Membership Inference Attack Susceptibility of Clinical Language Models

Abhyuday Jagannatha[1], **Bhanu Pratap Singh Rawat**[1], **Hong Yu**[1,2]
[1]College of Information and Computer Sciences,University of Massachusetts Amherst
[2]Department of Computer Science,University of Massachusetts Lowell
{abhyuday, brawat hongyu}@cs.umass.edu

Abstract

Deep Neural Network (DNN) models have been shown to have high empirical privacy leakages. Clinical language models (CLMs) trained on clinical data have been used to improve performance in biomedical natural language processing tasks. In this work, we investigate the risks of training-data leakage through white-box or black-box access to CLMs. We design and employ membership inference attacks to estimate the empirical privacy leaks for model architectures like BERT and GPT2. We show that membership inference attacks on CLMs lead to non-trivial privacy leakages of up to 7%. Our results show that smaller models have lower empirical privacy leakages than larger ones, and masked LMs have lower leakages than auto-regressive LMs. We further show that differentially private CLMs can have improved model utility on clinical domain while ensuring low empirical privacy leakage. Lastly, we also study the effects of group-level membership inference and disease rarity on CLM privacy leakages.

1 Introduction

Large contextual word embedding models have been proposed and used for obtaining the state-of-the-art performance on various NLP tasks (Wang et al., 2018, 2019). The high performance and relatively lower sample complexity requirements achieved by domain-specific pretrained language models (LMs) (Peng et al., 2019; Devlin et al., 2018) have propelled their use in private data domains like healthcare and finance. A primary concern in deploying such models in a public setting is that models trained on sensitive private data may leak information about sensitive training samples. In application domains such as healthcare, such a leak may violate patient rights and cause potential harm to the participants of the study.

De-identification or obfuscation of the underlying text may not necessarily mitigate all privacy concerns since anonymized data can be identified using additional sources (Narayanan and Shmatikov, 2008). Also, de-identification of large text datasets is only feasible through automated de-identification, which may lead to imperfect results. Differential Privacy (DP) (Dwork et al., 2014) is a systematic approach to limiting worst-case training data leakage in the context of machine learning. Intuitively, DP limits training data leakage by limiting the effect of adding any random data sample to the training set. This is a general definition for estimating privacy budgets and is usually defined independently of the nature of the underlying data. As a result, DP may be overly conservative for systems using high dimensional input with lower rank data space. For instance, in NLP where a topically relevant sentence is a very small subset of all randomly constructed sentences, mitigating the risk for any random data sample may not provide an efficient privacy mechanism. Moreover, DP based privacy accountants (Jayaraman and Evans, 2019) vastly overestimate the privacy leakages of large neural network models and may be practically unusable.

On the other hand, empirical privacy leakage estimation may provide more accurate privacy estimates. However, training data extraction methods such as Carlini et al. (2020) may be model-specific and do not provide a standardized way to estimate privacy leakages across models architectures. Also, due to the high dimensional nature of inputs to LMs, inference attacks that demonstrate the potential of data leakages from such models are difficult to construct and study. This poses a problem for the study of privacy leakage in large LMs. Regulations like HIPPAA and EU 2016/679 have been enacted to ensure patient privacy rights. These laws dictate the use of patient data for secondary uses

like research. However, they do not provide explicit guidelines for data leakage through release of Machine Learning or NLP model files. Due to a lack of clarity, sensitive data models are mostly used *in-house*. Systematic privacy leakage studies for large NLP systems can inform and help policy decisions in this domain. This can accelerate the development of private, high utility NLP models.

In this work, we design general membership inference attacks (Shokri et al., 2017) that can be used to quantitatively estimate the discrepancy in a large LM's response to training and out-of-training (OOT) samples. Recent work shows that this discrepancy, which contributes to the generalization gap, is related to the model's privacy leakage (Yeom et al., 2018). We use white-box and black-box membership inference adversaries to establish a standardized framework for our study.

Our main focus in this study is to examine the susceptibility of CLMs to membership attacks and empirically estimate their privacy leakage. We also investigate DP methods to make these modes more private. To the best of our knowledge, this is the first attempt to study empirical privacy leakages in CLMs and use differentially private CLM training. Our main findings are as follows:

1. Large LMs can have higher empirical privacy leakages (7%) than smaller LMs (2%).
2. Randomly masked LMs have lower privacy leakages than autoregressive LMs.
3. $(\epsilon, \delta)-$DP training using DP-SGD (Dwork et al., 2014) can reduce empirical privacy leakages while ensuring increased model utility.
4. Group-level membership inference attacks lead to higher privacy leakage than sample-level.
5. Patients with rarer disease profiles may be more vulnerable to higher privacy leakages.

2 Background

Pre-trained Language Models: Language Models (LMs)[1] such as BERT, RoBERTa or GPT2 are pretrained using large unsupervised text corpora such as WikiText that contains text from multiple domains. These pre-trained LMs are used for extracting contextual embeddings which are used for multiple downstream NLP tasks. For domain specific NLP tasks such as MedNLI or emrQA, the LMs are further fine-tuned on text from a particular domain such as clinicalBERT (Alsentzer et al., 2019), bioBERT (Lee et al., 2020) and sciBERT (Beltagy et al., 2019).

Differential Privacy: $(\epsilon, \delta)-$DP (Dwork et al., 2014) can be used to limit the privacy leakage of machine learning models trained through stochastic gradient descent(SGD). A randomized mechanism $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ — in our case, this refers to the SGD training algorithm — is (ϵ, δ)-DP compliant if for all $S \subset \mathcal{R}$ and $x, x' \in D$

$$\Pr[\mathcal{M}(x) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(x') \in S] + \delta. \quad (1)$$

An ML model with (ϵ, δ)-DP compliant training limits the privacy leakage to a function of ϵ with failure probability δ. An important property of (ϵ, δ)-DP systems is ***Group Differential Privacy***. Group DP provides (ϵ, δ) values for k correlated inputs in a dataset. In such a case the privacy degrades to $(k\epsilon, ke^{(k-1)\epsilon}\delta)$ (Dwork et al., 2014). In private applications, ideally we would keep $\delta \leq 1/N$, for dataset size N and $\epsilon << 1$.

Membership Inference: Machine learning model parameters can be probed to extract sensitive information about the training data samples (Homer et al., 2008; Yeom et al., 2018). Shokri et al. (2017) introduce the membership inference attacks, which refer to a class of attack models that predict whether a given data sample was present in the training data for a trained model. The advantage of an attacker using membership inference attack depends on the attacker's ability to distinguish between the *target* model's response to training and out-of-training data samples.

Shokri et al. (2017) proposed a shadow model based attack that can try to understand differences in model's response for training data samples vs out-of-training(OOT) data samples. Computation and memory costs of shadow models scale with the number of model parameters and therefore, these attacks are not feasible for large CLM models such as BERT. Yeom et al. (2018) propose black-box attacks that achieve privacy leakage similar to shadow model attacks.

3 Data

We use three US hospital datasets, namely MIMIC-III, UMM (UMass Memorial Health Care)and VHA (Veterans Health Administration) Hospitals

[1] All useful notations are also provided together in Appendix B.5.

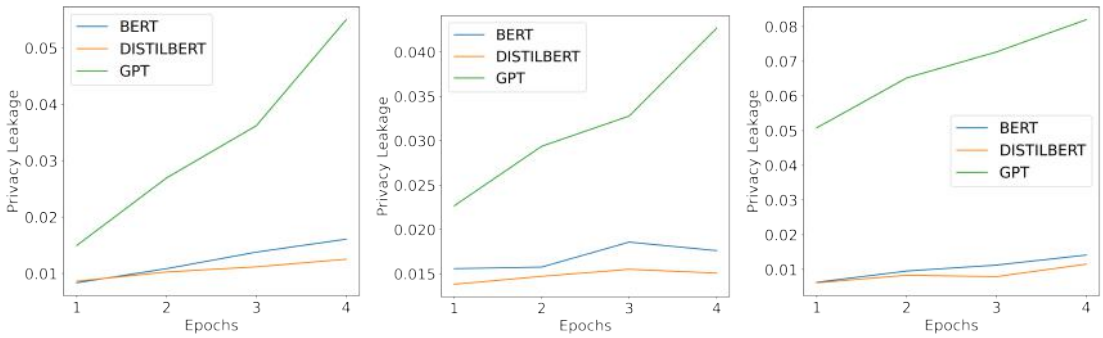


Figure 1: Sample-level privacy leakage using black box and white-box membership inference attacks for non-DP CLM models. Plots show results on the MIMIC-III data. Plot (a) shows S-BBA attack leakage. Plot (b) and (c) show S-AWBA and S-GWBA white box attack leakages respectively. The GPT2 model consistently has the highest leakage and DistilBERT has the lowest.

for our study. These datasets are used to train our CLMs and evaluate their privacy leakage. MIMIC-III (Johnson et al., 2016) is an publicly available dataset, while UMM and VHA data sources are private. Our main experiments are conducted with MIMIC-III and UMM datasets. Due to *in-house* computational constraints, we use VHA dataset to only evaluate privacy leakage for one of our CLM models. We extract EHR repositories from all three data sources along with patient and admission level information. We select all available patient notes from MIMIC-III. We use a manageable subset of patients from UMM and VHA repositories, detailed information provided in Appendix A.4.

In MIMIC-III, each patient's records are composed of a collection of EHRs of clinical visits or admissions. Each visit is composed of a group of EHRs that were recorded during that visit. Each EHR note is, in turn, a document that may be composed of multiple text samples. We use this hierarchical structure to study the effects of correlated samples on privacy leakage. In UMM and VHA , we could not use admission information. Therefore we treat each EHR record as an admission. Correlated samples can extend the notion of privacy to group-level privacy using the definition provided in Section 2. Dataset size statistics for all three datasets are provided in Table 1.

4 Methods

For our investigations, we use two main sets of methods. We use membership inference attacks to estimate empirical privacy leakage of a method. And we use differentially private stochastic gradient descent training (DP-SGD)(Abadi et al., 2016) to train DP clinical LM models.

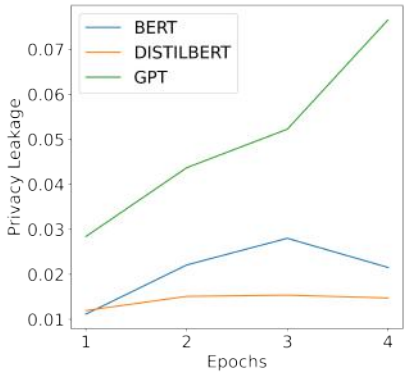


Figure 2: Admission-level membership attack (A-BBA) privacy leakage using MIMIC-III data for non-DP CLMs.

