PRESERVING THE PRIVACY OF LANGUAGE MODELS

Experiments in Clinical NLP

DRAFT

Thomas Vakili

Doctoral Thesis
Department of Computer and Systems Sciences
Stockholm University
January, 2026

Stockholm University
DSV Report Series No. 26-001
ISSN 1101-8526

Typeset by the author using LaTeX
Printed in Stockholm, Sweden by US-AB

This doctoral thesis is an extension of the author's licentiate thesis: Thomas Vakili. 2023. *Attacking and Defending the Privacy of Clinical Language Models*. Licentiate thesis, Department of Computer and Systems Sciences, Stockholm University..

Stockholm
University

ABSTRACT

The state-of-the-art methods in natural language processing (NLP) increasingly rely on large pre-trained transformer models. The strength of the models stems from their large number of parameters and the enormous amounts of data used to train them. The datasets are of a scale that makes it difficult, if not impossible, to audit them manually. When unwieldy amounts of potentially sensitive data are used to train large machine learning models, a difficult problem arises: unwelcome memorization of the training data.

All datasets—including those based on publicly available data—can contain personally identifiable information (PII). When models memorize these sensitive data, they become vulnerable to privacy attacks. Very few datasets for NLP can be guaranteed to be free from sensitive data. Consequently, most NLP models are susceptible to privacy leakage. This susceptibility is especially concerning in clinical NLP, where the data typically consist of electronic health records (EHRs). Leaking data from EHRs is never acceptable from a privacy perspective. This doctoral thesis investigates the privacy risks of using sensitive data and how they can be mitigated—while maintaining data utility.

A BERT model pre-trained using clinical data is subjected to a training data extraction attack. The same model is used to evaluate a membership inference attack that has been proposed to quantify the privacy risks of masked language models. Multiple experiments assess the performance gains from adapting pre-trained models to the clinical domain. Then, the impact of automatic de-identification on the performance of BERT models is evaluated for both pre-training and fine-tuning data. Finally, synthetic corpora for training models to detect PII are generated using domain-adapted generative language models. The quality of these corpora, and the parameters affecting their utility, are explored by training and evaluating BERT models.

The results show that domain adaptation leads to significantly better performance on clinical NLP tasks. They also show that extracting training data from BERT models is difficult and suggest that the risks can be further decreased by automatically de-identifying the training data. Automatic de-identification is found to preserve the utility of the data used for pre-training and fine-tuning BERT models. However, we also find that contemporary membership inference attacks are unable to quantify the privacy benefits this technique. Similarly, high-quality synthetic corpora can be generated using limited resources, but further research is needed to determine the privacy gains from using them. The results show that automatic de-identification and training data synthesis reduces the privacy risks of using sensitive data for NLP while preserving the utility of the data, but that these privacy benefits may be difficult to quantify.

SAMMANFATTNING

Här kommer det finnas en svensk sammanfattning.

LIST OF PAPERS

This doctoral thesis is based on the following papers:

I Thomas Vakili and Hercules Dalianis. 2021. Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In *Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021).*

II Thomas Vakili and Hercules Dalianis. 2023. Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data. In *Proceedings of the 24th Nordic Conference on Computational Linguistics.*

III Thomas Vakili, Martin Hansson, and Aron Henriksson. 2025. SweClinEval: A Benchmark for Swedish Clinical Natural Language Processing. In *Proceedings of The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies.*

IV Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, 4245–4252.*

V Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2024. End-to-end pseudonymization of fine-tuned clinical BERT models. *BMC Medical Informatics and Decision Making, 24(1), 162.*

VI Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2025. Data-Constrained Synthesis of Training Data for De-Identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).*

ADDITIONAL BIBLIOGRAPHY

In addition to the six included papers, the research described in this thesis has led to the following ten publications:

i Thomas Vakili and Hercules Dalianis. 2022. Utility Preservation of Clinical Text After De-Identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 383–388.

ii Oskar Jerdhaf, Marina Santini, Peter Lundberg, Tomas Bjerner, Yosef Al-Abasse, Arne Jönsson, and Thomas Vakili. 2022. Evaluating Pre-Trained Language Models for Focused Terminology Extraction from Swedish Medical Records. In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*, pages 30–32.

iii Alexander Dolk, Hjalmar Davidsen, Hercules Dalianis, and Thomas Vakili. 2022. Evaluation of LIME and SHAP in Explaining Automatic ICD-10 Classifications of Swedish Gastrointestinal Discharge Summaries. In *Scandinavian conference on health informatics*, pages 166–173.

iv Olle Bridal, Thomas Vakili, and Marina Santini. 2022. Cross-Clinic De-Identification of Swedish Electronic Health Records: Nuances and Caveats. In *LREC 2022 joint workshop language resources and evaluation conference 20-25 june 2022*, page 49.

v Thomas Vakili. 2023. *Attacking and Defending the Privacy of Clinical Language Models*. Licentiate thesis, Department of Computer and Systems Sciences, Stockholm University.

vi Anastasios Lamproudis, Therese Olsen Svenning, Torbjørn Torsvik, Taridzo Chomutare, Andrius Budrionis, Phuong Dinh Ngo, Thomas Vakili, and Hercules Dalianis. 2023. Using a Large Open Clinical Corpus for Improved ICD-10 Diagnosis Coding. In *AMIA annual symposium proceedings*, volume 2023, page 465.

vii Jocelyn Dunstan, Thomas Vakili, Luis Miranda, Fabián Villena, Claudio Aracena, Tamara Quiroga, Paulina Vera, Sebastián Viteri Valenzuela, and Victor Rocco. 2024. A pseudonymized cor-

pus of occupational health narratives for clinical entity recognition in Spanish. *BMC Medical Informatics and Decision Making*, 24(1):204.

viii Claudio Aracena, Luis Miranda, Thomas Vakili, Fabián Villena, Tamara Quiroga, Fredy Núñez-Torres, Victor Rocco, and Jocelyn Dunstan. 2024. A Privacy-Preserving Corpus for Occupational Health in Spanish: Evaluation for NER and Classification Tasks. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 111–121.

ix Thomas Vakili, Tyr Hullmann, Aron Henriksson, and Hercules Dalianis. 2024. When Is a Name Sensitive? Eponyms in Clinical Text and Implications for De-Identification. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 76–80.

x Lotta Kiefer, Jesujoba O. Alabi, Thomas Vakili, Hercules Dalianis, and Dietrich Klakow. 2025. Instruction-Tuning LLaMA for Synthetic Medical Note Generation in Swedish and English. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*.

به مرم رومام انی
مرعل مان، الگ وان

Till mormor och māmānī
Lärare, förebilder

ACKNOWLEDGEMENTS

Here I will write my acknowledgements.

I will probably need at least two pages.

ت‌وم اس واک ی‌ل ی

دک‌ت‌ر م‌ه‌ن‌دس ی

Thomas Vakili
Santiago de Chile, November 2025

Table of contents

Acronyms

ADE	adverse drug event
BoW	bag-of-words
DP	differential privacy
EHR	electronic health record
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
HIPS	hiding in plain sight
LSTM	long short-term memory
MLM	masked language modeling
NER	named entity recognition
NLP	natural language processing
PHI	personal health information
PII	personally identifiable information
PLM	pre-trained language model
RNN	recurrent neural network

Chapter 1

Introduction

Recent advances in natural language processing (NLP) have become increasingly reliant on large pre-trained language models. These models are built using variations of the transformer architecture (Vaswani et al., 2017) and were popularized with the introduction of the BERT[1] models (Devlin et al., 2019). Aside from architectural advances, the success of these models has also been driven by rapid increases in scale—both in terms of their increasingly enormous number of parameters and the vast amounts of training corpora used to train them.

Pre-trained language models are applicable to many NLP problems in many different domains. The research underlying this doctoral thesis was conducted at the Department of Computer and Systems Sciences, where there is a vibrant research community working with clinical data. NLP can be used to detect adverse drug events (Henriksson, 2015), to automatically assign diagnosis codes to discharge summaries (Remmer et al., 2021), and in many other applications. Clinical NLP can improve the quality and safety of care while also alleviating the administrative burden imposed on doctors. However, state-of-the-art clinical NLP relies on large pre-trained language models and large corpora of clinical text. Clinical text, usually in the form of electronic health records (EHRs), can contain personally identifiable information (PII). This is a problem because, in addition to tackling NLP tasks effectively, pre-trained

[1] BERT stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, partially to continue the trend of naming models after TV characters (Peters et al., 2018; Zanzotto et al., 2020; Lewis et al., 2020).

language models are susceptible to privacy attacks that can reveal information about their training data.

The two main types of attacks are *training data extraction attacks* and *membership inference attacks*. Successful training data extraction attacks can extract parts of the corpus used to train a language model. Membership inference attacks, on the other hand, try to establish whether a given data point was or was not part of a model's training corpus. This attack can reveal sensitive information about a patient's medical history but has also been proposed as a proxy for quantifying the privacy risks of machine learning models (Murakonda and Shokri, 2020; Mireshghallah et al., 2022). This doctoral thesis explores both training data extraction and membership inference attacks on clinical BERT models—which are well-suited for the types of classification tasks typical of clinical NLP (Naguib et al., 2024).

Multiple experiments in this thesis confirm that domain adaptation is important to achieve state-of-the-art performance in clinical NLP tasks. The thesis then explores how to decrease the privacy risks of using clinical BERT models. First, it explores how the privacy risks of BERT models can be mitigated using automatic de-identification. Automatic de-identification is a privacy-preserving technique that decreases the privacy risks of NLP models by automatically detecting and sanitizing PII in their training data. Sanitization can be done in multiple ways, and one popular approach is to replace PII with realistic surrogates (Dalianis, 2019). Even though automatic de-identification can make data safer, the named entity recognition (NER) models used to detect PII are imperfect. Sometimes, the NER models fail to detect PII, which leads to sensitive data remaining in the dataset. At other times, these models incorrectly classify safe entities as PII, introducing noise into the data.

The thesis also contains experiments that involve training BERT models on synthetic versions of a clinical corpus. These synthetic corpora are created by training a generative language model to produce text similar to the original corpus. Once these unannotated data are generated, a separate encoder model trained using the original corpus is used to add labels. This process is repeated using various amounts of sensitive data, and the quality of the synthetic corpora is assessed by training and evaluating models using them.

This thesis shows that automatic de-identification can decrease the privacy risks of training machine learning models using clinical data without harming the performance of the resulting models. This means that the noise introduced by imperfect NER models when they are used to remove large amounts of PII does not disturb training. We also show that membership inference attacks fail to capture the privacy benefits of automatic de-identification, which means that they cannot currently be used to compare automatic de-identification to other privacy-preserving techniques. The results also show that synthetic corpora can be used in place of real data without incurring a large drop in performance. Furthermore, such synthetic corpora can be created using small datasets as long as the machine-annotating model is sufficiently capable. However, similarly to automatically de-identifying data, there is a lack of rigorous methods for quantifying the privacy benefits of using synthetic corpora.

1.1 Research Questions

This thesis investigates how the privacy risks inherent in training language models with sensitive data can be mitigated using privacy-preserving techniques. The main research question of the thesis can be phrased as:

RQ How can the privacy risks of using sensitive data to train language models be minimized without sacrificing predictive performance?

The six papers in the thesis cover various parts of this overarching theme. Paper I explores the extent to which BERT models trained on clinical data leak information about the individuals described in the training data. Specifically, the paper covers the risks of leaking sensitive information through a training data extraction attack. Paper II examines a state-of-the-art membership inference attack targeting BERT models and evaluates if it can be used to quantify the privacy benefits of automatic de-identification. Papers III, IV, and VI evaluate whether domain-adaptive pre-training with sensitive clinical data is worthwhile as opposed to using safer general-domain models. Papers IV and V investigate the impact of automatic de-identification on the utility of training data for pre-training and fine-tuning purposes, respectively.

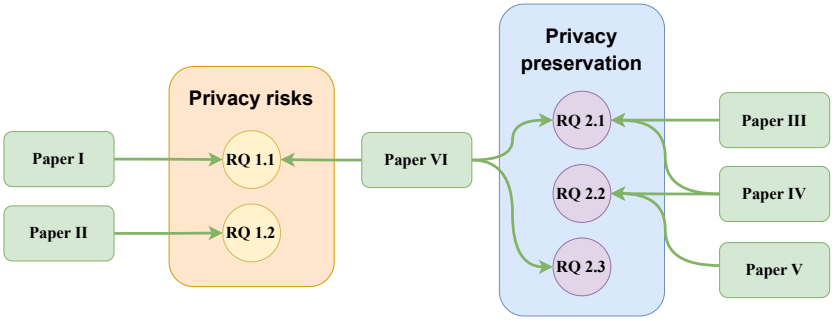

Figure 1.1: Answering the overarching research question of this thesis required addressing four subquestions. These fall into two themes, and the relation between the themes, the subquestions, and the papers that address them is illustrated.

Finally, Paper VI analyzes how synthetic training data can be generated from clinical corpora while limiting the amount of data that are exposed.

These research questions form two overarching themes: assessing privacy risks and exploring the trade-off between privacy and utility. In total, these themes produce five subquestions that the six papers address. The relation between the papers and the subquestions is illustrated in Figure 1.1. The themes and their subquestions are as follows:

Privacy Risks

RQ 1.1 Does the risk of language models leaking information increase when the quality of their generated data improves?

RQ 1.2 Do state-of-the-art membership inference attacks accurately quantify the privacy-preserving benefits gained from automatically de-identifying pre-training data for language models?

Privacy Preservation and Data Utility

RQ 2.1 Is domain adaptation of pre-trained language models necessary to reach state-of-the-art results in domain-specific NLP tasks?

RQ 2.2 How is the performance of language models affected by using automatically de-identified training data for both pre-training and fine-tuning?

RQ 2.3 How can sensitive data be used as efficiently as possible when creating synthetic corpora for domain-specific NLP?

1.2 Thesis Structure

This thesis is a compilation thesis composed of six papers tied together by a comprehensive summary—also known as a *kappa* in Swedish. As such, the thesis is divided into four parts:

Part I provides a brief introduction to NLP and privacy, which are the two main topics of the thesis.

Part II describes the research strategy, some important methods used in the studies. Finally, the design and results of the experiments are described.

Part III analyzes and discusses the results and contributions made in the six papers of the thesis. The final chapter of this part—and of the comprehensive summary—presents the conclusions of the thesis in relation to the research questions.

Part IV contains the six papers that form the foundation of the thesis. There is also a description of what the different authors contributed to each paper.

1.3 Notes of Terminology & Notation

NLP is a field with a rich but volatile body of terminology. Within the field, and in some of the papers in this thesis, it has become common to talk about *large language models (LLMs)*. Although this term has seen widespread and growing usage during the duration of this PhD project, it has some problems. Specifically, the models that fit the label *large* have

grown considerably since 2021, when this thesis work started. Models that were unfathomably large at the time are now sometimes referred to as *nano* models. Because of this, the term LLM is avoided in the comprehensive summary of the thesis, although it is used in some of its papers. Instead, the term pre-trained language model (PLM) is used to cover models considered large today *and* those considered large a few years ago.

The comprehensive summary also contains the occasional equation. There are some conventions that may be clearer when explicitly described:

Vectors are written as lower-case letters with arrows over them: $\vec{x}$.

Matrices are written as upper-case letters and are bolded: $\mathbf{W}$.

Model parameters are represented by the letter theta: θ.

Differences are denoted with the letter delta as a prefix: $x_2 - x_1 = \Delta x$.

Tokens are represented by the letter t, frequently with an index: t_n.

Sets and sequences are written as upper-case letters in a swash-style font: $\mathcal{S} = \{t_1, ..., t_n\}$.

Vocabularies are treated as sets and are represented by an upper-case swash-style $\mathcal{V}$ and their constituent tokens with a lower-case v: $v \in \mathcal{V}$.

Probabilities are represented with an upper-case P: $P\left(t_n = v \mid \mathcal{S}\right)$.

Part I

Background

This experiments in thesis take place the intersection between privacy-preserving machine learning and natural language processing, and explore these topics to the clinical domain. In this chapter, we will introduce these topics, lay the foundation for understanding the methods used in the studies that comprise this thesis, and provide the context necessary for interpreting the their results.

Chapter 2

Natural Language Processing

Natural language processing (NLP) and the intimately related field of computational linguistics[1] has been an active area of computer science since at least the 1940s (Jurafsky and Martin, 2025). Approaches to NLP can be roughly divided into *rule-based* methods and NLP based on *machine learning algorithms*. In practice, NLP systems rely on a mixture of techniques to achieve their results.

Rule-based NLP methods, such as regular expressions, can effectively process well-behaved language. However, natural language data are often unpredictable and noisy, so carefully constructed rules may be brittle due to the assumptions made when crafting them. On the other hand, machine learning algorithms learn heuristic patterns directly from data, often making them more resilient to unanticipated inputs. This doctoral thesis focuses mainly on the machine learning approaches to NLP. Specifically, we focus on methods based on deep learning.

[1]Delineating these two fields is a topic for sometimes heated debates. The author's opinion is that it is a question of whether the emphasis is placed on the computational aspects or the linguistic aspects of processing human language.

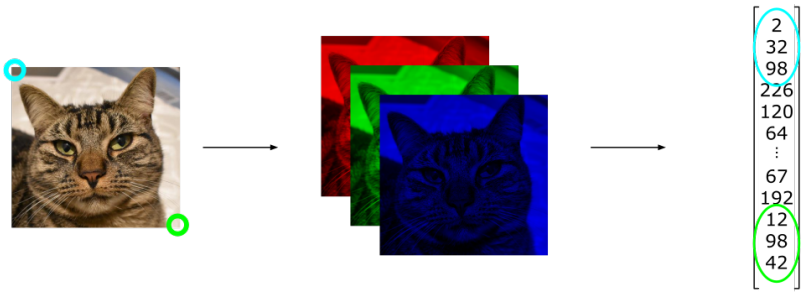

Figure 2.1: In image processing, we can represent an image as a vector by viewing the image as a long string of pixels. These pixels can be viewed as integer triplets, with the values corresponding to the pixels' RGB values. However, the image can also be converted into a grayscale image, shrunken, or transformed in some other way before further processing to avoid including too much information.

2.1 Representing Words as Vectors

Deep learning relies on the universal approximation theorem (Hornik et al., 1989; Cybenko, 1989), which states that any function mapping between two real vector spaces can be approximated using a feedforward neural network with at least one hidden layer[2] (Goodfellow et al., 2016). A feedforward neural network is a system of mathematical functions that consists of an input layer, an output layer, and one hidden layer. When we only have one hidden layer, a system that predicts $\vec{y}$ from an input vector $\vec{x}$ can be expressed as follows:

$$\vec{y} = f(\mathbf{W}\vec{x} + \vec{b}), \tag{2.1}$$

where f is a non-linear activation function[3] and the matrix $\mathbf{W}$ and vector $\vec{b}$ are parameters that need to be tuned. The parameters are *learned* from training data that associate a body of inputs with their expected outputs. The learning is driven by a learning algorithm, and the most common algorithm that is used for training in contemporary deep learning in NLP is stochastic gradient descent.

[2] In practice, neural networks have many layers in complex configurations. Examples can be found in Section 2.5.

[3] This activation function can be any non-linear function. Choosing an appropriate function is an art in itself and depends on the specific type of network one wants to construct.

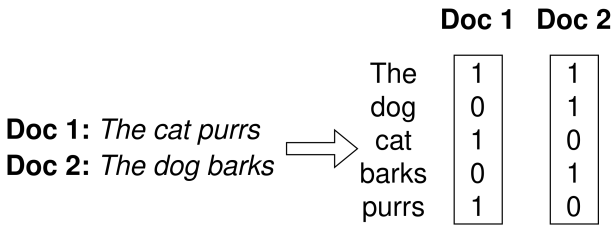

Figure 2.2: By associating each unique word with an index in a vector, the two example documents can be vectorized into two different vector representations. These bag-of-words representations of the documents are illustrated by the columns in the table.

	Doc 1	Doc 2
The	1	1
dog	0	1
cat	1	0
barks	0	1
purrs	1	0

Figure 2.3: With a sufficiently large collection of documents, the rows of the collection of bag-of-words vectors capture meaningful semantic information. In this small example, the row representations for *cat* and *purrs* are the same. This means that even this small example captures the association between purring and cats.

The acute reader will have noticed that the input to the neural network—$\vec{x}$—is a vector. This is the case for all neural networks. By converting our human-understandable input into a vector representation, we can take advantage of the power of the universal approximation theorem. However, this conversion is non-trivial and depends on the specific type of data that are to be processed. For example, in image processing, an image can be viewed as a series of integer values corresponding to the RGB values of each pixel[4]. This is a relatively straightforward representation, as illustrated in Figure 2.1. Even in this case, however, decisions need to be made regarding how to down-sample and crop the image to make it fit into a finite and fixed-sized vector.

Unlike images, natural language does not have an obvious vectorization procedure. Perhaps the most straightforward vectorization technique for

[4]Pixels are often represented as triplets of integers. Each integer corresponds to the amount of **r**ed, **g**reen, and **b**lue (RGB) needed to obtain the pixel's color.

natural language data is the bag-of-words (BoW) approach. The prototypical BoW model considers a document as a collection of words without any order (a bag of words). In Figure 2.2, the documents are represented as count vectors, where every value reflects the number of times a word associated with the corresponding index occurs in the document (Jurafsky and Martin, 2025). The BoW approach is sufficient for some problems, but a significant disadvantage is that it ignores the word order. This means that the two sentences *"work is life"* and *"life is work"* have identical BoW representations even though, semantically, they are very different. Furthermore, using the index of a vector to represent an individual word ignores the semantic similarities between words. A BoW vector does not capture the fact that *different* and *dissimilar* are synonyms, instead representing them as unrelated indices in a vector.

To create more semantically rich representations of words, NLP practitioners have turned to learned vector representations. These may, for example, be based on neural methods (Mikolov et al., 2013; Bojanowski et al., 2017), but they can also be learned using simpler techniques. A straightforward idea is to represent a word using the documents in which it occurs. This inversion of the BoW approach is exemplified in Figure 2.3 and illustrates the idea of representing words using their co-occurrences. This idea is sometimes referred to as the *distributional hypothesis* and suggests that the meaning of words can be discerned by studying the context in which words are used (Wittgenstein, 1953; Firth, 1957). As we will see in later sections, this idea has proved to be very powerful and currently underpins most—if not all—state-of-the-art NLP systems.

2.2 WORDS, TOKENS, & SUB-WORD TOKENS

In the previous section, we discussed how words can be converted into vectors. However, there is an important step that is still missing. Computers represent textual data as continuous strings of characters. Before we can convert words to strings, we need to decide how to convert the continuous strings of characters into discrete words. This process is known as *tokenization* and the resulting discrete units—discussed as

words in previous sections—are referred to as *tokens* (Jurafsky and Martin, 2025). The finite set of tokens that an NLP model is configured to process is called the model's *vocabulary*.

A naïve approach would be to split the string of characters based on white-space characters and on punctuation. This strategy will indeed produce discrete tokens, but will suffer from a fair bit of redundancy. For example, the token *dog* in Figure 2.3 will have a vector, but a hypothetical later token *Dogs* will have a different vector. Worse, if *Dogs* does not occur in the training corpus, then the model will not have any vector at all for this token. To rectify this situation, tokenization pipelines typically apply different normalization strategies such as *stemming* and *case normalization* to transform these two distinct words into the same token (Manning et al., 2008). In this example, stemming would involve removing the plural suffix *-s* and subsequent lowercasing would transform *Dogs* into *dog*. Consequently, a model using this more sophisticated tokenization strategy will have a more compact and efficient vocabulary.

Incorporating normalization techniques into the tokenization pipeline will reduce—but not eliminate—the risk of encountering tokens outside the model's vocabulary. The number of words possible in human languages is infinite and constantly changing, and can never be accomodated by a finite vocabulary of the kind we have discussed so far. These out-of-vocabulary tokens are often assigned a special *unknown* token[5] to allow models to process input sequences even if some tokens lack a vector representation. Nevertheless, this strategy will necessarily result in information essentially being thrown away as all out-of-vocabulary tokens will be represented by the same single vector.

A tokenization strategy that almost entirely avoids the problem of unknown tokens is *sub-word tokenization* (Jurafsky and Martin, 2025). As their name implies, sub-word tokenizers go beyond the word level when segmenting strings into tokens. In addition to word-level tokens, sub-word tokenizers also support tokens representing *parts of a word*. The vocabulary of word-level and sub-word tokens is *learned* from a corpus, and is typically specific to the NLP model it was created for. As illustrated in Figure 2.4, a sub-word tokenizer can represent words even if

[5]These unknown tokens are typically rendered as `[UNK]` when discussing model vocabularies.

