KART: Parameterization of Privacy Leakage Scenarios from Pre-trained Language Models

Anonymous ACL submission

Abstract

For the safe sharing pre-trained language models, no guidelines exist at present owing to the difficulty in estimating the upper bound of the risk of privacy leakage. One problem is that previous studies have assessed the risk for different real-world privacy leakage scenarios and attack methods, which reduces the portability of the findings. To tackle this problem, we represent complex real-world privacy leakage scenarios under a universal parameterization, *Knowledge, Anonymization, Resource, and Target* (KART). KART parameterization has two merits: (i) it clarifies the definition of privacy leakage in each experiment and (ii) it improves the comparability of the findings of risk assessments. We show that previous studies can be simply reviewed by parameterizing the scenarios with KART. We also demonstrate privacy risk assessments in different scenarios under the same attack method, which suggests that KART helps approximate the upper bound of risk under a specific attack or scenario. We believe that KART helps integrate past and future findings on privacy risk and will contribute to a standard for sharing language models.

1 Introduction

Recent natural language processing (NLP) has benefited from language models such as Transformer (Vaswani et al., 2017), GPT (Radford et al., 2018), and BERT (Devlin et al., 2019). However, we face a privacy risk when sharing language models since personal information in the pre-training data could be recovered from the models (Misra, 2019; Hisamoto et al., 2020; Carlini et al., 2019, 2021; Inan et al., 2021; Lehman et al., 2021; Vakili and Dalianis, 2021; Hoory et al., 2021).

No guidelines for publishing pre-trained language models have been established since we lack knowledge about the impact of a model release on privacy safety. This is in contrast to the data itself, for which standards of processing, sharing, and publishing have already been legislated in several countries (The United States Department of Health and Human Services, 2012; European Commission, 2018). Alternatively, model providers have published language models only when the pre-training data is free of sensitive personal information. In the biomedical domain, for example, BioBERT (Lee et al., 2019) and BioMegatron (Shin et al., 2020) are publicly available, both of which are pre-trained with biomedical articles. ClinicalBERT (Huang et al., 2019), BlueBERT (Peng et al., 2019), UTH-BERT (Kawazoe et al., 2021), and MS-BERT (D'Costa et al., 2020), which use manually anonymized clinical records, have also been published. However, EhrBERT (Li et al., 2019) and AlphaBERT (Chen et al., 2020) are also pre-trained with clinical records but have not been released.

How can we decide whether a language model is safe enough to share? Studies have assessed the privacy risk under various attacks and real-world privacy leakage scenarios, but it has been difficult to integrate them to estimate the upper bound of the risk. Moreover, no studies have directly compared the risk under the same definition of privacy leakage with different attack methods or scenarios. We attribute these shortcomings to a lack of methodology to clarify the presupposed scenarios.

To address the issue, we represent a privacy leakage scenario as a set of primary factors under the *Knowledge, Anonymization, Resource, and Target* (KART) parameterization, as in Figure 1. The primary factors are as follows: (i) *Prior knowledge* (K): information of the target people already known to the attacker; (ii) *Target information* (T): personal information that the attacker wishes to obtain; (iii) *Anonymization* (A): removal of personal information from the pre-training data; (iv) *Auxiliary resources* (R): resources used by the attacker other than the language model. We show that KART can simplify complex scenarios assumed in previ-

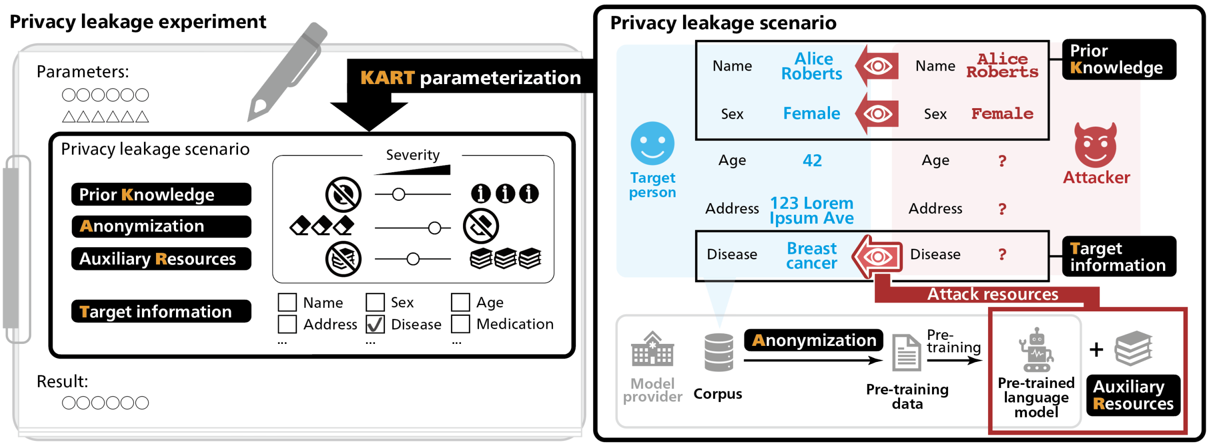


Figure 1: Scenario-aware privacy risk assessment using KART parameterization. Any privacy leakage experiment implicitly or explicitly assumes the scenario where an attacker who has *prior knowledge* about a target person attacks a pre-trained language model to obtain other personal *target information*. The attacker may also use *auxiliary resources*. The target information may or may not be in the pre-training data depending on the *anonymization*. Parameterizing the assumed scenario would improve the portability of the findings of the experiment.

ous studies and assist risk comparison in different scenarios or attacks.

Our contribution is an introduction of KART, which will enable comprehensive privacy risk assessment by improving scenario awareness and the portability of findings of past and future studies.

2 Related Work

2.1 Security of clinical records

Data is not secure after only deleting attributes that alone can determine from whom the data originated, such as names or IDs. The data can still be re-identified using remaining attributes and external data. Sweeney (2002) proposed the deletion or generalization of attributes to prevent re-identification.

Clinical records must be handled carefully as they contain sensitive health information that patients do not wish to spread unnecessarily (Fernández-Alemán et al., 2013; Mooney and Pejaver, 2018). Improper disclosure may lead to mental pain, biases in education and employment, and target marketing to vulnerable people (Price and Cohen, 2019). In the United States, the research use of health information is mainly regulated by the Federal Common Rule for human subject research and the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The HIPAA Privacy Rule[1] refers to sensitive clinical information as protected health information (PHI), such as clinical history, clinical test results, and genomes. General identifiers such as names and addresses are also PHI if linked with health information. The HIPAA Privacy Rule obligates the 18 identifiers listed in Appendix A to be removed from clinical records for the second usage (The United States Department of Health and Human Services, 2012).

2.2 Privacy attacks on language models

There are two types of prior study on the privacy risk in NLP according to the attack method.

The first is on attacks to obtain personal information in input texts from gradients or encoded representations (Zhu et al., 2019; Song and Raghunathan, 2020; Pan et al., 2020). The second is on attacks to restore personal information in the pre-training data from a model, which we review further in Section 4. They are divided into membership inference and model inversion. Membership inference is a prediction of whether a document was in the pre-training data (Shokri et al., 2017; Misra, 2019; Hisamoto et al., 2020). Model inversion is an estimation of specific attributes of target people (Fredrikson et al., 2015; Carlini et al., 2019, 2021; Inan et al., 2021; Lehman et al., 2021; Vakili and Dalianis, 2021; Hoory et al., 2021).

