rest of the paper is organized as follows. We start by presenting a comprehensive background in NLP and related works in Section II. We then present the proposed methodology to take into account the privacy assessment directly during the training of the model in Section III. Our exhaustive evaluation is reported Section IV before concluding Section V.

## II. Background

This section presents a comprehensive background on NLP (Section II-A), how hospitals leverage them for de-identifying clinical reports (Section II-B), the associated privacy leakages (Section II-C), and mitigation strategies (Section II-D).

### A. Natural language processing

Natural Language Processing (NLP) consists in understanding and processing textual data using machine learning models. The field underwent a breakthrough in 2017 with the advent of the Transformer [19]. This novel architecture revolutionized translation at the time. It consists of an encoder-decoder neural network with a parallel computing scheme which uses positional encoding and various attention mechanisms (see [20] for details). The objective of the encoder is to embed (i.e., turn into vectors) the input sentences. Each word is embedded into a latent space, by taking the whole sentence as context. For example, the word ”orange” in the sentences ”The orange house” and ”I ate an orange” will be turned into two different vectors. The encoder will have learnt to pay attention to the word ”house” in the first sentence and ”ate” in the second. The decoder then learns to translate these latent vectors into new sentences, for example to French. Each tokens (words or subwords) are predicted sequentially, by paying attention to the input and the previously predicted words. Thanks to this, the model knows to return ”maison” before ”orange” in the French translation.

Two trends followed in 2018 with BERT models [21] which focus on the encoder part of the Transformer and GPT models which use the decoder. The first are very efficient for classification tasks. They are pretrained on enormous unlabeled datasets to learn very complex embeddings of words. It is then possible to fine-tune such models on specific data and even to learn new tasks by adding just a few layers to the model. For instance, 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) which has learnt to imitate human expression. The outputed sentences are the most probable answers according to the model, based on probabilities it learnt by observing sentences in its huge training dataset.

They can be used to create chatbots such as chatGPT, which derives from a GPT-3 base. OpenAI then extended its training with supervised and reinforced phases to ensure answers are non-toxic and do not include fake news [22, 23].

Since then, there have been a myriad of new models inspired by the Transformer. For instance RoBERTa [24] which is a much larger version of BERT, or distilBERT [25] which uses knowledge distillation to produce a smaller size model. Models in other languages have also appeared such as CamemBERT [26] for French. Figure 1 maps the evolution of the transformer-based models. The NLP Cookbook [4] also surveys these different models and their specificities.

In our study, we will assess privacy leakages of NLP-models trained on medical reports through two use-cases: text classification and text generation. For this, we will use a BERT model and a distilGPT2 model.

### B. NLP for Privacy

In order to exploit or share their patients' information for research purposes while ensuring patient privacy, hospitals have started exploiting language models for de-identifying clinical reports [27, 6, 7, 28, 29]. De-identification of text-based clinical reports consist of the removal or replacement of personally identifying information from electronic health reports. Although strict anonymization is considered a very difficult task, this pseudonymization (e.g., identifying and replacing the personally identifying information by a plausible surrogate) makes it difficult to reestablish a link between the patient and its data and is considered enough protection for research purposes. Personal identification information is taken from a list which may vary from one country to another and includes for example address, date, birth date, hospital, patient id, email, visit id, last name, first name, phone, city, zip code and social security number.

The common pipeline of NLP models to achieve de-identification of clinical reports is similar to the one reported Figure 2 which depicted the workflow adopted by the Hospices Civils de Lyon (HCL) [7]. The base model (A) is a CamemBERT [26] which is a BERT model [21] specialized on French texts. This model specialization to French is done through a fill-mask task (i.e., random holes are added to the sentences and the model learns how to fill them). Following the same specialization process, the CamemBERT model is then fine-tuned on medical texts (A') to have a better statistical understanding of the language used in medical reports. Lastly, model A' is fine-tuned into B' to detect Personal identification tokens, using a manually labeled dataset.