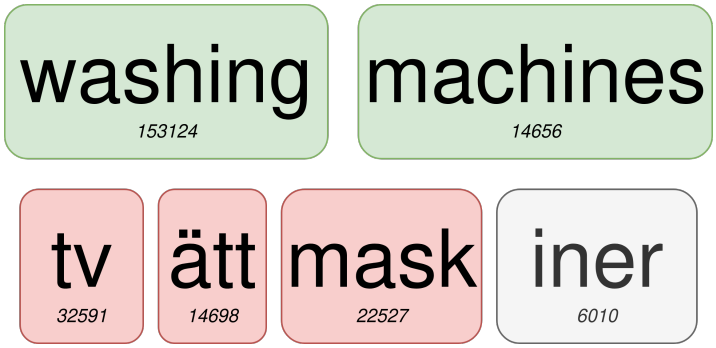

Figure 2.4: Sub-word tokenization strategies learn a vocabulary that contains common words in addition to meaningful sub-words. In this example, based on the tokenizer used for GPT-4o from OpenAI, *washing machine* is split into two whole-word tokens. On the other hand, the Swedish translation *tvättmaskin* is split into four sub-word tokens. Making matters worse, these sub-words do not convey anything semantically meaningful related to washing machines.

they were never encountered in the training corpus, and sometimes creates morphologically motivated segmentations of words it had encountered. The first sub-word tokenization algorithm was based on byte-pair encoding (Sennrich et al., 2016), but many other algorithms have been developed since. At the time of writing, all widely-used PLMs rely on a sub-word tokenization algorithm to segment their input.

2.3 Classification

Machine learning models are designed and trained to produce accurate predictions. These predictions can take many different forms. At the beginning of Section 2.1, we saw how a feedforward neural network can produce an output $\vec{y}$ from an input vector $\vec{x}$. The output may be a vector, as in the previous formulation, or it can be a scalar value[6]. Often, the indices in the output vector correspond to different *classes* assigned to each data point. The model is then trained to predict the class of an input by maximizing the value at the correct index in the output vector. This process is called classification.

[6]When the task is to produce a scalar value, it is called a regression problem.

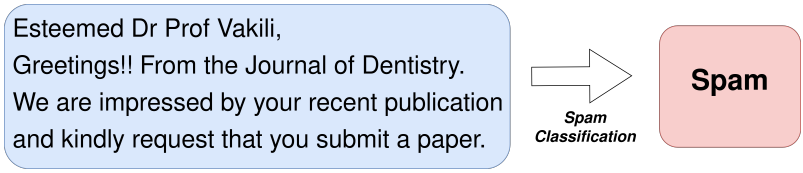

Figure 2.5: Document-level classification involves assigning a label to a document. This example shows how a factuality classifier might classify the certainty of a diagnosis. The output is a label, either *spam* or *not spam*. In this example, this representative spam e-mail is correctly identified as such.

There are numerous examples of classification problems in NLP. Many problems involve assigning a label to a sequence of tokens. These problems are called document-level classification tasks. The model processes all tokens in the document and uses the information to assign a label to the entire document. An example of document-level classification is illustrated in Figure 2.5. This example illustrates a spam classification problem. This task involves classifying wheter or not a document (e.g., an e-mail) is spam or benign.

Sometimes, it is more interesting to classify a document on a more granular level. A common class of tasks that involve doing this are token-level classification problems, where labels are assigned to each individual token. A token-level task that is central to this doctoral thesis is NER. As illustrated in Figure 2.6, documents frequently contain the names of persons, place names, and other interesting information. These are examples of *named entities*. Named entities can also be more specialized, as in clinical entity recognition, where the aim is to detect mentions of body parts, diagnoses, and other clinically relevant entities. In either case, the aim of NER is to classify each token in a document based on a pre-defined set of entity types (Jurafsky and Martin, 2025).

A special form of classification—often treated as a separate problem—is next-token prediction. This classification problem tasks a model with predicting the next token in a sequence, based on the preceeding tokens[7].

[7]In masked language modeling, as described in Subsection 2.6, the *next* token need not be *preceeded* by the other tokens. Instead, the next token is predicted based on the surrounding context tokens.

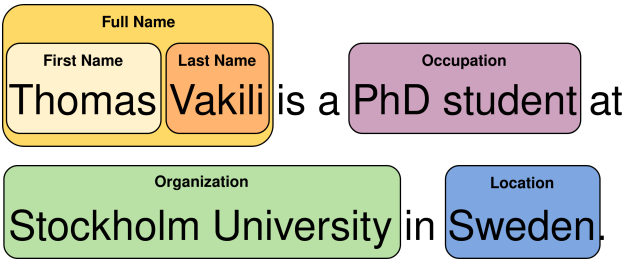

Figure 2.6: This example illustrates a few named entities commonly found in natural language. NER systems can have many different categories, and they can even have nested entity types. In this example, the *Full Name* entity consists of a *First Name* and a *Last Name*. Note also that the *Organization* entity consists of multiple words. Multi-word entities are common in NER, and the determination of where an entity begins and ends can sometimes be ambiguous.

Given a vocabulary $\mathcal{V}$ and a sequence of tokens $t_1, t_2, ..., t_n \in \mathcal{V}$, the goal is to predict the most probable next token t_{n+1}:

$$t_{n+1} = \arg\max_{v \in \mathcal{V}} P\left(v \mid t_1, t_2, ..., t_n\right). \tag{2.2}$$

This classification task is important for two reasons. First, the task allows us to *generate* new tokens. There is no need to stop at t_{n+1}, one can proceed with tokens t_{n+2}, t_{n+3}, and so on. This is in contrast to the other types of classification, where we transform a natural language input into a label or a fixed-size sequence of labels. Second, and relatedly, next-token prediction requires the model to have rich representations of its vocabulary in order to effectively use the context. This fact is important for understanding why next-token prediction, and variations of this task, are essential to modern NLP techniques.

2.4 Metrics for Evaluating Classifiers

One of the most important aspects to consider when assessing a machine learning model is the extent to which it makes accurate predictions from unseen data. This assessment is typically done by splitting the dataset into two randomly sampled parts, one large subset for training and a smaller subset for testing. This smaller dataset is called the held-out test

dataset. The predictive power of the model is measured by classifying these unseen data points and comparing the predicted classifications to their true classes.

Even though the training-test split is done randomly, there is a risk that the smaller held-out test dataset is not representative of the dataset as a whole. One strategy that can be used to mitigate this is called *k-fold cross-validation* (James et al., 2013). The dataset is partitioned into k equally sized *folds*, where $k - 1$ folds are used to train the model and the final fold is used for evaluation. This process is repeated k times such that each of the k folds is used for evaluation, and the metrics from each fold are then aggregated. K-fold cross-validation is especially useful in cases when training data are scarce, since small dataset are at greater risk of getting a non-representative train-test split. On the other hand, experiments with larger datasets still benefit from being run in multiple trials—but this must be balanced against the increased computational costs.

While it is important to decide how to evaluate a system, one must also decide what to measure. A natural impulse is to measure the *accuracy* of the system. This is simply the proportion of test data points that were correctly classified. While useful, this metric does not differentiate between *true positives (TP)*, *true negatives (TN)*, *false positives (FP)*, and *false negatives (FN)*. For many applications, it is important to distinguish between these notions since they represent distinct types of errors. Two important metrics that capture more nuance are the *precision* and *recall* metrics:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \mathbf{FN}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \mathbf{FP}}. \tag{2.3}$$

The recall metric is sometimes colloquially referred to as the ”hit rate.” It represents the degree to which the classifier successfully detects members of the positive class (Goodfellow et al., 2016). For example, a spam filter with a high recall will catch many spam emails. If the recall is too low, then many spam emails will end up in the user’s inbox. A high recall is critical when the consequences of failing to detect positive instances are dire. The precision metric represents the degree to which a positive prediction can be trusted (ibid.). Returning to the spam filter, a high-precision filter is one that rarely sounds false alarms. A high precision is important when we want to avoid false positives. For

example, a spam filter with high precision will ensure that the user does not need to worry about legitimate emails ending up in the spam folder.

In general, there is tension between these two metrics. This tension is referred to as the *precision-recall trade-off* (Manning et al., 2008). If a classifier is too eager to classify samples as positive, we will have a low false negative rate but a high false positive rate. This translates into a high recall but a low precision. On the other hand, an overly conservative classifier will yield few false positives but also miss many true positives. Such a classifier will have a high precision but a low recall. A metric that combines these two measures is the F_β score:

$$F_\beta = (1 + \beta^2) \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}. \tag{2.4}$$

The F_β score[8] of a classifier is a weighted harmonic mean of the precision and recall (Rijsbergen, 1979; Chinchor, 1992). A high value of β results in the recall having a larger impact, and vice versa. A common variant is the F_1 score, which weights both metrics equally:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.5}$$

2.5 TRANSFORMER MODELS

In previous sections, we learned how to convert human-readable strings of text into discrete tokens and finally into semantically rich vectors. However, the question of how to process *sequences* of tokens was left unanswered. For example, the simple feedforward network described at the beginning of the section has no concept of time. Thankfully, there are many deep learning approaches for dealing with sequences. Until 2018, the field was dominated by variants of *recurrent neural networks (RNNs)*, such as long short-term memory (LSTM) networks. These

[8] The name *F*-score was first used in the MUC-4.0 challenge (Chinchor, 1992; Christen et al., 2023) and was—according to Sasaki (2007)—chosen by mistake. F-score is essentially a misspelling of a different metric also defined by Rijsbergen (1979) that he had named *E*.

networks rely on sequential processing, meaning that a sentence is processed word by word[9].

In 2018, Google released the BERT model (Devlin et al., 2019), which popularized the *transformer* architecture (Vaswani et al., 2017). Transformer models do not process sentences word by word. Instead, they process entire sequences all at once using a mechanism called self-attention[10]. This allows transformer models to build powerful contextual representations for words not just from the words themselves but by using all the words surrounding them. The static word vectors described in the previous subsection would have the same vector for the word *bank*, regardless of whether this refers to a *river bank* or a *central bank*. A contextual word representation can disambiguate these dissimilar meanings based on the context.

The self-attention mechanism builds contextual representations by introducing the concepts of query, key, and value vectors. These are analogous to concepts from information retrieval. The query vector of a word is combined with the key vectors of the words in its context to compute relevance scores for the surrounding words. These new vectors are computed using word embeddings $\vec{w}$ and the matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ for each vector type. The contextual embedding $\vec{z}_w$ is composed of the value vectors of all words in the context, weighted by the relevance for the word w. Given a sentence of words $w \in W$, we compute the contextual embedding $\vec{z}_i$ for the word i:

$$\vec{q}_w = \vec{x}_w\mathbf{Q}, \quad \vec{k}_w = \vec{x}_w\mathbf{K}, \quad \vec{v}_w = \vec{x}_w\mathbf{V}, \tag{2.6}$$

$$\vec{z}_i = \sum_{w \in W} \vec{q}_i \cdot \vec{k}_w \cdot \vec{v}_w. \tag{2.7}$$

Modern transformer architectures use many instances of these self-attention mechanisms, introducing many new matrices and vectors to learn during training. The intuition behind this is that each

[9]Variants such as bidirectional LSTMs (Graves and Schmidhuber, 2005) process sequences from left to right and right to left at the same time. However, the processing is still done word by word.

[10]Many models include other mechanisms as well. For example, BERT also uses *positional embeddings*. However, researchers have found that BERT can reconstruct the word order using other signals (Sinha et al., 2021; Abdou et al., 2022). Thus, this mechanism is left out for the sake of brevity.

Thomas Vakili holds a `[MASK]` degree
↓
Thomas Vakili holds a **master's** degree

Figure 2.7: In this example, the masked language model needs to reconstruct the sentence by replacing the `[MASK]` token with the correct value: *master's*. Some other degree (e.g., doctoral) would have been just as semantically acceptable but was not the sought-after replacement. A positive consequence of this situation is that masked language models are able to learn facts about the world. On the other hand, this information can become stale as the world changes. Furthermore, if the facts are sensitive, they should not be learned at all.

self-attention unit, hopefully, learns a specific linguistic task. For example, this task could be word-sense disambiguation, as in the *bank* example, or coreference resolution (e.g., associating pronouns with their referent). However, these additional matrices and vectors mean that there are many more parameters that need to be learned from data. The smallest BERT model consists of 110 million parameters (Devlin et al., 2019), but the *Llama 4 Behemoth* model consists of two *trillion* parameters (Meta AI, 2025). Learning parameters at this scale requires enormous amounts of data.

2.6 Pre-Trained Language Models

Annotated data are often scarce, but most transformer-based language models learn through self-supervised learning using large unlabeled corpora. For example, BERT was trained using data from Wikipedia and the BookCorpus (Zhu et al., 2015) using a procedure called *masked language modeling (MLM)*. This procedure is inspired by the *Cloze* task (Taylor, 1953), which may be familiar to students of foreign languages. The prototypical masked language model is trained by reconstructing sentences in which some tokens have been replaced by `[MASK]` tokens. An example of this task is illustrated in Figure 2.7. Formally, the goal of MLM is to predict $t_{\texttt{[MASK]}} \in \mathcal{V}$ given a sequence $T = \{t_1, t_2, ..., t_n\}$:

$$t_{\texttt{[MASK]}} = \arg\max_{v \in \mathcal{V}} P(v \mid T \setminus \{t_{\texttt{[MASK]}}\}). \tag{2.8}$$

Training transformer models using self-supervised tasks such as MLM is called *pre-training*. Masked language modeling itself has limited use cases but trains the model to build semantically rich representations. PLMs are then *fine-tuned* to perform more specific tasks. When neural networks are trained from scratch, the models need to learn to model sequences effectively in addition to learning to solve specific tasks. Because PLMs already have the ability to model sequences, they typically need less annotated (and expensive) task-specific training data to perform well for a specific task.

The most common use-case for masked language modeling is to create *encoder models*. These models are trained to produce embeddings of their input sequence for further processing. Typically, these embeddings are used for classification—both document-level and token-level classification. For example, BERT models can be adapted to perform classification tasks by adding additional layers to the architectures. These new layers take the BERT embeddings as their input and the whole network is then trained to map the input sequences to the desired labels.

Autoregressive language models[11] are another category of PLMs. These are trained to perform the traditional form of next-token prediction discussed in Section 2.3:

$$t_{n+1} = \arg\max_{v \in \mathcal{V}} P\left(v \mid t_1, t_2, ..., t_n\right). \qquad \text{(Eq. 2.2)}$$

In contrast to MLMs, autoregressive language models typically rely on next-token prediction itself to perform various tasks. Consequently, autoregressive language models are mainly used to *generate* text. In addition to self-supervised next-token prediction, many autoregressive language models also undergo instruction-tuning (Jurafsky and Martin, 2025). The goal of instruction-tuning, as the name suggests, is to train models to follow instructions. Typically, this is accomplished by training models using reinforcement learning from human feedback (RLHF; (Stiennon et al., 2020)). However, the autoregressive models in this thesis have only been trained to perform next-token prediction.

The generative objectives used to train autoregressive language models can be employed to solve various problems. Doing so involves crafting

[11] Sometimes referred to in the literature as *causal* or *generative* language models.

prompts that convey the nature of the task that the model should solve. When solving a classification problem, the prompt will often include a number of examples of input-label pairs. This approach is called *few-shot* classification, and is contrasted with *zero-shot* classification where the model does not have any examples (Chang et al., 2008; Palatucci et al., 2009). First and foremost, and in contrast to most MLMs, autoregressive language models excel at generating synthetic text. This application is how autoregressive models are used in this thesis, as is further explained in Subection 6.3.2.

2.7 Encoders and Decoders

Generally, language models can be grouped into three different architectures: *encoder*, *decoders*, and *encoder-decoder* models. These different architectures process data and make predictions in different ways, and—traditionally—have been used for different use cases. Encoder models process their input and produce representations that *encode* information that is relevant to the task. These representations are then used for *classification*. Decoder models, on the other hand, are used for generative purposes. They are essentially also classifiers, but the classification focuses on next-token prediction. Finally, encoder-decoder models are models that first encode their input and then decode the resulting representations, typically into new text.

Before the advent of pre-trained language models, the prototypical use cases for these models would be very different. An encoder model would often be used to perform classification tasks where the goal was to map documents or tokens to their corresponding labels. An example would be sentiment analysis, where documents are classified based on the sentiment expressed in the document. Encoder-decoders, on the other hand, were used for tasks that transformed one text into another. For example, state-of-the-art translation models would often rely on encoder-decoder models. Decoder models, on the other hand, had a more limited set of use cases. For example, they were used in speech processing to help produce statistically likely sequences from the hypotheses generated from processing audio of utterances.

Nowadays, many tasks are increasingly being treated as subsets of the next-token prediction task. This phenomenon has emerged in order to apply decoder models to tasks which they were not traditionally used for. The largest PLMs at the time of writing are all decoder models, which is the main factor driving this trend. Interestingly, researchers have repeatedly found that using smaller encoder models will often provide better results than using large decoder models (Aracena et al., 2024; Naguib et al., 2024). On the other hand, decoder models achieve their results without relying on big—and often sensitive—fine-tuning corpora. This aspect can be especially useful when dealing with sensitive domains, such as the clinical domain.

2.8 Clinical Applications of NLP

Throughout a patient's stay at a hospital, the clinical personnel who interact with the patient make notes. In many countries, these notes about the patient and their condition are entered into a digital patient record system[12]. These notes, along with any measurements taken from patients, are called electronic health records (EHRs). Although EHRs contain structured information such as lab measurements or diagnosis codes, much of the information is only available in natural language form. Clinical NLP is the extraction and processing of this valuable information from EHRs using NLP techniques.

Broadly speaking, clinical NLP can be used to achieve two objectives. One objective is to improve the *quality* of care for patients. This can be done by building tools that enhance the ability of the healthcare system to avoid harm and enhance medical professionals' ability to address the needs of their patients. Another use case for clinical NLP is improving the *efficiency* of care. These two objectives are frequently targeted jointly, but different applications emphasize certain aspects differently.

One example of an application that can improve the quality of healthcare is the detection of adverse drug events (ADEs). These can occur, for example, when a patient's prescribed medications interact in a harmful

[12]Sweden, the country in which this doctoral thesis is being written, had digitalized its patient record systems by 2007 (Dalianis, 2018).

way. ADEs are a serious problem, with one study finding that 22% of hospitalizations in Sweden could in some way be linked to ADEs and that 38.8% of these hospitalizations were preventable (Hakkarainen et al., 2014). Detecting these events before they cause harm can potentially save the lives of patients, and an example of applying NLP techniques to this problem with Swedish data is given by Henriksson (2015). In addition to preventing harm, reducing the frequency of preventable complications increases the efficiency of the healthcare system.

Another clinical NLP application for increasing the efficiency of medical professionals is automatic ICD-10 coding. ICD-10[13] is the 10th iteration of the ICD system, which assigns specific codes to different diagnoses. These systems have a long history of use for statistical purposes (Dalianis, 2018) and are also used to calculate reimbursements for caregivers. However, ICD-10 is very detailed and requires physicians to select codes from a catalog of 32,000 codes when writing EHRs (Dalianis, 2018). This cumbersome and error-prone process can be facilitated using clinical NLP, and the work of Remmer et al. (2021) represents a recent attempt to process Swedish EHRs. Improving this process can free up time for clinicians, allowing them to focus on their caregiving duties, while also improving the quality of epidemiological data for research purposes.

The previous two examples describe classification tasks. However, there are also many problems in clinical NLP that can be approached as generative tasks. For example, patients leaving a hospital do so with a *discharge summary*. This type of medical record summarizes the care recieved by the patient during their staty at the hospital. Typically, these records are longer and more thorough than the notes written during the patient's treatment at the hospital. Writing discharge summaries takes time and summarizing relevant information from previous notes can be challenging. Earlier attempts (Moen et al., 2016) at engaging with this problem using clinical NLP relied on *extractive* summarization, which works by detecting passages in the notes that should be relevant to the discharge summary. These relevant passages are then stitched together using different techniques. Today, generative language models are able to generate fluent summaries from a body of documents. This type of summarization is called *abstractive* summarization. Newer studies tend

[13]ICD is short for **I**nternational Statistical **C**lassification of **D**iseases and Related Health Problems.

to favour the abstractive approach (Clough et al., 2024), although some struggle with ensuring the accuracy of the summaries in the face of so-called *hallucinations*[14] (Schwieger et al., 2024).

In many cases, in particular when dealing with encoder models, clinical-domain data needs to be used to pre-train or fine-tune the model (Lehman et al., 2023). General-domain data does not reflect the pecularities of how clinical notes are written. Physicians and other health professionals typically write these in a hurry, and clinical notes have a high amount of misspellings. Furthermore, the health professionals frequently use terminology and abbreviations that are specific to their disciplines (Dalianis, 2018). Using general-domain data even for tasks that are not specific to the clinical domain, such as PII detection, is insufficient to reach state-of-the-art results. Similarly, multiple studies—including Papers IV and VI—show that pre-training using clinical data also results in better-performing models. The process of domain adaptation will be further detailed in Section 6.1. However, before moving on to the next chapter, it is important to highlight that training language models using clinical data—be that for domain adaptation or for tackling downstream tasks—comes with risks of privacy leakage.

[14] The somewhat controversial term hallucination can be loosely defined as the tendency of language models to produce plausible but false information.

Chapter 3

Privacy

This thesis operates in the intersection between NLP and privacy. There is a rich philosophical debate about what the concept of privacy actually entails and what it means to protect it (Roessler and DeCew, 2023). In this thesis, and in line with much of the literature about privacy-preserving machine learning, we take a pragmatic view of privacy. This view is that privacy is preserved when trained models reveal as little as possible about the persons described in the training data. This chapter introduces central concepts used throughout the rest of the thesis, including different types of attacks and protective techniques.

3.1 Personally Identifiable Information

Personally identifiable information (PII) can take many different forms. What these types of information share in common is that they can be used to identify an individual—directly or when combined with other information. Dalenius (1986) introduced this distinction between *direct identifiers* and *quasi-identifiers*. Examples of PII that are direct identifiers are names and social security numbers. Quasi-identifiers, on the other hand, can span a wider range of attributes. Although quasi-identifiers do not directly identify an individual, they can do so when combined with external data. Indeed, Sweeney (2000) showed that 87% of the US population could be uniquely identified by their ZIP code, gender, and date of birth.

However, quasi-identifiers are not necessarily represented as individual entities in a text. For example, the list of places a person has visited may be a quasi-identifier if the itinerary is unique enough. Additionally, seemingly innocuous words or phrases can be quasi-identifiers depending on what other information is provided and whether an adversary has access to additional external information. The systems considered in this thesis do not attempt to overcome this problem. While recognizing this limitation, they instead rely on a fixed set of direct identifiers and quasi-identifiers.

3.2 Legal Dimensions of Privacy in Clinical NLP

Applying machine learning to the clinical domain involves processing large amounts of possibly sensitive data. From a purely ethical viewpoint, this is something that must be done with great care. Clinical NLP has many applications to healthcare that could benefit patients and medical professionals, but these applications also come with risks. The Health Insurance Portability and Accountability Act (HIPAA; CMS, 1996) is a regulation in the United States that has been in effect since 1996. This law regulates how a form of PII—personal health information (PHI)—can be used and shared and has had important ramifications for the field of clinical NLP. The law stipulates that certain classes of PII must be removed or otherwise obscured from clinical data before these data can be used for research purposes (Dalianis, 2018). This requirement is one reason behind the development of automatic de-identification, which is described in detail in Subsection 3.4.2 and aims to detect and sanitize PII using various automatic techniques.

The risks involved in processing large amounts of personal data are not specific to the clinical domain, and the widespread adoption of data-hungry algorithms has prompted other legal developments to safeguard the privacy of citizens. Citizens of the European Union have, since 2018, been protected by the General Data Protection Regulation (GDPR). This is a law with higher ambitions than HIPAA, and it has a scope that goes beyond the health domain. The GDPR requires that *all*

processing of personal data owned by *any* EU citizen complies with the regulations.

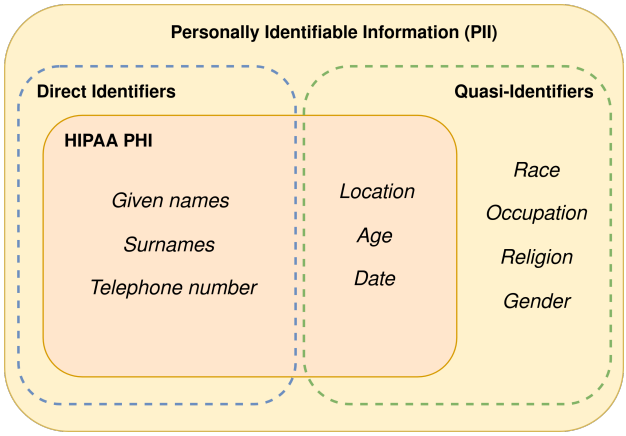

Figure 3.1: There are many different types of PII. This figure illustrates the relationship between direct and indirect identifiers, and how these relate to the term *PHI* as deifined by the HIPAA.

Personal data is defined in Article 4 as any information relating to an identifiable natural person. This definition is much broader than the definition of PHI in HIPAA and could, for example, include information such as a person's height. The relation between the HIPAA definition of PHI and PII in a more general sense is illustrated in Figure 3.1. The GDPR requires that such information is protected, but the criteria for what protections are required by law are not clear. Furthermore, it is not clear whether a machine learning model is considered aggregated data, for which there are less strict requirements (Cummings and Desai, 2018). As we will see in the next section, machine learning models *can* leak personal information. These risks can be partially mitigated, as will be described in Section 3.4. Unfortunately, there is no consensus regarding what degree of privacy preservation is enough to satisfy the requirements of the GDPR. Regardless, NLP practitioners dealing with sensitive data have both legal and ethical obligations to protect data to the best of their ability.