4.1 Membership Inference Attacks

Membership inference attacks are accomplished by exploiting the difference in a model's responses to training and OOT data samples. We use the *advantage* of membership inference attacks to estimate the empirical privacy leakage. The experimental framework for obtaining privacy leakage using a membership inference adversary is defined by (Yeom et al., 2018) as follows:

Membership Inference Experiment $\mathbf{Exp}^{M}(\mathcal{A}, A, n, D)$: Let $\mathcal{A}$ be an adversary, A be a learning algorithm, n be a positive integer, and D be a distribution over data points (x, y). The membership experiment in (Yeom et al., 2018) proceeds as follows:

- Sample $S \in D_n$.

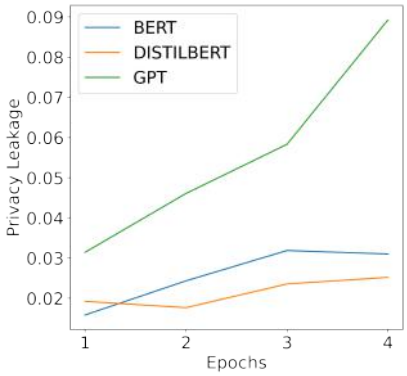


Figure 3: Patient-level membership attack (P-BBA) privacy leakage using MIMIC-III data for non-DP CLMs.

- Choose $b \leftarrow \{0, 1\}$ uniformly at random.
- Draw $z \in S$ if $b = 0$, or $z \in D$ if $b = 1$.
- $\text{Exp}^M(\mathcal{A}, A, n, D) = 1$ if $\mathcal{A}(z, A(S), n, D) = b$ and 0 otherwise. $\mathcal{A}$ must output either 0 or 1.

The membership advantage of the adversary Adv^M or alternatively the empirical privacy leakage observed in the experiment is

$$PL(\mathcal{A}) = Pr[\mathcal{A} = 0|b = 0] - Pr[\mathcal{A} = 0|b = 1]. \quad (2)$$

NLP models may be released with API or white-box access to the internal model. Therefore, we design both black-box and white-box adversaries for our study.

Samples For membership inference experiments in CLMs, the sample S consists of the input sentence and the constructed language modeling target. In case of an auto-regressive language model such as GPT-2, the input sequence is a fixed length token sequence and the target is the next token sequence. In case of MLM objective models like BERT, the input sequence is a randomly masked fixed length token sequence and the output is the relevant masked token values. It should be noted that masks are generated at random in MLM training procedures, so the masks generated for membership inference experiment will be different than those generated during training of the CLM.

4.1.1 Black-Box Adversary

We adapt the black box attack defined by Yeom et al. (2018) (*Adversary* 2). This attack model assumes access to the error probability density functions of the trained model, $f(\epsilon|b = 0)$ and $f(\epsilon|b = 1)$.

Sample-level Attack For a sample x, the output of the adversary $\mathcal{A}(x)$ is $argmax_{b\in(0,1)} f(\epsilon(x)|b)$. In context of language models with non-zero error means, we can define $f(\epsilon|b)$ using error mean and unit variances for both b=1 and 0. However, we use a simpler scheme that only assumes access to the mean of the training error. In preliminary experiments this choice did not reduce the attack advantage. We use the mean of training error μ_{tr} as the threshold to predict $\mathcal{A}(x)$. Our sample level attack is very similar to threshold adversary proposed by Yeom et al. (2018). A random sample x is predicted to be a training data sample if $e(x) < \mu_{tr}$ else it is predicted to be an OOT data sample. We denote this sample-level black-box membership inference attack as ***S-BBA***.

Group-level Attack To investigate group-level privacy leakages, we devise group-level membership inference attacks for our black-box access adversary. For group-level membership inference, we treat each group as a single data sample. To estimate the error of a group g we take a mean of all samples in that group. We use the same condition and threshold ($e(g) < \mu_{tr}$) as that in S-BBA for our group-level privacy attacks. The group-level attacks based on patient and admission level groups are referred as ***P-BBA*** and ***A-BBA*** respectively.

4.1.2 White-Box Adversary

With white-box access to the model, we can use hidden layer outputs, gradients and other model specific information to identify differences in model's behavior between training and out-of-training data samples. Based on preliminary experiments, gradient and attention values proved to be more informative than hidden layer outputs.

Attention-based white-box attack (S-AWBA): In this attack we use the self-attention outputs of CLMs as the input to the attack model. A k-head transformer layer with a sequence input length of n, produces kn^2 attention values. For a 12 layer LM such as BERT, with 12 attention heads for each transformer layer, the number of attention outputs is $144n^2$. For an input sequence of 512 tokens, a model like BERT will have around 37 million attention values. To reduce the size of attack model input, we instead use the "concentration" of the attention vector $a = \{a_i\}_0^n$ using

$$C(a) = \sum_{i=0}^{n} a_i \log(a_i). \quad (3)$$

$C(a)$ is higher when the attention is diffused across all token positions and lower when high attention values are focused on specific tokens. We obtain one $C(a)$ value for each token on each attention layer. To further reduce the size of the attention information, we use four estimates computed over the sentence length $\{C(a)_j\}_0^n$, namely mean, median, 5-percentile and 95-percentile values. Computation of $C(a)$ values and aggregation over the sentence provides us with a compressed 144×4 dimensional attention representation for any input sample. We use this attention representation as an input to the attention-based attack model.

Gradient-based white-box attack(S-GWBA): Gradient values have been previously (Nasr et al., 2019) used in membership inference attack models against DNNs. Intuitively, the gradient values of a training sample's loss are expected to have lower absolute value than an OOT data sample's loss derivative. Membership inference attacks using gradient values from large LM models pose similar computational challenges as the attention-based attacks. An LM model, such as BERT has 110 million parameter vectors. As a result, it produces 110 million partial derivatives. We aggregate the gradient values for each neural network layer by taking the squared-norm of all parameters in that layer. For a model like BERT, this results in a 202 dimensional aggregate. We use this gradient representation as an input for the attack model.

4.2 Differentially Private Model Training

We use differentially private SGD (DP-SGD) proposed by Abadi et al. (2016) to train our $(\epsilon, \delta)-$DP models. DP-SGD uses per-sample gradient clipping to bound the effect of each sample and uses additive Gaussian noise to inject noise into the mini-batch gradients. In our experiments, we use the noise standard deviation (std) as a hyperparameter and denote it as σ. We use Rényi Differential Privacy (RDP) framework (Mironov, 2017) to estimate the total privacy loss due to the DP-SGD training.

5 Experiments and Results

We run our experiments on non-DP and DP CLMs that are trained on EHR datasets described in Section 3. DP and non-DP CLMs use SGD and DP-SGD training procedures. Training details are provided in Appendix A. We use BERT, DistilBERT and GPT2 base models to initialize the CLM training. These CLMs are used in the following sections to study the patterns of privacy leakage across datasets and models.

Dataset	#Token	#Admission	#Patient
MIMIC-III	695M	57k	46k
UMM	1.33B	1184k	58k
VHA	8.31B	2019k	35k

Table 1: Number of tokens, admission and patients in the three datasets. The number of tokens are computed using BERT-base-cased tokenizer. We do not have admissions information for UMM and 3, so we consider one EHR note per admission.

5.1 What is the sample level empirical privacy leakage of CLM models?

We examine privacy leakage for clinical-domain tuned CLMs using both white-box and black-box membership inference attacks on text snippets from electronic health records. MIMIC-III and UMM data are used for the main set of experiments. Due to *in-house* computational constraints for VHA , we were unable to replicate all experiments on VHA data. However we do provide black box inference attack based PL results for BERT models in Appendix B. The general trends discussed in the following sub-sections hold true also for the limited results from VHA .

Experimental Design: Formally, we estimate the leakage by calculating S-BBA, S-AWBA and S-GWBA adversary advantage for non-DP CLMs. To evaluate the privacy leakage using Equation 2, we use the membership experiment defined in Section 4.1. We draw a sample z from the training or test data for b = 0 or 1, respectively. Estimation of $\Pr[\mathcal{A} = 0|b = 0]$ and $\Pr[\mathcal{A} = 0|b = 1]$ use training and test data samples. White-box attacks use a multi-dimensional vector of gradient or attention aggregates. To estimate the difference in attention and gradient values for training and OOT data sample, we train a logistic regression model (A) on 20% split of the train and test datasets. Recall from section 4.1 that A is used to predict b for a data sample z. The remaining 80% data samples are used to estimate $\Pr[\mathcal{A} = 0|b = 0]$ and $\Pr[\mathcal{A} = 0|b = 1]$ for provided $\mathcal{A}$.

Results: Privacy leakage plots using MIMIC-III for all non-DP CLMs are provided in Figures 1. The privacy leakage of all models for all epochs re-

Model /σ	2	1	0.1	1e-2	1e-3	1e-4	1e-20
GPT2	0/.009	0/.014	0/.014	0/.014	0/.017	4e-3/.016	8e-3/.018
BERT	9e-2/.011	2e-3/.011	3e-3/.010	1e-3/.011	3e-4/.017	3e-3/.016	5e-3/.016
DistilBERT	0/8e-3	0/5e-2	0/9e-2	0/8e-3	1e-5/.010	1e-3/.010	0/.012

Table 2: Empirical privacy leakage for DP models trained on MIMIC-III data using S-BBA and S-AWBA attacks. Black box S-BBA always leads to lower PL than white box S-AWBA attack. PL values reported in this table are the maximum PL values over all three epochs of these models. Underlined models have RDP computed $\epsilon < 1$. All other models have $\epsilon > 1$. RDP accounting results are in Appendix B.

mains lower than 10%. DistilBERT model that has the lowest number of parameters amongst all three models, on average has the lowest sample level privacy leakage. This behaviour is consistent across all attack methods, both black-box and white-box. A consistent trend in our results is that the white-box access attacks lead to higher PL values. This is expected. Gradient based white-box attack is the most effective attack. The leakiest model is GPT2 with privacy leakages is as high as 7%. On average, all PL values increase with the number of epochs.

5.2 Can we train (ϵ, δ)-DP CLM models with improved clinical domain performance?

Models with 7% PL, as shown in the previous section, should be further studied before releasing them in the public domain. To this end, we use DP-SGD training methods to enforce privacy constraints. We then study the trade-off between the strictness of privacy constraints and model utility.

Experimental Design: We use the DP-SGD and RDP for DP model training and privacy accounting as defined in Section 4.2. To limit the group based privacy degradation in ϵ calculation, we limit the number of samples from each patient to 50. We use Gaussian mechanism with σ ranging from 2.0 to 1e-4 in the DP-SGD algorithm. We also use a negligible σ of 1e-20 to ablate the effect of additive noise from DP-SGD training.