2.3 Generalization in privacy risk evaluation

It is difficult to estimate the upper bound of the privacy risk from pre-trained language models covering numerous privacy leakage scenarios. Several universal evaluation methods have been introduced.

Carlini et al. (2019) proposed *exposure*, the frequency with which natural language generation (NLG) reproduces canary sequences added to the pre-training data. Inan et al. (2021) used the per-

[1] https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

Table 1: Examples of the possible values of the K, A, R, and T factors in KART parameterization.

Factor	Value	Scenario
K	$\varnothing$	The attacker has no prior knowledge
	$\{(\mathit{age}, \mathit{sex})_p\}_{p \in \mathcal{P}}$	The attacker already knows the age and sex of the target people
A	f_{HIPAA}	The pre-training data is anonymized under the HIPAA Privacy Rule
	id	The pre-training data is not anonymized at all
R	$\varnothing$	The attacker has no resource other than $\mathcal{M}$
	$\{\widetilde{\mathcal{D}}_{\text{train}}\}$	The attacker has obtained a corpus with a similar distribution to the pre-training data
	$\{\mathit{full\ name}_{\overline{p}}\}_{\overline{p} \in \overline{\mathcal{P}}}$	The attacker knows the full name of the non-target people in the pre-training data
T	$\{\mathit{address}_p\}_{p \in \mathcal{P}}$	The attacker predicts the address of the target people (model inversion)
	$\{(\mathit{full\ name}, \mathit{address})_p\}_{p \in \mathcal{P}}$	The attacker predicts the full name and address of the target people together (model inversion)
	$\{\mathbb{1}[d \in \mathcal{D}_{\text{train}}]\}_{d \in \mathcal{D}}$	The attacker predicts whether a document $d \in \mathcal{D}$ was in the pre-training data (membership inference)

plexity with which a model $\mathcal{M}$ outputs exact substrings of the pre-training data. If the substrings derive from a single person and $\mathcal{M}$ gives a far lower perplexity than other models, the disclosure can be actual privacy leakage rather than coincidental.

Differential privacy (Dwork, 2006) is a constraint on privacy risk integrated with privacy mechanisms. In machine learning, a model $\mathcal{M}$ pre-trained on a dataset D is ϵ-differentially private if, for any adjacent dataset D' different from D in a single record, the probability that a model $\mathcal{M}'$ pre-trained on D' is distinguished from $\mathcal{M}$ never exceeds the upper bound defined by ϵ. Model providers can determine the value of ϵ beforehand and ensure privacy by perturbing a model correspondingly during training (Abadi et al., 2016; Kerrigan et al., 2020). Differential privacy is mathematically guaranteed to be robust to any prior knowledge of the attacker. Hoory et al. (2021) assessed the practical privacy risk of an ϵ-differentially private language model using the *exposure* metric (Carlini et al., 2019).

3 KART Parameterization

We propose KART to characterize a privacy leakage scenario with four primary factors. Refer to Table 1 for examples of the values of each factor.

Personal information We denote personal information as $\mathit{category}_{\mathit{person}} = \mathit{value}$. For example, "$\mathit{full\ name}_{p_0}$ = Alice Roberts" means that the full name of the person p_0 is Alice Roberts. We denote the universal set of all personal information in the world as I_U: $I_U = \{\mathit{category}_{\mathit{person}}\}_{\forall \mathit{person}, \forall \mathit{category}}$.

Let $\mathcal{D}_{\text{private}}$ denote a corpus, all of whose documents are used for pre-training, and let $\mathcal{P}$ be the set of people in $\mathcal{D}_{\text{private}}$. We denote the set of the personal information in $\mathcal{D}_{\text{private}}$ as $I_{\mathcal{D}_{\text{private}}}$ $(\subset I_U)$.

***A* factor: anonymization** The A factor is an operation, which we denote as a, to anonymize $\mathcal{D}_{\text{private}}$ to build pre-training data. Examples of a are the complete manual deletion of personal information under the HIPAA Privacy Rule ($a = f_{\text{HIPAA}}$), automated anonymization, and no operation at all as long as the model is strictly kept private ($a = \text{id}$).

We denote the pre-training data as $\mathcal{D}_{\text{train}} = a(\mathcal{D}_{\text{private}})$ and the set of the remaining personal information as $I_{\mathcal{D}_{\text{train}}} = a(I_{\mathcal{D}_{\text{private}}})$.

Let $\mathcal{M}$ be the pre-trained language model. $\mathcal{M}$ may memorize some of the personal information in $a(I_{\mathcal{D}_{\text{private}}})$ during the pre-training. We denote such memorization as m and the set of the memorized information as $m(a(I_{\mathcal{D}_{\text{private}}}))$. $m(a(I_{\mathcal{D}_{\text{private}}}))$ is a subset of $a(I_{\mathcal{D}_{\text{private}}})$, which is a subset of $I_{\mathcal{D}_{\text{private}}}$: $m(a(I_{\mathcal{D}_{\text{private}}})) \subseteq a(I_{\mathcal{D}_{\text{private}}}) \subseteq I_{\mathcal{D}_{\text{private}}}$.

The attacker can obtain all of $m(a(I_{\mathcal{D}_{\text{private}}}))$ in the worst case, but the personal information not memorized by $\mathcal{M}$ will not leak.

***K* and *T* factors: prior knowledge and target information** The K factor is the set of prior knowledge about the target people that is already known to the attacker such as full name, age, or sex. We denote the set as I_K. I_K is a subset of I_U: $I_K \subseteq I_U$.

The T factor is the set of personal information that the attacker aims to obtain, which we note as I_T. This greatly affects the definition of the privacy risk since it determines which pieces of information in the pre-training data are considered in the privacy leakage. I_K and I_T are disjoint: $I_K \cap I_T = \varnothing$.

In model inversion, I_T is a subset of I_U: $I_T \subseteq I_U$. In membership inference, I_T is existence or absence of arbitrary documents in the pre-training data: $I_T = \{\mathbb{1}[d \in \mathcal{D}_{\text{train}}]\}_{d \in \mathcal{D}}$.

The notion of the *K* factor rests on the fact that the more prior knowledge available on a person, the more identifiable the person becomes after privacy attacks. Suppose that the attacker aims to obtain the disease names of a person p_1 and has revealed that "diabetes" was in the pre-training data:

$$I_T = \{diseases_{p_1}\},\ \{\text{diabetes}\} \subset m(a(I_{\mathcal{D}_{\text{private}}})).$$

The attacker does not yet know the diseases of p_1 since it is unclear who is diabetic in the pre-training data. However, if the attacker knows the full name of p_1 ($I_K = \{\textit{full name}_{p_1}\}$) and the model $\mathcal{M}$ associates "diabetes" with the full name, the attacker can infer that p_1 is diabetic. This is expressed as below with $\widehat{x}$ denoting the prediction of x:

$$\widehat{diseases}_{p_1} = \{\text{diabetes}\},\ \ \widehat{I_T} = \{\widehat{diseases}_{p_1}\}.$$

Even if the attacker has prior knowledge about people who are *not* the target of the attack, we do not include such knowledge in the *K* factor but the *R* factor, as discussed in the next subsection.