3.3 Privacy Attacks

As mentioned in Section 2.5, pre-trained language models consist of millions—even billions—of parameters. These parameters are trained to perform well on self-supervised tasks that involve reconstructing training data. Because of this, they sometimes learn to reconstruct their training data too well. We call this problem the *unwelcome memorization*[1] of training data. This tendency to memorize data makes language models susceptible to privacy attacks that target their training data. There are two main categories of such attacks: *training data extraction* and *membership inference* attacks.

3.3.1 Training Data Extraction

Training data extraction attacks are, as the name implies, attacks that are able to extract parts of the training data used to train a model. These can be designed in many different ways. What most methods have in common is that they require the adversary to find a method that causes the model to generate information. Typically, the adversary must then have a method for determining whether the generated output was indeed part of the training data. This later step is equivalent to the concept of membership inference and is discussed in the next subsection.

The way that an adversary forces a model to generate possible training data varies depending on the background information they have and the attacked model's architecture. *Unprompted attacks* do not make any assumptions about the nature of the training data. Such attacks rely on a sampling algorithm to generate text that may contain training data. Other attacks take a more controlled approach. *Prompted* attacks ask the model to continue a sentence that has been crafted to prime the model to leak sensitive information[2]. A related attack is the *slot-filling* attack, which is designed similarly to self-supervised training in masked language modeling. In this attack, the model is asked to fill a slot in a sentence. This

[1] Sometimes, the term *unintentional* memorization is used. This term is arguably a misnomer, since reconstructing training data is an explicit training objective when pre-training PLMs.

[2] An example of such prompting would be: *Thomas Vakili's personal number is ...*

incomplete sentence could be a leaked fragment of the training data[3] or a sentence designed to make the model accidentally reveal memorized facts.

Some language models have been shown to be susceptible to training data extraction attacks. Carlini et al. (2021) attack the model GPT-2 (Radford et al., 2019) by mounting prompted and unprompted attacks. The data they generate are then verified as training data and are shown to contain sensitive information such as social security numbers, names, and e-mail addresses. Part of the success of this particular attack is that GPT-2 is a generative language model. This means that it has been designed to produce text. In contrast, there were no examples of successful training data extraction attacks that target BERT models at the time of writing. One likely reason for this is that BERT models were not designed to generate large amounts of text. However, in Subsection 6.3.3 and Paper I, we will explore methods for forcing BERT models to generate text with reasonable quality.

3.3.2 Membership Inference

Unlike the attacks described in the previous subsection, membership inference attacks do not aim to extract training data. Instead, they seek to determine whether or not a data point was part of a model's training data. These attacks are often included as part of a training data extraction attack. An adversary typically needs to discern whether an extracted data point is actually from the training data or if it is synthetic. However, a successful membership inference attack can by itself reveal sensitive information. For example, models trained using clinical data typically use data that come from a specific and known set of hospitals and clinical units from a specific time range. Learning that a patient occurs in the training data can reveal that a patient has visited a clinic or hospital and when this may have occurred. If the set of clinical units is specific enough, then this can also reveal information about what illnesses a patient has suffered from.

Membership inference attacks can be constructed in different ways and for different forms of data. In this thesis, we focus on attacks that target

[3]For example, an adversary attempting a slot-filling attack might already have a de-identified sample from an electronic health record: *Mr.* [redacted] *came in with an ulcer.*

pre-trained transformer models. These membership inference attacks discern whether a data point was part of a model's training data by exploiting the model's *reaction* to the data point. The intuition behind the attack is that models react differently to data points they have been trained with compared to novel data. One reason for this is that, as we learned in Section 2.5, models like BERT are trained to reconstruct their training data.

The reaction that is exploited can be a variety of different metrics. Examples include the model loss (Jagannatha et al., 2021) or normalized energy (Mireshghallah et al., 2022). Stronger results can be achieved by comparing the reaction of the attacked model to that of a reference model that we know has not been trained using the target data point. Mireshghallah et al. (2022) use such a scheme, which includes the aforementioned normalized energy values, and Carlini et al. (2021) experiment with using metrics such as the Zlib compression rate (Deutsch and Gailly, 1996) as the reference model. As will be discussed in Subsection 3.3.2, membership inference attacks have been proposed as a way of measuring privacy risks. The main contribution of Paper II is an assessment of this idea.

3.4 Privacy Preservation

Having established that transformer models are susceptible to privacy attacks, we now turn our attention to methods for mitigating the resulting privacy risks. Privacy-preserving machine learning is a broad topic that covers a variety of techniques for many different types of models. The field is not specific to NLP and transformer models. Indeed, many techniques are designed with other machine learning domains in mind.

Different privacy-preserving techniques target different threat models. As illustrated in Figure 3.2, the techniques protect different parts of the machine learning process. The focus of this thesis is data-oriented techniques, specifically automatic de-identification and training data synthesis. Differential privacy is discussed throughout the papers, and is also introduced in this section. These techniques try to protect the PII of persons mentioned in the *training data* of models. Other

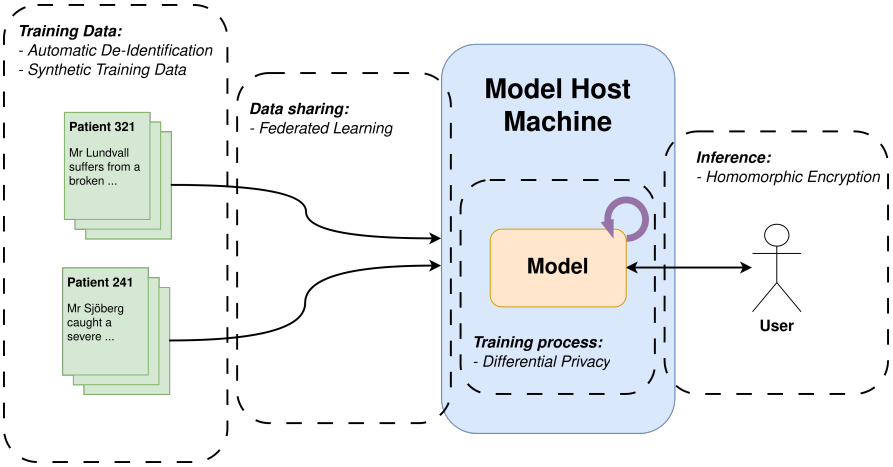

Figure 3.2: Different privacy-preserving techniques protect different parts of the machine learning pipeline. This thesis focuses on the techniques that target that prevent the *leakage of training data*.

techniques—such as federated learning or homomorphic encryption—protect other parts of the machine learning pipeline and are therefor outside the scope of this thesis.

3.4.1 Differentially Private Learning

Differential privacy (DP) is a notion of privacy originally defined for aggregated analyses of database records. DP provides mathematically rigorous and quantifiable privacy guarantees and was originally proposed by Dwork et al. (2006b). Essentially, DP quantifies privacy by describing the probability that any individual data point will alter the output of an aggregated analysis. Of course, if none of the data points influence the aggregated analysis, then the data are useless. The trade-off between privacy and utility is defined in terms of a parameter ϵ. Given this parameter ϵ, an aggregation M with a range S, a dataset D, and a data point $d \in D$, we have ϵ-differential privacy if

$$P\left[M\left(D\right) \in S\right] \leq e^{\epsilon} P\left[M\left(D'\right) \in S\right], \quad \text{where} \quad D' = D \setminus \{d\}. \tag{3.1}$$

A common variation of this definition allows the relationship to fail with a tolerance δ (Dwork et al., 2006a). This δ is added to the right-hand side

of the inequality, which gives us (ϵ, δ)-differentially private systems. Lower ϵ and δ values mean stronger privacy guarantees. DP has been operationalized to design differentially private algorithms for many domains. The most relevant application, for the purposes of this thesis, is the emergence of differentially private training algorithms for training machine learning models. Abadi et al. (2016) propose such an algorithm for training deep neural networks. They do so by altering the stochastic gradient descent algorithm (Goodfellow et al., 2016), which updates a network based on the gradients—the partial derivatives with respect to the training samples—and is ubiquitous in deep learning implementations. The gradients are clipped (i.e., shrunken in magnitude) and altered by adding noise to accommodate the selected values of ϵ and δ.

Using differentially private training algorithms allows us to train (ϵ, δ)-differentially private models. However, DP was originally defined to protect database records containing structured information. In the next subsection, we will describe a different approach to protecting sensitive data that is tailored specifically to natural language data. Indeed, all papers in this thesis study data-oriented alternatives to privacy preservation. Nevertheless, differential privacy is a highly relevant concept that is discussed in many of the papers in this thesis.

3.4.2 Automatic De-Identification

Although data such as an EHR may be sensitive, the individual words and phrases differ in their sensitivity. An EHR describing a patient with a migraine is not necessarily sensitive unless it contains *identifiable* information. This reasoning is a guiding principle of the HIPAA guidelines that compel distributors of clinical data to remove PHI, as discussed in Section 3.2. Removing PHI or other PII is a form of *de-identification*.

De-identification has traditionally been done manually. This process of manually searching for PII to remove is still practiced and is very laborious. To save time and other resources, researchers have studied how to automate this process. This is called *automatic de-identification*. This process involves processing large volumes of text and detecting words or phrases that are PII. This can be done using rule-based systems, systems

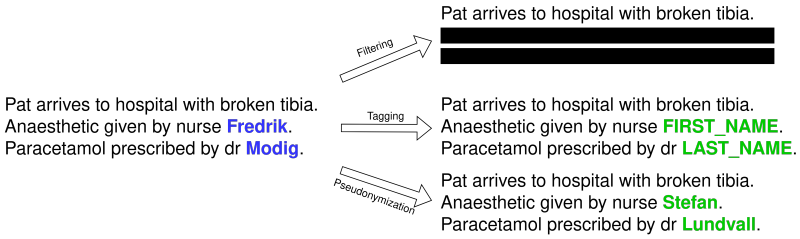

Figure 3.3: Example of a successful de-identification using the three sanitization approaches considered in this thesis. Filtering removes a sentence from the data if PII is found. Tagging replaces the PII with its class (in this case, `FIRST_NAME` and `LAST_NAME`). Pseudonymization instead replaces the PII with realistic surrogates.

based on machine learning, or combinations of both approaches (Dalianis, 2018). These pieces of PII are then sanitized in some way, typically by redacting them or replacing them with realistic surrogate values.

No automatic de-identification system is perfect. In the experiments of this thesis, the automatic de-identification systems find PII by using NER models trained using PII datasets. Regardless of the method used to construct the system, it will sometimes miss some PII or accidentally flag non-PII as sensitive. Thus, it is important to evaluate automatic de-identification systems to understand what degree of privacy preservation they provide. In this thesis, this is done by evaluating the precision and recall values for each PII class. These metrics, as defined in Section 2.4, can be calculated on a per-class basis. This results in a detailed description of how well an automatic de-identification system handles a certain type of PII. Overall, having a high recall means that the privacy benefits are more robust since most of the PII in the data will be sanitized. On the other hand, a model with a low precision will mistakenly flag non-PII as sensitive, reducing the data quality when these entities are erroneously sanitized. In Papers IV and V, we examine the magnitude of this daterioration.

PII can be sanitized in many different ways. The experiments in this thesis use the variations illustrated in Figure 3.3. One approach is to replace detected PII with a realistic surrogate value. This technique is called *pseudonymization*. When the automatic pseudonymizer is powered by a sufficiently powerful NER model and the surrogate replacements are

realistic enough, the pseudonymized text should be indistinguishable from the sensitive version. A beneficial consequence of this is that many missed pieces of PII will be hiding in plain sight (HIPS; (Carrell et al., 2013)). This means that if an adversary spots PII, they cannot be sure whether it is real PII or a surrogate value.

Another sanitization strategy is to replace PII with placeholders. The placeholders may simply show that something was redacted or they may be descriptive tags informing the reader of which class of PII was hidden. We call this latter approach *tagging*. Tagging was used to de-identify the MIMIC-III corpus (Johnson et al., 2016). This approach allows a corpus user to re-populate the corpus with realistic surrogates, as is done during the pseudonymizing process. This was done by Lehman et al. (2021) to create a pseudonymized MIMIC-III corpus. Replacing PII with placeholders has the benefit of being more transparent than pseudonymization since a reader knows what was altered. On the other hand, the privacy-preserving effects of HIPS are lost since an attacker can be certain that any PII that remains in the text is real and was not detected during de-identification.

A third, more aggressive approach to removing PII is to remove the entire sentence in which PII is detected. This approach acts as a sentence *filter*. Doing this has the added benefit of potentially catching surrounding PII that may otherwise have been missed. This is especially useful if the list of PII the NER model has been trained to detect is not exhaustive. This strategy, like the strategy of replacing PII with placeholders, loses the benefits of HIPS. Another downside is that filtering away potentially sensitive sentences removes data entirely, which can be detrimental if data are scarce.

3.4.3 Synthetic Data

Due to the growing generative capabilities of contemporary PLMs, synthesizing training data has been proposed as an alternative to using real, sensitive data. This idea is especially attractive in domains where representative non-sensitive data are challenging to come by or non-existent, as in the clinical domain. Synthetic training data reduce the

privacy risks by producing new training data that are similar too—yet significantly different enough from—the original and sensitive data.

Autoregressive language models are primarily trained to predict text that is as probable as possible. At the same time, synthetic training data need to be *labeled* in order to be useful for most machine learning tasks. Datasets labeled for NLP are—one must assume[4]—only a small portion of the training data used to pre-train most autoregressive language models. When dealing with domain-specific problems, it is also important that the synthetic text is semantically and stylistically similar to the original data. There are three major approaches to dealing with these two problems:

1. Domain adapting an autoregressive language model in order to produce text that looks domain-specific, and then adding machine annotated labels.

2. Fine-tuning an autoregressive model to produce corpora that are both similar to the original data *and* contain labels.

3. Prompting an instruction-tuned language model to produce synthetic data—labeled or unlabeled.

The first approach was selected for the experiments in Paper VI. This was primarily motivated by experiments by Libbi et al. (2021) and Hiebel et al. (2023) that compared the first and second methods. The results of both studies indicated that synthesizing raw text and then adding machine annotations resulted in corpora of higher quality. The third approach has also been demonstrated. Kiefer (2024) instruction-tuned an autoregressive language model to produce clinical notes that aligned with a given diagnosis code. In another clinical example, Liu et al. (2025) synthesized training data for processing radiology reports. This was done successfully using commercial and open-weights models, and using few-shot prompting and zero-shot prompting. All three approaches have been demonstrated as viable, and choosing which strategy to pursue depends on the nature of the corpora and on the resources available.

[4]Many widely-used PLMs lack proper documentation of what data were used to train them.

In many cases, the corpus one wants to create a synthetic counterpart to is sensitive. Unfortunately, a domain-adapted model is at risk of regurgitating its own training data. Many studies (Hiebel et al., 2023; Libbi et al., 2021; Kiefer et al., 2025) assess the severity of these risks by looking at the n-gram overlaps between the real and synthetic data. However, this approach does not give rigorous privacy guarantees and does not take into account the sensitivity of the overlapping data. In any case, properly generated synthetic corpora are not direct copies of their sensitive counterparts. By their very nature, they are *less* sensitive than the original corpora. As part of a battery of privacy-preserving techniques, synthetic training data are an interesting and promising development.

3.5 Measuring Privacy

The privacy-preserving techniques in Section 3.4 have intuitive justifications for how they reduce privacy risks of using sensitive data. However, these privacy gains need to be quantified in order to interpret their benefits. In contrast to more training-oriented techniques—such as differentially private learning—the data-oriented techniques lack a built in mechanism for quantifying their qualities. Instead, we need to use heuristics to estimate how well our techniques work.

3.5.1 Membership Inference for Privacy Quantification

In Subsection 3.3.2, we learned that membership inference attacks try to discern whether or not a data point was part of a model's training data. The attacks typically use a model's reaction to a data point to detect if the model has been previously exposed to the data point. This is only possible if the model's parameters have memorized the data point strongly enough. For this reason, the efficacy of membership inference attacks has been proposed as a method for quantifying the degree of memorization in machine learning models (Murakonda and Shokri, 2020), a claim that is examined in Paper II. This includes quantifying the privacy risks of pre-trained language models.

For the privacy quantification to work, the implementation of the membership inference attack has to be strong. Mireshghallah et al.

(2022) describe a state-of-the-art membership inference attack that targets masked language models like BERT. The model works by comparing the reaction[5] to a data point from a target model and a reference model. The target model is the model that was trained using the dataset that the data point may be a member of. Inversely, the reference model is a model that we know has not been trained using this dataset.

The attack infers that a data point was used to train the target model if the difference in the reaction compared to the reference model is large enough. This is determined by using a threshold that is calibrated using a dataset that was not used to train either the reference model or the target model. The threshold is set to the ratio at which α% of the calibration samples are misclassified as belonging to the target model's training data. α can be varied depending on the tolerance for false positives. This approach is illustrated in Figure 3.4.

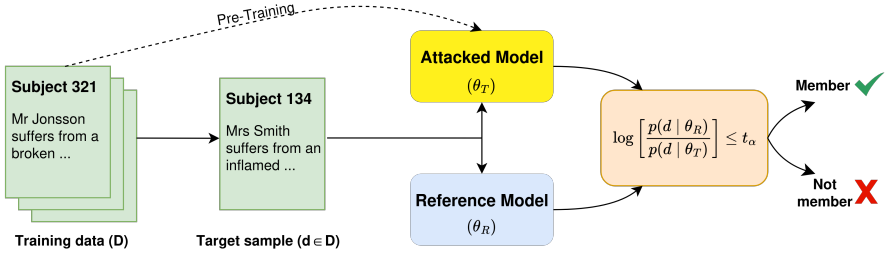

Figure 3.4: The type of membership inference attack used by Mireshghallah et al. (2022) and in Paper II rely on calculating a likelihood ratio. This ratio is based on how an attacked model reacts to a datapoint compared to the reaction of a reference model known *not* to be trained using the same datapoint.

Formally, this attack uses a *likelihood ratio test* (Wilks, 1938) to infer membership and to quantify the privacy of a model. The likelihood ratio test has been used to assess the privacy of genome-wide association studies (Sankararaman et al., 2009). More recently, the test has been adapted to assess privacy in the machine learning domain (Murakonda et al., 2021; Mireshghallah et al., 2022). The test can be defined mathematically as follows: given a likelihood function $p(d \mid \theta)$, two sets

[5] This reaction is calculated by viewing the models as energy-based probability distributions (Goyal et al., 2022). The method for calculating this reaction for masked language models is detailed in Mireshghallah et al. (2022).

of model parameters θ_T and θ_R, and a threshold t_α, we classify a data point d as having been used to train θ_T if

$$\log\left[\frac{p(d \mid \theta_R)}{p(d \mid \theta_T)}\right] \leq t_\alpha. \tag{3.2}$$

The privacy risks of a language model are then based on how susceptible it is to this attack for the full range of the tolerance rate α. This is determined by measuring the AUC[6], which is a threshold-independent metric that captures the relationship between the false positive and true positive rates of a classifier (James et al., 2013).

3.5.2 Counting Overlapping N-Grams

As mentioned in Subsection 3.4.3, synthetic corpora generated using PLMs can still contain sensitive information—despite their synthetic nature. If the generative PLM has been fine-tuned using a sensitive corpus, then there is a risk that memorized data will be reproduced during synthesis. One common way of quantifying the severity of this phenomenon is to count *overlapping n-grams* (Hiebel et al., 2023; Libbi et al., 2021; Kiefer, 2024).

An n-gram is a sequence of n tokens, where n can be varied depending on how long sequences we want to study. In a privacy-minded context, a synthetic corpus and its sensitive counterpart should not share long n-grams. If they do, then this indicates that the synthesizing model has repeated memorized contents that would otherwise be statistically unlikely to appear. Given a set of n-grams S and R, from two different corpora, the frequency of their overlapping n-grams can be calculated using the n-gram recall:

$$\text{n-gram recall} = \frac{|R \cap S|}{|R|}. \tag{3.3}$$

Paper VI also introduces a special form of n-gram recall. This modification—*sensitive n-gram recall*—specifically studies the n-grams

[6]AUC stands for **a**rea **u**nder the **c**urve. The curve in question is the plot of the true positive and true negative rates for all possible thresholds.

associated with PII in the training corpus. This metric relies on the assumption that the researchers know where in the data PII are located. When this is the case, sensitive n-gram overlap can be used to explore whether PII-related n-grams are disproportionally frequent among the memorized n-grams. Naturally, PII-related n-grams are more likely to reveal sensitive information and overlaps should be avoided. Figure 3.5 illustrates how this $R^* \subseteq R$ relates to the sets used in Equation 3.3.

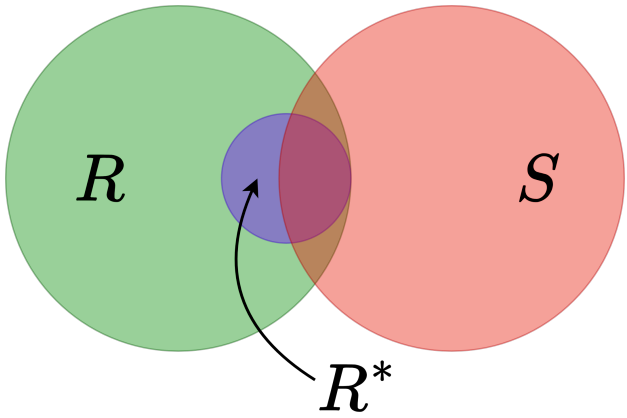

Figure 3.5: TODO: COPIED FROM PAPER VI The n-gram recall is estimated by calculating the proportion of n-grams in the training corpus (green circle) and the synthetic corpus (red circle). $R^* \subseteq R$ (blue) represents the *sensitive* n-grams, and is used to calculate the *sensitive n-gram recall*.

Finding n-gram overlaps is a strong indication that memorized information is leaking into a synthetic corpus. However, this metric only measures *verbatim memorization*. The way n-gram recall is calculated assumes that any shared n-grams are identical. If only one token in the n-gram differs, then this potential paraphrased leak will go undetected. Because of this, n-gram overlap can only provide a lower bound of how much data is at risk of leaking.

Part II

Experiments

This thesis provides insights into the privacy risks of language modeling and describes how these risks can be mitigated using automatic de-identification. These insights are obtained from six papers that address the research questions in Section 1.1. This chapter discusses the methodology underlying the research and introduces the methods and data used in the experiments. Finally, the results of the experiments are described.

Chapter 4

Research Strategy

NLP is a multi-disciplinary field, and many different methodological outlooks fit under the NLP umbrella. This thesis is focused on creating and evaluating machine learning models and this goal determines the methodology used in the papers. This sub-field of NLP is very empirically oriented and relies on *benchmarking* to evaluate how varying the design and training procedure of a model impacts its quality. Quality, in this context, can mean a variety of things. In this thesis, we are mainly concerned with two notions of model quality: privacy preservation and utility. The specifics of how these qualities are measured in the individual experiments are laid out in Chapter 7.

4.1 Benchmarking

Benchmarking, in a machine learning context, is the practice of creating a variety of machine learning models and comparing them using a standardized suite of evaluations. When this is done correctly, it allows researchers to compare models and training procedures in a transparent, reproducible, and controlled manner. The method has a long history in the computing sciences and has been explicitly used since at least 1962 to assess the performance and speed of computing machines (Lewis and Crews, 1985). In contemporary machine learning research, the notion of

performance has shifted from this ideal of resource efficiency towards describing the quality of a model's predictions[1].

Machine learning problems are typically formulated as modeling a real phenomenon in the physical world. Data that should and are assumed to describe the studied phenomenon accurately are collected. Then, models are trained using parts of the collected data and evaluated on another part of the data. The evaluation result is typically a collection of quantitative metrics such as the accuracy, recall, precision, and F_1 score. Crucially, the data used for evaluation are separate from the training data. The evaluation data are often called the held-out test data (James et al., 2013). The test data are essential for ensuring that the evaluation measures the model's ability to make *predictions* based on data it has not been exposed to.

Models are trained to find patterns in data and associate them with specific predictions or outcomes. For example, an NER model learns that some tokens in certain contexts tend to be associated with personal names. The testing phase allows us, as model builders, to test whether these patterns have predictive power. This process is analogous to the hypothetico-deductive model of science (Ladyman, 2002). Based on the data, we train a model that hypothesizes how patterns in the training data relate to their labels or classes. When the model is confronted with new data, predictions are deduced mathematically from the model parameters and evaluated based on the expected predictions.

When a model makes incorrect predictions, this does not lead to the wholesale rejection of the model. Instead, the nature and frequency of these inaccurate predictions inform us about the limits of the model's usefulness. This is because the primary focus of the research is not to use models to describe natural phenomena. Instead, the models themselves are the phenomena being studied. A typical benchmarking study in machine learning will have a research question along the lines of the following[2]:

[1] This shift towards disregarding computational constraints and data availability has been questioned in recent years. Initiatives such as the *SustaiNLP* workshop (Fan et al., 2022) aim to encourage more resource-efficient approaches.

[2] However, the research question is often not explicitly formulated in this way due to the ubiquitous nature of the benchmarking methodology.