Results: We see a decline in empirical privacy leakage across all models for all attack methods when we apply a Gaussian noise within the range {1e-4, 2.0}. An empirical privacy leakage for a sample-level white-box attack (S-AWBA) and black box attack (S-BBA) are provided in Table 2. There is a significant decrease in empirical privacy leakage. Most models exhibit less than 1% (often less than 0.1%) privacy leakage. We do see increased privacy leakage through attention access in S-AWBA, however, these leakages are around 1-2% and are negligible for DistilBERT model. Empirical privacy leakages for all models-attack combinations are provided in the Appendix B for MIMIC-III and UMM data. One consistent pattern across all results show that models with $\sigma \in \{2.0, 1.0, 0.1, 0.01\}$ have negligible privacy leakages for all sample and group membership inference attacks. For $\sigma \in \{0.001, 0.0001, 1e-20\}$ we see more than 1% privacy leakages in S-AWBA and S-GWBA. The most successful attack on non-DP models is conducted using S-GWBA, providing PL as high as 2% for GPT2. Group membership inference attacks are advantageous even for a DP model. Group PL for models are as high as 5% for $\sigma = 1e-20$ (S-GWBA). In conclusion, group-level membership inference attacks and white box attacks are able to extract more than 1% privacy leakages from our DP models for very low epsilon values. However, most models with $\sigma \leq 1e-3$ show negligible privacy leakage. Results for other models on MIMIC-III and UMM datasets are available in Appendix B. They show similar behaviour. To understand the trade-off between σ and model utility, we compute the LM loss and MedNLI performance for each non-DP model epoch. Model performances for first two epoch BERT models (DP-SGD, epoch=1) on MIMIC-III data are provided in Table 3. All models with $\sigma \leq 0.01$ have improved MLM test loss as compared to BERT-base-cased. $\sigma \in \{0.01, 0.001\}$ provides MLM loss lower than the BERT-base-cased model while ensuring that empirical PL from all attack methods are kept below 1%. MedNLI experiments and results for MIMIC-III BERT models are provided in Appendix A and B respectively. We observe that most MedNLI performances for DP BERT models show improved model utility with accuracy between BERT-base and ClinicalBERT.

Model / Epochs	1	2
Non-DP BERT	0.59	0.57
DP BERT(2.0)	6.15	7.02
DP BERT(1.0)	5.65	6.76
DP BERT(0.1)	3.82	4.86
DP BERT(1e-2)	2.00	2.33
DP BERT(1e-3)	1.18	1.23
DP BERT(1e-4)	0.85	0.86
DP BERT(1e-20)	0.75	0.76

Table 3: DP BERT (σ) denotes BERT-base-cased models trained on MIMIC-III using DP-SGD with noise std σ. MLM loss is obtained by validating the model on test split.Untrained BERT-base-cased model's MLM loss on MIMIC-III test is **3.49.** DP models for $\sigma \leq 1e-3$ have improved model utility due to DP training. For non-DP models increasing epochs increases model utility, whereas for DP model utility decreases with increased epochs.

5.3 Does group level information provide more advantage to the membership inference attacker ?

To study the phenomenon of privacy degradation due to correlated data samples, we use admission and patient level membership inference attacks as defined in Section 4.1.

Experimental Details: We use the patient and admission ids provided in MIMIC-III dataset to construct the groups. For UMM data, we did not have access to admission level information, so each EHR is treated as one admission. The patient ids are used to create patient groups. The threshold used for the A-BBA and P-BBA attack adversaries are the same as that computed for S-BBA attack.

Results: Group level privacy attacks for MIMIC-III and UMM are provided in Figures 2 and 3. Comparing these plots with Figure 1 we observe an increase in the empirical privacy leakage for all models when group-level averages are used. The performance of larger group samples (patient-level aggregation) is higher than smaller group samples (admission or EHR level aggregation). This gap is more pronounced for UMM (figure in Appendix) where the EHR level group is much smaller than the patient-level group.

6 Discussion

Leakage in Non-DP models: Our results are consistent with existing efforts connecting overfitting and privacy leakage (Yeom et al., 2018). Large LM models pretrained with dataset sizes in millions tend to have lower generalization gaps compared to supervised deep neural net models trained on more limited, expensive labelled data. Therefore we see a maximum sample level privacy leakage of around 7% for non-DP CLMs. Additionally, losses like MLM use a random mask to produce a self-supervised sample. This random mask adds an additional layer of obfuscation, thereby reducing the empirical privacy leakage. GPT2 model, which uses an auto-regressive loss has higher privacy leakage than an equal sized MLM model. DistilBERT LM model that has the smallest network size, shows the lowest privacy leakage as expected.

Trade-offs with privacy budgets: For theoretical privacy budgets computed using RDP accountant, only DP-SGD models with $\sigma \in 1, 2$ provide an $\epsilon < 1$. All other models with $\sigma < 1.0$ have ϵ values greater than 1, with the lower values like $\sigma = 1e-3$ having ϵ values as high as 10^4. ϵ computed for all models are provided in Appendix B. The privacy leakage predicted by RDP is a vast overestimate of our empirical PL values. All models with $\sigma > 1e-3$ show negligible empirical privacy leakage for both black-box and white-box attacks. Privacy budgets computed through DP accountants are vastly conservative and in the current form are not useful for deep neural net based NLP models. Based on our experiments, we find that keeping σ higher than 1e-3 may prevent non-trivial privacy leakages from white-box membership inference and black-box group membership inference attacks. Conversely we see high model utility (close to non-DP model utility) for low σ values (Table 3). According to group DP, for a group size of k, ϵ values should degrade to $50 \times \epsilon$ and δ should degrade even more. While group level membership inference attacks increase the privacy leakage, they do not show such significant privacy degradation. This may be partially due to the nature of the EHR dataset, that has large text spans of standard instructions and copy-pasted text. This text results in several nearly-identical data samples across patients in both training and OOT data sets. It is also possible that our group-level attack is too simplistic, and more sophisticated methods are needed.

Practical Attack Scenarios: As discussed in Section 1, membership inference attacks against an NLP system may not be applicable in a practical

scenario due to the high dimensional input space and the lower rank semantic space of sentences. Our attacks can be used practically when partial information about the patient, such as a snippet of their EHR, is known to the attacker. Nevertheless, membership inference attacks serve as a good standardized framework for comparing relative privacy leakage of models and study the effects of enforcing privacy constraints on them.

Privacy for Rare Disease Patients: To understand if patients the behaviour of privacy leakage for outlier patients (or those with rarer disease profiles), we examine the black-box patient membership inference attacks in more detail for MIMIC-III trained CLMs. We group patients into buckets based on the rarity of their disease profiles. We use the coded ICD (International Classification of Diseases) entries for each patient to estimate the probability of that patient's disease profile. Details for the probability estimation are provided in Appendix A.

We divide the patients into 100 buckets based on log-normalized ranges of disease profile probability. We discard buckets with less than 10 patients. We calculate privacy leakage using Equation 2 for each bucket individually. We use the same threshold used in S-BBA attack. We observe that for most non-DP models with epoch > 1, ***bucket's average probability is inversely correlated to privacy leakage***. Patients with rarer disease profiles tend to exhibit higher privacy leakage. The relevant graphs are provided in Appendix B.

7 Related Work

Pre-trained Language Models are extensively used in current state-of-the-art pipelines for different NLP tasks. These models are generally pre-trained using large unsupervised text from multiple domains (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019). For domain-specific tasks such as MedNLI or emrQA, these LMs are further fine-tuned on domain-specific data which could be private and not publicly available (Alsentzer et al., 2019; Lee et al., 2020; Beltagy et al., 2019).

The language models when fine-tuned on private data and released publicly are prone to privacy attacks and can leak individual training examples (Carlini et al., 2020). This poses even higher risks when the models are trained on private clinical data as it may lead to leaking sensitive patient information. *Membership inference attack* (Hisamoto et al., 2020; Nasr et al., 2019) is the most common privacy attack, where the adversary tries to predict whether a particular example was used for training the model or not. *Model inversion attack* reconstruct representative views of the training examples. (Carlini et al., 2020) showed that large LMs such as GPT-2 can memorize sentences and using extraction attack were able to extract verbatim sequences from the training set including identifiable information (public) such as names, phone numbers and email addresses. To counter these attacks, differential private training techniques (Abadi et al., 2016; McMahan et al., 2018) are used to train deep learning models. These differentially private training techniques can lead to a reduction in model accuracy (Jayaraman and Evans, 2019) and increase the pre-training time, which is quite significant for large LMs.

8 Conclusion and Future Work

Our experiments show that large LM models exhibit lower privacy leakages compared to supervised DNNs studied by Jayaraman and Evans (2019); Shokri et al. (2017); Yeom et al. (2018).

BERT and GPT2 do still exhibit non-trivial privacy leakages of up to 7% for white-box attacks. In private datasets like EHR repositories, this suggests potential susceptibility to training data extraction attacks. While the exact privacy leakage requirements are subjective to the application, we should strive for lower than 1% PL values in clinical domain. Our experiments with DP-SGD training show that when used with low σ values, it can reduce empirical privacy leakages to less than 1% while maintaining improvements in model utility during training. We also show that patients with rarer diseases may exhibit higher privacy leakages. Our work represents the first standardized comparison of privacy leakages in commonly used LM architectures.

Future Work: This work is the first step in studying privacy properties of CLMs, and has several future directions. A major research direction is to propose better model-agnostic and model-specific membership inference and training data extraction attacks. Our evaluation framework, along with a publicly available MIMIC-III is well suited to provide a testing framework for privacy attacks. Our future work involves establishing such a benchmark that provides the DP and non-DP models to researchers who already have access to MIMIC-

III. An automated testing framework that computes patient-level privacy leakage can be established using our defined data splits. Another research direction is the use of data-dependent DP methods such as PATE (Papernot et al., 2018), which may be able to provide better (ϵ, δ) values. Use of PATE for training large LMs is prohibited by the high computational cost of multiple-teacher training in PATE, and further study to reduce the computational complexity of such methods is required.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy.

Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167.

Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1895–1912.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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 preprint arXiv:1907.11692*.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private recurrent language models. In *International Conference on Learning Representations*.

Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE.

Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*.

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. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

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)*, pages 3–18. IEEE.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE.