Literature shows that the resulting models are able to outperform usual de-identification techniques which only use regular expressions and manual rules to remove private tokens (i.e., up to 0.99 of F1-Score with NLP models compared to 0.85 with manual rules only [6]). Although these models can represent a very efficient clinical text de-identification tool, only few of these studies evaluate the utility loss related to the de-identification and none of them evaluate the potential privacy leakage of the NLP models itself due to memorization of training data [10]. This latter attack surface is detailed in the next subsection.

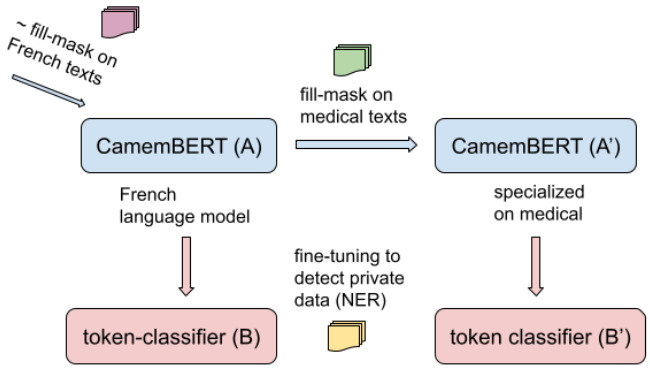


Fig. 2. Workflow by the Hospices Civils de Lyon (HCL) for de-identifying clinical reports.

### C. Privacy for NLP

Large Language Models (LLM) are trained on very large datasets. For instance, training chatGPT required scraping the Internet for years. Consequently, numerous personal data such as individuals' addresses have been used during the training. BERT models on the other hand are usually fine-tuned for specific tasks with domain-oriented data. In the medical domain, datasets usually include sensitive patient records. In both of these cases, the concern is that models may leak information from the training data after their deployment which represent an important privacy leakage.

In our study, we look into privacy leakages of NLP models by firstly analyzing how models memorize specific data (Section II-C1), then how textual information can be extracted from these models (Section II-C2), and finally a more common attack in privacy called Membership Inference Attack (MIA, Section II-C3).

*1) **Data memorization**:* Machine learning models are expected to extract trends from the training data in order to generalize to new data. Rare data or "outliers" on the other hand are usually not supposed to be memorized. A model memorizing rare data does not only negatively impact the utility but also privacy. Indeed, the more you learn on a small subset of individuals, the higher the information leakage since you can more easily pinpoint this information to its source. For example, we expect chatGPT to know Harry Potter's address (which can be found on numerous pages online) but not the reader's address (which should be nonexistent or at least hard to find online).

It turns out it is possible to measure this undesirable memorization, coined *counterfactual memorization* in [12]. To do so on any data, you must compare the performance of a model trained on a dataset with that data, to a second model trained without. This is computationally expensive to do for every data, so counterfactual memorization is actually computed

with an empiric expectation: we create multiple copies of the dataset and train many models on different subsets. Each data will have models it was trained on and models it was not. We can then compute the expected memorization:

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

where $\text{score}(M_D, x)$ is the score for $x$ of the model trained with the dataset $D$. Both terms will cancel out for common data (their removal has no impact) but may give a high difference for rare data. Data points with memorization above a certain threshold will be considered at risk.

*2) **Data extraction**:* Data extraction is a type of attack which aims to use the model to reconstruct information from the original data [11]. This attack mainly concerns text-generation models, such as GPT. These models are trained to output text based on what they saw during training. However, we do not expect the model to be a basic parrot and repeat the sentences it saw exactly. It is all the more a concern if the data it repeats is sensitive. It turns out it was the case of GPT-2 for instance from which individuals' names and addresses could be extracted [30].

In [31], the term $k - extractibility$ is used to denote sequences that can be extracted from the model when prompted by an input sequence of length $k$. The lower $k$ is, the easier it is to extract the sequence. We thus expect a model to have the highest $k$ possible on private queries.

*3) **Membership inference**:* Membership Inference Attack (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 instance if a hospital has trained a model to detect cancer and you learn your colleague's data was used in the training. You will have indirectly learned that he is probably suffering from cancer.