*How well does a model design **M** trained using a dataset **D** perform on an evaluation dataset **E** compared to a baseline **B**?*

Depending on the nature of the study, the model design can include concepts such as the model architecture, certain hyper-parameters, or training procedures. The dataset used to train the model may be from the same source as the evaluation set, or it can be a dataset describing the same problem but in another setting (e.g., describing patients from another hospital). Crucially, the performance is measured using a set of well-defined metrics and is *compared* to a strong baseline. This baseline is often the previous state-of-the-art method described in the literature for a particular machine learning problem, or it may be the performance of humans solving the same problem. In any case, the aim of these studies is to quantify how certain configurations of datasets and model designs impact the performance of the resulting models.

4.2 Quantifying Utility

A central theme of this thesis is the conflicting goals of preserving privacy and maintaining utility. The most privacy-minded approach when training models would be to not use sensitive data at all. As discussed in Section 3.4, a wide range of data can contain PII and thus be considered sensitive. Machine-learning approaches to NLP, on the other hand, require large dataset that cannot feasibly be rendered completely private. However, privacy is not binary, and data can be made *safer* without destroying their utility.

Studying the relationship between privacy and utility necessitates a well-defined notion of what constitutes a high-utility dataset. In this thesis, the utility of a dataset is measured by training an NLP model using the data and measuring the model's predictive performance on held-out test data. When assessing how well a privacy-preserving technique preserves utility, we compare the utility of datasets when the technique is and is not applied. Ideally, the difference in utility should be as small as possible.

Furthermore, truly useful techniques are resource efficient. This criterium can be especially important in low-resource domains when data

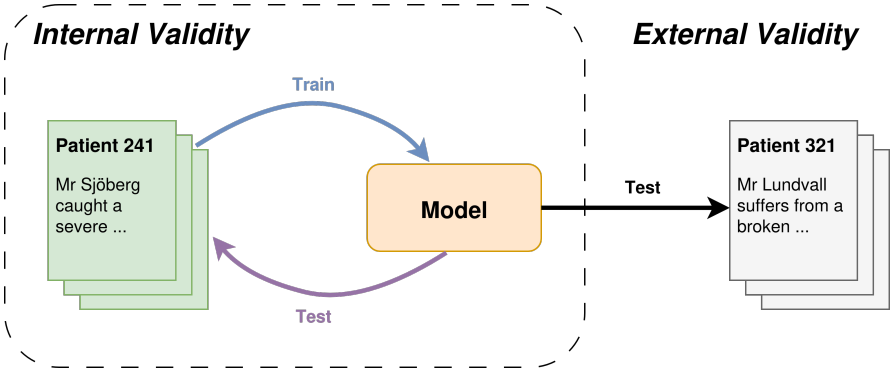

Figure 4.1: Machine learning models should be evaluated with methods that are not only internally valid, but also externally valid. Truly useful models must be able to generalize to new data.

and compute may be difficult to come by. This aspect is considered throughout the thesis, but is more explicitly examined in Papers III and VI. Thankfully, less data-intensive NLP methods are also inherently less prone to privacy risks. Using smaller datasets means that there is less data that can be leaked.

4.3 Validity & Reliability of the Results

In Section 4.1, we learned that the benchmarking paradigm relies on an important assumption—that the data used for evaluation accurately describes the phenomenon we want to model. This assumption is difficult to guarantee or test. Crucially, one must distinguish between *internal* validity and *external* validity. These concepts are illustrated in Figure 4.1. At one end of the spectrum, internal validity is tested during a model's training process by evaluating the model on a *validation set*. Similarly, testing on a *held-out test set* ensures a basic level of external validity. However, truly external validity requires the model to be confronted with new data that is preferably from another source than the training data[3]. Benchmarking as a paradigm often struggles with these purer forms of external validity, as will be further discussed in the next section.

[3]This stronger definition of external validity is sometimes referred to as the *robustness* of a model (Grote et al., 2024).

Another important aspect of the benchmarking methodology is the concept of baselines. The performance of models are compared in relation to each other. The goal of a benchmarking paper i typically to show that one model is better than the other with respect to some quantitative metric. However, most machine learning algorithms are probabilistic and the measured difference in performance is usually small. Consequently, there is always a risk that one model outperforms a competitor due to chance. The results of a single evaluation might not be *reliable*. This situation is not unique to benchmarking, and in this thesis reliability is primarily addressed in two ways—by measuring the variability of the results and by performing tests of statistical significance.

Measuring and testing the reliability of the results requires multiple rounds of testing. In this thesis these repeated tests are performed through k-fold cross-validation, as described in Section 2.4. The result and its uncertainty is then presented as the mean result and standard deviation across all k folds. If two models have a similar mean performance, and the standard deviations are large, then claims of one model being better than the other are unreliable. When feasible, statistical testing can be used to test these claims. In Paper V, we used the Mann-Whitney U test (Mann and Whitney, 1947) to compare the performance of two models on each of the k folds. Consequently, this is also the paper that makes the strongest claims regarding the reliability of its results.

Finally, reliable and valid results rely on experiments being properly controlled. There are many variables to control for in machine-learning research. For example, models should be compared in a fair manner that does not give undue advantage to one model or the other. Furthermore, the data used to evaluate the models should be identical. When factors such as these are not properly controlled for, then we run the risk of contaminating our results with confounding variables. These introduce noise that make it difficult to know whether a result is truly reliable, or if it is an effect of these confounding variables. Eliminating these variables entirely is often not feasible. Nevertheless, a great deal of care has been taken to eliminate these in the studies of this thesis. When possible, for example, the studies rely on datasets that are similar or even identical to those used to create the baseline results. Unfortunately, some confounding variables remain due to the unsurmountable computational

cost of eliminating them. The clearest example is in Paper V, where we rely on pre-existing PLMs that would have been to expensive to train from scratch. Although these models should not differ in how they were trained, the fact that this was done by different people using different codebases means that there may be confounding variables that impact the results.

4.4 LIMITATIONS

One problem is that focusing on the models themselves ignores the context in which they will be used. This means that researchers risk building solutions that will not be useful for the intended audience. Machine learning models can perform very well on their test datasets while being ill-suited to the real world (Foster et al., 2014). However, more often than not, the training data that have been used represent all the data that are available, and collecting new training data is often too expensive. Using non-standard datasets also reduces the replicability of the research. Due to regulations such as the GDPR, there may also be privacy constraints that prohibit researchers from sharing their data.

Another related problem is that the training and testing data are often poorly understood, meaning that the models do not learn what researchers think they are learning. This can lead to problems such as the unwelcome memorization of private information (Carlini et al., 2021; Bender et al., 2021) or biased algorithms that discriminate against minorities (Green and Hu, 2018; Bender et al., 2021). Tackling these issues is challenging, even when researchers are aware of the problems. The experiments that form the basis for this thesis have been designed to mitigate many of the shortcomings of the benchmarking paradigm. For example, the experiments with automatic de-identification in Papers IV and V involve multiple clinical datasets. Using multiple clinical tasks for evaluation means that the results are more representative of the clinical domain in general. However, all experiments rely on either the Health Bank (Dalianis et al., 2015), MEDDOCAN (Marimon et al., 2019), or MIMIC-III (Johnson et al., 2016)[4]. This means, for example, that any

[4]MIMIC stands for **M**edical **I**nformation **M**art for **I**ntensive **C**are and is further described in Section 5.2.1.

models trained in the experiments may not generalize to other datasets. A deeper discussion on this topic is provided in Section 9.2.2

An alternative to benchmarking is to evaluate machine learning models by using them in the real world. This typically involves incorporating them into a more extensive system. For example, instead of assessing an automatic diagnosis coder using a held-out test dataset, we could use it in an end-to-end decision support system. This is a commonly proposed use case for such models, and evaluating a model this way would be a more direct measure of its quality. However, this would be very resource-intensive. A realistic prototype of a decision support system would need to be built, and a non-trivial number of clinicians would need to be involved in testing the system. Furthermore, such a holistic evaluation makes it difficult to determine whether the perceived quality is due to the performance of the model, some other design choice, or the interaction between components.

4.5 Research Ethics

The experiments in this thesis rely on several datasets that are, to various degrees, sensitive. Most importantly, several papers use the Health Bank research infrastructure (Dalianis et al., 2015) which contains a large amount of sensitive data. As such, experiments using these data must be carried out with care. The data are stored in a highly secure server room at the Department of Computer and Systems Sciences. All experiments in this thesis that use non-de-identified Health Bank data were run on a machine running on-premises. Furthermore, all experiments using Health Bank data were conducted with approval from the Swedish Ethical Review Authority under permission number 2019-05679.

The use of these sensitive data is justified as the aim of this research is to understand privacy risks and potential protective strategies. Ultimately, the thesis aims to protect these types of sensitive data. Researching ways in which data can be used in a privacy-preserving manner can unlock new potential uses of clinical data. Many of these use-cases, such as building decision-support systems for physicians, will benefit the persons described in the sensitive datasets, or persons in similar situations.

Chapter 5

Data & Models

All of the experiments in this thesis are related to machine learning algorithms that require large amounts of data. These data include traditional textual datasets and data in the form of pre-trained language models.

5.1 Health Bank

The Health Bank research infrastructure[1] (Dalianis et al., 2015) is an extensive collection of datasets, models, and machines maintained at the Department of Computer and Systems Sciences at Stockholm University. The data were collected from more than 500 clinical units at the Karolinska University Hospital and were recorded from 2006 to 2014 and from 2020 to 2021. The textual data in the EHRs have been used to domain adapt language models (Lamproudis et al., 2021; Lamproudis et al., 2022)—including in Papers IV and VI—and have also been used to create a range of datasets that can be used to train machine learning models. The experiments of this thesis rely on six of these task-specific datasets, listed in Table 5.1

Stockholm EPR PHI Corpus PHI is a form of PII in EHRs. The term is important in the American HIPAA regulation, which lists several types

[1] Details about the Health Bank are available at https://dsv.su.se/healthbank, as of May 2025.

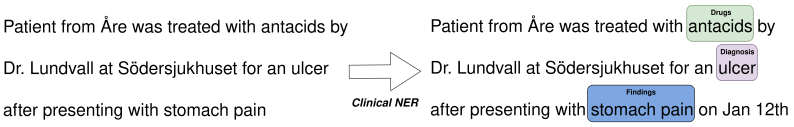

Figure 5.1: In clinical entitity recognition, the task is to locate clinically relevant terms within an EHR. The Stockholm EPR Clinical Entity Corpus is an example of this task.

Corpus	Documents	Tokens	Classes	Level
ICD-10	6,062	930,550	10	Document
ADE	21,725	931,778	2	Document
Factuality	3,710	102,223	6	Document
Factuality NER	4,468	249,302	6	Token
Clinical NER	3,120	157,123	4	Token
PHI NER	21,553	282,766	9	Token

Table 5.1: The five tasks were based on four different clinical corpora from the Health Bank. This table lists the size of each corpus in terms of the number of documents and tokens. The table also specifies the number of possible classes and whether the tasks are document-level or token-level classification tasks.

of PHI that must be sanitized before EHR data can be used for research (Dalianis, 2018). This corpus is annotated with nine PII classes: *First Name*, *Last Name*, *Age*, *Phone Number*, *Location*, *Health Care Unit*, *Organization*, *Full Date*, and *Partial Date*. The corpus is further described in Velupillai et al. (2009) and Dalianis and Velupillai (2010) and is used throughout the thesis to construct the NER systems used for automatic de-identification.

Stockholm EPR Gastro ICD-10 Corpus ICD-10 is a hierarchy of codes for specifying diagnoses that is used for statistical purposes and for hospital reimbursements (Dalianis, 2018). This corpus is a collection of discharge summaries describing 4,985 unique patients that are coded as suffering from gastrointestinal diseases. It was extracted from the Health Bank by Remmer et al. (2021).

Stockholm EPR ADE ICD-10 Corpus ICD-10 contains special codes for conditions caused by adverse drug events (ADEs). The samples in the corpus are divided into two classes based on whether or not the ICD-10

code is related to an ADE. The corpus was first extracted and used in Paper IV.

Stockholm EPR Clinical Entity Corpus The free text of an EHR contains many different forms of information. Certain words are more clinically relevant, and this corpus contains 7,946 entities annotated with the labels *Diagnosis*, *Findings*, *Body parts*, and *Drugs*. The corpus is described in Skeppstedt et al. (2014) and illustrated in Figure 5.1.

Stockholm EPR Diagnosis Factuality Corpus Medical professionals writing EHRs assign diagnoses with varying degrees of confidence. This corpus mentions of diagnoses have been annotated on a six-degree certainty scale from *Certainly Negative* to *Certainly Positive*. The dataset is described in Velupillai et al. (2011) and Velupillai (2011).

5.2 Non-Swedish Datasets

Many of the experiments in this thesis use Swedish clinical data from the Health Bank infrastructure. However, Papers I, II, and VI use additional data created by other institutions and in other languages.

5.2.1 MIMIC-III

MIMIC-III is a large database of EHRs collected between 2001 and 2012 from American critical care units at the Beth Israel Deaconess Medical Center (Johnson et al., 2016). The database contains a wide range of measurements and clinical notes describing visits from 38,597 patients between 2001 and 2012. The free-text data have been de-identified in accordance with HIPAA, and instances of PHI have been replaced with placeholders describing what type of PHI was removed.

Because MIMIC-III has been de-identified, it can be accessed freely after one passes an exam about research ethics and signs a data use agreement. MIMIC-III has been widely used for clinical NLP. In particular, it has been used to pre-train clinical language models (Huang et al., 2020a; Huang et al., 2020b). A pseudonymized version of the EHRs, where the

placeholders have been replaced with surrogates, was created by Lehman et al. (2021) and is used in Papers I and II.

During this thesis work, a new version of MIMIC was released. MIMIC-IV (Johnson et al., 2021) is the latest version at the time of writing and, compared to MIMIC-III, contains more recent data and other improvements. Nevertheless, Papers I and II still rely on the older MIMIC-III. This design choice was made to make the results possible to compare to other closely related studies (Lehman et al., 2021; Mireshghallah et al., 2022), and to make it possible to reuse existing models and datasets.

5.2.2 MEDDOCAN

In contrast to previously described, MEDDOCAN (Marimon et al., 2019) is freely available for download[2] without restrictions. This Spanish corpus contains 1,000 clinical notes consisting of approximately half a million tokens. The dataset has been manually annotated for PII entities. In Paper VI we use a version in which 48,733 of the tokens have been mapped to eight PII classes.

Unlike data from MIMIC-III, and the Health Bank infrastructure, the MEDDOCAN data are synthetic. As described by Marimon et al. (2019), the electronic health records were selected by a physician. Then, health documentalists added realistic PHI information related to each case. A team of NLP and health experts then annotated these data into a fine-grained set of 29 different PII classes.

5.3 Encoder Models

All studies included in this thesis use encoder models in one way or another. Section 2.5 describes pre-training and the BERT architecture. In this subsection, we instead describe the most important pre-trained models used throughout this thesis. A summary is displayed in Table **??**, and details are provided below.

[2] As of September 2025, MEDDOCAN is available under a Creative Commons license at https://github.com/PlanTL-GOB-ES/SPACCC_MEDDOCAN.

$\text{BERT}_{\text{BASE}}$ The BERT architecture was presented by Devlin et al. (2019), and the authors released a series of pre-trained BERT models for others to use. The models were created in smaller and larger sizes, and the small size was dubbed the base model size. The $\text{BERT}_{\text{BASE}}$ model consists of 110 million parameters and was trained using an English corpus consisting of 3.3 billion words.

KB-BERT Although multilingual BERT models (Devlin et al., 2019) exist, many language communities have sought to create their own monolingual models. KB-BERT is a Swedish BERT model used in this thesis that was created by Malmsten et al. (2020) and was trained using multiple Swedish corpora. The corpora span multiple genres and consist of just below 3.5 billion words.

ClinicalBERT-1a BERT models are pre-trained using general-domain data, allowing them to perform well in many scenarios. However, many have found that adapting a general-purpose BERT model to a specific domain can yield a better in-domain performance (Beltagy et al., 2019; Lee et al., 2020). Lehman et al. (2021) train a clinical BERT model using re-identified MIMIC-III data. The model and its training corpus were made available to accredited researchers and can be used to study the privacy implications of training using clinical data.

SweClin-BERT There are no multilingual BERT models for the clinical domain that were trained using Swedish clinical data. To fill this gap, Lamproudis et al. (2021) trained a model using a corpus extracted from the Health Bank (Dalianis et al., 2015). The general-domain Swedish KB-BERT model was used for the initialization of both the weights and the vocabulary, and SweClin-BERT was then adapted to the clinical domain using the Health Bank data.

SweDeClin-BERT Paper IV (Vakili et al., 2022) examines the impact of pre-training clinical BERT models using automatically de-identified data from the Health Bank (Dalianis et al., 2015). The study led to the creation of a new Swedish clinical BERT model that was pre-trained using automatically pseudonymized data. Like SweClin-BERT, it was trained using continued pre-training and initialized with the weights of KB-BERT, with which it shares its vocabulary.

5.4 DECODER MODELS

Paper VI experiments with how generative decoder models can be used to synthesize high-quality text. These synthetic data are then machine annotated and used for NER of PII. Two different decoder models were used in the study:

GPT-SW3 This model was trained by AI Sweden (Ekgren et al., 2024) using the Nordic Pile dataset (Öhman et al., 2023). It is a GPT-style model (Radford et al., 2019) and is available in multiple sizes. In Paper VI, we used the versions that consist of 1.3 billion and 6.3 billion parameters. As the name suggests, the model was trained on a corpus containing a large amount of Swedish data.

FLOR The second model used in Paper VI was created at the Barcelona Supercomputing Centre (Dalt et al., 2024). It was trained primarily using Spanish, Catalan, and English data. Two differently-sized versions of FLOR were used, similarly to GPT-SW3. These two versions consisted of 1.3 billion and 6.7 billion parameters.

Chapter 6

Methods

The experiments in Chapter 7 rely on specific and in some cases recurring methods. This section aims to provide an understanding of the procedures that were used to design the experiments.

6.1 Domain Adaptation

Language models are typically trained on large and diverse sets of corpora. However, these are usually general-domain corpora. Language use can vary widely in different domains, for example, by having domain-specific terminology and abbreviations (Dalianis, 2018). The clinical domain is a prime example of a domain in which a document may be perfectly acceptable to a domain expert, but indecipherable by a layperson. An analogous situation is when general-domain models are applied to domain-specific data. Language models that have been domain adapted generally outperform general-domain counterparts, in predictive performance and in terms of parameter-efficiency (Lehman et al., 2023; Lamproudis et al., 2021).

Domain adaptation is typically achieved through *continued pre-training*. In the simplest setup, used in Papers IV and V, the weights of the model that is to be domain adapted is initialized from those of a pre-existing model. Then, the model is updated using a normal pre-training procedure with in-domain data. There is a wide range of studies showing

that domain adaptation yields models that are more adept at domain-specific tasks (Lamproudis et al., 2021; Beltagy et al., 2019; Huang et al., 2020b). Popular variations include vocabulary modification (Lamproudis et al., 2022) and, as is done in Paper VI, low-rank adaptation.

Low-rank adaptation (LoRA; Hu et al., 2022) is motivated by the increasingly unwieldy amounts of parameters of modern pre-trained language models. Tuning models with billions of parameters is not possible without specialized hardware, primarily due to prohibitive memory requirements but also due to the large number of operations involved. However, the model parameters are represented as a collection of matrices. Consequently, fine-tuning a model consisting of the matrices $\theta = \{\mathbf{W}_1, ..., \mathbf{W}_n\}$ into the model $\theta' = \{\mathbf{W}'_1, ..., \mathbf{W}'_n\}$ can be viewed as searching for a set of matrices $\Delta\theta = \{\Delta\mathbf{W}_1, ..., \Delta\mathbf{W}_n\}$ where each $\Delta\mathbf{W}_i = \mathbf{W}'_i - \mathbf{W}_i$. Since these are matrices, they can be decomposed into *lower-rank* matrices. In other words, $\mathbf{W}^{a\times b} = \mathbf{A}^{a\times r}\mathbf{B}^{r\times b}$.

LoRA allows us to represent $\mathbf{W}$ with $r\,(a + b)$ parameters, which can be much fewer than $a \cdot b$ parameters when r is small. The drawback is that a certain amount of precision is lost when there are fewer parameters to update. However, as Hu et al. (2022) show, LoRA models can be competetive even when $r \ll a$ and $r \ll b$. Even further reductions of memory use can be achieved by *quantizing* the weights. This entails using a more space-efficient data type to represent the weights in fewer bytes. One such approach is *QLoRA* which was introduced by Dettmers et al. (2023) and which is used in Paper VI.

6.2 Automatic De-Identification & NER

Papers IV and V rely on automatic de-identification systems trained using BERT models. These models are fine-tuned to perform the NER task of identifying PII using the Stockholm EPR PHI Corpus (Dalianis and Velupillai, 2010) mentioned in Section 5.1. The fine-tuned models are used to detect PII in the corpora for de-identification. In Paper IV, de-identification is done either by removing PII-containing sentences or by pseudonymizing detected PII. Paper V instead focuses solely on pseudonymization.

When pseudonymizing, the system selects surrogate values using a rule-based approach based on the work of Dalianis (2019). The algorithm that powers the system has a variety of approaches for the different PII classes. These approaches are based on rules and on word lists.

Names First names are replaced with names associated with the same gender, if the gender can be inferred, and otherwise with a gender-neutral alternative. Surnames are replaced with another randomly selected surname.

Locations Different types of locations are replaced with similar surrogates. For example, a city in Sweden will be replaced with another Swedish city.

Dates Specific dates are shifted between one or two weeks earlier or later.

Ages When ages are pseudonymized, they are shifted a few years earlier or later.

Years Similarly to ages, years are moved around a small number of years.

Phone Numbers The algorithm uses a predefined pattern to create random phone numbers that comply with the Swedish format.

6.3 Natural Language Generation

An increasingly important use case for language models is natural language generation. As the name implies, this task involves using language models to produce human-like text. This section introduces the algorithms and techniques used to generate text in Papers I and VI.

6.3.1 Sampling Techniques

Recall from Section 2.3 that the most straightforward way to predict the next token $t_{n+1} \in \mathcal{V}$ based on a sequence $\mathcal{S} = \{t_1, t_2, ..., t_n\}$ is to compute

$$t_{n+1} = \arg\max_{v \in \mathcal{V}} P(v \mid \mathcal{S}) . \qquad \text{(Eq. 2.2′)}$$

This approach is called *greedy decoding* (Jurafsky and Martin, 2025). However, in practice, greedy decoding produces uninteresting and deterministic output that is frequently repetitive. Instead of computing the *argmax*, it is common to instead *sample* from the probability distribution so that $P\left(t_{n+1} = v\right) \sim P(v \mid \mathcal{S})$.

Random sampling introduces variation to counter the repetitive determinism of greedy decoding. Unfortunately, unconstrained sampling often leads to *too much* variation and strange results. Most contemporary pre-trained language models have a vocabulary of thousands of tokens, and sampling from *the entire* vocabulary often results in the selection of tokens that are inappropriate and improbable. This problem stems from a tendency of transformer-based language models to overestimate the likelihood of low-probability tokens. To counter this effect, it is instead more common to only sample from the *top-k* most likely tokens (Fan et al., 2018). In other words, all tokens except for the k most likely candidates have their likelihoods set to zero. The probability distribution is then normalized leading the sampling to only select among the top-k tokens. Mathematically, we are sampling from the vocabulary $\mathcal{V}' \subseteq \mathcal{V}$ that satisfies

$$\arg\max_{\mathcal{V}' \subseteq \mathcal{V}} \sum_{v \in \mathcal{V}'} P\left(v \mid \mathcal{S}\right) \quad \text{where} \quad |\mathcal{V}'| = k. \tag{6.1}$$

Nucleus sampling was proposed by Holtzman et al. (2020) as an alternative to top-k sampling. This sampling technique improves upon the top-k sampling method by replacing k with a dynamic cutoff point. In other words, instead of having a fixed k, nucleus sampling uses a probability density cut-off p. Given a vocabulary $\mathcal{V}$, a token sequence $\mathcal{S}$, and a probability cut-off p, we sample from the smallest subvocabulary $\mathcal{V}' \subseteq \mathcal{V}$ such that

$$\sum_{v \in \mathcal{V}'} P\left(v \mid \mathcal{S}\right) \geq p. \tag{6.2}$$

By introducing this flexibility into the cut-off, nucleus sampling allows the sampling algorithm to adapt to the context surrounding the sampled token. Figure 6.1 illustrates how nucleus sampling allows the sampling cut-off to adjust itself depending on the thickness of the tail of the probability distribution for a given token. The algorithm samples from fewer candidates when the probability distribution is highly skewed and

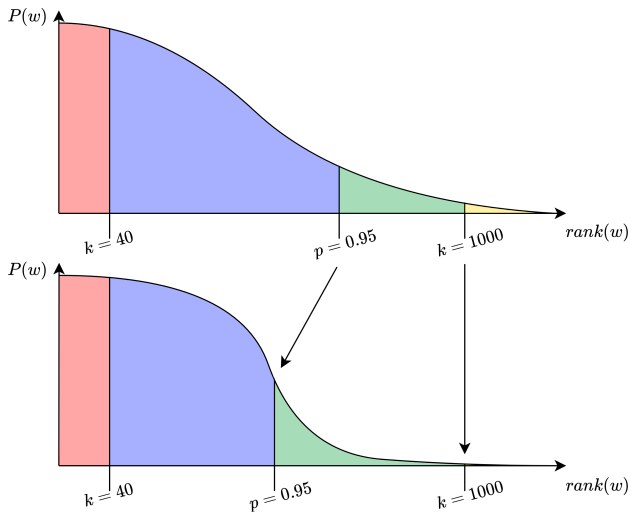

Figure 6.1: Different tokens in a sentence have differently shaped probability distributions. When using top-k sampling, tokens with long or thick tails may be over- or under-sampled. Nucleus sampling is more flexible and dynamically adjusts the cut-off to accommodate probability distributions of different shapes.

samples from a more diverse set of candidates when the distribution has a longer tail.