***R* factor: auxiliary resources** The *R* factor is the set of resources other than $\mathcal{M}$ that are available to the attacker and that can indirectly aid the disclosure of the target information. Possible examples are language models other than $\mathcal{M}$, corpora, and personal information of *non-target* people.

Suppose again that the attacker guesses the diseases of the person p_1. An extreme example is that the model provider has released the model $\mathcal{M}$ whose pre-training data is not anonymized at all ($a = \text{id}$), and the attacker has access to $\mathcal{D}_{\text{public}}$, the anonymized version of the pre-training data:

$$R = \{\mathcal{D}_{\text{public}}\},\ \mathcal{D}_{\text{public}} = a'(\mathcal{D}_{\text{private}}),\ a' = f_{\text{HIPAA}}.$$

In this scenario, $\mathcal{D}_{\text{public}}$ does not cause privacy leakage alone but may raise its risk because the attacker only has to use $\mathcal{M}$ to fill $\mathcal{D}_{\text{public}}$ with the personal information removed from $\mathcal{D}_{\text{private}}$.

Another possible scenario is that the attacker has part of the pre-training data $\mathcal{D}^{\overline{\mathcal{P}}}_{\text{train}}$ where only non-target people are mentioned but cannot access the other part $\mathcal{D}^{\mathcal{P}}_{\text{train}}$:

$$R = \{\mathcal{D}^{\overline{\mathcal{P}}}_{\text{train}}\}.$$

In this scenario, the attacker may train a new language model $\mathcal{M}_{\text{shadow}}$ with $\mathcal{D}^{\overline{\mathcal{P}}}_{\text{train}}$ and attack $\mathcal{M}_{\text{shadow}}$ repeatedly to see how $\mathcal{M}_{\text{shadow}}$ reacts under the existence or absence of specific personal information in $\mathcal{D}^{\overline{\mathcal{P}}}_{\text{train}}$. This may enable the training of a classifier that can receive the reactions of $\mathcal{M}$ to the attacks and predict the existence or absence of specific target information in $\mathcal{D}^{\mathcal{P}}_{\text{train}}$.

Even without $\mathcal{D}^{\overline{\mathcal{P}}}_{\text{train}}$, similar attacks are possible if the attacker has a corpus $\widetilde{\mathcal{D}}_{\text{train}}$ that is irrelevant to $\mathcal{D}_{\text{train}}$ but has a very similar distribution:

$$R = \{\widetilde{\mathcal{D}}_{\text{train}}\}.$$

Personal information can also be the *R* factor. Suppose the attacker already knows the full name and diseases of non-target people $\overline{\mathcal{P}}$, who are mentioned in $\mathcal{D}_{\text{private}}$ but not the target of the attack:

$$R = \{(\textit{full name}, \textit{diseases})_{\bar{p}}\}_{\bar{p} \in \overline{\mathcal{P}}}.$$

This provides the attacker with positive samples of name-disease pairs in $\mathcal{D}_{\text{private}}$. The model $\mathcal{M}$ might act slightly differently to positive name-disease pairs and randomly generated negative name-disease pairs. Thus, the attacker may train a classifier to predict whether arbitrary name-disease pairs are present in $\mathcal{D}_{\text{private}}$ (Lehman et al., 2021). We assume that the prior knowledge on non-target people fits the *R* factor better than the *K* factor since it provides an indirect clue for the privacy attack unlike the knowledge on target people.

Summary In any privacy leakage scenario, the attacker attempts to obtain the target information I_T and attacks the pre-trained language model $\mathcal{M}$ to obtain the personal information in the pre-training data $I_{\mathcal{D}_{\text{private}}}$ using the prior knowledge I_K and resources R:

$$\mathcal{M}, I_K, R \xrightarrow{\text{attack}} \widehat{I_T}$$
$$\widehat{I_T} \subseteq \{\widehat{category}_{person} | category_{person} \in m(a(I_{\mathcal{D}_{\text{private}}}))\}.$$

Various privacy leakage scenarios can be represented as a combination (I_K, a, R, I_T), where each value corresponds to the *K*, *A*, *R*, and *T* factors.

4 KART-based Review of Related Work

We overview privacy leakage scenarios in prior studies with our proposed KART parameterization, which is outlined in Table 2.

Misra (2019) and Hisamoto et al. (2020) dealt with membership inference. They assessed the risk of a language model disclosing whether a specific document was in the training data ($I_T = \{\mathbb{1}[d \in \mathcal{D}_{\text{train}}]\}_{d \in \mathcal{D}}$). Misra (2019) discussed a GPT-1 model pre-trained with a public corpus and fine-tuned with a private corpus. Hisamoto et al. (2020) examined a Transformer model pre-trained with a private corpus. Both studies simulated the case that the attacker can access part of the pre-training data ($R = \{\mathcal{D}^{\overline{\mathcal{P}}}_{\text{train}}\}$) and trains a classifier to distinguish documents in and out of $\mathcal{D}_{\text{private}}$.

Table 2: KART parameterization of the simulated privacy leakage scenarios in previous studies.

Study	I_K	a	R	I_T	Attack method
Misra (2019)	$\varnothing$	id	$\{\mathcal{D}^{\overline{\mathcal{P}}}_{\text{train}}\}$	$\{\mathbb{1}[d \in \mathcal{D}_{\text{train]}}\}_{d\in\mathcal{D}}$	Membership inference
Hisamoto et al. (2020)	$\varnothing$	id	$\{\mathcal{D}^{\overline{\mathcal{P}}}_{\text{train}}\}$	$\{\mathbb{1}[d \in \mathcal{D}_{\text{train]}}\}_{d\in\mathcal{D}}$	Membership inference
Carlini et al. (2019)	$\varnothing$	id	$\varnothing$	$\{\textit{credit card number}_p\}_{p\in\mathcal{P}}$	NLG*
	$\varnothing$	id	$\varnothing$	$\{\textit{social security number}_p\}_{p\in\mathcal{P}}$	NLG*
Carlini et al. (2021)	$\varnothing$	id	$\varnothing$	I_U	NLG*
Inan et al. (2021)	$\varnothing$	id	$\varnothing$	I_U	NLG*
Lehman et al. (2021)	$\{(\textit{full name}, \textit{sex})_p\}_{p\in\mathcal{P}}$	id	$\varnothing$	$\{\textit{diseases}_p\}_{p\in\mathcal{P}}$	Language modeling*
	$\{(\textit{full name}, \textit{sex})_p\}_{p\in\mathcal{P}}$	id	$\{(\textit{full name}, \textit{sex}, \textit{diseases})_{\overline{p}}\}_{\overline{p}\in\overline{\mathcal{P}}}$	$\{\textit{diseases}_p\}_{p\in\mathcal{P}}$	Classification*
	$\varnothing$	id	$\{\textit{full name}_{\overline{p}}\}_{\overline{p}\in\overline{\mathcal{P}}}$	$\{\textit{full name}_p\}_{p\in\mathcal{P}}$	Classification*
	$\{\textit{first name}_p\}_{p\in\mathcal{P}}$	id	$\varnothing$	$\{\textit{last name}_p\}_{p\in\mathcal{P}}$	Language modeling*
	$\{\textit{last name}_p\}_{p\in\mathcal{P}}$	id	$\varnothing$	$\{\textit{first name}_p\}_{p\in\mathcal{P}}$	Language modeling*
Vakili and Dalianis (2021)	$\varnothing$	id	$\varnothing$	$\{(\textit{full name}, \textit{diseases})_{\overline{p}}\}_{\overline{p}\in\overline{\mathcal{P}}}$	NLG*
	$\{\textit{sex}_p\}_{p\in\mathcal{P}}$	id	$\varnothing$	$\{(\textit{full name}, \textit{diseases})_{\overline{p}}\}_{\overline{p}\in\overline{\mathcal{P}}}$	NLG*

* Model inversion.