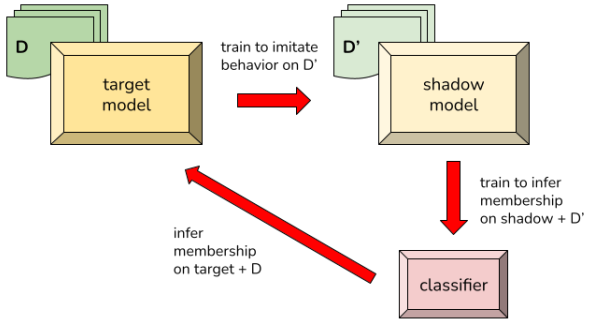


Fig. 3. A Membership Inference Attack (MIA) leverages auxiliary dataset $D'$ to build shadow models used to train a classifier to infer a piece of data's membership in the training information.

There are various techniques that can be used to achieve an MIA. One of them consists in using shadow models [32]. Shadow models are trained to imitate the target model's behavior on an auxiliary dataset with a similar distribution to the original. An adversary model (i.e., a classifier) is then trained to infer membership from these shadow models. This attack is depicted on Figure 3.

### *D. Mitigation strategies*

The most popular approach which helps to mitigate privacy risks is Differential Privacy (DP).

DP is a mathematical property that a model must verify in order to leak as little information as possible. This property imposes the model to learn a bounded 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 $\epsilon$ the *privacy budget*. The lower $\epsilon$ 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 [18]. The idea is to apply DP during the training phase by clipping the gradient updates and adding centered noise at each step. An illustration of the process is given on Figure 4. DP is known to significantly decrease the accuracy of the model.

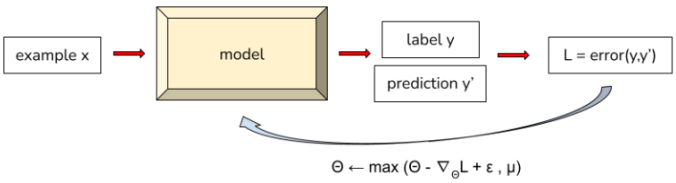


Fig. 4. Pipeline to learn a model with Differential Privacy (DP) in order to make the participation of an individual indistinguishable to an observer accessing the output of the model.

## III. Assessing privacy in the training phase

The process of training a model requires choosing hyper-parameters (e.g, model layers, learning rate, number of epochs) which will be used during the learning phase of the model. Choosing optimal hyper-parameter requires running multiple trials with different values for the parameters and evaluating them with cross-validation. This evaluation is often only controlled by the accuracy of the model. We argue that taking into account the privacy assessment during this step is a required methodology to provide a good utility and privacy trade-off.

Figure 5 depicts the pipeline including the proposed methodology where the new building blocks are in red. Once the raw data is processed, the hyper-parameter optimization begins. Both the accuracy and the privacy leakages are evaluated to judge the model.