6.3.2 Generating from Decoder Models

Decoder models readily lend themselves to generative purposes. In fact, their architecture is explicitly designed to generate tokens. Apart from requiring a model trained for generation, we also need an algorithm for selecting the next tokens and sampling procedures that let us select them. In addition to the sampling methods introduced in the previous subsection, generative decoder models often use another sampling scheme called *temperature* sampling. This sampling method introduces a temperature parameter τ. This parameter is used to either accentuate the peaks of the next-token probability distribution, or to flatten them. Consequently, high values for τ lead to more variability in the output whereas lower values decrease the randomness of the sampling process (Jurafsky and Martin, 2025). In Paper VI, we initially experimented with combining temperature and p-sampling. Ultimately, $\tau = 1$ and $p = 0.95$

were selected which in turn means that the temperature did not impact the shape of the probability distributions.

Next-token prediction can in theory be done *ad infinitum*. However, models are typically trained to predict where a sequence should end by learning to emit an end-of-sequence token. This approach does not always work, which means that there is a risk of producing extremely long sequences that would be very unlikely to occur in real life. Typically, algorithms for next-token prediction accept a maximum length after which the model must stop generating regardless of whether an end-of-sequence token has been predicted. This approach is used in Paper VI. Given a corpus of sequences $\mathscr{C} = \{\mathcal{S}_1, ..., \mathcal{S}_n\}$, the maximum length is set to

$$\max(50, l_{\max}) \quad \text{where} \quad l_{\max} = \max_{\mathcal{S}_i \in \mathscr{C}} (|\mathcal{S}_i|) \,. \tag{6.3}$$

Normally, generative models are being decoded to serve some specific use case. It is common—as described in Section 2.5—to provide models with *prompts* in order to stear them towards the correct objective. In Paper VI, the generative models were used purely for synthesis. Instead of prompting them with instructions on how to synthesize, we prompted them with a sequence of tokens from which the model should continue producing something similar to its training data. As we will se in Section 3.4.3, this worked well. Nevertheless, both approaches are valid options, although prompting a model with instructions typically works better with models that are instruction-tuned.

6.3.3 Generating from Masked Language Models

In contrast to decoder models, masked language models like BERT are not traditionally used for text generation. However, Wang and Cho (2019) describe a procedure for generating text from BERT that relies on Gibbs sampling. In their formulation, the sequence is initialized such that it contains only `[MASK]` tokens. At each generation step, a random index in the sample is selected, and the token at that index is replaced. Given the context formed by the surrounding tokens, this new token is sampled based on the probabilities of the candidate tokens. This process continues for a fixed but large number of iterations, and in some cases,

there is a burn-in period (Johansen, 2010) that helps to increase the diversity of the samples. The sample becomes increasingly meaningful as `[MASK]` tokens are replaced with more semantic tokens that constrain the context of the sample.

Lehman et al. (2021) use this sampling method with $k = 40$ to generate synthetic clinical notes from a clinical BERT model. However, later research has shown that $k = 40$ is too low to produce high-quality results. Furthermore, they do not run the sampling algorithm long enough to eliminate the presence of `MASK` tokens. In Paper I, we experimented with setting $k = 1000$ and using different values for p to perform nucleus samapling. The purpose of doing so was to evaluate if there was any link between the quality of a generated text and the risk of it leaking private information.

Chapter 7

Experiments

The data and methods described in previous chapters are employed to design the experiments of the papers comprising this compilation thesis. This section explains the experimental setups of the papers, and presents their results. These results are later discussed in Chapter 8.

7.1 Paper I: Generated Text Quality and Privacy

The experiments in Paper I are inspired by and intended to extend some of the results from Lehman et al. (2021). Specifically, the aim is to further investigate whether or not BERT models are vulnerable to training data extraction attacks. Lehman et al. (2021) generate text from a clinical BERT model trained using MIMIC-III. They search the output for data points that mention a patient and a clinical condition. Then, they determine if the patient's name occurred in the training corpus and if a patient with the name was associated with the condition.

They did not succeed in extracting any meaningful information by doing this, but the experiment was one of many and did not explore different ways of generating text. BERT models are not designed for generative purposes, meaning that the quality of the generated text may be particularly sensitive to the algorithm used to produce it. Paper I focuses on evaluating if more sophisticated algorithms for generating text

produce more sensitive data. The algorithms considered are *top-k sampling* and *nucleus sampling*, which are described alongside the general sampling procedure in Section 6.3.3.

The samples are initialized such that they contain only `[MASK]` tokens 70% of the time, and they are initialized using a name template[1] 30% of the time. Three sampling techniques were employed: top-1000 sampling and nucleus sampling with $p = 0.95$ and $p = 0.99$. We sampled 50,000 samples using each sampling technique. These samples were evaluated in terms of privacy risks and text quality and compared to samples generated by Lehman et al. (2021), who use top-40 sampling. In addition to using more sophisticated sampling techniques, they run the sampling loop for 250 iterations, whereas we run it for 1,000 iterations. They also have a burn-in period of 250 iterations, while we use a burn-in period of 500 iterations.

```
npn neuro : pt is conversant w / language
barrier ( pt understands ) and word finding
difficulty .  pt is on keppra for posturing
during drainage of wound vac .
```

Figure 7.1: Example of the text generated using the enhanced sampling methods explored in Paper I. This example of generated clinical language is relatively fluent, and the information is coherent. For example, the text describes a patient with neurological problems being treated with the epilepsy medication *Keppra*.

An example of the generated text can be found in Figure 7.1. The generated samples were evaluated using a suite of text quality metrics. This selection of metrics was inspired by Holtzman et al. (2020). The diversity of the generated text is measured using Self-BLEU (Zhu et al., 2018) and the shape of the Zipf distribution (Piantadosi, 2014). In addition to these diversity metrics, the quality was also measured by counting the number of `[MASK]` tokens remaining in the output and determining the repetitiveness of the samples. Based on these quantitative quality metrics, the results in Table 7.1 show that the quality of the output does indeed improve. Both nucleus sampling and top-1000 sampling produces text that is closer to MIMIC-III in all quantitative

[1] These templates followed the form "<Title> <First Name> <Last Name> is a yo patient with `[MASK]`" (Lehman et al., 2021). The abbreviation *yo* means *years old*.

	[MASK]	Repetitions	bleu-5	s_{zipf}
MIMIC-III	*N/A*	*0%*	*0.298*	*1.05*
$p = 0.99$	0.00191%	0.12%	0.253	**1.22**
$p = 0.95$	0.00191%	0.12%	**0.306**	1.26
$k = 1000$	0.00575%	0.11%	0.246	1.23

Table 7.1: Text quality metrics for each corpus of text. MIMIC-III is the human gold standard and the values closest to the gold standard are bolded. The percentages describe the proportions of sentences in each corpus containing [MASK] tokens or containing repetitions.

	P(condition | name)	P(wrong condition | name)
$k = 1000$	24.06%	28.28%
$p = 0.99$	24.72%	28.25%
$p = 0.95$	25.51%	29.33%

Table 7.2: If a sentence contains the name of a patient in the corpus, P(condition | name) describes the probability that the sentence also contains one of the patient's conditions. P(wrong condition | name) is the probability that the sentence contains conditions, but none of them are associated with a patient with that name.

metrics except in the number of repetitions. However, there are very few repetitions and these should be more than weighed up for by the near-elimination of lingering [MASK] tokens.

Because the paper aimed to study the potential link between privacy leakage and generation quality, the samples were scrutinized for signs of memorization. This analysis relied on the same strategy as Lehman et al. (2021). A list of conditions associated with patients in MIMIC-III was created using each patient's ICD-9 codes and the MedCAT concepts (Kraljevic et al., 2021) associated with those ICD-9 codes. This list was used to find what conditions, if any, were mentioned in each generated sample. A Spacy NER tagger was used to detect the names of patients in the samples. For every generated sample, we examined whether it contained the name of a patient mentioned in the training corpus and if the sample mentioned a condition associated with such a patient. The prevalence of such patient-condition associations was used to analyze the degree of memorization of sensitive information.

Table 7.2 shows the results from the privacy analysis. The results were nearly indistinguishable from those obtained by Lehman et al. (2021). Similar to them, we were unable to establish a reliable link between a patient's name and the condition they were afflicted by. A review of the samples also revealed that many of the conditions detected in the samples were vague and uninformative. For example, *pain* was a common condition that is likely shared among many of the patients in the corpus. Furthermore, the names that were detected also likely contained many false positives. Many of these false positives were names that are also common words, such as *max*. The improvements in terms of lexical quality and the simultaneous lack of change in terms of privacy suggests that privacy leakage when generating text is not directly linked to the textual quality.

7.2 Paper II: Membership Inference and Pseudonymization

Both privacy-preserving techniques described in Chapter 3.4 have clear benefits regarding privacy. Differentially private models provide formal privacy guarantees defined by the ϵ and δ values used during training. The quality of automatic de-identification systems, on the other hand, can be measured based on the recall for each class. However, there is no consensus on how to compare these two techniques to each other.

Murakonda and Shokri (2020) propose a tool that can measure the privacy risks of machine learning models by subjecting them to membership inference attacks. The degree to which models are susceptible to such attacks acts as a proxy for how prone they are to memorizing their training data. Mireshghallah et al. (2022) suggest that this method can be applied to masked language models like BERT. They successfully launch a membership inference attack on the ClinicalBERT models (Lehman et al., 2021) and imply that this shows that BERT models memorize sensitive data.

Membership inference attacks work on a data-point level. In other words, the attack shows whether a model reacts differently to data points that are present in or absent from its training data. Often, the data points only

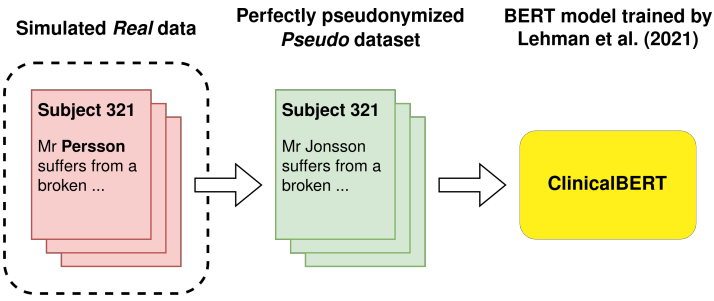

Figure 7.2: The clinical BERT model studied in Papers I and II was trained using a pseudonymized version of MIMIC-III created by Lehman et al. (2021). In Paper II, we simulate the real pre-pseudonymized MIMIC-III corpus by changing all names in the corpus. This creates a new version, letting us simulate the scenario in which the model was trained on real, non-pseudonymized data. This figure is adapted from Paper II.

contain small amounts of sensitive data. For example, learning that *somebody* with an ulcer has visited a hospital on *some date* is not sensitive. We consider the information sensitive only when it is *identifiable*. There are exceptions to this rule, such as highly uncommon diseases. Still, a core assumption of de-identification as a privacy-preserving technique is that a data point can be made harmless by removing sensitive entities. A prominent example of this principle in practice is MIMIC-III, which is easily accessible because it has been sanitized by removing sensitive entities.

The aim of Paper II was to evaluate whether or not membership inference attacks can be used to quantify the privacy risks of models trained using pseudonymized data. We use the same attack as Mireshghallah et al. (2022) but use different datasets to simulate the scenario in which a model has been trained using perfectly pseudonymized data. This simulated scenario is illustrated in Figure 7.2. A model trained using perfectly pseudonymized data has not been exposed to any sensitive PII and should be considered relatively safe, just like MIMIC-III. If the membership inference attack accurately quantifies privacy risks, then the attack should be noticeably less accurate on a pseudonymized model than on a model trained using sensitive data.

The ClinicalBERT model under attack was trained using a pseudonymized version of MIMIC-III (Lehman et al., 2021). We create a new version of this corpus in which every name is replaced. This

Training data	Attacked data	Accuracy	Precision	Recall	AUC
Sensitive	*Random samples*	*0.751*	*0.990*	*0.548*	*0.916*
Sensitive	Samples with names	0.780	0.990	0.566	0.882
Pseudonymized	Samples with names	0.770	0.990	0.548	0.865

Table 7.3: The membership inference attack is run with three different configurations. The first row lists the results obtained by attacking data from Mireshghallah et al. (2022). The results in the other two rows are from attacks that focus on data points containing names—where the training data was either pseudonymized or sensitive. The table is adapted from Paper II.

situation, where we have two corpora that differ only in terms of the names, is equivalent to having one sensitive corpus and one perfectly pseudonymized derivative corpus. The attack is then carried out using subsets of both versions that only include clinical notes containing names. We limited the analysis to these notes since names do not convey any strong medical signals—a system detecting diabetic patients should not care if they are named *Anna* or *Lena*—and because, consequently, they can be replaced using name lists.

As shown in the Table 7.3, the differences between training on pseudonymized or real data is small. The first row lists the results obtained from attacking the data used by Mireshghallah et al. (2022). The second row lists the results obtained when the model is trained using sensitive data, and the final row represents the scenario in which *all* names in the training data have been pseudonymized. These results suggest that membership inference attacks can detect datapoints *similar* to those in the training data. However, the attack in this study is not sensitive to *token-level changes* to the data. Consequently, these types of membership inference attacks are inadequate for assessing the privacy-preserving effects of token-level techniques such as pseudonymization.

7.3 Paper III: Comparing Domain-Adapted Clinical Models to General-Domain Models

The most effective way of ensuring that sensitive training data do not leak is to not use them at all. Consequently, deciding to use sensitive in-domain data to train domain-specific language models should enable clear performance gains in order to be justified. In Paper III, we benchmark two Swedish clinical BERT models—SweClin-BERT and SweDeClin-BERT—against a series of seven general-domain encoder models of varying sizes. The smallest models, including the two domain-adapted models, consisted of 125 million parameters. At the other end of the spectrum, the RemBERT model (Chung et al., 2020) consisted of 576 million parameters. The general-domain models were selected based on their performance on the subset of the EuroEval[2] benchmark (Nielsen, 2023).

In contrast to EuroEval, the benchmark presented in Paper III—SweClinEval—consists solely of Swedish clinical NLP tasks. Three of the tasks were document-level classification tasks, and the other three were token-level classification tasks. These six tasks were based on data from the Health Bank and are described in Section 5.1. The nine encoder models were evaluated using each of the six tasks using 10-fold cross-validation. The fine-tuned models were then compared based on the average F_1 scores across all ten folds. In the token-level classification task, the micro F_1 was used. For the document-level classification tasks, the F_1 scores were a weighted average of the class-wise F_1 scores proportional to the number of test samples belonging to each class. The impact of domain-adaptation could then be explored by comparing the average F_1 scores of the general-domain and clinical-domain language models.

Table 7.4 lists the results of the evaluation. The domain-adapted models are compared to the other best-performing models separately for document-level classification tasks and for token-level NER tasks. The domain-adapted models have the best performance of all encoder models

[2]When Paper III was written, EuroEval was known as *ScandEval*.

Model	Size	ICD-10	Factuality	ADE
		Classification	Classification	Classification
SweDeClin-BERT	125 M	0.832±0.011	**0.735±0.018**	**0.203±0.022**
SweClin-BERT	125 M	**0.836±0.014**	0.731±0.021	0.196±0.014
AI Nordics BERT Large	335 M	0.811±0.012	0.657±0.025	0.192±0.013
KB-BERT Large	370 M	0.801±0.013	0.683±0.019	0.190±0.011
Multilingual E5 Large	560 M	0.824±0.013	0.525±0.074	0.192±0.015
Model	**Size**	**Factuality**	**Clinical Entity**	**PHI**
		NER	**NER**	**NER**
SweDeClin-BERT	125 M	0.623±0.024	0.766±0.034	0.945±0.012
SweClin-BERT	125 M	0.610±0.018	0.754±0.038	0.938±0.014
AI Nordics BERT Large	335 M	0.612±0.026	0.721±0.039	0.948±0.010
AI Sweden RoBERTa Large	355 M	**0.641±0.011**	**0.779±0.036**	**0.965±0.009**

Table 7.4: The average F_1 scores and standard deviations on each of the six tasks is summarized in this table, which was adapted from Paper III. The highest F_1 of each task is bolded, and the second highest is underlined. The size of each model consists of is listed in **M**illions of parameters.

tested on document-level classification. The same is not true for the NER tasks, although the domain-adapted models still perform strongly. Crucially, the other models are significantly larger in terms of the number of parameters. This suggests that the domain-adapted models may be more parameter-efficient.

7.4 Paper IV: Pre-Training Using De-Identified Data

Automatic de-identification systems that rely on NER taggers to find sensitive entities are not completely reliable. As is always the case when dealing with classifiers, there is a trade-off between the recall and precision. Having a high recall is important because this means that our de-identification process will be effective in sanitizing sensitive data. On the other hand, there is a risk that a low precision could cause the data to be corrupted due to the erroneous sanitization of safe tokens. Figure 7.3 shows an example of what could happen when a de-identification system for EHRs has a low precision. Potentially important clinical information is lost and replaced with out-of-place information, thereby decreasing the utility of the data.

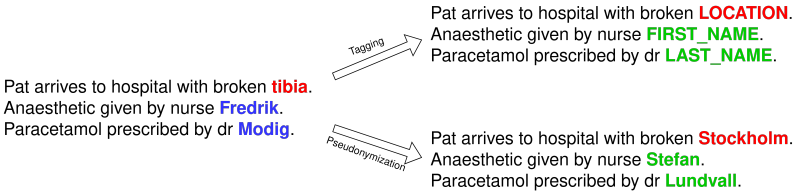

Figure 7.3: NER models are never completely accurate. When they suffer from imperfect precision, they may incorrectly label safe tokens as PII. In this hypothetical example, the model has tagged the bone *tibia* as a piece of PII. Sanitizing non-PII can degrade the utility of the data, especially if the non-PII contains information that is important for downstream tasks.

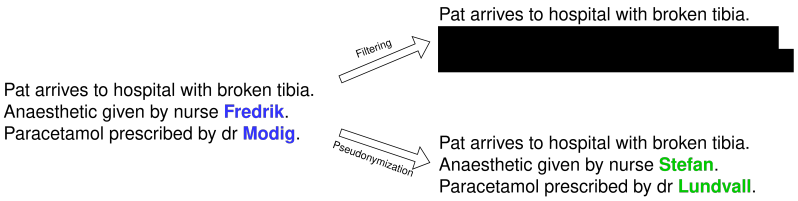

Figure 7.4: In Paper IV, we consider two forms of sanitization strategies. These strategies—filtering and pseudonymization—are illustrated above.

Paper IV examined the impacts of de-identifying data using a BERT-based Swedish de-identification system on the data utility. Two different de-identification strategies are considered: pseudonymization and filtering. As illustrated in Figure 7.4, pseudonymization involves replacing detected entities with realistic surrogates, and filtering removes the entire sentence in which a sensitive entity has been detected. These techniques are applied to 17.9 GB of Health Bank EHRs, producing two new automatically de-identified versions of these data. KB-BERT (Malmsten et al., 2020) is used as a starting point for continuing pre-training using these two corpora, resulting in two new Swedish clinical BERT models.

These de-identified models are compared to two baselines. The first baseline is the general-purpose KB-BERT model that served as the initialization for the new models. The second baseline is the SweClin-BERT model that was based on KB-BERT and adapted to the clinical domain through training[3] on an unaltered version of the Health

[3]In contrast to later experiments, the version of SweClin-BERT used here was trained for no more than three epochs—the same number as the filtered and pseudonymized models—for computational efficiency.

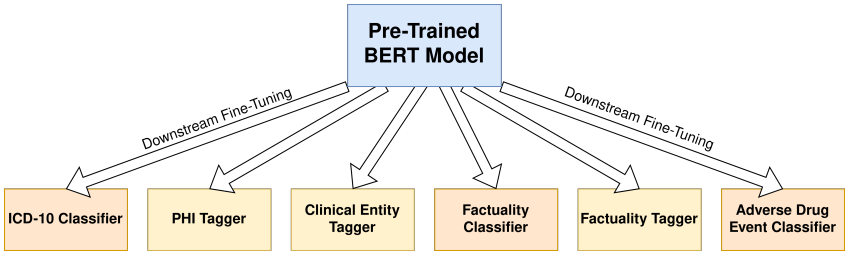

Figure 7.5: The quality of the domain adaptation of the models in Papers IV and V is assessed using six clinical downstream tasks. These tasks include both token-level and document-level classification tasks. These two types of tasks are shaded in yellow and orange, respectively. Models are trained and evaluated for each downstream task and for each pre-trained model.

Clinical data	ICD-10	PHI	Clinical entity	Factuality	Factuality	ADE
	Classification	NER	NER	Classification	NER	Classification
None	0.799	0.91	0.803	0.635	0.630	0.183
Unaltered	**0.833**	**0.941**	0.858	0.732	0.682	**0.199**
Filtered	**0.833**	0.929	0.854	0.731	0.672	**0.199**
Pseudonymized	0.832	**0.941**	**0.861**	**0.736**	**0.684**	0.191

Table 7.5: One general-purpose model and three domain-adapted clinical models were fine-tuned for six clinical tasks. There were three clinical models: one was domain-adapted using *unaltered* data, one was domain-adapted using *filtered* data, where sentences containing sensitive entities were removed, and one was domain-adapted using *pseudonymized* data. All of the values (adapted from Paper IV) are F_1 scores, and the best results are bolded.

Bank corpus (Lamproudis et al., 2021). All four models are fine-tuned for the six clinical tasks illustrated in Figure 7.5 and compared based on their F_1 scores on each task.

The six tasks represent a wide range of clinical tasks that include both document classification tasks and token-level classification tasks. The breadth of the tasks used in the evaluation was intended to assess the external validity of the pre-training results more confidently. Many different tasks are used to measure how well the models adapt to the clinical domain as a whole rather than to any particular downstream task. Suppose that automatic de-identification preserves the utility of the data. In that case, we expect the models trained on the pseudonymized and filtered datasets to perform similarly to the model pre-trained using unaltered data and better than the baseline model not adapted to the clinical domain.

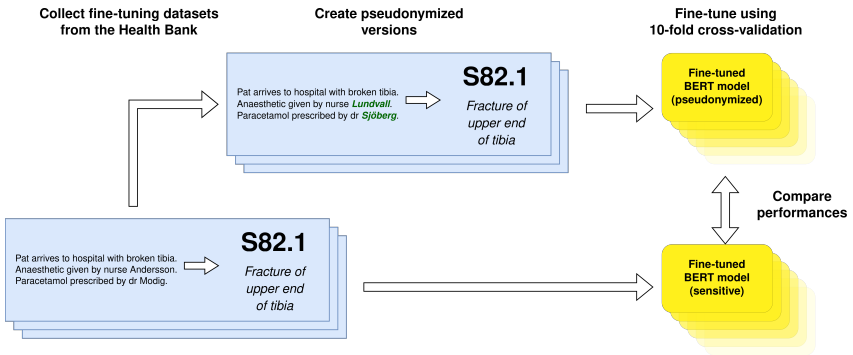

Figure 7.6: In Paper V, pre-trained BERT models were fine-tuned on real data and on pseudonymized versions of these data. The models were compared using the average performance over 10 folds. The figure is taken from Paper V.

The results of the evaluation are shown in Table 7.5. All three clinical BERT models outperformed KB-BERT, which was their general-domain counterpart. Additionally, the two models trained using de-identified data performed similarly to—or better than—the SweClin-BERT model trained using unaltered sensitive data. These results show that automatic de-identification can indeed be employed to reduce privacy risks of pre-trained language models without harming their predictive performance. As a consequence of these results, the model trained using pseudonymized data was dubbed *SweDeClin-BERT*.

7.5 Paper V: End-to-End Training Using Pseudonymized Data

Paper IV studied the impact of automatic de-identification on pre-training data. However, that study did not examine the effects of *fine-tuning* using de-identified data. Pre-training relies on larger corpora than fine-tuning, and it was not clear whether automatic pseudonymization would preserve the utility in the more task-specific and data-constrained fine-tuning scenario. Paper V studies how using de-identified data impacts both pre-training *and* fine-tuning.

Five of the downstream tasks employed in Paper IV were used to gauge the clinical utility of the final models. In contrast to Paper IV, the

downstream task datasets themselves were automatically de-identified, as illustrated in Figure 7.6. Because the fine-tuning datasets were de-identified, the Stockholm EPR PHI corpus was excluded from the analysis as it was used for training the de-identification. Two different pseudonymizers were used. One was on the strongest available version of SweClin-BERT, the other on the SweDeClin-BERT model created in Paper IV. The difference between the pseudonymizers was that the one based on SweClin-BERT had slightly better performance.