A Experimental Details

A.1 Training Details

We use the same procedure as used by clinicalBERT and bioBERT to train our non-DP models. We split the total number of patients for each data source into a 70-30 train-test split. The dataset statistics for each data source is presented in Table 1. BERT, DistilBERT and GPT2 are trained on all notes from training split using non-DP SGD based training. These models are referred to as non-DP models. The learning rate of $1e-5$ is used and the models are trained for 4 epochs.

We use DP-SGD with a gradient clipping threshold of 1 and varying σ values to produce $(\epsilon, \delta)-$DP models with varying ϵ budgets. Allowed failure probability δ is kept constant at $1e-6$. To limit the privacy degradation due to group-level privacy losses, we limit each patient's contribution to the training dataset to 50 randomly selected samples. To manage additional computational complexity introduced by per sample gradient clipping, we run DP-SGD training for 3 epochs instead of 4 in the case of non-DP models. We refer to these models as DP models in our paper.

A.2 Estimation of Patient Disease Profile Probability

ICD coding uses a standardized vocabulary (ICD9 or ICD10) to document the relevant disease for a patient. For an admission, a patient may be coded with multiple ICD codes depending on the presentation of their symptoms and the course of diagnosis. To simplify our analysis, we assume that each ICD code is attributed independently. As a result the probability of observing a patient with $\mathbf{c} = \{c_1, ...c_k\}$ codes is

$$p(\mathbf{c}) = \prod_{\mathbf{c}} p(c_i) \prod_{\mathcal{C} \setminus \mathbf{c}} (1 - p(c_i)). \quad (4)$$

Here $\mathcal{C}$ is the set of all ICD codes and $\mathbf{c}$ is the set of ICD codes for a patient. Since we assume that each ICD code is attributed independently, maximum likelihood estimate of any code c' is

$$p(c') = \frac{\#\, of\ patients\ with\ code\ c'}{\#\, of\ total\ patients}. \quad (5)$$

The same procedure can be followed for admission or note level analysis.

A.3 MedNLI Experiment Details

All the model training details for MedNLI experiments are provided in Table 4. All the models were trained over 8 seeds and their final performance was calculated by averaging over the top three performance values. MedNLI is a classification dataset and hence accuracy is used as an evaluation metric for all models.

MedNLI Dataset	
Train datapoints	$12,626$
Test datapoints	$1,421$
Learning Rate	$5e-5$
Batch Size	12
Max Seq. Len.	200
Epochs	3

Table 4: Training details for MedNLI experiments.

A.4 Training Data Details

MIMIC-III (Johnson et al., 2016) is a publicly available critical care database made available via https://mimic.physionet.org/.

Hospital 2 data is collected by selecting patients who are suffering from any form of cardiovascular diseases or patients who are suffering from any cancerous diseases. The patients were selected with the help of ICD9 and ICD10 codes mentioned in their structured medical data. The ICD codes for cardiovascular diseases were used according to https://www.questdiagnostics.com/dms/Documents/Other/PDF_MI4632_ICD_9-10_Codes_for_Cardio_38365_062915.pdf. The ICD codes for cancerous diseases were used according to https://seer.cancer.gov/tools/conversion/2014/ICD9CM_to_ICD10CM_2014CF.pdf

VHA data was collected by randomly selecting patients who had hospital visits during the year 2020. All 2020 EHRs for the selected patients were extracted to produce the dataset.

B Detailed Results and Plots

B.1 RDP accounting for DP-SGD models

We use RDP (Mironov, 2017) accountant [2] to calculate ϵ values for $\delta = 1e-6$. For MIMIC-III DP BERT models (epoch 1) with σ values of 2.0, 1.0, 0.1, the ϵ values are 0.223, 0.626, 403. For

[2] https://github.com/tensorflow/privacy

epoch 3 the ϵ values are 0.223, 0.628, 733. Models with $\sigma < 0.1$ have very large ϵ values. Since RDP accounting is independent of model accounting, ϵ values for DistilBERT and GPT2 models show similar behavior. Group DP ϵ values can be obtained by multiplying ϵ with group size ($k = 50$ for our experiments).

The RDP calculated ϵ values for UMM BERT models (epoch 1) with σ values of 2.0, 1.0, 0.1, are 0.219, 0.553, 391. For epoch 3 the ϵ values are 0.220,0.553, 619. Similar to MIMIC-III ϵ values, models with $\sigma < 0.1$ have very large ϵ values.

B.2 Detailed Results for MIMIC-III Dataset

In this section we provide privacy leakage, and model utility results for non-DP and DP models trained on MIMIC-III dataset.

B.2.1 Privacy Leakage Results

Figures 5, 6 and 4 show sample level black box (S-BBA) attack privacy leakage for BERT, DistilBERT and GPT2 models.

Figures 8, 9 and 7 show admission level black box (A-BBA) attack privacy leakage for BERT, DistilBERT and GPT2 models.

Figures 11, 12 and 10 show patient level black box (P-BBA) attack privacy leakage for BERT, DistilBERT and GPT2 models.

Figures 14, 15 and 13 show S-AWBA attack privacy leakage for BERT, DistilBERT and GPT2 models.

Figures 17, 18 and 16 show S-GWBA attack privacy leakage for BERT, DistilBERT and GPT2 models.

B.2.2 LM Loss Results

Figure 19 and 20 provide masked language modeling loss for allBERT and distilBERTDP models, trained on MIMIC-III. Figure 21 shows autoregressive LM loss for GPT2 DP models.

B.2.3 Privacy Leakage based on Patient Profiles

In this section we provide detailed results for correlations between patient privacy profiles and privacy leakage. We divide MIMIC-III admissions and patients into buckets based on the log normalized probability of their disease profile. The probability of an admission or patient's disease profile is obtain using the procedure outlined in Section A.2. Plots of correlation between a bucket's (admission-level) privacy leakage and its patient disease probability are in Figure 22, 23 and 24. Trends for patient level disease profile and privacy leakage are same as shown in the aforementioned plots.

B.2.4 MedNLI Results for MIMIC-III Models

In Table 5, we see that Non-DP BERTmodel achieves an accuracy of 0.8023 on MedNLI dataset (Romanov and Shivade, 2018) and BERT-base-cased achieved a performance of 0.776 per (Alsentzer et al., 2019). All the DP trained model achieved a performance higher than BERT-base-cased model and quite close to Non-DP trained BERTmodel which shows the utility of fine-tuning models with differentially private training technique.

B.3 Detailed Results for UMM Dataset

We see similar privacy leakage pattern for UMM Dataset as MIMIC Dataset. We observe highest privacy leakage for Non-DP GPT2 model using gradient-based white box attack (S-GWBA) as shown in Fig. 32.

Fig. 27 show sample level black box attack (S-BBA) privacy leakages for BERT, DistilBERT and GPT2 models.

Fig. 28 show hospital admission level black box attack (A-BBA) privacy leakages for BERT, DistilBERT and GPT2 models.

Fig. 29 show patient level black box attack (P-BBA) privacy leakages for BERT, DistilBERT and GPT2 models.

Fig. 30 show negative log likelihood loss for BERT, DistilBERT and GPT2 models.

Fig. 31 show sample level attention-based white box attack (S-AWBA) privacy leakages for BERT, DistilBERT and GPT2 models.

Fig. 32 show sample level gradient-based white box attack (S-GWBA) privacy leakages for BERT, DistilBERT and GPT2 models.

B.4 Results for VHA Dataset

Due to computational and access constraints for VHA dataset, we were only able to evaluate black-box-attack on BERT model for VHA data. The patient and sample-level black-box privacy leakages are provided in Table 6. We see patterns consistent with MIMIC-III and UMM results. Group-level privacy leakage is higher than sample-level leakage, and increasing epochs increases the privacy leakage.

B.5 Useful Notations

LM: Language Model
CLM: Clinical Language Model
DP: Differentially Private
RDP: Rényi Differential Privacy
PL: Privacy Leakage
OOT: out-of-training
A-BBA: Admission level Black Box Attack
S-BBA: Sample level Black Box Attack
P-BBA: Patient level Black Box Attack
S-GWBA: Sample level Gradient-based White Box Attack
S-AWBA: Sample level Attention-based White Box Attack
MedNLI: A Natural Language Inference Dataset for the Clinical Domain (Romanov and Shivade, 2018)

		BERTAccuracy		
	σ	Epoch 1	Epoch 2	Epoch 3
Non-DP	-	0.7922	0.7938	0.8023
DP	1e-4	0.8095	0.7978	0.8086
	1e-3	0.7903	0.8037	0.7952
	1e-2	0.7933	0.8069	0.7926
	1e-1	0.8055	0.7954	0.7861
	1.0	0.7926	0.8023	0.7936
	2.0	0.7962	0.8015	0.7929
	1e-20	0.7884	0.8098	0.7896

Table 5: Results for all three epochs of DP and Non-DP BERTmodels. Gradient clip value for DP models is set to 1 and group samples are limited to 50. The Non-DP trained BERT model at Epoch 3 is similar to clinicalBERT model (Alsentzer et al., 2019) which has the reported performance of **0.808** on MedNLI dataset. The BERT-base-cased model which is not finetuned on MIMIC-III data achieved an accuracy of **0.776** per (Alsentzer et al., 2019).

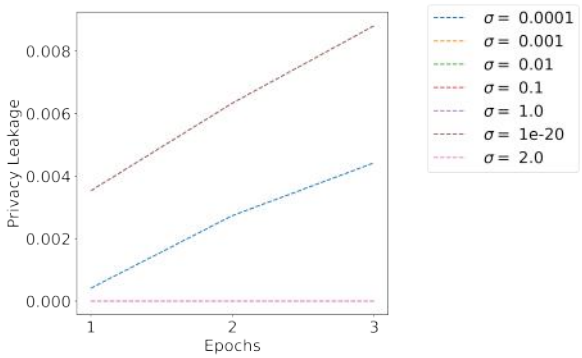


Figure 4: Privacy leakage for GPT2 models with varying σ values using MIMIC-III data. Leakage obtained using sample level black box attack (S-BBA). Gradient clip value is 1, group samples are limited to 50.

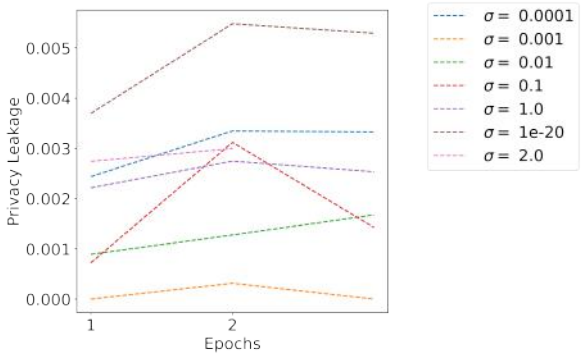


Figure 5: Privacy leakage for BERT models with varying σ values using MIMIC-III data. Leakage obtained using sample level black box attack (S-BBA). Gradient clip value is 1, group samples are limited to 50.