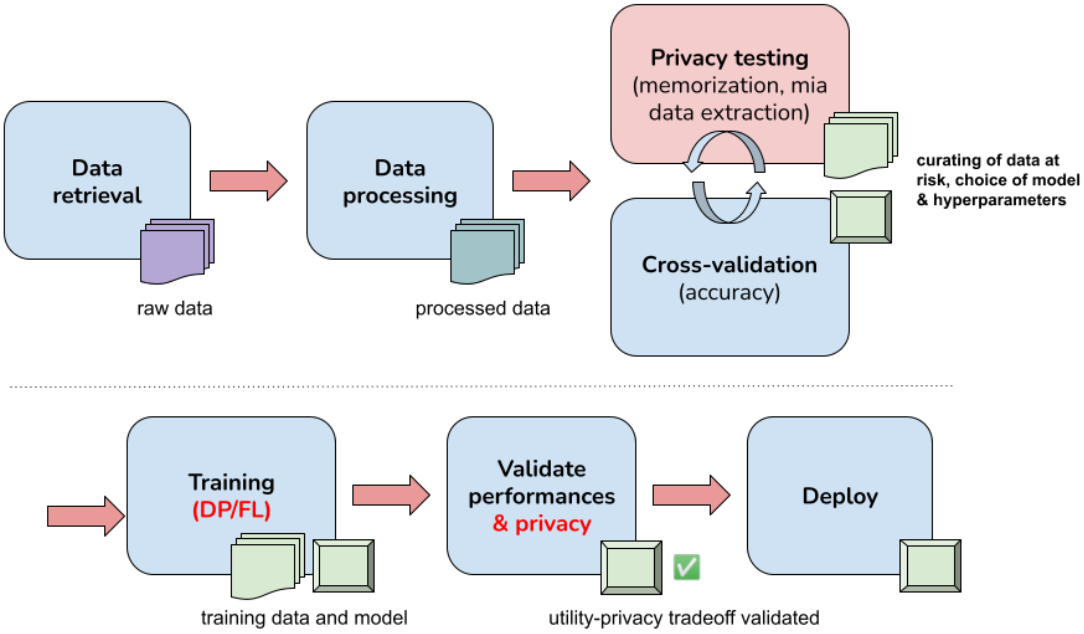


Fig. 5. Taking into account both the accuracy of the model and the privacy leakage to cure sensitive information subject to counterfactual memorization and to fix the hyper-parameters of the model provides a better utility and privacy trade-off than mitigation strategy such as using Differential Privacy.

Specifically, the privacy leakages are evaluated through counterfactual memorisation, inference of membership and data extraction. Identification of counterfactually memorized data is useful to cure sensitive data that are at risk in the model. Measuring the membership inference and the data extraction as well as the accuracy of the model are then used to evaluate the impact of the considered hyper-parameters. The training of the model can be done through Differential Privacy (DP) and also Federated Learning (not covered in this paper).

This optimization process once finished will return clean data and the hyper-parameter values that are best suited for the model to achieve the best accuracy-privacy tradeoff.

## IV. Empirical evaluations of privacy risk

This section reports an exhaustive evaluation of the privacy risks related to NLP models trained on medical data and the impact of DP to mitigate the risks. We consider a real use-case and setup (Section IV-A) before quantifying both the different privacy risks and the impact of mitigations (Section IV-B and Section IV-C, respectively).

### A. Experiment setups

To conduct the privacy risk assessment on sensitive data, we considered the BLUE dataset [33]. This dataset includes the Hallmarks of Cancer corpus of around 1,000 documents which consists of medical texts in English labeled according to 10 types of cancer. By investigating the dataset, we can extract meaningful information: labels are not equally represented, the number of words and the number of characters are around 250 and 1,600 respectively, and the number of unique words is around half of that number, which means half of the words in a single text are unique on average. The classification task (i.e., identification of the type of cancer from the reports) on such a dataset is thus a very complex task.

For the models, we considered DistilGPT2, BERT and DistilBERT from Hugging Face [34]. DistilGPT2 is an English-language generative model pre-trained with the supervision of the smallest version of GPT-2. Like GPT-2. DistilBERT, in turn, is a smaller, faster, cheaper and lighter version of BERT.

### B. Exhaustive privacy evaluations

We started by evaluating the counterfactual memorization of a BERT model fine-tuned with the Hallmarks of Cancer corpus for the classification of types of cancer. The expected memorization is reported Figure 6. The resulting distribution is centered around 0 (which corresponds to no counterfactual memorisation) but five texts have a memorization higher than 0.5 (which corresponds to high memorization). Among those "counterfactuals memorization", $60\%$ have a larger size than average (character length, number of words and number of unique words). Particular attention must therefore be paid to information with a larger than average size in the choice of counterfactual information to evaluate.

We then evaluated data extraction. To achieve that we fine-tuned a distilGPT2 model for text-generation on the medical texts. For each text, we sampled random subsequences of 4 different lengths (from $10\%$ to $75\%$ of the minimal text size). We then checked if part of the output of the model was present verbatim in the original dataset. The result of the extraction is depicted Figure 7.

The correlation between the number of extracted data and the prompt size is not exactly what we expected : longer prompts do indeed yield more extraction but there seems to be a soft spot with a prompt size of $0.25\%$.