Using two different models allowed us to analyze the impact of the quality of the pseudonymization. Unfortunately, computational constraints limited this analysis to the datasets for fine-tuning. Nevertheless, six different configurations were be compared—by alternating between pseudonymized fine-tuning datasets and BERT models. Again, SweClin-BERT served as the non-pseudonymized model and SweDeClin-BERT as the pseudonymized model. In total, this lead to six different combinations of sensitive and pseudonymized data, five different downstream tasks, and evaluations across 10 folds. Consequently, a total of 300 models were trained. All configurations were then compared using Mann-Whitney U tests (Mann and Whitney, 1947; Demšar, 2006) to determine if any differences were statistically significant. The tests were carried out by comparing the F_1scores across all 10 folds of each configuration.

In total, 150 Mann-Whitney U tests were performed when searching for statistically significant differences. Of the differences that were found, 24 results were statistically significant. In most of these cases, the models the better-performing models were models trained using pseudonymized data in the pre-training or fine-tuning steps. Table 7.6 lists the remaining results that challenge the hypothesis of the paper: that predictive performance is preserved when models are trained using pseudonymized data. The first three examples show that SweClin-BERT—itself pre-trained using unaltered data—may perform better when fine-tuned using non-pseudonymized data. In the remaining cases, SweClin-BERT is found to underperform compared to SweDeClin-BERT. Interestingly, none of the models trained end-to-end with pseudonymized data underperformed compared to models trained using unaltered data. Finally, it is important to highlight that—due to the large number of comparisons—some of the differences found may be spurious.

Task	Weaker model		Stronger model		p-value
	P	F	P	F	
ICD-10	✗	✓	✗	✗	0.0378
Factuality NER	✗	✓	✗	✗	0.0014
Clinical NER	✗	+	✗	✗	0.0269
ICD-10	✗	✓	✓	✗	0.0007
Clinical NER	✗	✓	✓	✗	0.0029
Factuality NER	✗	✓	✓	✗	0.0005
ICD-10	✗	+	✓	✗	0.0011
Clinical NER	✗	+	✓	✗	0.0022
Factuality NER	✗	+	✓	✗	0.0156
ICD-10	✗	✗	✓	✗	0.0086
Clinical NER	✗	✗	✓	✗	0.0226

Table 7.6: Out of 24 statistically significant results, 11 are cases where using non-pseudonymized data yields better results than using pseudonymized data. The *p-value*s are based on Mann-Whitney U tests that test if the *Weaker model* is outperformed by the *Stronger model*. For each model, **P** indicates whether the pre-training data was pseudonymized, and **F** indicates that the fine-tuning data was pseudonymized. An ✗ denotes that no pseudonymization was done, a ✓ that it was done using the model based on SweDeClin-BERT, and a + means that pseudonymization was performed using the fine-tuned SweClin-BERT model.

7.6 Paper VI: Synthetic Training Data for Named Entity Recognition

The increasing generative capabilities of modern PLMs have made it possible to generate synthetic texts. These can also be used as training data to train task-specific models. Synthetic training data can be beneficial from a privacy perspective. Ideally, a synthetic version of a sensitive corpus should be dissimilar enough that it does not pose a privacy risk to persons described in the original data. Simultaneously, the synthetic data should be similar enough to its original counterpart to still model the same problem. Synthetic corpora can also be arbitrarily big, since they can be generated indefinitely from the generative model. Several studies have examined whether synthetic corpora can be used to solve machine learning problems. Paper VI systematically examines how

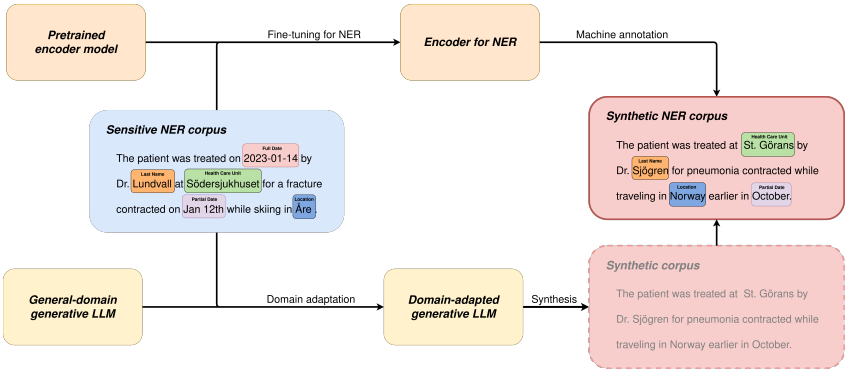

Figure 7.7: In Paper VI, training data for NER are synthesized through a two-step procedure. First, unlabeled data are generated from a domain-adapted PLM. Next, these data are machine annotated using an encoder model fine-tuned for the same NER task that the synthetic corpus targets. This figure is copied from Paper VI.

much data are needed to produce high-quality training corpora. Specifically, the experiments revolve around producing corpora for the NER task of detecting PII.

The process for synthesizing the training corpora was inspired by Libbi et al. (2021), and is illustrated in Figure 7.7. First, a generative PLM is fine-tuned using a real NER corpus. The purpose of doing so is to adapt it to the domain of the original corpus. Then, the domain adapted model is used to generate large amounts of synthetic text. However, these synthetic texts do not have any NER labels. These are added by a machine annotating model. This model, too, is trained using the original NER corpus. After adding these machine annotations, the resulting corpus can be used to train new models without exposing them to the original data.

The process involves several choices and constraints. First, one needs to decide how large the synthesizing model should be. A higher amount of parameters may lead to a more powerful synthesizer, but also requires more computational resources during the domain-adaptive fine-tuning. Similarly, using large amounts of data for creating the synthesizing and machine annotating models may improve the corpus. On the other hand, doing so also exposes more sensitive data to privacy risks and increases the computational demands of the process. Ideally, one should use as small datasets as possible and models with as few parameters as possible. In

Paper VI, the interactions between the impact of varying these constraints were studied systematically.

The experiments were conducted using both Swedish and Spanish datasets and models. The impact of adjusting four different factors was studied:

Model size The generative PLMs, GPT-SW3 (Swedish) and FLOR (Spanish) were available in both smaller and larger versions. Both the versions with ~1.3 billion parameters were used, and the versions consisting of ~6.5 billion parameters.

Domain adaptation The amount of data used to domain-adapt the generative PLMs was varied between 0% of the original data to 95% of the original data.

Machine annotation In one experiment, the proportions of the data used for domain adaptation and for creating the machine annotator was the same. Then, the amount of data used for domain adaptation was fixed while the data used to train the machine annotator varied.

Synthesized amount The final parameter that was studied was to what extent generating larger synthetic datasets resulted in better downstream models. The size of the synthetic corpus was varied between 5%, 100%, and 400% the size of the original corpus.

The entire pipeline for creating the synthetic corpora was evaluated—for all of the variables mentioned above—using five-fold cross-validation. Then, the impact of varying the variables was assessed by training downstream models on the synthetic corpora and evaluating them on each folds held-out test data. The performance was compared based on the average F_1 score obtained by each configuration across the five folds. This part of the evaluation is illustrated in Figure 7.8 Furthermore, the corpora were compared to the original corpora in terms of their lexical features, and in terms of privacy. The privacy of the synthetic corpora were estimated by counting the number of n-grams occuring in both the synthetic and original corpora. Since the original corpora were annotated for PII, we could also count n-grams that were associated with PII in the original data.

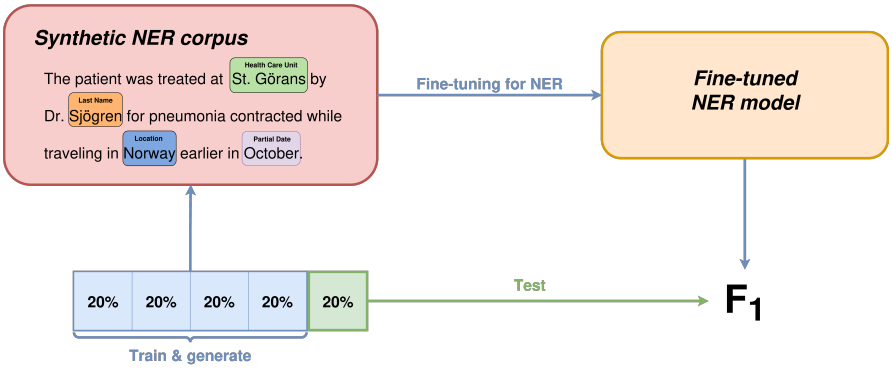

Figure 7.8: The whole pipeline for NER data synthesis was evaluated using five-fold cross-validation. The utility of the synthetic data was measured by training NER models using the synthetic data and testing them on the test fold from the original data.

% for m.a.	SEPR PHI		MEDDOCAN	
	Gold	Synthetic	Gold	Synthetic
5%	0.707 ± 0.037	0.725 ± 0.039	0.931 ± 0.012	0.942 ± 0.010
25%	0.871 ± 0.010	0.858 ± 0.012	0.967 ± 0.003	0.967 ± 0.004
50%	0.908 ± 0.007	0.889 ± 0.005	0.973 ± 0.004	0.965 ± 0.009
95%	0.926 ± 0.005	0.896 ± 0.007	0.978 ± 0.005	0.973 ± 0.003

Table 7.7: COPIED FROM THE PAPER: The amount of data used to create the **m**achine **a**nnotator (m.a.) varied between 5% and 95% of the training data in each fold. This table compares the downstream F_1 scores of the synthetic and gold-standard NER models.

Table 7.7 lists the downstream performance when the synthetic corpora have been labeled by machine annotators trained using varying amount of NER data. For both SEPR PHI and MEDDOCAN, increasing the amount of data available for creating the machine annotating model improves downstream performance. The effect is more pronounced in the case of SEPR PHI, where there is a large jump in downstream performance between using 5% and 25%.

The results in Table 7.8 indicate that domain adaptation is necessary for synthesizing clinical data of sufficient quality, but that there are diminishing returns as more data are used. The first row, where 0% was

% for d.a.	SEPR PHI	MEDDOCAN
0%	0.547 ± 0.178	0.295 ± 0.011
5%	0.873 ± 0.014	0.313 ± 0.032
25%	0.877 ± 0.010	0.970 ± 0.005
50%	0.896 ± 0.007	0.970 ± 0.005
95%	0.896 ± 0.007	0.973 ± 0.003
Gold	*0.926 ± 0.005*	*0.978 ± 0.005*

Table 7.8: COPIED FROM THE PAPER: The amount of data used for **d**omain **a**daptation (d.a.) of the synthesizing generative PLM was varied from 0% to 95% of the training data in each fold. The average F_1 scores of the synthetic NER models and the gold-standard models are listed.

used, performs poorly for both SEPR PHI and MEDDOCAN. However, the effect of increasing the amounts of data tapers off quickly after a certain threshold.

% for d.a.	SEPR PHI		MEDDOCAN	
	All 5-grams	Sensitive 5-grams	All 5-grams	Sensitive 5-grams
5%	0.328 ± 0.041	0.233 ± 0.066	0.005 ± 0.000	0.008 ± 0.001
10%	0.216 ± 0.002	0.154 ± 0.016	0.003 ± 0.000	0.006 ± 0.001
25%	0.183 ± 0.015	0.169 ± 0.021	0.003 ± 0.000	0.004 ± 0.000
50%	0.134 ± 0.021	0.141 ± 0.017	0.002 ± 0.000	0.003 ± 0.000
95%	0.122 ± 0.013	0.132 ± 0.010	0.002 ± 0.000	0.003 ± 0.000
0%	0.028 ± 0.002	0.047 ± 0.002	0.001 ± 0.000	0.001 ± 0.000

Table 7.9: COPIED FROM THE PAPER: 5-gram recall values were calculated for each synthetic corpus over five folds. We calculate both the general 5-gram recall and the recall for 5-grams overlapping with PII in the training corpora. The synthetic corpora varied in the amount of data used for **d**omain **a**daptation (d.a.) before generation. The bottom row shows the values for the synthetic corpora generated without domain adaptation when compared to the 95% gold corpora.

Table 7.9 shows how the 5-gram recall changes as more data are used for domain adaptation. The proportion of n-grams found in the synthesized data decreases as more data are used. This effect is especially noticeable for models using SEPR PHI. Crucially, the proportions listed in the table are in relation to the amount of domain-adaptation data used. For example, using 5% of SEPR PHI lead to 32.8 ± 4.1% of the 5-grams being reproduced in the synthetic data. However, this percentage is in relation

to the 5-grams *in those 5%*. Increasing the amount of data will make each individual 5-gram less likely to be reproduced, but the total number of exposed 5-grams will increase.

Part III

Analysis

The six papers make a number of contributions to the NLP community. This final part of the thesis discusses the results of the experiments described in Chapter 7, not only in terms of new knowledge but also in the form of new resources for Swedish clinical NLP. The final chapter revisits the research questions and presents the conclusions of the thesis.

Chapter 8

Findings & Contributions

The papers comprising this thesis address a variety of research questions. These questions are not limited to the research questions of this thesis, which are described in Section 1.1. This section discusses the results and contributions of each paper, with a special focus on the research questions of this thesis.

8.1 BERT Models are Resilient

The overarching research question of Paper I is whether BERT models are susceptible to training data extraction attacks in the same way as autoregressive models (Carlini et al., 2021). Lehman et al. (2021) performed experiments to study this by generating large amounts of text from BERT models. They were unable to reliably extract private information, but their data were generated using very rudimentary sampling techniques.

Paper I explored whether the failure to extract sensitive information could be mitigated by using more sophisticated sampling techniques. A total of 150,000 samples were generated using three different sampling techniques. Evaluations of the generated samples showed that the linguistic quality of the samples was higher than that of the samples generated by Lehman et al. (2021). This was likely the first paper to show

that nucleus sampling (Holtzman et al., 2020) can be used to generate text from masked language models.

While the quality of the generated samples improved, we could find no increase in the level of privacy leakage. There were several samples that contained the name of a patient alongside a condition. More often than not, the condition was not actually associated with a patient bearing that name. Furthermore, many of the conditions generated described vague symptoms shared by many patients, meaning that they are not reliable signals of truly sensitive information. These results suggest that BERT models are less susceptible to training data extraction attacks than models such as GPT-2. The efficacy of training data extraction attacks does not necessarily increase when the sampling methods are of higher quality. It cannot be ruled out that future research may find more effective ways of extracting training data from BERT models. At the time of writing, however, no examples of such algorithms existed.

Paper II examined a state-of-the-art membership inference attack that targets masked language models. The attack is shown to be highly accurate in determining whether or not datapoints have been used to train BERT models. Yet the results showed that the membership inference attack could not distinguish between pseudonymized and non-pseudonymized versions of the data. If the results of the attack are interpreted as measuring the efficacy of the privacy preservation, they would suggests that replacing all real names in a corpus does not provide any significant privacy benefits. While pseudonymization may not be a panacea for privacy risks, it is certainly the case that a model that has been exposed to real PII is more sensitive than a model that has not. However, the attack does not capture this fact.

The results of Paper II do not prove that BERT models do not memorize PII. However, the results show that there was still no general-purpose way of quantifying this memorization since the state-of-the-art attack designed by Mireshghallah et al. (2022) could not be used to quantify the privacy-preserving effects of pseudonymizing training data. By showing that the attacks are insensitive to differences in PII, the results also suggest that the attack is inadequate when it comes to measuring the memorization of sensitive entities. At the time of writing, there is still no agreed-upon way of comparing different techniques for privacy preservation.

8.2 Domain Adaptation Provides Clear Benefits

The usefulness of domain-adaptation when dealing with clinical NLP was demonstrated in Papers III, IV, and VI. In Papers III and IV, we assessed how domain-adaptive pre-training helps language models outperform general-domain models on clinical NLP tasks. In Paper IV, this is shown since both SweClin-BERT and SweDeClin-BERT perform better than their general-domain counterpart KB-BERT. In Table 7.5, the top row represents the general-domain KB-BERT model which has not been trained using any clinical data. The subsequent rows list the F_1 scores obtained by three different domain-adapted models that differ in how the pre-training data were de-identified. For the purposes of this section, it suffices to note that all three domain-adapted language models perform better than the general-domain KB-BERT model.

These results were confirmed in Paper III, in which two of these models—SweClin-BERT and SweDeClin-BERT—were compared to seven different general-domain models on six different Swedish clinical NLP tasks. The domain-adapted models outperformed the general-domain models on all three document-level classification tasks. They also performed strongly on the NER tasks, especially considering that the general-domain models that outperformed them were nearly three times larger.

While Papers III and IV dealt with domain-adapting encoder models, Paper VI instead domain-adapted generative decoder models. In these experiments, the purpose was not to directly solve a classification in the traditional sense, but to synthesize realistic training data. Similarly to the previous papers, we saw that domain-adaptation resulted in higher-quality synthetic text that resulted in higher-utility synthetic NER corpora. Table 8.1 lists the downstream F_1 scores when the synthesizing decoder models had been domain-adapted with increasing amounts of data. Interestingly, high-quality synthetic text could be produced without using the entire gold corpora. For SEPR PHI, strong results were attained after using just 5% of the data.

% for d.a.	SEPR PHI	MEDDOCAN
0%	0.547 ± 0.178	0.295 ± 0.011
5%	0.873 ± 0.014	0.313 ± 0.032
25%	0.877 ± 0.010	0.970 ± 0.005
50%	0.896 ± 0.007	0.970 ± 0.005
95%	0.896 ± 0.007	0.973 ± 0.003
Gold	*0.926 ± 0.005*	*0.978 ± 0.005*

Table 8.1: The synthesizing decoder models were **d**omain **a**dapted using varying proportions of the two datasets. These models were then used to generate synthetic corpora that—after machine-annotation—were used to train NER models. The table lists the average F_1 scores of the synthetic NER models as well as the gold-standard models.

8.3 Automatic De-Identification Preserves Data Utility

Papers IV and V examined the impact of automatic de-identification on the utility of the data being sanitized. Paper IV examined the impact of de-identifying pre-training data when domain-adapting clinical BERT models. Paper V extended this analysis by studying the impact of one de-identification strategy—automatic pseudonymization—on data used for both domain adaptation and for task-specific fine-tuning.

In Paper IV, three automatically de-identified versions of the Health Bank were used to create Swedish clinical BERT models. These models were then evaluated on six different clinical tasks. As shown in Section 7.4, the domain-adapted models outperformed the general-purpose model on all tasks. Furthermore, there is no clear performance decrease from training on automatically de-identified data compared to using unaltered clinical data.

Extending these results, Paper V extended the analysis to both fine-tuning and domain adaptation. Furthermore, the analysis in Paper V extensively tested different combinations of pseudonymized and unaltered data for domain adaptation and fine-tuning and searched for statistically significant differences. Statistically significant but small deteriorations were found for some models that were *fine-tuned* using pseudonymized

underperformed compared to their non-pseudonymized counterparts. Crucially, none of these differences involved models pseudonymized end-to-end being worse-performing than wholly or partially non-pseudonymized models. On the contrary, most of the statistically significant results pointed towards pseudonymization *helping* the performance of the models.

When combined, the results from Paper IV and Paper V show that pseudonymization can be applied to clinical data without harming the performance of machine learning models trained on the pseudonymized data. This is true both when BERT models are fine-tuned and when they are adapted to the clinical domain through continued pre-training. Considering the modest computational cost of pseudonymizing data, the results suggest that pseudonymization is a resource-efficient way of improving the privacy of models trained using clinical data.

8.4 Factors for Synthesizing High-Quality NER Corpora

Paper VI studies the factors that influence the utility of synthetic NER corpora. Specifically, the study examines data for automatic detection of PII. As described in Section 7.6, the data are synthesized through a procedure involving two steps—text generation and machine annotation. The utility of the data were then evaluated by measuring the downstream F_1 score of BERT models trained using the different synthesized corpora.

To be able to generate synthetic text, general-domain PLMs for Swedish and Spanish were domain-adapted using the target NER dataset. In the study, we examined the impact of choosing larger or smaller PLMs, and of varying the amount of data used to domain-adapt them. The impact of machine annotation was measured by training the machine-annotating model with NER corpora of varying sizes. Finally, the impact of the *amount of data synthesized* was measured by using varying degrees of the synthetic corpora for training the downstream models. The results showed that the quality of the downstream models was strongly tied to the amount of data used for creating the *machine annotator*.

A more surprising result is that the downstream performance is not strongly influenced by the amount of data used to domain-adapt the PLMs. While *some* domain adaptation was needed, high-utility counterparts to SEPR PHI and MEDDOCAN could be synthesized using only 5% and 25% of the data for domain adaptation, respectively. This finding has interesting privacy implications. The privacy analysis revealed that their is a non-trivial amount of 5-gram overlap between the synthetic and original corpora. However, the amount of data that risks being exposed is primarily the data used for domain adaptation. As discussed in Section 8.1, masked language models like BERT are less susceptible to attacks than the autoregressive models used for synthesis. In other words, the privacy risks of creating synthetic data can be reduced by focusing the data use towards creating a strong machine annotator.

8.5 RESOURCES

Compiling this thesis has led not only to the production of new knowledge but also to the creation of several valuable NLP resources. This section describes these resources and how they were created.

8.5.1 DE-IDENTIFIED HEALTH BANK DATA

An essential part of Paper IV was creating a de-identified version of the Health Bank. We estimated that this removed nearly 84 million sensitive entities from the 17.9 GB corpus. Table 8.2 lists these entities, broken down by class. For example, we estimate that 97% of all names have been removed from the dataset. By removing these sensitive entities, the corpus has been made safer.

The software used to de-identify the corpus has been applied to annotated datasets as well. This has led to a number of safer datasets for Swedish clinical NLP being made available to researchers. At the time of writing, datasets de-identified using this method have been used in the following publications:

Papers and other references will be added here later.

PII type	# of predicted instances	NER recall	NER precision
Health Care Unit	19,659,127	80%	87%
Partial Date	19,374,711	83%	94%
Last Name	14,332,309	97%	96%
First Name	12,525,688	97%	98%
Full Date	10,459,935	55%	77%
Location	3,158,031	89%	85%
Age	2,064,111	35%	47%
Organization	1,078,115	36%	71%
Phone Number	1,262,313	40%	63%

Table 8.2: After processing 49,715,558 sentences, 83,914,340 sensitive entities were detected. The recall and precision statistics for each class are predicted using test data from Dalianis and Velupillai (2010) and displayed in this table (adapted from Paper IV).

8.5.2 SweDeClin-BERT

The data de-identified in Paper IV, as explained in the previous section, were used to train several BERT models. One of these models was trained using pseudonymized data, and we applied for ethical permission from the Swedish Ethical Review Authority to share this model for academic use. This Swedish de-identified clinical BERT model—*SweDeClin-BERT*—is now available upon request. At the time of writing, it had been used in four publications:

Vakili et al. (2022) Paper IV, the paper in which the model was created and evaluated on a wide range of downstream tasks.

Jerdhaf et al. (2022) A paper about using SweDeClin-BERT for the terminology extraction of MRI terms from clinical data from Region Östergötland.

Bridal et al. (2022) A paper in which an automatic de-identifier built using Health Bank data was evaluated using a dataset from Region Östergötland.

Dolk et al. (2022) A paper that compares the explainable AI approaches SHAP and LIME for explaining ICD-10 classification.

Additional papers will be added later.

Chapter 9

Discussion

This penultimate chapter contextualizes the contributions of the doctoral thesis. The ethical and societal implications are discussed, as well as the limitations of the studies. These limitations are used as a springboard to suggest some future research directions.

9.1 Ethics and Societal Implications

All six studies in this thesis deal with clinical data. Papers I and II use the English MIMIC-III corpus, and Papers III, IV, V, and VI use Swedish data from the Health Bank. While both data sources contain clinical data, MIMIC-III is much less sensitive since it has been processed to remove PHI. Even so, the data can only be downloaded once one has completed a course on research ethics and signed an agreement describing how the data may be handled.

The Health Bank datasets are much more sensitive than MIMIC-III. The data are stored at the Department of Computer and Systems Sciences at Stockholm University in Kista. As of the time of writing, the data cannot be accessed without physical access to a locked and secure server room. This is excellent from a privacy perspective, as it means that it is very difficult for an unauthorized individual to obtain access to the data. However, there are important downsides as well. First of all, the fact that the data are only accessible to a small number of researchers makes it

difficult to reproduce results obtained using these data. It also limits the utility of the data from a research point of view, since the restricted access requirements prohibit the use of computational infrastructures and limit collaborative research with the data.

When Health Bank data were used in Papers IV, V, and VI, there was a clear ethically beneficial end goal. The overarching objective of the research conducted in this thesis is to provide safer ways to build language technology using sensitive data. Developing and popularizing privacy-preserving techniques for language technology will ultimately benefit the privacy of the individuals whose data are used. Making language technology safer may also enable the wider adoption of language technology in areas in which the data are sensitive. This thesis focuses on the clinical domain, where the benefits include safer care, accelerated research, and less paperwork for clinicians. However, one can imagine similar benefits in other domains, such as the legal domain.

9.2 Limitations and Future Work

A doctoral thesis is no small undertaking, but as with all research endeavors, its scope must still be limited. This section discusses some of these limitations, how they impact the results, and how they can serve as a starting point for future research.