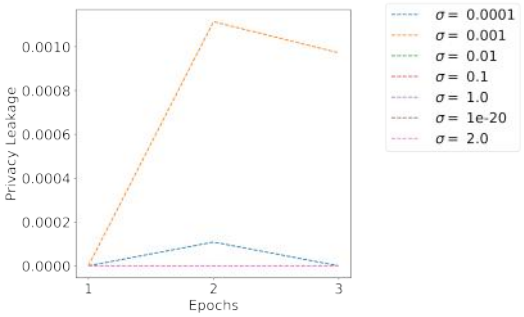


Figure 6: Privacy leakage for DistilBERT models with varying σ values using MIMIC-III data. Leakage obtained using sample level black box attack (S-BBA). Gradient clip value is 1, group samples are limited to 50.

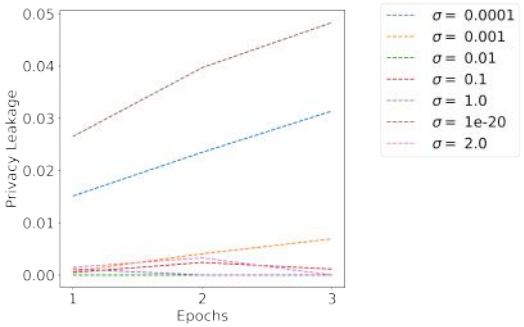


Figure 7: Group Privacy leakage for GPT2 models with varying σ values using MIMIC-III data. Leakage obtained using admission level black box attack (A-BBA). Gradient clip value is 1, group samples are limited to 50.

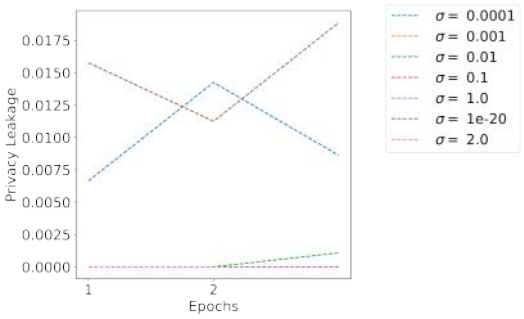


Figure 8: Group privacy leakage for BERT models with varying σ values using MIMIC-III data. Leakage obtained using admission level black box attack (A-BBA). Gradient clip value is 1, group samples are limited to 50.

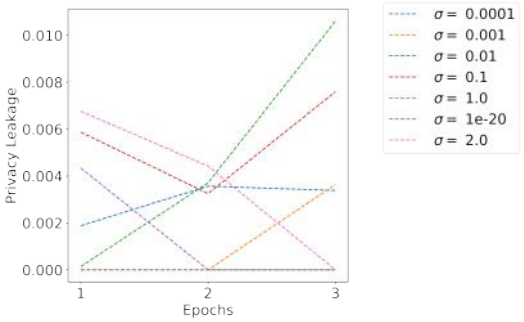


Figure 9: Group privacy leakage for DistilBERT models with varying σ values using MIMIC-III data. Leakage obtained using admission level black box attack (A-BBA). Gradient clip value is 1, group samples are limited to 50.

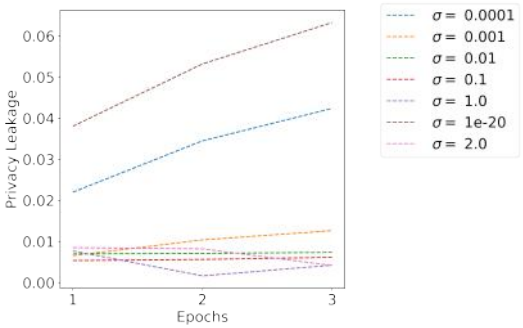


Figure 10: Group privacy leakage for GPT2 models with varying σ values using MIMIC-III data. Leakage obtained using patient level black box attack (P-BBA). Gradient clip value is 1, group samples are limited to 50.

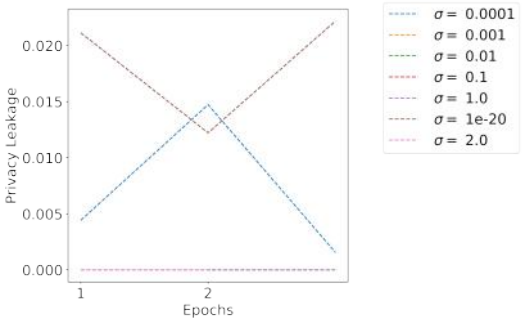


Figure 11: Group privacy leakage for BERT models with varying σ values using MIMIC-III data. Leakage obtained using patient level black box attack (P-BBA). Gradient clip value is 1, group samples are limited to 50.

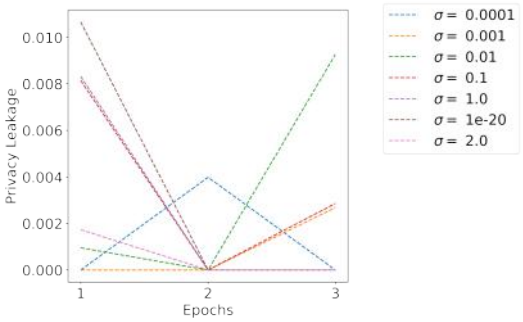


Figure 12: Group privacy leakage for DistilBERT models with varying σ values using MIMIC-III data. Leakage obtained using patient level black box attack (P-BBA). Gradient clip value is 1, group samples are limited to 50.

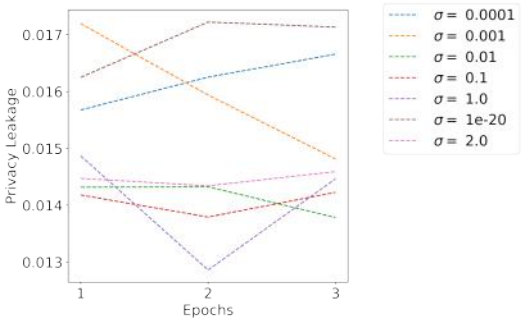


Figure 13: Privacy leakage for GPT2 models with varying σ values using MIMIC-III data. Leakage obtained using Attention based white box attack (S-AWBA). Gradient clip value is 1, group samples are limited to 50.

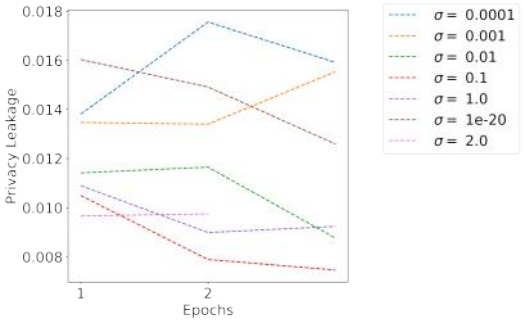


Figure 14: Privacy leakage for BERT models with varying σ values using MIMIC-III data. Leakage obtained using Attention based white box attack (S-AWBA). Gradient clip value is 1, group samples are limited to 50.

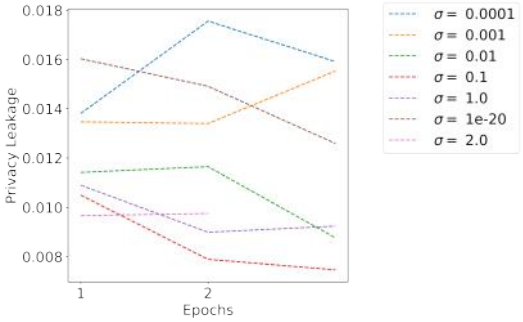


Figure 15: Privacy leakage for DistilBERT models with varying σ values using MIMIC-III data. Leakage obtained using Attention based white box attack (S-AWBA). Gradient clip value is 1, group samples are limited to 50.

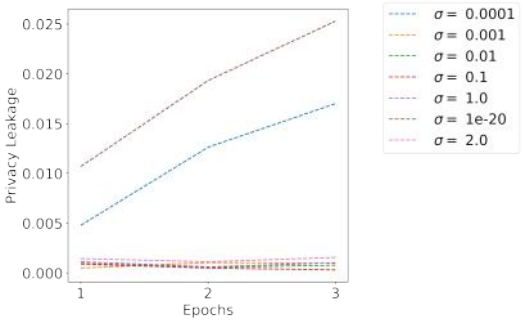


Figure 16: Privacy leakage for GPT2 models with varying σ values using MIMIC-III data. Leakage obtained using Gradient based white box attack (S-GWBA). Gradient clip value is 1, group samples are limited to 50.

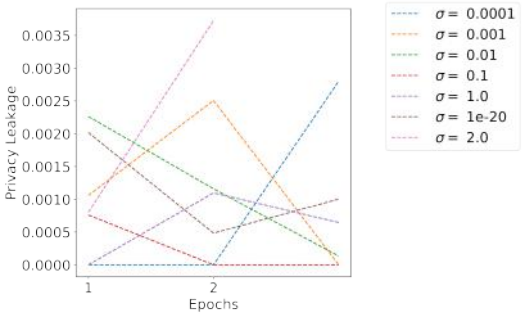


Figure 17: Privacy leakage for BERT models with varying σ values using MIMIC-III data. Leakage obtained using Gradient based white box attack (S-GWBA). Gradient clip value is 1, group samples are limited to 50.

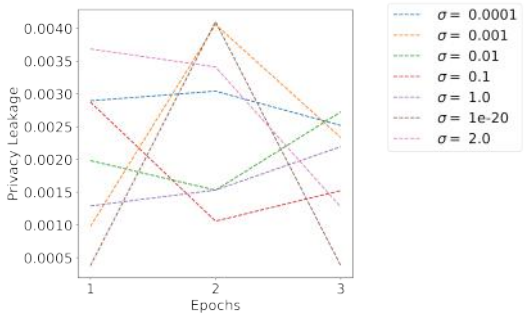


Figure 18: Privacy leakage for DistilBERT models with varying σ values using MIMIC-III data. Leakage obtained using Gradient based white box attack (S-GWBA). Gradient clip value is 1, group samples are limited to 50.