Carlini et al. (2019), Carlini et al. (2021), and Inan et al. (2021) simulated NLG to restore substrings of the pre-training data without prior knowledge or auxiliary resources ($I_K = \varnothing, R = \varnothing$). Carlini et al. (2019) provided a case study of risk assessment the disclosure of for credit card numbers or social security numbers. Carlini et al. (2021) and Inan et al. (2021) considered the disclosure of any substring in the training data to be a privacy breach. That is, the attacker places no limits on the target information ($I_T = I_U$).

Lehman et al. (2021) and Vakili and Dalianis (2021) assumed that an attacker cross-refers to multiple personal information to reveal diseases of patients. The attacker cannot learn about a target person if the model only outputs a disease, but the attacker can associate the disease with the person if the model relates the disease to another attribute such as a full name. This problem formulation has a drawback in that the range of the target information is limited. However, it is advantageous because it covers scenarios where the model does not output an exact substring of the pre-training data but a similar one, or where the model outputs multiple pieces of personal information non-adjacently within a sentence. Such privacy leakage has not been examined in other studies. Lehman et al. (2021) examined a scenario where the attacker already knows the full name and sex of the target people ($I_K = \{(\textit{full name}, \textit{sex})_p\}_{p\in\mathcal{P}}$) and uses them as an NLG prompt to disclose diseases ($I_T = \{\textit{diseases}_p\}_{p\in\mathcal{P}}$). They also simulated the case that the attacker knows the name, sex, and diseases of non-target people in the pre-training data ($R = \{(\textit{full name}, \textit{sex}, \textit{diseases})_{\overline{p}}\}_{\overline{p}\in\overline{\mathcal{P}}}$) and trains a classifier to predict whether arbitrary people appeared in the pre-training data. Vakili and Dalianis (2021) simulated predictions of the full name and diseases together ($I_T = \{(\textit{full name}, \textit{diseases})_p\}_{p\in\mathcal{P}}$) using no prior knowledge ($I_K = \varnothing$) or the sex of the target people ($I_K = \{\textit{sex}_p\}_{p\in\mathcal{P}}$), although the risk was not directly compared.

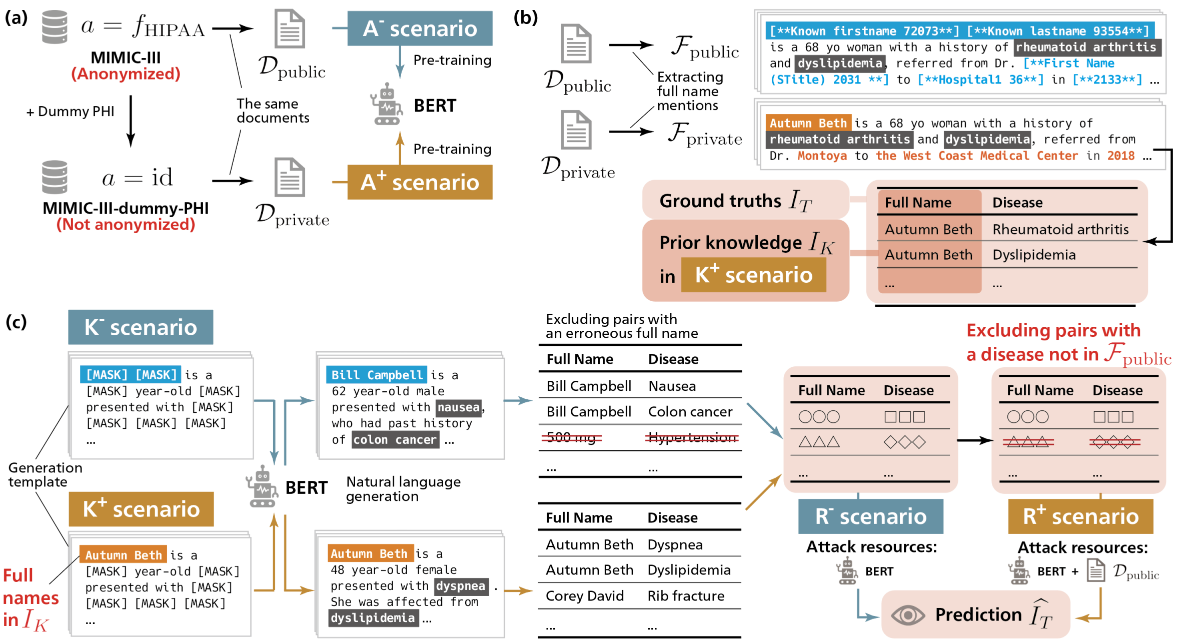
This KART-based review shows that previous studies have covered various scenarios and attacks. These are difficult to directly compare at present but may be integrated in a future meta-analysis.

5 Experimental Demonstration

We demonstrate scenario-aware privacy risk assessment and compare the privacy risk among scenarios under the same definition of privacy leakage.

5.1 Scenarios

We prepared eight privacy leakage scenarios. In all the scenarios, a pre-trained BERT model $\mathcal{M}$ is published whose pre-training data comprises clinical records in a hospital. An attacker exploits $\mathcal{M}$ to estimate the full name and diseases of the patients in the pre-training data. The scenarios, however, have different details that may affect privacy risks. We name each scenario using "K^+," "K^-," "A^+," "A^-," "R^+," and "R^-" because they are represented


(a)
$a = f_{\text{HIPAA}}$
MIMIC-III
(Anonymized)
+ Dummy PHI
$a = \text{id}$
MIMIC-III-dummy-PHI
(Not anonymized)
$\mathcal{D}_{\text{public}}$
The same documents
$\mathcal{D}_{\text{private}}$
A⁻ scenario
Pre-training
BERT
Pre-training
A⁺ scenario
(b)
$\mathcal{D}_{\text{public}}$
$\mathcal{D}_{\text{private}}$
$\mathcal{F}_{\text{public}}$
Extracting full name mentions
$\mathcal{F}_{\text{private}}$
Ground truths I_T
Prior knowledge I_K in K⁺ scenario
(c)
K⁻ scenario
Generation template
K⁺ scenario
Full names in I_K
BERT
Natural language generation
Excluding pairs with an erroneous full name
Excluding pairs with a disease not in $\mathcal{F}_{\text{public}}$
R⁻ scenario
Attack resources:
BERT
R⁺ scenario
Attack resources:
BERT + $\mathcal{D}_{\text{public}}$
Prediction $\hat{I}_T$

[**Known firstname 72073**] [**Known lastname 93554**] is a 68 yo woman with a history of rheumatoid arthritis and dyslipidemia, referred from Dr. [**First Name (STitle) 2031 **] to [**Hospital1 36**] in [**2133**] ...