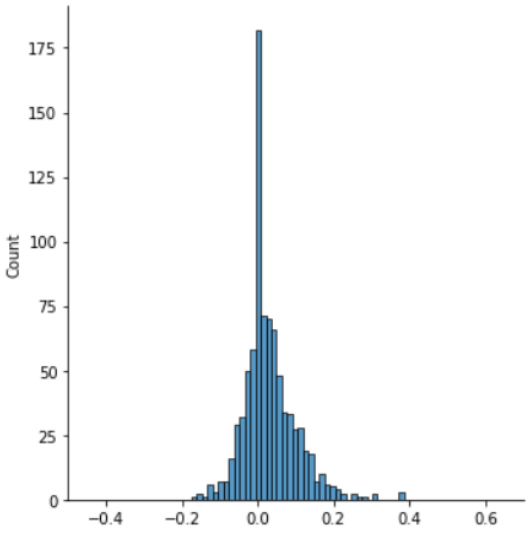


Fig. 6. Data at the far right of the distribution (far from 0) are counterfactual memorized.

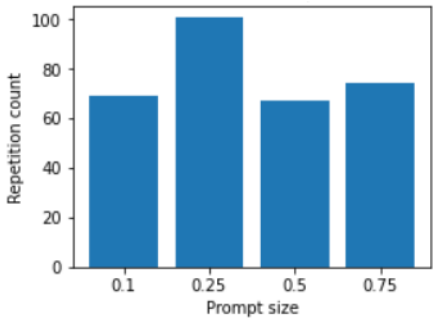


Fig. 7. Extracted count for each experiment

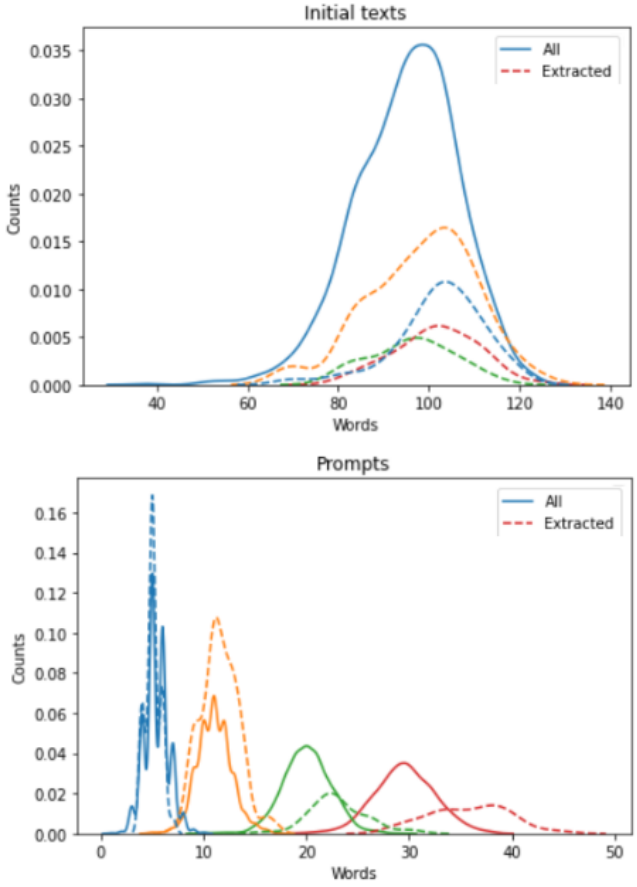


Fig. 8. Distributions of the number of words for original texts and prompts. Different colors correspond to the different prompt sizes. Extracted distributions are shifted to the right.

Moreover, we find after further analysis that the longer texts in terms of number of words are the ones more extracted, as illustrated Figure 8: the distributions of the number of words are shifted to the right for extracted sentences.

Finally, we evaluated the risk of membership inference on a BERT model trained for classification. This classifier is the target model under MIA. We trained a shadow model in the same manner as the target model, but on a different and controlled training set, integrating both data part of the training of the target model (train data) and data which are not part of the training of the target model (test data). Since a learning model has a better confidence score on a data item which has already been seen during its training, this confidence can be leveraged to infer membership (Section II-C3). In practice, the confidence score for train and test data will become more separable as training advances. This evolution in the confidence score of the model from 1 epoch to 5 and 9 epochs is illustrated in Figure 9 and at convergence in Figure 10. After a sufficient number of epoch, we can see part of the train data are clearly distinguishable. We can therefore use an XGBoost classifier to find a threshold to easily infer membership for these data points. However, passed above this threshold, it becomes difficult to predict whether or not a data point belongs to the training data.