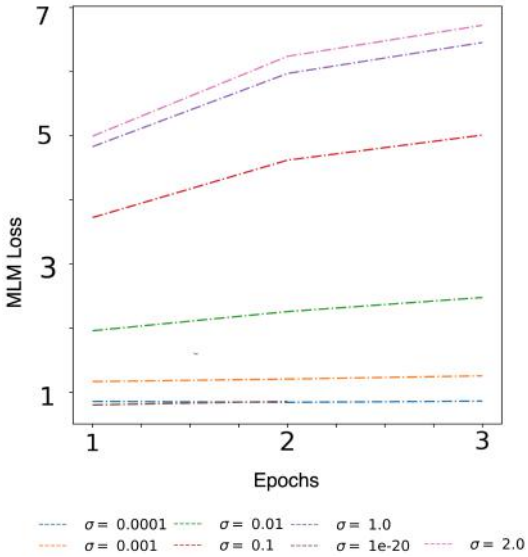


Figure 19: Test data LM loss for DP BERT models with varying σ values using MIMIC-III data. Gradient clip value is 1, group samples are limited to 50. Loss of BERT-base-cased model that is not trained on clinical data is 3.49. All DP models with less than 3.49 loss plots have increased model utility due to DP training. Non-DP CLM model loss is 0.60,0.57,0.55,0.53 for epoch 1 - 4. Non-DP mlm loss is lower than all DP models.

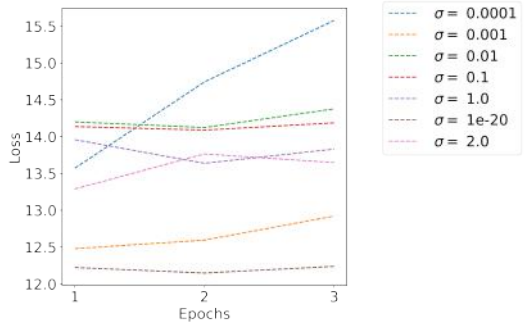


Figure 20: Test data LM loss for DP DistilBERT models with varying σ values using MIMIC-III data. Gradient clip value is 1, group samples are limited to 50. Loss of DistilBERT-base model that is not trained on clinical data is 3.49. All DP models with less than 3.49 loss plots have increased model utility due to DP training. Non-DP CLM model loss is 0.63,0.59,0.58,0.58 for epoch 1 - 4. Non-DP mlm loss is lower than all DP models.

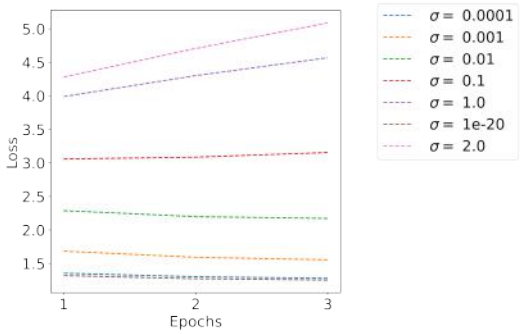


Figure 21: Test data LM loss for GPT2 models with varying σ values using MIMIC-III data. Gradient clip value is 1, group samples are limited to 50. Loss of GPT2-base model that is not trained on clinical data is 3.49. All DP models with less than 3.49 loss plots have increased model utility due to DP training. Non-DP CLM model loss is 1.2,1.19,1.17,1.16 for epoch 1 - 4.

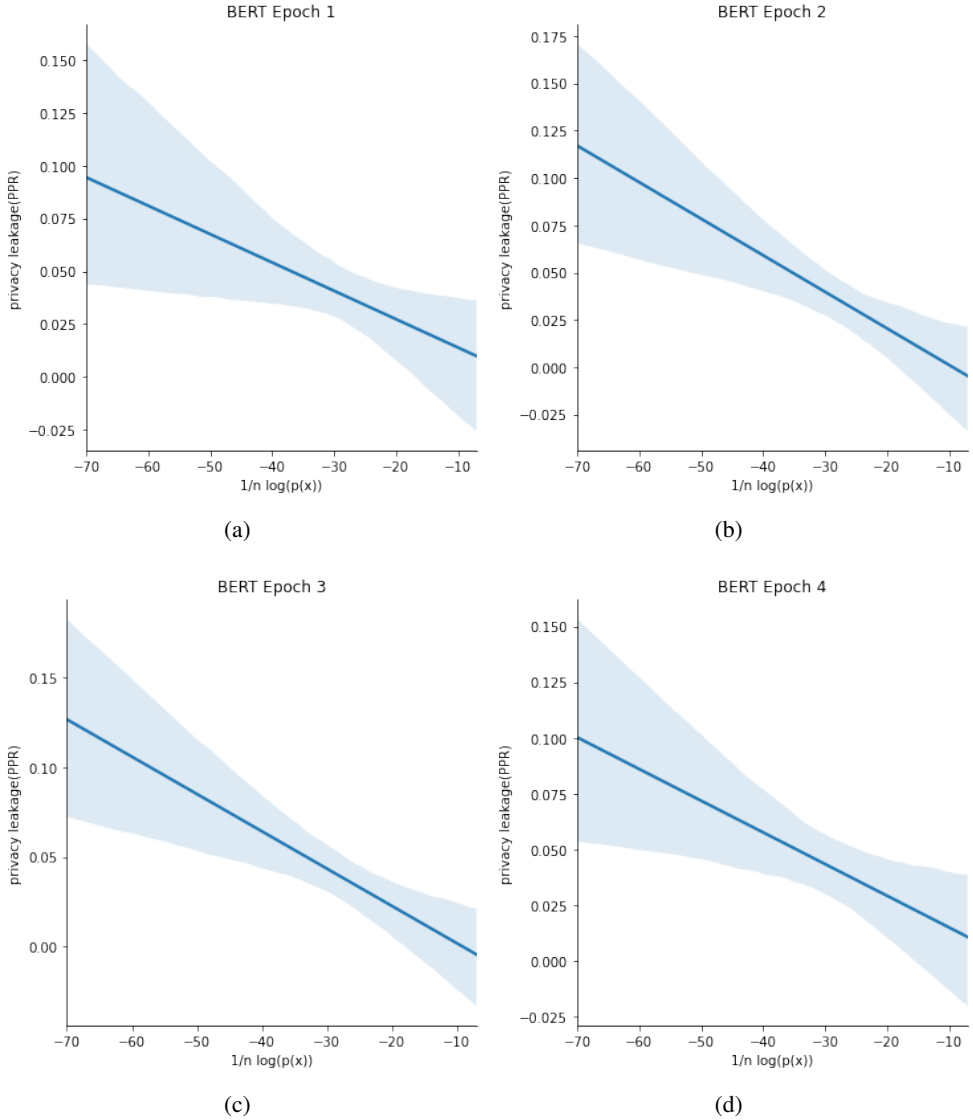
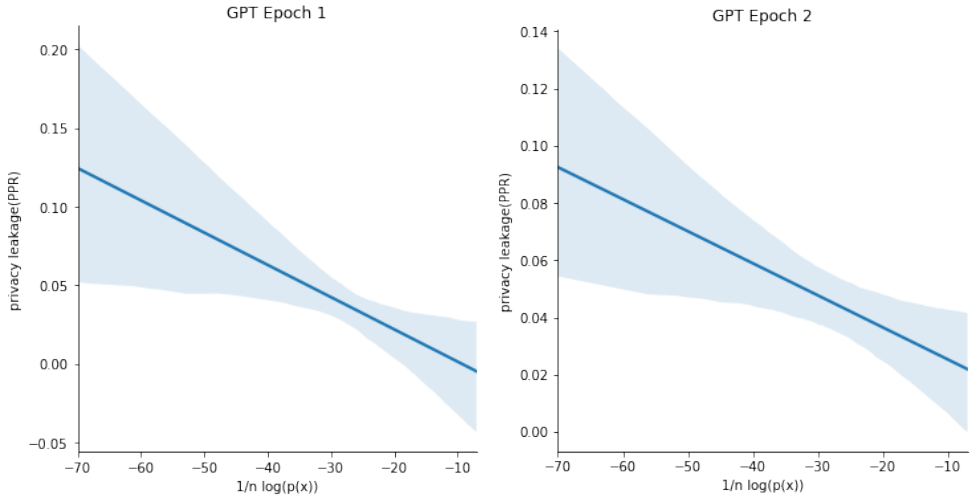
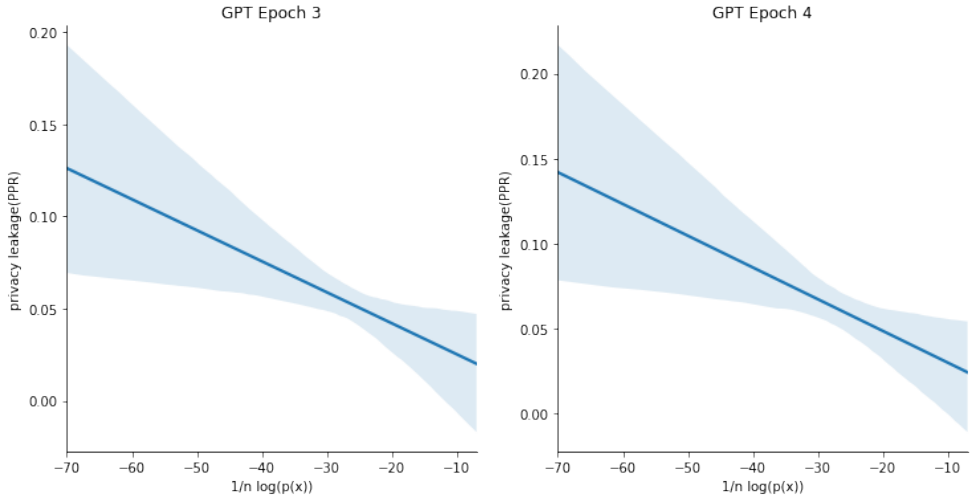


Figure 22: Privacy leakage vs log-normalized probability plots for Non-DP trained BERT models on MIMIC-III data. We see the negative correlation between probability of disease profile (admission-level) and privacy leakage.



(a) (b)



(c) (d)

Figure 23: Privacy leakage vs log-normalized probability plots for Non-DP trained GPT2 models on MIMIC-III data. We see the negative correlation between probability of disease profile (admission-level) and privacy leakage.