Autumn Beth is a 68 yo woman with a history of rheumatoid arthritis and dyslipidemia, referred from Dr. Montoya to the West Coast Medical Center in 2018 ...

Full Name	Disease
Autumn Beth	Rheumatoid arthritis
Autumn Beth	Dyslipidemia
...	...

[MASK] [MASK] is a [MASK] year-old [MASK] presented with [MASK] [MASK] [MASK] [MASK] ...

Autumn Beth is a [MASK] year-old [MASK] presented with [MASK] [MASK] [MASK] [MASK] ...

Bill Campbell is a 62 year-old male presented with nausea, who had past history of colon cancer ...

Autumn Beth is a 48 year-old female presented with dyspnea . She was affected from dyslipidemia ...

Full Name	Disease
Bill Campbell	Nausea
Bill Campbell	Colon cancer
~~500 mg~~	~~Hypertension~~
...	...

Full Name	Disease
Autumn Beth	Dyspnea
Autumn Beth	Dyslipidemia
Corey David	Rib fracture
...	...

Full Name	Disease
○○○	□□□
△△△	◇◇◇
...	...

Full Name	Disease
○○○	□□□
~~△△△~~	~~◇◇◇~~
...	...

Figure 2: Overview of the privacy leakage experiment. (a) The model provider publishes a BERT model. Its pre-training data is anonymized in the A^- scenarios ($\mathcal{D}_{\text{public}}$) but not in the A^+ scenarios ($\mathcal{D}_{\text{private}}$). (b) The attacker aims to reveal name-disease pairs present in "full name mentions" in $\mathcal{D}_{\text{private}}$. (c) Attack with NLG using the pre-trained BERT model. Different templates are used in the K^+ and K^- scenarios. Predictions are refined differently when $\mathcal{D}_{\text{public}}$ is available (R^+ scenarios) or unavailable (R^- scenarios) to the attacker.

by different K, A, and R factors:

$$K^+ : I_K = \{\textit{full name}_p\}_{p\in\mathcal{P}}, \quad K^- : I_K = \varnothing,$$
$$A^+ : a = f_{\text{HIPAA}}, \qquad A^- : a = \text{id},$$
$$R^+ : R = \{f_{\text{HIPAA}}(\mathcal{D}_{\text{private}})\}, \quad R^- : R = \varnothing$$

X^+ and X^- are higher and lower severity choices for the X factor, respectively.

$K^+A^+R^+$ scenario Let $\mathcal{D}_{\text{private}}$ be the clinical records. They are used with no anonymization to pre-train a BERT model $\mathcal{M}$ from scratch ($\mathcal{D}_{\text{train}} = a(\mathcal{D}_{\text{private}})$, $a = \text{id}$). In addition to $\mathcal{M}$, the model provider also publishes a corpus $\mathcal{D}_{\text{public}}$, which is composed of the same documents as $\mathcal{D}_{\text{private}}$ but anonymized under the HIPAA Privacy Rule ($\mathcal{D}_{\text{public}} = a'(\mathcal{D}_{\text{private}})$, $a' = f_{\text{HIPAA}}$). An attacker estimates the past or present diseases of N patients $\mathcal{P} = \{p_1, ..., p_N\}$, all of whose full name is already known to the attacker. This scenario can be parameterized as follows: $I_K = \{\textit{full name}_p\}_{p\in\mathcal{P}}$, $a = \text{id}$, $R = \{\mathcal{D}_{\text{public}}\}$, and $I_T = \{\textit{diseases}_p\}_{p\in\mathcal{P}}$.

$K^+A^+R^-$ scenario The same as $K^+A^+R^+$ except that $\mathcal{D}_{\text{public}}$ is unavailable: $I_K = \{\textit{full name}_p\}_{p\in\mathcal{P}}$, $a = \text{id}$, $R = \varnothing$, and $I_T = \{\textit{diseases}_p\}_{p\in\mathcal{P}}$.

$K^+A^-R^+$ scenario The same as $K^+A^+R^+$ except that the pre-training data is anonymized under the HIPAA Privacy Rule: $I_K = \{\textit{full name}_p\}_{p\in\mathcal{P}}$, $a = f_{\text{HIPAA}}$, $R = \{\mathcal{D}_{\text{public}}\}$, and $I_T = \{\textit{diseases}_p\}_{p\in\mathcal{P}}$.

$K^-A^+R^+$ scenario The same as $K^+A^+R^+$ except that the attacker does not know the full name of the patients. Note that the attacker must guess the full name and diseases together since the prediction of the diseases alone does not often reveal the subject: $I_K = \varnothing$, $a = \text{id}$, $R = \{\mathcal{D}_{\text{public}}\}$, and $I_T = \{(\textit{full name}, \textit{diseases})_p\}_{p\in\mathcal{P}}$.

$K^+A^-R^-$ scenario The same as $K^+A^-R^+$ except that $\mathcal{D}_{\text{public}}$ is unavailable: $I_K = \{\textit{full name}_p\}_{p\in\mathcal{P}}$, $a = f_{\text{HIPAA}}$, $R = \varnothing$, and $I_T = \{\textit{diseases}_p\}_{p\in\mathcal{P}}$.

$K^-A^+R^-$ scenario The same as $K^-A^+R^+$ except that $\mathcal{D}_{\text{public}}$ is unavailable: $I_K = \varnothing$, $a = \text{id}$, $R = \varnothing$, and $I_T = \{(\textit{full name}, \textit{diseases})_p\}_{p\in\mathcal{P}}$.

$K^-A^-R^+$ scenario The same as $K^-A^+R^+$ except that the pre-training data is anonymized under the HIPAA Privacy Rule: $I_K = \varnothing$, $a = f_{\text{HIPAA}}$, $R = \{\mathcal{D}_{\text{public}}\}$, and $I_T = \{(\textit{full name}, \textit{diseases})_p\}_{p\in\mathcal{P}}$.

$K^-A^-R^-$ scenario The same as $K^-A^-R^+$ except that $\mathcal{D}_{\text{public}}$ is unavailable: $I_K = \varnothing$, $a = f_{\text{HIPAA}}$, $R = \varnothing$, and $I_T = \{(\textit{full name}, \textit{diseases})_p\}_{p\in\mathcal{P}}$.

This choice of scenarios is only an example and does not cover all real-world privacy leakage, but it may still help estimate the upper bound of risk since the $K^+A^+R^+$ scenario is favorable to the attacker.

For simplicity, we collectively refer to the scenarios with a parameter value X^+ or X^- as "X^+

Table 3: Results of the comparison of the privacy risk from a pre-trained language model in different scenarios.