In this first scenario, the attacker does not know the training time of the target model. So the threshold it finds is not optimal. Our attack yielded an MIA accuracy of 0.56. which is quite low. Actually, the highest accuracy we can obtain is training is 0.6 (highest training accuracy). We can see on Figure 10 why that is the case: the attacker can easily identify test data (0.95 precision), but is not as efficient on train data (0.54 precision).

The attack can actually be improved by giving more information to the adversary, for example the true label of the data. With this information, we can build a decision tree with thresholds on each label.

More importantly, when applying the first MIA scenario but only on counterfactual data, we obtain an accuracy of 0.8. This shows counterfactuals are at higher risk of being exploited by adversaries. This hints that removing these data is a good privacy measure before training the final model.

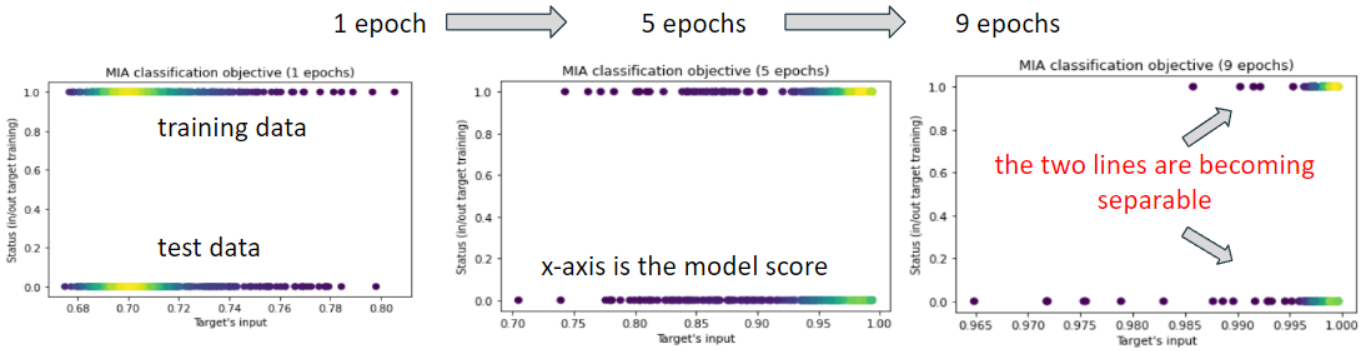


Fig. 9. Distribution of model's scores during training. The lines for training and test data become separable

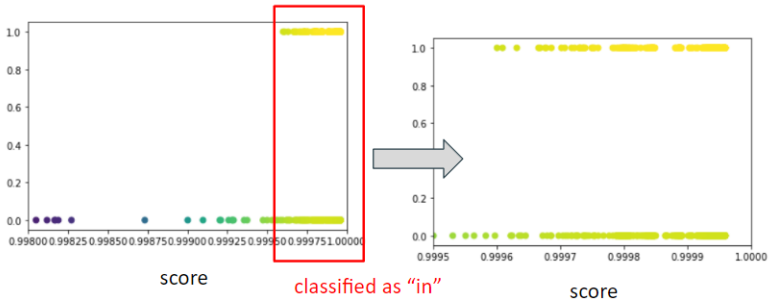


Fig. 10. MIA attack: high prediction on "out" data but lower on "in"

### C. Mitigations

We tried implementing mitigation strategies to reduce the previously illustrated risks. Figure 11 shows the evolution of accuracy with and without DP for the classification task on BERT. Unfortunately, DP reduces drastically the accuracy of the model. In that simulation, $\epsilon$ is at 600 which is already too high for privacy concerns (we usually expect $\epsilon$ to be between 1 and 10). We decided to simplify the task by only looking at the two most common labels. Then, we obtained a vanilla accuracy of 0.97, a DP accuracy of 0.60 and an epsilon of 200, which is slightly more acceptable. We attacked both models with an MIA which returned 0.59 accuracy for the vanilla model and 0.55 for the DP model. This shows DP can indeed mitigate the risk to a certain extent, but it not an acceptable measure for a relevent accuracy-privacy tradeoff.