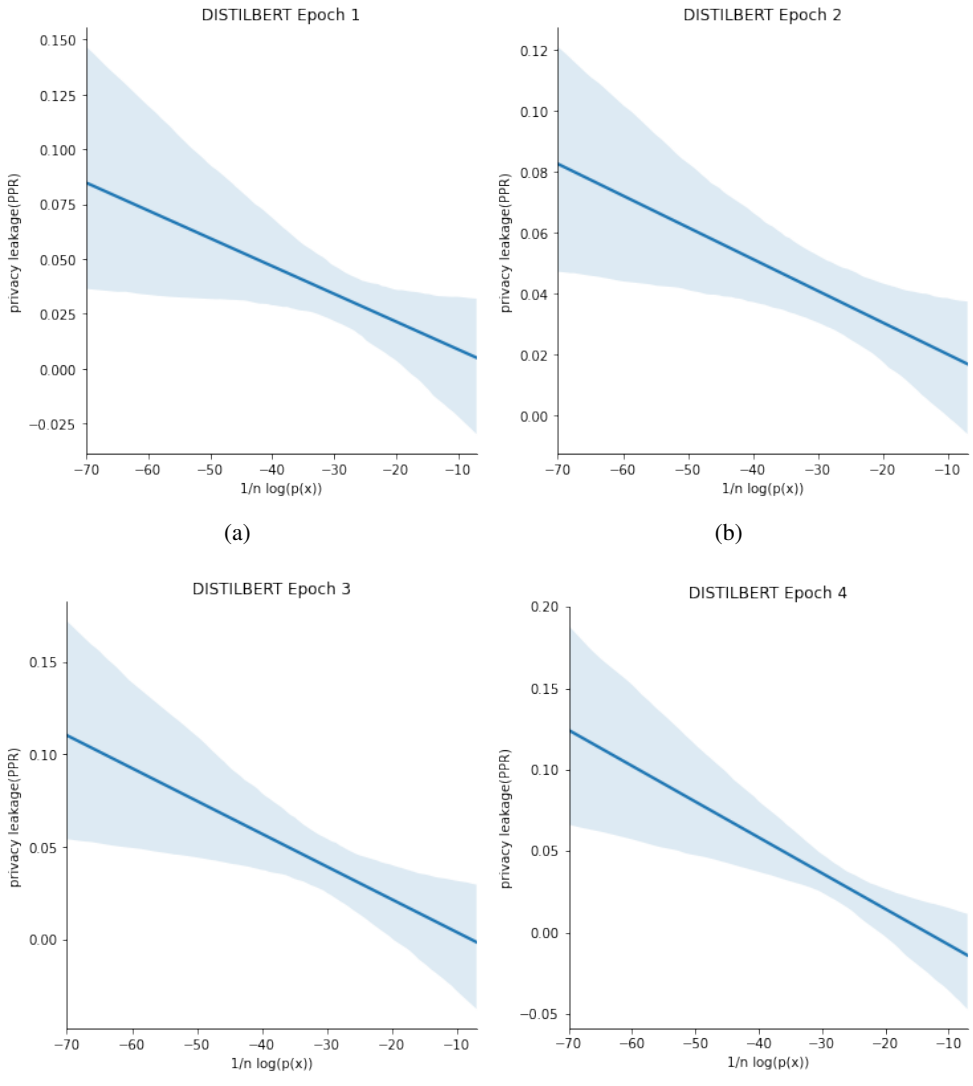


Figure 24: Privacy leakage vs log-normalized probability plots for Non-DP trained DistilBERT models on MIMIC-III data. We see the negative correlation between probability of disease profile (admission-level) and privacy leakage.

BERT epoch	S-BBA 1	P-BBA
1	1.27	0.92
2	1.61	1.42
3	1.89	2.09
4	2.03	2.91

Table 6: Sample and patient-level black-box attack privacy leakage for BERT model using VHA data.

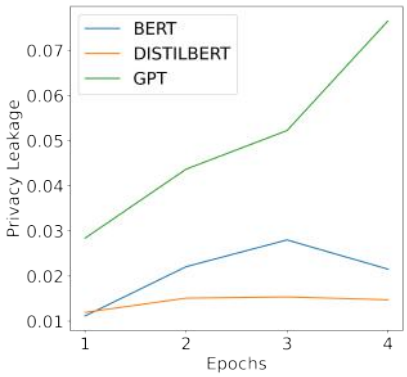


Figure 25: Black box attack privacy leakage for MIMIC-III data with hospital admission level aggregate.

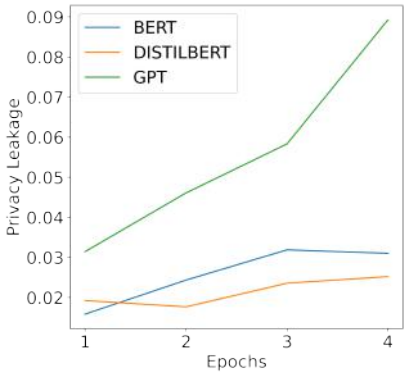


Figure 26: Black box attack privacy leakage for MIMIC-III data with patient level aggregate.

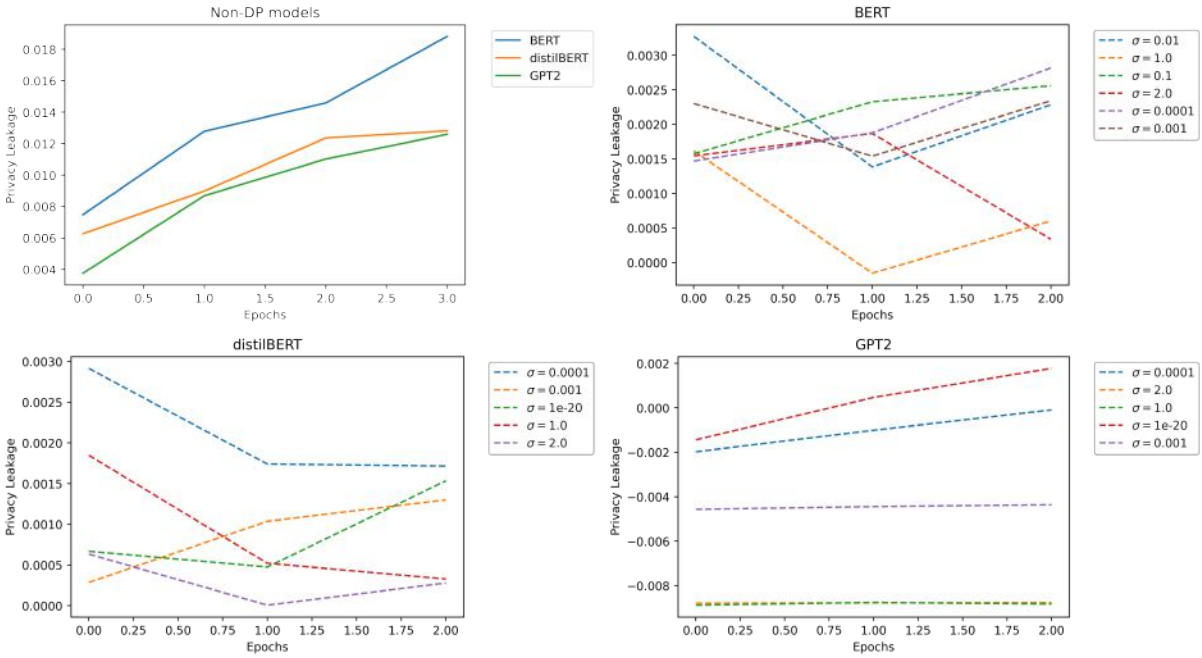


Figure 27: Sample-level Black box attack (S-BBA) privacy leakage using Hospital 2 data for Non-DP and DP trained variants of BERT, DistilBERTand GPT2. For all DP models, the Gradient clip value is set to 1 and group samples are limited to 50.

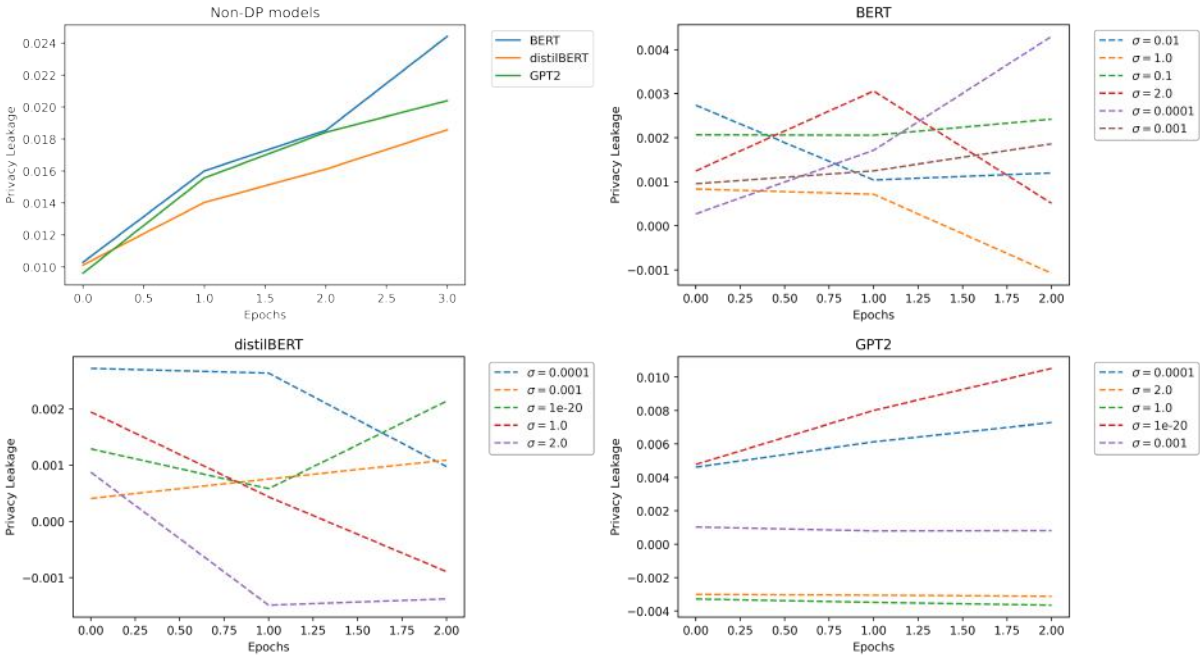


Figure 28: Admission-level Black box attack (A-BBA) group privacy leakage using Hospital 2 data for Non-DP and DP trained variants of BERT, DistilBERTand GPT2. For all DP models, the Gradient clip value is set to 1 and group samples are limited to 50.