Anchor scenario			Weakened scenario			Privacy risk margin
Name	Name-disease pairs per 10k generations (correct/valid)	Privacy leakage ratio*	Name	Name-disease pairs per 10k generations (correct/valid)	Privacy leakage ratio*	
$K^+A^+R^+$	351 / 10,049	3.50%	$K^+A^+R^-$ $K^+A^-R^+$ $K^-A^+R^+$	351 / 27,239 306 / 9,721 0 / 101	1.29% 3.14% 0.00%	**2.21%** **($\geq$0%)** **0.36%** **($\geq$0%)** **3.50%** **($\geq$0%)**
$K^+A^+R^-$	351 / 27,239	1.29%	$K^+A^-R^-$ $K^-A^+R^-$	306 / 25,644 0 / 249	1.19% 0.00%	**0.10%** **($\geq$0%)** **1.29%** **($\geq$0%)**
$K^+A^-R^+$	306 / 9,721	3.14%	$K^+A^-R^-$ $K^-A^-R^+$	306 / 25,644 0 / 0	1.19% NA**	**1.95%** **($\geq$0%)** NA**
$K^-A^+R^+$	0 / 101	0.00%	$K^-A^+R^-$ $K^-A^-R^+$	0 / 249 0 / 0	0.00% NA**	**0.00%** **($\geq$0%)** NA**
$K^+A^-R^-$	306 / 25,644	1.19%	$K^-A^-R^-$	0 / 0	NA**	NA**
$K^-A^+R^-$	0 / 249	0.00%	$K^-A^-R^-$	0 / 0	NA**	NA**
$K^-A^-R^+$	0 / 0	NA**	$K^-A^-R^-$	0 / 0	NA**	NA**

*Calculated with the actual numbers of pairs, not with the values in the table. **No valid name-disease pairs were generated.

scenarios" or "X^- scenarios," respectively.

5.2 BERT pre-training

We pre-trained two BERT models, each of which was repeatedly used in A^+ and A^- scenarios, respectively. Two sets of pre-training data were made by sampling the same 100k clinical records from MIMIC-III (Johnson et al., 2016), which comprises clinical records anonymized under the HIPAA Privacy Rule, and "MIMIC-III-dummy-PHI," which we built by adding dummy personal information to MIMIC-III[2]. This method eliminated the risk of a real-world privacy breach. The samples from MIMIC-III-dummy-PHI corresponded to $\mathcal{D}_{\text{train}}$ in the A^+ scenarios and $\mathcal{D}_{\text{private}}$ in all the scenarios. The samples from MIMIC-III were used as $\mathcal{D}_{\text{train}}$ in the A^- scenarios and $\mathcal{D}_{\text{public}}$ in the R^+ scenarios. See Appendices B.1, B.2, and C.1 for more details.

5.3 Model inversion attack

5.3.1 Gold standard target information

The privacy attack is a prediction of name-disease pairs $I_T = \{(n_i, d_{i,j})\}_{1\leq i\leq N, 1\leq j\leq N_i}$, where n_i and $d_{i,1}, d_{i,2}, \ldots$ denote the full name and diseases of the i-th patient, respectively. We made the gold standard from $\mathcal{D}_{\text{private}}$ as in Figure 2. First, we extracted all "full name mentions," five consecutive sentences beginning with the patient demographics "(*first name*) (*last name*) is a (*age*) year old (*sex*)." Then, we chose full name mentions $s_1, \ldots, s_N$ so that every full patient name in $\mathcal{F}_{\text{private}} = \{s_i\}_{1\leq i\leq N}$ was unique and encoded as two tokens by the BERT tokenizer. For each s_i, we extracted and normalized disease names into a controlled unique identifier (CUI) in the UMLS metathesaurus (Bodenreider, 2004) using MetaMap 2020 (Aronson, 2001). We built I_T by pairing a full name n_i and CUIs $d_{i,1}, d_{i,2}, \ldots$ identified in s_i.

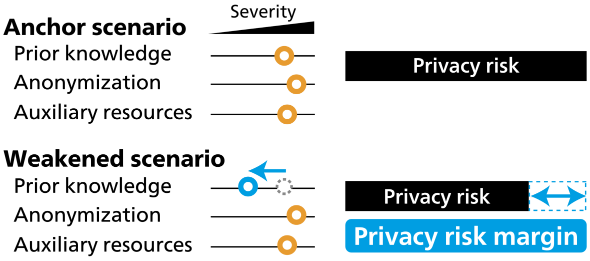


Figure 3: Risk comparison between anchor and weakened scenarios. All the primary factors of the weakened scenario are the same or less severe than those of the anchor scenario. The weakened scenario may result in a zero or positive privacy risk margin.

5.3.2 Name-disease estimation

To simulate attacks as in Figure 2, we generated at least 10k documents with the BERT model, extracted name-disease pairs, and made predictions $\widehat{I_T}$ by choosing "valid" pairs and excluding erroneous ones. Appendices B.3 and B.4 give details.

5.3.3 Risk comparison between scenarios

We measured the privacy risk in each scenario with the *privacy leakage ratio*, the ratio of the number of correct predictions to that of valid ones ($|I_T \cap \widehat{I_T}|$ / $|\widehat{I_T}|$). Then, we compared the risk in two scenarios differing in one factor as in Figure

[2] The implementation is at `https://github.com/[*** masked ***]/[*** masked ***]`

3. For each pair, we referred to the more severe scenario as the *anchor scenario* and the other as the *weakened scenario*, and computed the *privacy risk margin*, the pairwise difference of privacy leakage ratio.

5.4 Results and analysis

As shown in Table 3, the privacy risk margin was greater than or equal to zero for all the scenario pairs where the margin could be calculated, suggesting that the scenario parameterizations coincided with the resulting privacy risk.

The upper bound of risk under this attack method could be approximated by the privacy risk ratio in the $K^+A^+R^+$ scenario. Its magnitude may be small because the privacy leakage ratio in the $K^+A^+R^+$ scenario may be mostly contributed to by random guesses rather than actual disclosures of personal information, given the small privacy risk margin between the $K^+A^+R^+$ and $K^+A^-R^+$ scenarios.

6 Discussion

We have introduced KART, a simple parameterization to clarify assumed privacy leakage scenarios during the risk assessment of sharing language models.

The estimation of the upper bound of privacy risk requires a wide coverage of real-world privacy leakage scenarios. Our KART-based review first simply clarified the scenarios dealt with in previous studies and their variety, suggesting the difficulty in direct comparison. Although we have not yet successfully integrated their findings, KART may provide a novel meta-analysis method for gaining comprehensive knowledge in the future. We have also shown that KART helps risk comparison in different scenarios under the same attack method or vice versa. Applying KART prior to every privacy leakage experiment would improve the comparability of future studies. KART should also spotlight scenarios that have not been fully explored.

The privacy risk margin was always zero or positive between anchor and weakened scenarios in our experiment. This may not always be consistent since privacy leakage occurs stochastically. However, if the scenario severity often coincides with the privacy risk margin in future studies, risk assessment may be streamlined by focusing on the most severe scenario possible under a given attack.