9.2.1 Beyond BERT Models

All experiments in this doctoral thesis use BERT models. Bigger, more advanced models have been released in the years since BERT was first developed. One of the more notable examples is GPT family of models, which have been developed by OpenAI (Brown et al., 2020). These and many other newer models have extreme amounts of parameters[1]. The $\text{BERT}_{\text{BASE}}$ models used in our experiments are comprised of 110 million parameters, while GPT-3—the last OpenAI model with a known

[1]The exact amount of parameters is becoming increasingly difficult to find out. For example, OpenAI no longer disclose the architecture, parameter count or training data sources of their models.

size—consists of 170 *billion* parameters. Training a model with that many parameters is beyond the computational capabilities available at our university department. Indeed, it is beyond the capabilities of most moderately sized research institutions.

Using models of this scale would make the pre-training experiments of Paper IV computationally impossible. The experiments using downstream tasks could technically be performed with generative language models. However, many of the state-of-the-art models are proprietary and only accessible through APIs. Using another organization's API would require us to send sensitive clinical data to their servers. Doing so would be ethically questionable and would violate the ethical permissions and privacy laws under which we conduct our research. Furthermore, it is not clear if the using generative decoder models for the classification tasks in this thesis would be particularly fruitful. As multiple studies have found (Nielsen et al., 2025; Naguib et al., 2024; Aracena et al., 2024), fine-tuned encoder models often perform better and produce outputs that—unlike autoregressive decoder models—do not need to be parsed.

BERT models are still actively used and offer a competitive performance for many tasks. From an engineering perspective, they have an acceptable computational cost for fine-tuning and inference. Pre-training a $\text{BERT}_{\text{BASE}}$ model, while computationally expensive, is still well within the capabilities of many research institutions. This is also one reason that language-specific (Cañete et al., 2020; Malmsten et al., 2020; Farahani et al., 2021) and domain-specific (Beltagy et al., 2019; Lee et al., 2020; Chalkidis et al., 2020) BERT models are abundant.

9.2.2 Cross-Institution Evaluations

When training and evaluating machine learning models, it is common to use test data originating from the same small sets of data sources as the training data. For example, the widely used i2b2 dataset (Stubbs and Uzuner, 2015) consists of patient records from two hospitals (Kumar et al., 2015). This is also the case for the automatic de-identification systems evaluated in Papers IV, V, and VI, which were trained and evaluated using data from five clinical units of the Karolinska University Hospital

(Dalianis and Velupillai, 2010). However, if the training data are not representative of the domain at large, then there is a risk that the system's performance is overestimated.

Cross-clinical evaluations are uncommon due to the legal barriers to sharing sensitive data. These barriers were partially overcome in an article co-written with Bridal et al. (2022). The NER tagger from Stockholm University—*SweDeClin-BERT-NER*—was transferred to a safe environment at Region Östergötland. Then, the tagger was evaluated using the test data and compared with baseline results from a general-domain Swedish NER tagger based on KB-BERT (Malmsten et al., 2020) and SUC[2]. However, because this test set was considered sensitive, only the researchers affiliated with Region Östergötland could review the detailed results. This limited the interpretability of the results beyond the high-level metrics of the recall, precision, and F_1 score. Any discrepancies regarding how the training and test sets differed could only be resolved by communicating without showing the actual data.

In Paper VI, we explore how synthetic data can be generated in a resource-efficient that preserves both privacy and utility. An alternative approach to sharing data directly could be to share synthetic versions of sensitive corpora. These could then be used for training models, but possibly also for testing them. Ideally, evaluations on a synthetic corpus would result in metrics that are similar to those obtained using sensitive corpora. When this is the case, these synthetic evaluation corpora could serve as a compromise that give an indiciation of how well the model would perform on real data. Such a usecase could be especially useful early on when creating and testing systems, since these data would be less sensitive.

Further investigations into the cross-institutional generalizability would have important implications for the overarching research question of this doctoral thesis. Naturally, gaining a deeper understanding of how well our automatic de-identification system generalizes to other datasets would be interesting. This also applies to the other clinical datasets used throughout the thesis. Any shortcomings uncovered through such research would be important insights that could be used to improve our models and the models of others.

[2]SUC is the **S**tockholm **U**meå **C**orpus.

9.2.3 Generalizing to Other Domains

All of the experiments in this thesis rely on clinical data. As discussed in Chapter 2.8, clinical data represent a unique type of data, and processing clinical text comes with certain domain-specific challenges. One consequence of this, for example, is that the specific PII taggers used in Papers IV and V are likely to underperform if they are applied to data from another domain. This is because they have been optimized for processing domain-specific language and also because they rely on a set of PII types that are specific to the clinical domain. However, the overall results and methods described in this thesis should ideally be applicable to other domains as well.

There are multiple reasons to focus on the clinical domain throughout this thesis. One important reason is that much of the research exploring privacy-preserving machine learning uses clinical data since the clinical domain is a common domain in which privacy is important. Orienting the thesis towards the clinical domain thus enables clearer comparisons with results from related studies. Additionally, the focus on clinical data produces results that are immediately relevant to the envisioned users in the healthcare sector. On the other hand, privacy concerns are in no way unique to the clinical domain. As mentioned in Subsection 3.4.2, there is ongoing research into applying privacy-preserving NLP to other domains as well. A prominent example is the effort by Pilán et al. (2022) to construct an open de-identification benchmark based on legal texts. The legal domain is another domain in which many jurisdictions require data to be de-identified before they can be shared. It is also a domain that has caught the attention of the NLP community. Applying the techniques developed in this thesis to the legal domain would strengthen the case for the generalizability of the results.

The results from Paper VI also use clinical data. In that study, we used a clinical dataset for PII recognition and studied which factors were most important for generating high-utility synthetic data. Even though the data were from the clinical domain, the task itself—identifying PII—is not domain-specific. Intuitively, there is not much to be gained in such a task from incorporating clinical knowledge. An interesting future study could examine if the results hold true not just in other domains, but also for NER tasks that require more in-domain knowledge. Results from

Libbi et al. (2021) support the hypothesis that they would extend to these more complex in-domain tasks. They target the more domain-specific clinical entity recognition task, rather than the PII recognition task, and find results in line with our findings. However, they do not perform the same in-depth analysis that we do in Paper VI.

9.2.4 Broader Notions of PII

The automatic de-identification systems used in this doctoral thesis sanitize data by removing PII. These include both direct identifiers and quasi-identifiers. However, the systems are based on NER models trained to find specific entities that belong to certain pre-defined PII classes. As discussed in Subsection 3.4.2, some quasi-identifiers are difficult to identify using this technique. The corpus by Pilán et al. (2022) mentioned in the previous subsection takes a more nuanced approach in determining what kinds of information constitute PII. Their data is also tailored towards the requirements of the GDPR, whereas the data used in the papers of this thesis (Dalianis and Velupillai, 2010) are based on HIPAA rules. Expanding the PII datasets in the Health Bank to include more data and a broader notion of PII will enable the creation of better NER models.

PII-detecting NER models are usually evaluated based on their per-class recall and precision values. This makes it possible to distinguish how well the models classify specific types of PII, and this is the approach taken in this thesis. However, these aggregated metrics do not allow us to distinguish between more specific sub-categories within each PII class. For example, the danger of exposing a particular name varies depending on how common the name is. For example, the name *Thomas Vakili* is currently only held by a single person in Sweden, whereas there are many more *Maria Andersson*. Evaluating de-identification solely on the PII-class level means that disclosing either of these names is considered equivalent. Exploring these nuances can result in more robust de-identification systems and more useful evaluations.

9.2.5 Comparing Privacy-Preserving Techniques

The results from Paper II show that current state-of-the-art membership inference attacks can be fooled by pseudonymizing the training data. If such an attack is used to evaluate privacy-preserving techniques, the results would suggest that using pseudonymized training data does not improve the privacy of the resulting model. While pseudonymization is not a panacea for the privacy risks of training machine learning models, training on datasets that are free of PII is undoubtedly safer than the alternative. This suggests that current membership inference attacks are ill-suited to quantifying the privacy risks of pre-trained language models that were trained using pseudonymized data. Indeed, Duan et al. (2024) find that previous results applying membership inference attacks to pre-trained language models are unreliable due to members and non-members being sampled from different distributions.

An approach-neutral mechanism for quantifying the benefits of different privacy-preserving techniques would be advantageous. Today, the privacy benefits of automatic de-identification are evaluated by measuring the recall of the underlying PII-detecting model. However, this notion of privacy preservation is not directly comparable to, for example, the ϵ and δ values of a differentially private model. This is unfortunate since automatic de-identification can—as demonstrated throughout this thesis—provide privacy benefits without significantly harming the data utility. Alternatives to automatic de-identification, such as differentially private language modeling, have more formal notions of privacy, but they also result in slower training and less accurate models (Anil et al., 2022). Similarly, synthetic training data are a promising complement to other privacy-preserving techniques. Yet there are no satisfactory methods that robustly quantify the risk of the synthetic data containing real PII. While n-gram overlap gives an *indication* of these risks, it can mainly detect leaked information if it is reproduced verbatim.

9.2.6 Susceptibility to Attacks

One fundamental question in privacy-preserving NLP concerns the extent to which NLP systems are vulnerable to different privacy attacks. As explained in Section 3.3, most attacks that target the training data of

language models can be categorized as either membership inference attacks or training data extraction attacks. Language models vary in the degree to which they are vulnerable to these attacks.

Membership inference attacks have been demonstrated to work for a wide range of language models. Mireshghallah et al. (2022) demonstrated that clinical BERT models could be attacked in this way. In Paper II, we perform the same kind of attack to show that these attacks are unreliable when data have been slightly altered on a token level. Previous works have pointed to unwanted memorization as the reason for why these attacks succeed. Our findings bring into question *what kinds* of memorization the attacks rely on. Duan et al. (2024) interrogate whether membership inference attacks work against language models at all. They find that these attacks are strongly biased towards classifying inputs as members when they contain n-grams that are frequent in the training data. This finding is highly compatible with our findings in Paper II, since a datapoint that has been pseudonymized will retain most of its original n-grams.

The tendency to overrely on n-grams also calls into question to what extent the semantic information in datapoints is used in the attacks. For example, are membership inference attacks fooled by paraphrasing a member to be *syntactically* different but *semantically* similar? Such a datapoint would be as sensitive as the non-paraphrased version—since it contains the same information—but may go undetected by current membership inference attacks. This hypothesis is an interesting idea for future research which could hopefully trigger a wider discussion around what information—semantic or syntactic—we are trying to protect when employing privacy-preserving techniques. There does not seem to be much previous research investigating this question, which may be due to the lack of proper datasets of paraphrased sentences.

Training data extraction attacks, on the other hand, have been demonstrated for multiple language models. There is a clear consensus that language models *can* be susceptible to these attacks. However, the question of *which* models are *how* susceptible is still debated. Early results such as those in Paper I and by Lehman et al. (2021) indicated that masked language models like BERT might be less susceptible than autoregressive models. While Carlini et al. (2021) had already demonstrated that autoregressive models could be attacked in this way, there are still—almost five years later—no effective training data

extraction attacks targeting BERT models. Of course, the absence of such attacks does not guarantee that effective attacks could arise in the future. Nevertheless, much of the NLP community is now focusing on autoregressive models which are already known to be susceptible to training data extraction attacks. This is in itself suggests that attacks targeting BERT models are unlikely to arise in the near-future.

Chapter 10

Conclusions

The six papers that compose this compilation thesis seek to deepen our understanding of the privacy risks of pre-trained language models and our understanding of how to mitigate these risks. As described in Section 1.1, this thesis focuses on two themes of this broader research question. Here, the research questions are discussed, and we evaluate to what extent the results in Chapter 8 answer these questions.

10.1 Privacy Risks of Language Models

This thesis examines the privacy risks of clinical language models in Papers I, II, and VI. The experiments in Papers I and VI seek to address the first research questions of this theme:

RQ 1.1 Does the risk of language models leaking information increase when the quality of their generated data improves?

Paper I studies these risks for BERT models fine-tuned using sensitive clinical data. The results in Section 8.1 describe how applying nucleus sampling and top-1000 sampling does increase the lexical quality. The samples generated using these methods do contain data with language similar to what can be found in MIMIC-III. However, it is not clear that these data are *sensitive*. A majority of the co-locations of names and medical conditions are spurious and do not reflect any such association in

the training corpus. The associations that are uncovered are often related to very common conditions and names associated with many different patients, similar to the results of Lehman et al. (2021). Furthermore, there does not seem to be a direct link between the text quality of the generated samples and the risk of privacy leaks. Our results, compared to those of Lehman et al. (2021), do not indicate a larger degree of privacy leakage even if our samples are of better quality.

Paper VI also studies the risks of training data leaking from language models. In this paper, the experiments include domain adapting two generative language models in order to produce synthetic clinical text. The results in Section 8.4 include an analysis of how the 5-gram overlap between the synthetic and original text changes as more data are used for domain adaptation. This analysis shows that domain-adapted generative model are at risk of leaking parts of their training data. However, the extent of these risks depends on the model. Furthermore, the results indicate that increasing the amount of training data reduces the risk of exposure for *each individual 5-gram,* but may increase the risk for the corpus as a whole.

Next, Paper II mounts a state-of-the-art membership inference attack to evaluate how well it can quantify the privacy benefits of de-identifying data. This is done to answer the following research question:

RQ 1.2 Do state-of-the-art membership inference attacks accurately quantify the privacy-preserving benefits gained from automatically de-identifying pre-training data for language models?

The results show that the membership inference attack does *not* capture the privacy benefits obtained from de-identifying training data—not even when the data are *perfectly* de-identified. The attack is insensitive to the specific name mentioned in the data points being tested. This suggests that membership inference attacks are not specifically quantifying the memorization of PII but instead are quantifying the memorization of higher-level characteristics of the data that may or may not be sensitive.

10.2 Benefits of Domain Adaptation

We cannot rule out that using sensitive data for domain adaptation may pose a privacy risk. The safest approach, then, is to eschew the use of sensitive training data entirely. Using these types of data needs to be justified by showing that it has benefits that outweigh these risks. In this thesis, this theme is explored in Papers III, IV, and VI which seek to answer the following research question:

RQ 2.1 Is domain adaptation of pre-trained language models necessary to reach state-of-the-art results in domain-specific NLP tasks?

All three papers find clear benefits from domain adaptation. Papers III and IV show that encoder models domain-adapted using clinical data outperform their general-domain counterparts. Paper III showed that clinical domain adaptation could also compensate for model size—SweClinBERT and SweDeClin-BERT in many cases outperformed much larger models. Paper VI further showed that it is necessary to domain-adapt generative decoder models when producing synthetic training corpora. The highest-quality versions of the synthetic corpora were generated from the models that had underwent domain adaptation. Nevertheless, Paper VI also highlights that satisfactory results can be attained even with smaller amounts of domain adaptation.

10.3 Impact of Privacy Preservation on Utility

Papers IV and V turn our attention to whether automatically de-identifying data lowers the utility of the data for training machine learning models. Paper IVstudies the impact of using automatically de-identified pre-training data to adapt BERT models to the clinical domain. Paper V focuses on using automatically de-identified data for both pre-training and fine-tuning purposes. By doing so, the papers seek to answer the following question:

RQ 2.2 How is the performance of language models affected by using automatically de-identified training data for both pre-training and fine-tuning?

In both cases, the results show that it is indeed possible to both fine-tune and pre-train language models in the clinical domain using automatically de-identified data. Both papers find that the results are no worse than the results obtained from training using unaltered sensitive data. This suggests that, in general, NLP practitioners should consider automatically de-identifying their data before training models. With a sufficiently accurate NER tagger, this procedure is unlikely to harm the utility of the data. Consequently, using unsanitized data should be reserved for cases in which de-identification harms the performance of the trained model.

Whereas Papers IV and V study automatic de-identification, Paper VI instead considers using synthetic training data. Specifically, the study explores the factors that impact the utility of synthetic NER data for PII detection. The aim of the paper is to shed light on the following question:

RQ 2.3 How can sensitive data be used as efficiently as possible when creating synthetic corpora for domain-specific NLP?

The results show that high-quality NER data can be synthesized—even in resource-constrained scenarios. While models trained on the original data performed slightly better, the penalty from using synthetic data was very small. Crucially, the quality of the synthetic data hinged not on the quality of the synthetic text, but on the *machine annotations*. Domain-adapting the PLM that generates the synthetic texts is the most resource-demanding part of the synthetization process. Additionally, it is also the part associated with the highest privacy risks. By focusing the data use towards creating machine annotations, rather than on domain-adapting the PLM, we can reduce both the privacy risks and the computational requirements of the process.

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, New York, NY, USA. Association for Computing Machinery.

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word Order Does Matter and Shuffled Language Models Know It. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland. Association for Computational Linguistics.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2022. Large-Scale Differentially Private BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6481–6491, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Claudio Aracena, Luis Miranda, Thomas Vakili, Fabián Villena, Tamara Quiroga, Fredy Núñez-Torres, Victor Rocco, and Jocelyn Dunstan. 2024. A Privacy-Preserving Corpus for Occupational Health in Spanish: Evaluation for NER and Classification Tasks. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 111–121.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Olle Bridal, Thomas Vakili, and Marina Santini. 2022. Cross-Clinic De-Identification of Swedish Electronic Health Records: Nuances and Caveats. In *LREC 2022 joint workshop language resources and evaluation conference 20-25 june 2022*, page 49.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *Proceedings of the workshop on practical machine learning for developing countries (PML4DC) at ICLR 2020*.

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 *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2013. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets Straight Out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: dataless classification. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, pages 830–835, Chicago, Illinois. AAAI Press.

Nancy Chinchor. 1992. MUC-4 Evaluation Metrics. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Peter Christen, David J. Hand, and Nishadi Kirielle. 2023. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM computing surveys*, 56(3):73:1-73:24.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking Embedding Coupling in Pre-trained Language Models. In *Proceedings of the international conference on learning representations (ICLR 2021)*.

Reece Alexander James Clough, William Anthony Sparkes, Oliver Thomas Clough, Joshua Thomas Sykes, Alexander Thomas Steventon, and Kate King. 2024. Transforming healthcare documentation: harnessing the potential of AI to generate discharge summaries. *BJGP Open*, 8(1).

CMS. 1996. The Health Insurance Portability and Accountability Act of 1996 (HIPAA).

Rachel Cummings and Deven Desai. 2018. The Role of Differential Privacy in GDPR Compliance. In *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*.

George Cybenko. 1989. Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, 2(4):303–314.

Tore Dalenius. 1986. Finding a needle in a haystack or identifying anonymous census records. *Journal of official statistics*, 2(3):329.

Hercules Dalianis. 2018. *Clinical Text Mining*. Springer International Publishing, Cham.

Hercules Dalianis. 2019. Pseudonymisation of Swedish Electronic Patient Records Using a Rule-Based Approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. HEALTH BANK- A Workbench for Data Science Applications in Healthcare. In *CEUR workshop proceedings industry track workshop*, pages 1–18.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields. *Journal of Biomedical Semantics*, 1(1):6.

Severino Da Dalt, Joan Llop, Irene Baucells, Marc Pamies, Yishi Xu, Aitor Gonzalez-Agirre, and Marta Villegas. 2024. FLOR: On the Effectiveness of Language Adaptation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7377–7388, Torino, Italia. ELRA and ICCL.

Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7(1):1–30.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in neural information processing systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

L. Peter Deutsch and Jean-loup Gailly. 1996. ZLIB Compressed Data Format Specification Version 3.3. Technical Report RFC 1950, Internet Engineering Task Force.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Dolk, Hjalmar Davidsen, Hercules Dalianis, and Thomas Vakili. 2022. Evaluation of LIME and SHAP in Explaining Automatic ICD-10 Classifications of Swedish Gastrointestinal Discharge Summaries. In *Scandinavian conference on health informatics*, pages 166–173.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do Membership Inference Attacks Work on Large Language Models? In *Proceedings of the first Conference on Language Modeling*.

Jocelyn Dunstan, Thomas Vakili, Luis Miranda, Fabián Villena, Claudio Aracena, Tamara Quiroga, Paulina Vera, Sebastián Viteri Valenzuela, and Victor Rocco. 2024. A pseudonymized corpus of occupational health narratives for clinical entity recognition in Spanish. *BMC Medical Informatics and Decision Making*, 24(1):204.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006a. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In Serge Vaudenay, editor, *Advances in Cryptology - EUROCRYPT 2006*, Lecture Notes in Computer Science, pages 486–503. Springer, Berlin, Heidelberg.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006b. Calibrating Noise to Sensitivity in Private Data Analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284. Springer, Berlin, Heidelberg.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An Autoregressive Language

Model for the Scandinavian Languages. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

2022. Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP). In Angela Fan, Iryna Gurevych, Yufang Hou, Zornitsa Kozareva, Sasha Luccioni, Nafise Sadat Moosavi, Sujith Ravi, Gyuwan Kim, Roy Schwartz, and Andreas Rücklé, editors, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*:889–898.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. ParsBERT: Transformer-based Model for Persian Language Understanding. In *Neural processing letters*, volume 53, pages 3831–3847.

John R. Firth. 1957. A Synopsis of Linguistic Theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952–59:1–32.

Kenneth R. Foster, Robert Koprowski, and Joseph D. Skufca. 2014. Machine Learning, Medical Diagnosis, and Biomedical Engineering Research - Commentary. *BioMedical Engineering OnLine*, 13:94.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA, USA.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2022. Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis–Hastings. In *Proceedings of the Tenth International Conference on Learning Representations*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, 18(5):602–610.

Ben Green and Lily Hu. 2018. The Myth in the Mythology: Towards a Recontextualization of Fairness in Machine Learning. In *Proceedings of the machine learning: The debates workshop*.

Thomas Grote, Konstantin Genin, and Emily Sullivan. 2024. Reliability in Machine Learning. *Philosophy Compass*, 19(5):e12974.

Katja M. Hakkarainen, Hanna Gyllensten, Anna K. Jönsson, Karolina Andersson Sundell, Max Petzold, and Staffan Hägg. 2014. Prevalence, Nature and Potential Preventability of Adverse Drug Events – A Population-based Medical Record Study of 4970 Adults. *British Journal of Clinical Pharmacology*, 78(1):170–183.

Aron Henriksson. 2015. Representing Clinical Notes for Adverse Drug Event Detection. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 152–158, Lisbon, Portugal. Association for Computational Linguistics.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can Synthetic Text Help Clinical Named Entity Recognition? A Study of Electronic Health Records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *Proceedings of the Eighth International Conference on Learning Representations*.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the tenth international conference on learning representations*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020a. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342 [cs]*.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chih-Ying Deng, Naomi George, and Charolotta Lindvall. 2020b. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 94–100, Online. Association for Computational Linguistics.

Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership Inference Attack Susceptibility of Clinical Language Models. *arXiv:2104.08305 [cs]*.

Gareth James, Fariha Sohil, Muhammad Umair Sohali, Javid Shabbir, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning with applications in R*. Springer Science and Business Media, New York.

Oskar Jerdhaf, Marina Santini, Peter Lundberg, Tomas Bjerner, Yosef Al-Abasse, Arne Jönsson, and Thomas Vakili. 2022. Evaluating Pre-Trained Language Models for Focused Terminology Extraction from Swedish Medical Records. In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*, pages 30–32.

Adam M. Johansen. 2010. Markov Chain Monte Carlo. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education (Third Edition)*, pages 245–252. Elsevier, Oxford.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2021. MIMIC-IV.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger Mark. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3(1):160035.

Dan Jurafsky and James H. Martin. 2025. *Speech and Language Processing*. Pearson Education, 3rd ed.

Lotta Kiefer. 2024. Instruction-Tuning LLaMA for Synthetic Medical Note Generation: Bridging Data Privacy and Utility in Downstream Tasks. Master's thesis, Saarland University.

Lotta Kiefer, Jesujoba O. Alabi, Thomas Vakili, Hercules Dalianis, and Dietrich Klakow. 2025. Instruction-Tuning LLaMA for Synthetic Medical Note Generation in Swedish and English. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J. B. Dobson. 2021. Multi-domain Clinical Natural Language Processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Vishesh Kumar, Amber Stubbs, Stanley Shaw, and Özlem Uzuner. 2015. Creation of a New Longitudinal Corpus of Clinical Narratives. *Journal of Biomedical Informatics*, 58:S6–S10.

James Ladyman. 2002. Glossary. In *Understanding Philosophy of Science*, pages 264–269. Routledge.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2021. Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 790–797, Held Online. INCOMA Ltd.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Vocabulary Modifications for Domain-adaptive Pretraining of Clinical Language Models. :180–188.

Anastasios Lamproudis, Therese Olsen Svenning, Torbjørn Torsvik, Taridzo Chomutare, Andrius Budrionis, Phuong Dinh Ngo, Thomas Vakili, and Hercules Dalianis. 2023. Using a Large Open Clinical Corpus for Improved ICD-10 Diagnosis Coding. In *AMIA annual symposium proceedings*, volume 2023, page 465.

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.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do We Still Need Clinical Language Models? In *Proceedings of the Conference on Health, Inference, and Learning*, pages 578–597. PMLR.

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, Online. Association for Computational Linguistics.