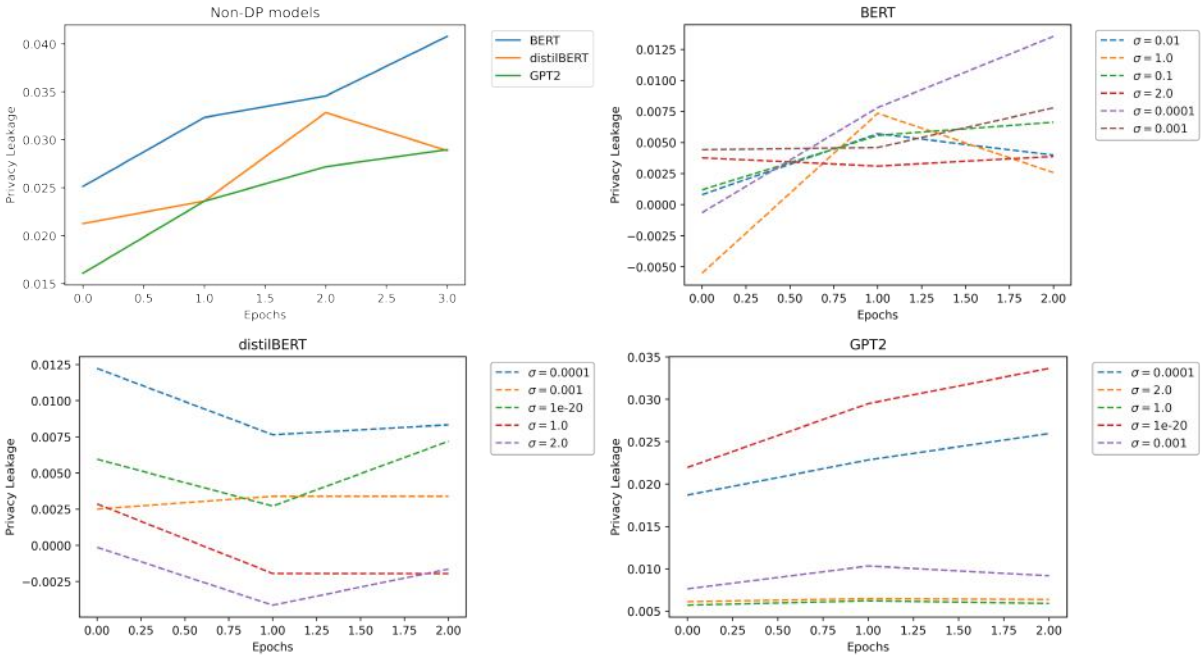


Figure 29: Patient-level Black box attack (P-BBA) group privacy leakage using Hospital 2 data for Non-DP and DP trained variants of BERT, DistilBERTand GPT2. For all DP models, the Gradient clip value is set to 1 and group samples are limited to 50.

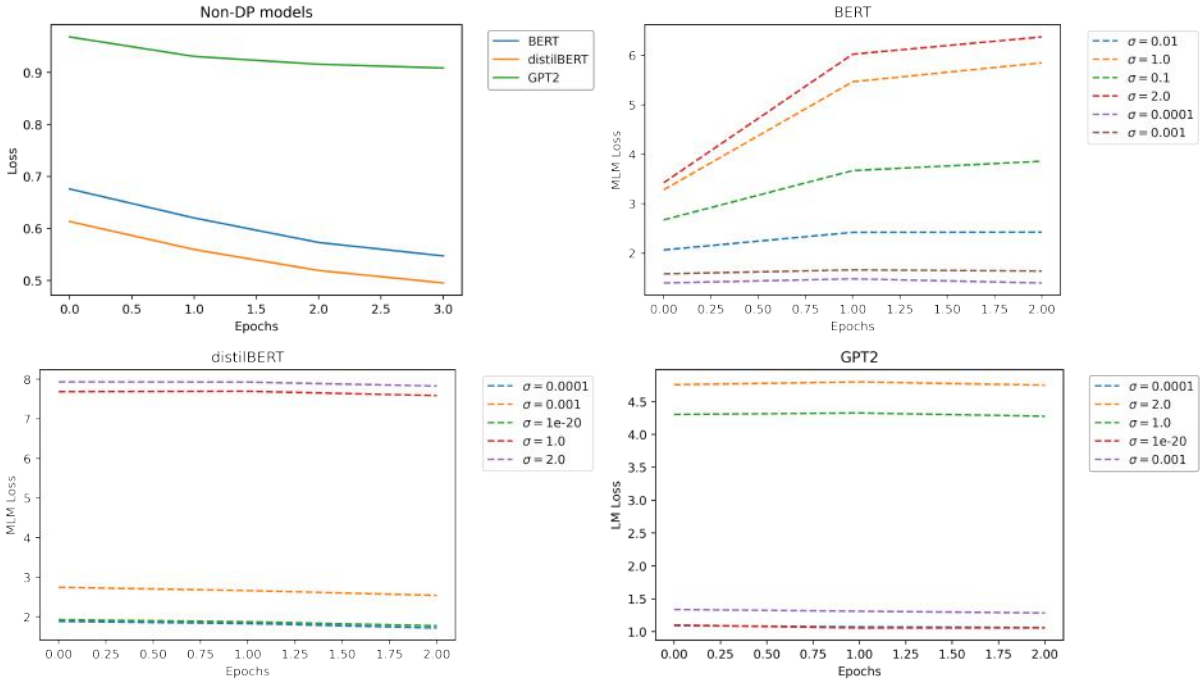


Figure 30: Negative Log Likelihood Loss for Non-DP CLMs and their DP trained variants with varying σ values. For all DP models, the Gradient clip value is set to 1 and group samples are limited to 50.

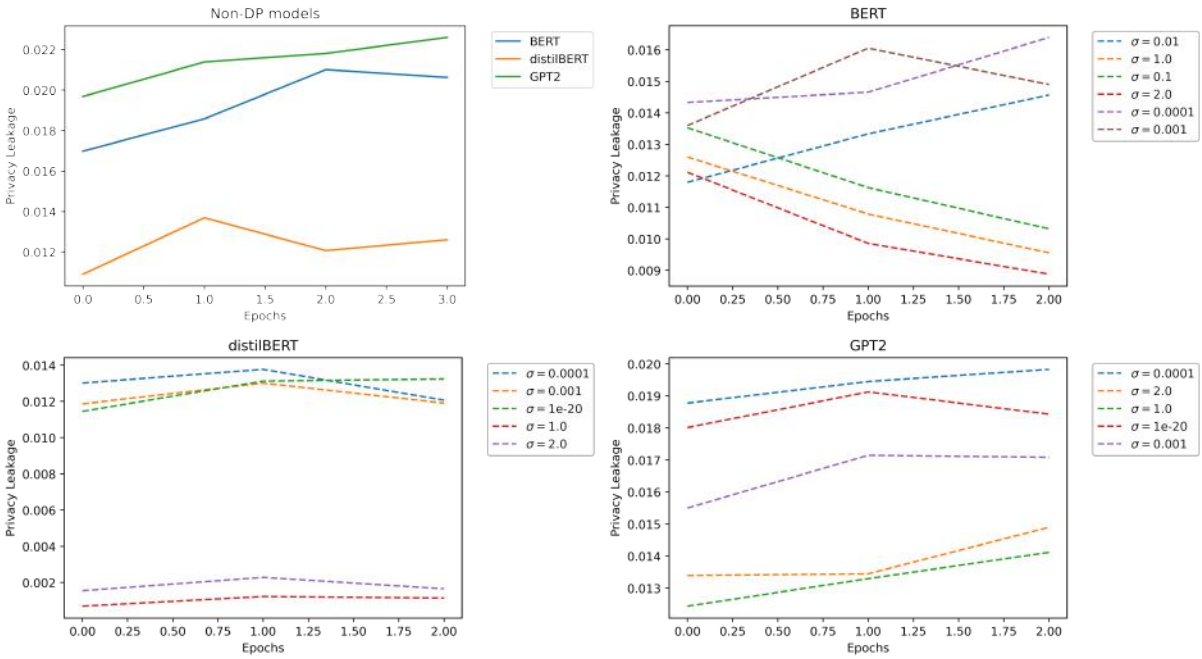


Figure 31: Privacy leakage using Attention-based white-box attack (S-AWBA) for Hospital 2 at sample level. Models shown are Non-DP and DP trained variants of BERT, DistilBERTand GPT2. For all DP models, the Gradient clip value is set to 1 and group samples are limited to 50.

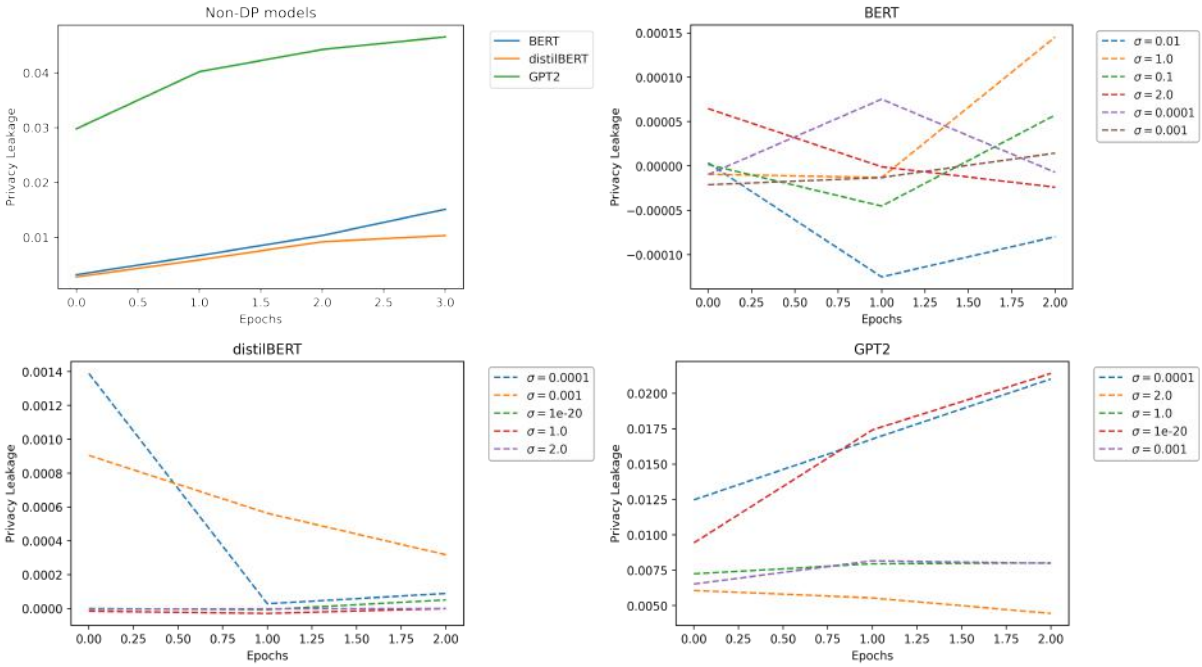


Figure 32: Privacy leakage using Gradient-based white box attack (S-GWBA) for Hospital 2 at sample-level. Models shown are Non-DP and DP trained variants of BERT, DistilBERTand GPT2. For all DP models, the Gradient clip value is set to 1 and group samples are limited to 50.