We assume that it is worthwhile simulating specific practical scenarios one by one in privacy risk assessment. KART alone does not provide a universal risk score valid in all scenarios. However, it has been unclear whether a single universal privacy risk metric covering all scenarios is possible. For example, Carlini et al. (2019) and Inan et al. (2021) proposed universal metrics of the upper bound of risk, but they focused on the disclosure of the exact substring of the pre-training data and did not cover other types of privacy leakage (Lehman et al., 2021; Vakili and Dalianis, 2021). Differential privacy (Dwork, 2006) is a strong framework to ensure privacy, but the relationship between ϵ and its real-world impact should still be examined (Hoory et al., 2021). Moreover, differential privacy in NLP is usually defined as the indistinguishability of two models pre-trained with datasets differing in one document. It is unclear whether the safety is robust to any privacy leakage, such as that caused by the generation of paraphrases of substrings of the dataset. Wagner and Eckhoff (2018) pointed out the existence of numerous privacy metrics and provided a guide to making suitable choices depending on the problem setting, which is probably a feasible means of privacy risk assessment.

It may be sufficient to always enforce a complete manual anonymization of pre-training data before sharing models for safe management. However, privacy risk assessment is still valuable, because language models can be accidentally made public if stolen or mistakenly put into public storage, as has happened to electronic health records (Myers et al., 2008). Moreover, the risk assessment may provide knowledge that may decrease future anonymization costs. Several off-the-shelf systems can effectively anonymize clinical records automatically (Heider et al., 2020), whose performance may be sufficient to publish language models pre-trained on anonymized corpora.

We expect KART to promote a wide range of future studies on privacy risks since it does not rely on domain-specific concepts.

7 Conclusion

It has been a challenge to assess the upper bound of the risk of sharing language models considering various real-world privacy leakage scenarios. We have proposed KART to simply parameterize complex scenarios. KART is expected to improve the portability of past and future privacy risk assessments and contribute to formulating privacy guidelines on language models.

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*, CCS '16, pages 308–318.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings AMIA Symposium*, pages 17–21.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–270.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Yen-Pin Chen, Yi-Ying Chen, Jr-Jiun Lin, Chien-Hua Huang, and Feipei Lai. 2020. Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation. *JMIR Med Inform*, 8(4):e17787.

Alister D'Costa, Stefan Denkovski, Michal Malyska, Sae Young Moon, Brandon Rufino, Zhen Yang, Taylor Killian, and Marzyeh Ghassemi. 2020. Multiple sclerosis severity classification from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 7–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming*, pages 1–12.

European Commission. 2018. 2018 Reform of EU data protection rules.

José Luis Fernández-Alemán, Inmaculada Carrión Señor, Pedro Ángel Oliver Lozoya, and Ambrosio Toval. 2013. Security and privacy in electronic health records: a systematic literature review. *J Biomed Inform*, 46(3):541–562.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 1322–1333.

Paul M. Heider, Jihad S. Obeid, and Stéphane M. Meystre. 2020. A Comparative Analysis of Speed and Accuracy for Three Off-the-Shelf De-Identification Tools. *AMIA Jt Summits Transl Sci Proc*, 2020:241–250.

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.

Shlomo Hoory, Amir Feder, Avichai Tendler, Sofia Erell, Alon Peled-Cohen, Itay Laish, Hootan Nakhost, Uri Stemmer, Ayelet Benjamini, Avinatan Hassidim, and Yossi Matias. 2021. Learning and evaluating a differentially private pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1178–1189.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Huseyin A. Inan, Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. 2021. Training data leakage analysis in language models. *arXiv preprint arXiv:2101.05405*.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035.

Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. 2021. A clinical specific BERT developed using a huge japanese clinical text corpus. *PLOS ONE*, 16(11):1–11.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. In *Proceedings of the 2nd Workshop on Privacy in NLP*, pages 39–45.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959.

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, and Hong Yu. 2019. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Med Inform*, 7(3):e14830.

Vedant Misra. 2019. Black box attacks on transformer language models. In *ICLR 2019 Debugging Machine Learning Models Workshop*.

Stephen J. Mooney and Vikas Pejaver. 2018. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annu Rev Public Health*, 39:95–112.

Julie Myers, Thomas R. Frieden, Kamal M. Bherwani, and Kelly J. Henning. 2008. Ethics in public health research: privacy and public health at risk: public health confidentiality in the digital age. *Am J Public Health*, 98(5):793–801.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1471–1488.

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 18th BioNLP Workshop and Shared Task*, pages 58–65.

W. Nicholson Price and I. Glenn Cohen. 2019. Privacy in the age of medical big data. *Nat Med*, 25(1):37–43.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706.

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.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS ’20, page 377–390.

Latanya Sweeney. 2002. K-anonymity: A model for protecting privacy. *Int J Uncertain Fuzziness Knowl-Based Syst*, 10(5):557–570.

The United States Department of Health and Human Services. 2012. Guidance on de-identification of protected health information. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, 14(5):550–563.

Thomas Vakili and Hercules Dalianis. 2021. Are clinical BERT models privacy preserving? The difficulty of extracting patient-condition associations. In *Association for the Advancement of Artificial Intelligence Fall 2021 Symposium in HUman partnership with Medical Artificial iNtelligence (HUMAN.AI)*, pages 30–36.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Adv Neural Inform Process Syst*, volume 30, pages 5998–6008.

Isabel Wagner and David Eckhoff. 2018. Technical privacy metrics: A systematic survey. *ACM Comput Surv*, 51(3).

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36.

Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Adv Neural Inform Process Syst*, volume 32, pages 14774–14784.

A 18 HIPAA Identifiers

Under the HIPAA Privacy Rule, clinical records for the second usage must meet either of the two conditions: (i) Experts determine that the clinical records are anonymized properly and that there is little risk of disclosing the subject of the information, or (ii) a set of specific identifiers (18 HIPAA identifiers) regarding the subjects of the information and their relatives, employers, and household members is removed from the clinical records (The United States Department of Health and Human Services, 2012). Table 4 the lists 18 HIPAA identifiers.

Table 4: 18 identifiers to be masked under the HIPAA Privacy Rule.

(A)	Names
(B)	All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people and (2) the initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people are changed to 000
(C)	All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
(D)	Telephone numbers
(E)	Fax numbers
(F)	Email addresses
(G)	Social security numbers
(H)	Medical record numbers
(I)	Health plan beneficiary numbers
(J)	Account numbers
(K)	Certificate/license numbers
(L)	Vehicle identifiers and serial numbers, including license plate numbers
(M)	Device identifiers and serial numbers
(N)	Web Universal Resource Locators (URLs)
(O)	Internet Protocol (IP) addresses
(P)	Biometric identifiers, including finger and voice prints
(Q)	Full-face photographs and any comparable images
(R)	Any other unique identifying number, characteristic, or code, except as permitted

B Details of Privacy Leakage Experiment

B.1 Pre-training Data

MIMIC-III (Johnson et al., 2016) is a publicly available dataset including over 2M clinical records of patients in the intensive care unit of Beth Israel Deaconess Medical Center. The clinical records are divided into 15 categories, including discharge summaries and progress notes. The clinical records are anonymized under the HIPAA Privacy Rule by replacing PHI incorporated into the HIPAA 18 identifiers with de-identification placeholders. To make MIMIC-III-dummy-PHI, we replaced the placeholders with dummy PHI. Dummy hospital names were randomly sampled from the i2b2 2006 dataset (Uzuner et al., 2007), and the other dummy identifiers were randomly generated with `Faker`.[3]

Next, we built two sets of pre-training data by sampling the same documents from MIMIC-III-dummy-PHI and MIMIC-III. The sampling was a random choice of 50% of the clinical records and further extraction of discharge summaries and progress notes, which left us with around 100k documents.