On the other hand, we evaluated the counterfactual memorization in order to cure the original training data.

We trained the BERT model for 9 epochs with and without the counterfactuals, and found that the MIA accuracy dropped from 0.57 to 0.51, without a drop of performance. We also trained a distilBERT in the same manner and reduced the MIA risk from 0.57 to 0.53. Both architectures returned a similar number of counterfactuals (1 more for distilBERT). We repeated the same procedure with 13 epochs of training which increased the MIA risk as expected, with no significant improvement of accuracy.

Altogether, this shows that removing counterfactually memorized data, and carefully choosing parameters such as model

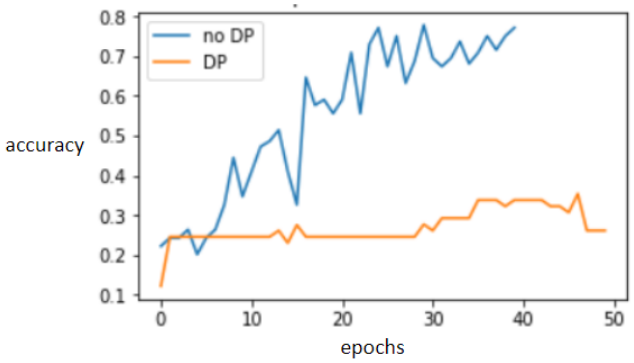


Fig. 11. DP-SGD: accuracy drops drastically ($\epsilon$ is at 600 which is high). The number of epochs correspond to mini-batch epochs.

size and number of epochs can improve the privacy-utility trade-off. There are of course other hyper-parameters to look at such as dataset size, learning rate and other model-specific parameters. which could also affect this trade-off between utility and privacy. Therefore, it is important to perform these assessments empirically on each dataset and model to be deployed for real-world applications.

## V. Conclusion

This paper presents a methodology to mitigate privacy leakages from NLP models directly during the training phase while maintaining its accuracy. By taking into account all the risks of privacy leakage (i.e., counterfactual memorization,

membership inference, and data extraction), the training data can be efficiently cured from sensitive information subject to counterfactual memorization and hyper-parameters of the model can be calibrated to provides a better utility and privacy trade-off than mitigation strategy such as using DP.

Given the sensitive nature of the data used for training many NLP models such as in the medical field, we believe that it is necessary to change practices to better take into consideration the risks linked to privacy. After having exhaustively presented the different risks of privacy leakages, we have illustrated these leakages through a use-case using fine-tuned NLP models with medical documents.

## Acknowledgement

This work has been supported by the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR, the Trusty-IA project.

## References

[1] X. Chen, J. Lin, and Y. An, "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)*, 2022, pp. 1801–1808.

[2] H. Yeo, E. Khorasani, V. Sheinin, I. Manotas, N. P. An Vo, O. Popescu, and P. Zerfos, "Natural language interface for process mining queries in healthcare," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 4443–4452.

[3] Q. Wei, X. Zuo, O. Anjum, Y. Hu, R. Denlinger, E. V. Bernstam, M. J. Citardi, and H. Xu, "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)*, 2022, pp. 2821–2827.

[4] S. Singh and A. Mahmood, "The NLP cookbook: Modern recipes for transformer based deep learning architectures," *CoRR*, vol. abs/2104.10640, 2021. [Online]. Available: https://arxiv.org/abs/2104.10640

[5] J. et al., "Deidentification of free-text medical records using pre-trained bidirectional transformers." 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34350426/

[6] X. Tannier, P. Wajsbürt, A. Calliger, B. Dura, A. Mouchet, M. Hilka, and R. Bey, "Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse," 2023.