B.2 BERT Pre-training

We pre-trained an uncased BERT-base model from scratch using $\mathcal{D}_{\text{train}}$. We did not fine-tune the BERT model provided by Devlin et al. (2019), which was pre-trained with BooksCorpus and Wikipedia, in order to avoid noise in the privacy leakage experiment. This was because it would be difficult to determine whether the full names and disease names output by the BERT model were disclosures from clinical records or just a reproduction of BooksCorpus or Wikipedia.

$\mathcal{D}_{\text{train}}$ was preprocessed in almost the same way as ClinicalBERT (Huang et al., 2019), but we did not delete digits. All the de-identification placeholders were removed. Owing to limitations in computational resources, we pre-trained the BERT model for 1M steps with the maximum length set to 128. The other hyperparameters were the same as in ClinicalBERT: learning rate, 2e-5; batch size, 64.

B.3 Privacy Attack Strategy

Step 1: Name-disease pair generation In the K^{-} scenarios, the full patient name and diseases were estimated simultaneously. We generated L documents $x_1, ..., x_L$ by filling `[MASK]` tokens of a 128-token-length template. The template contained four masking spans such as "`[CLS] (name-masking)` is a `(age-masking)` year old `(sex-masking)`

[3] `https://github.com/joke2k/faker`

presented with `(disease-masking) [SEP]`." In practice, the name-, age-, sex-, and disease-masking spans consisted of two, one, one, and 116 consecutive `[MASK]` tokens, respectively. Refer to Appendix B.4 for details of the method used to fill the blanks.

For each generated document x_l, we used the content filling the name-masking span as the prediction of the full patient name $\widehat{n_l}$. We also automatically extracted CUIs $\widehat{d_{l,1}}, \widehat{d_{l,2}}, \ldots$ using MetaMap 2020. Finally, we collected name-disease pairs $\{(\widehat{n_l}, \widehat{d_{l,m}})\}_{1 \leq l \leq L}$ from the L generated documents.

In the K^+ scenarios, we obtained name-disease pairs similarly except that the name-masking span in the template was replaced with a randomly sampled full patient name $n_i \in I_T$ in each generation. This was because the attacker was supposed to already know the full patient names and only had to estimate the diseases.

Step 2: Prediction refinement The BERT output for the name-masking span sometimes made no sense as a person's full name. We excluded such "invalid" name-disease pairs where the predicted full name $\widehat{n_l}$ did not match any of the full names listed in the `Faker` library.

In the R^- scenarios, the remaining "valid" name-disease pairs were used as the prediction $(\widehat{I_T})$:

$$\widehat{I_T} = \{(\widehat{n_l}, \widehat{d_{l,m}}) \mid \widehat{n_l} \in \textit{valid full names}\}_{1 \leq l \leq L}.$$

In the R^+ scenarios, name-disease pairs were further excluded if their CUI had no corresponding disease name in $\mathcal{F}_{\text{public}}$:

$$\widehat{I_T} = \{(\widehat{n_l}, \widehat{d_{l,m}}) \mid \widehat{n_l} \in \textit{valid full names}, \\ \widehat{d_{l,m}} \in \mathcal{F}_{\text{public}}\}_{1 \leq l \leq L}.$$

This is because the attacker, who has access to $\mathcal{F}_{\text{public}}$, can assume that predictions are probably incorrect if they contain diseases that are absent from $\mathcal{F}_{\text{public}}$.

Owing to limitations in computational resources and time, we obtained predictions for R^+ scenarios by refining corresponding ones for R^- scenarios. For example, we made predictions $\widehat{I_T}$ for the $K^+A^+R^+$ scenario by reusing name-disease pairs obtained in Step 1 in the $K^+A^+R^-$ scenario and then following Step 2.

B.4 Details of Natural Language Generation

Our NLG method is based on the Markov chain Monte Carlo method following Wang and Cho (2019). First, we designated the positions in the template where the `[MASK]` tokens are initially placed as "writable positions." Then, we repeated 1,000 iterations to randomly select one of the writable positions and to overwrite the word in that place with a new word. The new word was chosen randomly on the basis of the distribution given by masked language modeling in the first 250 iterations (burn-in period). Subsequently, we used a top-100 sampling strategy by setting the probability to zero for all of the words outside the top-100 posterior probabilities. The batch size was set to 32.

C Complementary Results

C.1 Performance of BERT Models in Downstream Task

We examined the performance of the pre-trained BERT models used in our experiment in the MedNLI task (Romanov and Shivade, 2018) to evaluate how well they were pre-trained. We fine-tuned the two BERT models used in A^+ and A^- scenarios and the off-the-shelf model released by Devlin et al. (2019).

For each model, we calculated the validation accuracy for each learning rate $\in \{2e-5, 3e-5, 4e-5, 5e-5\}$ and each number of epochs $\in \{2, 3, 4\}$, and used the combination that maximized the validation accuracy for the test set. The batch size was set to 16.

Table 5 shows the results. The performance of the off-the-shelf BERT model was comparable to that reported by Alsentzer et al. (2019). Our BERT model in A^- scenarios outperformed the off-the-shelf model and our BERT model in A^+ scenarios achieved similar performance. Our BERT models benefited from being pre-trained with MIMIC-III, the source from which the premise sentences in MedNLI were extracted. However, our models are the same as the off-the-shelf model in that they are pre-trained for 1M steps from scratch and are disadvantageous for a much smaller pre-training corpus (120M vs 3,300M words). We assume that the BERT models used in our privacy leakage experiments are well pre-trained.

C.2 Privacy Leakage Experiment

We show the actual numbers of generated documents, valid name-disease pairs, and correct name-disease pairs in each scenario in Table 6, which is complementary to Table 3. For the efficient use of

Table 5: Performance of the BERT models used in our experiment and the model provided by Devlin et al. (2019) in the MedNLI test set.

Model	Hyperparameters		Test acc.
	Learning rate	Epoch	
Ours (A^+ scenarios)	3e-5	4	72.29%
Ours (A^- scenarios)	2e-5	3	77.50%
Devlin et al. (2019)	3e-5	3	76.09%

Table 6: Actual numbers of generated documents, valid name-disease pairs, and correct name-disease pairs in each scenario.

Scenario	Generated documents	Name-disease pairs	
		Valid	Correct
$K^+A^+R^+$	10,016	10,065	352
$K^+A^+R^-$	10,016	27,283	352
$K^+A^-R^+$	10,016	9,737	306
$K^+A^-R^-$	10,016	25,685	306
$K^-A^+R^+$	111,968	1,127	0
$K^-A^+R^-$	111,968	2,789	0
$K^-A^-R^+$	310,016	0	0
$K^-A^-R^-$	310,016	0	0

computational resources, we increased the number of generations only for the four scenarios in which no correct name-disease pairs were obtained, but this did not change the result. See also Appendix B.3.