[7] A. Richard, F. Talbot, and D. Gimbert, "Anonymisation de documents médicaux en texte libre et en français via réseaux de neurones," in *Plate-forme Intelligence Artificielle 2023 (PFIA2023) - Journée Santé & IA*. Starsbourg, France: Association française pour l'Intelligence Artificielle (AfIA) and Université de Strasbourg and Association française d'Informatique Médicale (AIM), Jul. 2023. [Online]. Available: https://hal.science/hal-04139391

[8] B. Dura, C. Jean, X. Tannier, A. Calliger, R. Bey, A. Neuraz, and R. Flicoteaux, "Learning structures of the french clinical language:development and validation of word embedding models using 21 million clinical reports from electronic health records," 2022.

[9] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel, "Taxonomy of risks posed by language models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 214–229. [Online]. Available: https://doi.org/10.1145/3531146.3533088

[10] E. Lehman, S. Jain, K. Pichotta, Y. Goldberg, and B. Wallace, "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*. Online: Association for Computational Linguistics, Jun. 2021, pp. 946–959. [Online]. Available: https://aclanthology.org/2021.naacl-main.73

[11] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos, "Privacy and security issues in deep learning: A survey," *IEEE Access*, vol. 9, pp. 4566–4593, 2021.

[12] C. Zhang, D. Ippolito, K. Lee, M. Jagielski, F. Tramèr, and N. Carlini, "Counterfactual memorization in neural language models," *CoRR*, vol. abs/2112.12938, 2021. [Online]. Available: https://arxiv.org/abs/2112.12938

[13] K. Gu, E. Kabir, N. Ramsurrun, S. Vosoughi, and S. Mehnaz, "Towards sentence level inference attack against pre-trained language models," *PoPETS*, vol. 2023, p. 62–78, 2023.

[14] R. Panchendrarajan and S. Bhoi, "Dataset reconstruction attack against language models," 2021.

[15] A. Jagannatha, B. P. S. Rawat, and H. Yu, "Membership inference attack susceptibility of clinical language models," 2021.

[16] F. Mireshghallah, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, and R. Shokri, "Quantifying privacy risks of masked language models using membership inference attacks," 2022.

[17] Y. Wang, N. Xu, S. Huang, K. Mahmood, D. Guo, C. Ding, W. Wen, and S. Rajasekaran, "Analyzing and defending against membership inference attacks in natural language processing classification," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 5823–5832.

[18] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning

with differential privacy," oct 2016. [Online]. Available: https://doi.org/10.1145%2F2976749.2978318

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[20] J. Alammar. ((2018)) The illustrated transformer. [Online]. Available: Retrieved from https://jalammar.github.io/illustrated-transformer/

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[22] OpenAI, "Gpt-4 technical report," 2023.

[23] F. Faal, "Reinforcement learning for mitigating toxicity in neural dialogue systems," Ph.D. dissertation, Concordia University, 2022.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[26] L. Martin, B. Müller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," *CoRR*, vol. abs/1911.03894, 2019. [Online]. Available: http://arxiv.org/abs/1911.03894

[27] X. Yang, T. Lyu, C.-Y. Lee, J. Bian, W. R. Hogan, and Y. Wu, "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)*, 2019, pp. 1–3.

[28] T. Hartman, M. D. Howell, J. Dean, S. Hoory, R. Slyper, I. Laish, O. Gilon, D. Vainstein, G. Corrado, K. Chou *et al.*, "Customization scenarios for de-identification of clinical notes," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–9, 2020.

[29] Y. Tchouka, J.-F. Couchot, M. Coulmeau, D. Laiymani, P. Selles, A. Rahmani, and C. Guyeux, "De-identification of french unstructured clinical notes for machine learning tasks," 2022.

[30] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," *CoRR*, vol. abs/2012.07805, 2020. [Online]. Available: https://arxiv.org/abs/2012.07805

[31] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," 2023.

[32] R. Shokri, M. Stronati, and V. Shmatikov, "Membership inference attacks against machine learning models," *CoRR*, vol. abs/1610.05820, 2016. [Online]. Available: http://arxiv.org/abs/1610.05820

[33] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets," 2019.

[34] H. Face, "The platform where the machine learning community collaborates on models, datasets, and applications." [Online]. Available: https://huggingface.co