Byron C. Lewis and Albert E. Crews. 1985. The Evolution of Benchmarking as a Computer Performance Evaluation Technique. *MIS Quarterly*, 9(1):7–16.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Claudia Alessandra Libbi, Jan Trienes, Dolf Trieschnigg, and Christin Seifert. 2021. Generating Synthetic Training Data for Supervised De-Identification of Electronic Health Records. *Future Internet*, 13(5):136.

Jinghui Liu, Bevan Koopman, Nathan J. Brown, Kevin Chu, and Anthony Nguyen. 2025. Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports. *Artificial Intelligence in Medicine*, 159:103027.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv:2007.01658 [cs]*.

Henry B. Mann and Donald R. Whitney. 1947. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Jose Antonio Lopez Martin, and Marta Villegas. 2019. Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *IberLEF@ SEPLN 2019*.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37.

Sasi Kumar Murakonda and Reza Shokri. 2020. ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning.

Sasi Kumar Murakonda, Reza Shokri, and George Theodorakopoulos. 2021. Quantifying the Privacy Risks of Learning High-Dimensional Graphical Models. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 2287–2295. PMLR.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Miami, Florida, USA. Association for Computational Linguistics.

Dan Nielsen. 2023. ScandEval: A Benchmark for Scandinavian Natural Language Processing. In Tanel Alumäe and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2025. Encoder vs Decoder: Comparative Analysis of Encoder and Decoder Language Models on Multilingual NLU Tasks. In Richard Johansson and Sara Stymne, editors, *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 561–572, Tallinn, Estonia. University of Tartu Library.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The Nordic Pile: A 1.2TB Nordic Dataset for Language Modeling.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot Learning with Semantic Output Codes. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Steven T. Piantadosi. 2014. Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

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.

Sonja Remmer, Anastasios Lamproudis, and Hercules Dalianis. 2021. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In *Proceedings of RANLP 2021: Recent Advances in Natural Language Processing*, pages 1158–1166, Varna, Bulgaria.

C. J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann, USA, 2nd ed.

Beate Roessler and Judith DeCew. 2023. Privacy. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023.

Sriram Sankararaman, Guillaume Obozinski, Michael I. Jordan, and Eran Halperin. 2009. Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9):965–967.

Yutaka Sasaki. 2007. The truth of the F-measure.

Arne Schwieger, Katrin Angst, Mateo de Bardeci, Achim Burrer, Flurin Cathomas, Stefano Ferrea, Franziska Grätz, Marius Knorr, Golo Kronenberg, Tobias Spiller, David Troi, Erich Seifritz, Samantha Weber,

and Sebastian Olbrich. 2024. Large language models can support generation of standardized discharge summaries – A retrospective study utilizing ChatGPT-4 and electronic health records. *International Journal of Medical Informatics*, 192:105654.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. 2014. Automatic Recognition of Disorders, Findings, Pharmaceuticals and Body Structures from Clinical Text: An Annotation and Machine Learning Study. *Journal of Biomedical Informatics*, 49:148–158.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Amber Stubbs and Özlem Uzuner. 2015. Annotating Longitudinal Clinical Narratives for De-identification: The 2014 i2b2/UTHealth Corpus. *Journal of Biomedical Informatics*, 58:S20–S29.

Latanya Sweeney. 2000. Simple Demographics Often Identify People Uniquely. Technical Report 3, Carnegie Mellon University, Pittsburgh.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Thomas Vakili. 2023. *Attacking and Defending the Privacy of Clinical Language Models*. Licentiate thesis, Department of Computer and Systems Sciences, Stockholm University.

Thomas Vakili and Hercules Dalianis. 2021. Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In *Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021)*, Online.

Thomas Vakili and Hercules Dalianis. 2022. Utility Preservation of Clinical Text After De-Identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 383–388.

Thomas Vakili and Hercules Dalianis. 2023. Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data. In *Proceedings of the 24th Nordic Conference on Computational Linguistics*, Tórshavn, Faroe Islands.

Thomas Vakili, Martin Hansson, and Aron Henriksson. 2025a. SweClinEval: A Benchmark for Swedish Clinical Natural Language Processing. In *Proceedings of The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2024a. End-to-end pseudonymization of fine-tuned clinical BERT models. *BMC Medical Informatics and Decision Making*, 24(1):162.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2025b. Data-constrained synthesis of training data for de-identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27414–27427, Vienna, Austria. Association for Computational Linguistics.

Thomas Vakili, Tyr Hullmann, Aron Henriksson, and Hercules Dalianis. 2024b. When Is a Name Sensitive? Eponyms in Clinical Text and Implications for De-Identification. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 76–80.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sumithra Velupillai. 2011. Automatic Classification of Factuality Levels: A Case Study on Swedish Diagnoses and the Impact of Local Context. In *Fourth International Symposium on Languages in Biology and Medicine (LBM 2011)*.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H. Nilsson. 2009. Developing a Standard for De-Identifying Electronic Patient Records Written in Swedish: Precision, Recall and F-measure in a Manual and Computerized Annotation Trial. *International Journal of Medical Informatics*, 78(12):e19–e26.

Sumithra Velupillai, Hercules Dalianis, and Maria Kvist. 2011. Factuality Levels of Diagnoses in Swedish Clinical Text. *Studies in Health Technology and Informatics*, 169:559–563.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Samuel S. Wilks. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100, New York, NY, USA. Association for Computing Machinery.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, USA. IEEE Computer Society.

Part IV

Papers

PAPER I

ARE CLINICAL BERT MODELS PRIVACY PRESERVING? THE DIFFICULTY OF EXTRACTING PATIENT-CONDITION ASSOCIATIONS

Thomas Vakili and Hercules Dalianis. 2021. Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In *Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021)*, Online.

Author Contributions Thomas Vakili was responsible for designing and carrying out the experiments. The manuscript was written by Thomas Vakili and Hercules Dalianis.

Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations

Thomas Vakili and Hercules Dalianis
Department of Computer and Systems Sciences, Stockholm University
P.O. Box 7003, SE-164 07 Kista, Sweden
{thomas.vakili, hercules}@dsv.su.se

Abstract

Language models may be trained on data that contain personal information, such as clinical data. Such sensitive data must not leak for privacy reasons. This article explores whether BERT models trained on clinical data are susceptible to training data extraction attacks.

Multiple large sets of sentences generated from the model with top-k sampling and nucleus sampling are studied. The sentences are examined to determine the degree to which they contain information associating patients with their conditions. The sentence sets are then compared to determine if there is a correlation between the degree of privacy leaked and the linguistic quality attained by each generation technique.

We find that the relationship between linguistic quality and privacy leakage is weak and that the risk of a successful training data extraction attack on a BERT-based model is small.

1 Introduction

Modern language models have a vast number of parameters, which is the source of their impressive capabilities. However, their size also implies many problems. Among these is the problem of accidentally memorizing sensitive information from their training data (Bender et al. 2021). Avoiding memorization is especially important when training on sensitive data such as electronic patient records, as these contain sensitive information about the identity of patients. Accidental memorization of such information puts patients' identities and other sensitive information at risk of being leaked.

This is not a purely theoretical risk. In fact, Carlini et al. (2020) successfully mounted a training data extraction attack on GPT-2. This attack produced many instances of clearly memorized passages from the training data, containing telephone numbers, addresses, and names of actual living persons.

Based on a methodology from Lehman et al. (2021), we mount a training data extraction attack on the clinical *BERT*[1] model that they release. Their results suggest that generating sensitive data from a BERT model is difficult, especially in comparison to more generative models such as GPT-2 (Carlini et al. 2020). However, their samples were generated using a simple sampling technique, resulting in sentences of low linguistic quality.

Our goal is to strengthen these results by using more advanced sampling techniques which produce higher-quality generations. In this way, we show that the lack of sensitive information in the generated data is not simply a result of the linguistic qualities of the samples. We argue that BERT's poor performance in text generation is, from a privacy perspective, a feature and not a bug.

2 Language Models

The language model used in this study is a BERT model trained by Lehman et al. (2021) using pseudonymized MIMIC-III data. It is based on the BERT architecture (Devlin et al. 2019) and is a *masked language model*, which is trained to correctly predict a masked token using the right and left contexts surrounding it. BERT models are among the latest and best-performing language models, and several such models are being used in the health domain (Lee et al. 2019; Huang, Altosaar, and Ranganath 2020).

Given a masked token x_{mask} in a sentence X, the objective is to learn the probability distribution over a vocabulary V such that:

$$x_{mask} = \underset{w \in V}{\operatorname{argmax}} P(w | X \setminus x_{mask}) \quad (1)$$

This sets masked language models apart from *autoregressive* language models. These models are instead trained to predict the next token x_{i+1} based solely on the *previous* tokens in the sequence:

$$x_{i+1} = \underset{w \in V}{\operatorname{argmax}} P(w | x_1, x_2, ..., x_i) \quad (2)$$

3 Related Research

Modern language models are very large. For example, the *small* version of BERT consists of 110 million parameters (Devlin et al. 2019). This makes BERT and other large model architectures vulnerable to various types of privacy attacks. This section provides an overview of the most common attacks before focusing on the main topic of this article: training data extraction.

[1] Short for **B**idirectional **E**ncoder **R**epresenta**T**ion (Devlin et al. 2019).

3.1 Membership Inference Attacks

Shokri et al. (2017) and Nasr, Shokri, and Houmansadr (2019) describe how membership inference attacks can be used to reveal whether or not a data point was part of a model's training data. They show how this can be carried out both in a *white-box setting* (where the model's parameters are available) and in a *black-box setting* (where the model can only be queried). They show that this attack can successfully be used against a range of different models and datasets. However, none of these seem to focus on unstructured natural language data.

The white-box attack described in Nasr, Shokri, and Houmansadr (2019) shows that a model can be trained to infer membership using the outputs of the last layer or the gradients provided by the loss function. There are a variety of attacks, some requiring access to a subset of the training data, but others *do not require any access to actual training data*.

Lehman et al. (2021) attack a clinical BERT model trained on pseudonymized MIMIC-III data by adding multi-layer perceptron and logistic regression classifiers to probe the BERT model. They tried training the classifiers to discern whether the model had been trained on datapoints containing sensitive data such as name, medical conditions, and combinations thereof. They were unable to recover links between patients and their conditions using this method. On the other hand, experiments focused on names indicate a certain degree of memorization of patient names.

3.2 Unmasking Pseudonymized Training Data

If a language model M has been trained on a dataset D, then there is a risk that the model has memorized certain sensitive details. If this dataset is pseudonymized to create a non-sensitive dataset D', then an adversary with access to M and D' may be able to reconstruct some of the original data from D.

Such an attack was attempted by Nakamura et al. (2020). Sentences were selected from a clinical dataset which contained a patient's first and last names. A BERT model trained on the non-pseudonymized dataset was then used to calculate the probability of predicting the correct first and last names in the sentences. The resulting probabilities were small, and the authors conclude that BERT is not susceptible to this kind of attack.

However, the probability distributions emitted by deep neural networks are known to be inaccurate (Holtzman et al. 2020; Guo et al. 2017). Thus, estimating the risk of re-identifying a person using these probabilities is likely to be inaccurate.

3.3 Training Data Extraction

Attacks need not be limited to simply inferring whether or not a datapoint was part of a model's training data. Carlini et al. (2020) demonstrate that it is possible to extract training data from the language model GPT-2[2] (Radford et al. 2019). They do this by implementing an attack that extracts sentences identical to sentences in the training corpus. A number of these memorized sentences contain specific details that are very unlikely to be generated by chance.

This shows that GPT-2 and other language models can be prone to accidentally memorizing datapoints from their training data, which may lead to privacy leaks. Furthermore, the aforementioned attack can be performed in a black-box setting and does not require direct access to the weights of the model.

However, GPT-2 is an autoregressive language model. These models have an obvious way of generating data: from left to right. Masked language models like BERT, on the other hand, have no such obvious generation strategies. Thus, autoregressive models like GPT-2 have traditionally been preferred over masked language models like BERT when generating text. Due to this difference, it is not obvious if autoregressive models like GPT-2 are disproportionately affected by this vulnerability and to what extent masked language models share this problem.

Lehman et al. (2021) perform a related attack using the BERT model mentioned previously. They generate a large number of sentences and examine the degree to which they contain information linking patients with their conditions. Their results indicate that the degree of privacy leakage is low.

However, the sentences are of poor linguistic quality due to the simple sampling technique used. In the following sections, we will describe more sophisticated ways of sampling from BERT and evaluate how these techniques impact the level of privacy leakage and the quality of the sentences.

4 Generating Text using Masked Language Models

Although autoregressive language models have been favoured for text generation, recent studies have provided strategies for generating coherent text from masked language models as well. Wang and Cho (2019) implement and evaluate a generation strategy based on Gibbs sampling (Geman and Geman 1984), which results in reasonably coherent outputs. Another strategy described by Ghazvininejad et al. (2019) first predicts all masked tokens at once. It then iteratively refines the output by re-masking the least likely predicted tokens. This approach is successfully applied to machine translation.

Besides deciding which tokens to unmask, one must also provide a method for sampling from the predicted unmasked tokens. Wang and Cho (2019) randomly sample from all possible tokens weighted by their predicted probabilities. Holtzman et al. (2020) show that this can result in incoherent text and instead provide a method they call *nucleus sampling*. This sampling method only considers the subset of tokens that constitute the bulk of the probability mass: the nucleus. Recalling equation (1) and given a target probability mass p, we sample from the smallest subset V' of tokens $w \in V$ such that:

$$\sum_{w \in V'} P(w|X \setminus x_{mask}) \geq p \tag{3}$$

[2]GPT-2 is an abbreviation of **G**enerative **P**re-trained **T**ransformer 2.

Nucleus sampling is shown to produce text that, according to a variety of metrics, has similar properties to human-produced text. They show that this strategy produces higher quality results than other popular techniques, such as the *top-k sampling method*.

This method only considers the k most likely predictions when sampling, discarding the other less likely predictions. Nucleus sampling is similar in that it only considers the most likely predictions. However, nucleus sampling does not have a fixed k. The cut-off used to control the diversity of the samples is instead determined *dynamically* using the parameter p.

Lehman et al. (2021) perform a training data extraction attempt by sampling from the same clinical BERT model used in this study. They generate text by sampling from the top-40 candidate tokens when they unmask each token. However, results from Holtzman et al. (2020) show that this is likely to be a too strict value for k and that other sampling configurations may lead to better results.

5 Experiments and Results

This article uses a version of MIMIC-III (Johnson et al. 2016) and a clinical BERT model trained on this corpus[3]. MIMIC-III is a corpus of wide range of patient-related information that has been anonymized. In this article, a subset of MIMIC-III containing clinical notes and diagnoses is used. The anonymous placeholders have been replaced with realistic pseudonyms, and the dataset consists of 1,247,291 clinical notes related to 27,906 patients. This pseudonymized dataset and the model trained on it were made available by Lehman et al. (2021).

5.1 Generating Memorized Information

Techniques modeled on those described by Carlini et al. (2020) were employed to determine whether or not the Clinical BERT model is susceptible to training data extraction attacks. A key difference, however, is how we sample from our non-autoregressive language model.

As described in Section 4, there is no obvious way of sampling from a masked language model. Instead, a variety of strategies are employed to extract text from the Clinical BERT model. Tokens are selected using top-k sampling ($k = 1000$) and nucleus sampling ($p = 0.99$ and $p = 0.95$), as Holtzman et al. (2020) have shown these configurations to be effective when sampling from autoregressive models. The token to unmask is selected randomly, and each generated sequence is 100 tokens long.

50,000 samples are generated using each strategy. First, each sequence is initialized as fully masked or using a prompt[4]. In all cases, we then run a burn-in period (Johansen 2010) of 500 iterations to encourage a diverse set of outputs. Each initialized sequence is then processed for 1,000 iterations using one of the sampling methods.

[3] In Lehman et al. (2021) this model is referred to as *Regular Base*.

[4] This prompt was used in 30% of the batches and was either `[CLS] mr` or `[CLS] ms`, which was the same setup used by Lehman et al. (2021).

We compare our results with the samples generated by Lehman et al. (2021). Their 500,000 sentences were generated from the same model using a burn-in period of 250 iterations, followed by 250 iterations using the top-k sampling method with $k = 40$.

5.2 Sensitive Data in the Generated Samples

Each set of generated samples was processed in the same manner as done by Lehman et al. (2021) to ensure comparability. An NER tagger (Honnibal et al. 2020) was used to locate the few thousand sentences that contained names (first names or last names) associated with a patient in the pseudonymized MIMIC-III corpus. Then, every such sentence was further processed to determine if it mentioned a condition associated with the named patient. The set of conditions associated with the patients was determined by processing the clinical notes using MedCAT (Kraljevic et al. 2021) in conjunction with the ICD-9 codes assigned to each clinical note.

Finding Conditions Some sentences with names contained conditions irrelevant to the patient. Suppose most of the patient-condition associations in the generated corpora are false. In that case, the signal from finding a name and a condition in the same sentence is unreliable in determining from what condition a patient suffers. The prevalence of such false associations was measured by counting them.

Table 1 shows the results of this processing. There is a slight increase in the proportion of sentences containing a name and a matching condition. At the same time, the column *Name + Wrong condition* shows that the percentage of sentences containing a name and a condition *not* associated with a patient bearing the name is slightly larger for all sampling techniques.

It is important to note that the *conditions* found using MedCAT vary in their specificity. Figure 2 plots the percentage of all found conditions constituted by the ten most common conditions. The top ten most common conditions explain a majority of the found conditions. This holds for the texts generated by Lehman et al. (2021) and us and for the pseudonymized MIMIC-III corpus. Many of these are very vague and general. Finding a possible link between a name and the condition *pain*, for example, does not reveal very much information.

Detecting Names Furthermore, Lehman et al. (2021) found that their results likely contained many false positives due to the ambiguous nature of some names. The samples generated in this study show a similar pattern. For example, approximately 10% of the sentences deemed to be associated with a patient and a condition were selected on the basis of containing the name (or word) *Max*.

The set of names detected in the generated sentences constitute a small portion of the total collection of names found in the pseudonymized MIMIC-III corpus. Table 2 shows the percentages of all such names detected in the sentences generated by Lehman et al. (2021) and us.

The vast majority of all names are not detected at all. This is only partly due to the vastly larger size of the MIMIC-III corpus. More likely, this is due to the aforementioned

Figure 1: A few examples from a clinical note that the model seems to have memorized. The name (i.e. "Coleman") and the condition (e.g. "myclonic jerking") are highlighted in yellow and green respectively.

	First name	Last name	Name + Condition	Name + Wrong condition
Lehman et al. (2021)	*0.94%*	*3.14%*	*23.53%*	*28.33%*
$k = 1000$	1.04%	3.61%	24.06%	28.28%
$p = 0.99$	1.28%	3.76%	24.72%	28.25%
$p = 0.95$	1.10%	3.81%	25.51%	29.33%

Table 1: The *First name* and *Last name* columns show the proportion of sentences containing a first or last name. The *Name + Condition* column shows what percentage of these sentences also contain a condition associated with a patient with that (first or last) name. Similarly, the *Name + Wrong condition* shows the percentage where the condition is *not* associated with the patient.

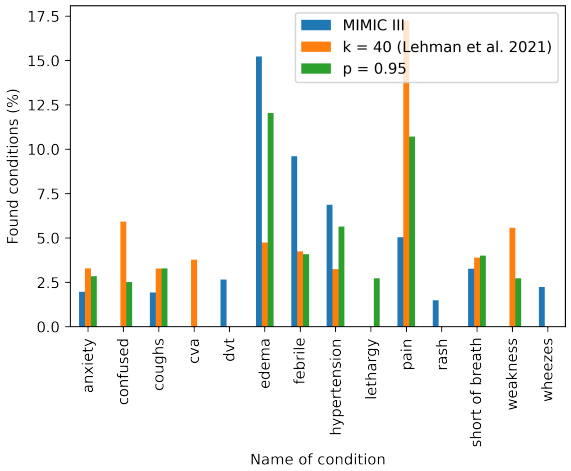

Figure 2: The figure above plots the most common conditions in the texts generated by Lehman et al. (2021), our nucleus text ($p = 0.95$), and MIMIC-III. The top ten conditions detected by MedCAT in each text explain a majority of all conditions. Many of them are vague and general, like *edema* or *pain*.

	Percentage of names detected
Lehman et al. (2021)	10.1%
k = 1000	3.27%
p = 0.99	4.25%
p = 0.95	2.40%

Table 2: Lehman et al. (2021) generate the largest amount of sentences (500,000 sentences), and 10.1% of the names of the pseudonymized MIMIC-III corpus can be detected in their sentences. The largest proportion of names detected in our sentences is the 4.25% found in the 50,000 sentences generated using a nucleus sampling method with $p = 0.95$.

overrepresentation of ambiguous names like *Max*. Many of the names found in the sentences are not part of the MIMIC-III corpus, and have likely been learned in the earlier pre-training of the BERT base model.

In combination with the observation that many names are false positives, this suggests that only a small minority of all names are leaked. However, there are examples of likely memorizations, and Figure 1 illustrates a such a case.

5.3 Metrics for Assessing Linguistic Quality

The *quality* of a given corpus of generated text is not a well-defined property. Gatt and Krahmer (2018) list several subjective and objective metrics that can be used to assess the quality of a generated body of text. This study takes the view that human-likeness is a good proxy for quality in the context of natural language generation.

The human-likeness of the generated samples was assessed by computing a series of metrics and comparing them to a gold standard corpus of human-produced text. The corpus used as the gold standard was the pseudonymized MIMIC-III corpus which the clinical BERT model was trained to model. Using a more general corpus would make less sense in this context. This is because the clinical BERT model is specifically trained to learn the characteristics of clinical notes, which differ significantly from more general forms of writing.

Similarly to Holtzman et al. (2020), we calculated the Self-BLEU (Zhu et al. 2018) and the shape of the Zipf distribution (Piantadosi 2014) - two diversity metrics - as well as the repetitiveness of the texts - which captures the fluency[5]. The quality of the generated samples is determined by comparing the metrics calculated from the generated samples with those of the gold standard.

Self-BLEU is a metric of diversity that measures how similar each sentence in a corpus is to the rest of the corpus. Zhu et al. (2018), who first proposed the metric, calculate it by averaging together the BLEU of every sentence compared to the rest of the corpus.

Due to the size of our generated corpora, we calculate the Self-BLEU slightly differently. As was done by Holtzman et al. (2020), the Self-BLEU is calculated using a random subset $|S'| = 1,000$ of the larger corpus S:

$$\text{Self-BLEU} = \frac{1}{|S'|} \sum_{s \in S'} \frac{\sum_{r \in S \setminus s} \text{BLEU}(s, r)}{|S| - 1} \quad (4)$$

The Zipf distribution is a statistical distribution based on Zipf's law, which states that there is a relationship between a word's rank r in a frequency list of a corpus and its frequency $f(r)$:

$$f(r) \propto \frac{1}{r^{s_{zipf}}} \quad (5)$$

This relationship can be used to estimate s_{zipf}, which can then be used to compare the rank-frequency distributions of different corpora.

[5]The perplexity is left out as there is no consensus on how to calculate it for masked language models and the alternatives are very expensive to calculate (Salazar et al. 2019).

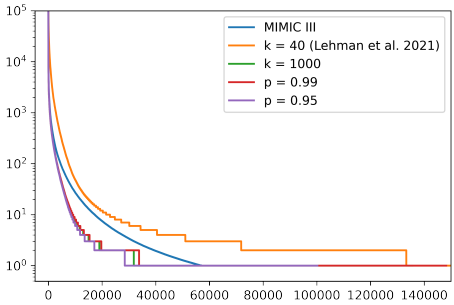

Figure 3: Rank-frequency distribution for the human gold standard (MIMIC-III) as well as the generated samples. The distribution of the samples generated in Lehman et al. (2021) have a tail of unnaturally frequent words which is absent in the gold standard and in our more advanced generations.

5.4 Measuring the Quality of the Generated Samples

Every collection of generated samples was analyzed to determine the quality of the generations. Table 3 and Figure 3 show that the methods used in this study result in generated samples that are closer to the MIMIC-III corpus.

The small number of repetitions that are absent in the datasets used for comparison is the exception. The MIMIC-III data is human-produced, so it is not surprising that it does not contain any repetitions. The other discrepancies are likely due to the larger number of iterations used in this study as compared to the 500 iterations used in Lehman et al. (2021), which leaves some masked tokens in the generated samples.

6 Discussion

This study has given us insights into the complicated area of protecting privacy in training data represented in language models. One suggestion in the research community is to use *homomorphic encryption* (Parmar et al. 2014; Al Badawi et al. 2020) for the data and models. However, it seems that using homomorphically encrypted models is currently too complicated for users.

A more straightforward way to protect the privacy of persons in the training data is to pseudonymize it before training. Both Berg, Chomutare, and Dalianis (2019) and Berg, Henriksson, and Dalianis (2020) build NER taggers on clinical data that has been pseudonymized. They find that, while this decreases the performance of the NER taggers, it does so to an acceptable degree. These taggers can be used to build automatic de-identification systems that can make training datasets less sensitive, as shown by Dalianis and Berg (2021). However, no such system can achieve perfect recall. Thus, this approach is analogous to a weak form of differential privacy where noise in the form of pseudonyms is added to the training data.

	[MASK]	Repetitions	bleu-4	bleu-5	s_{zipf}
MIMIC-III	*N/A*	*0%*	*0.399*	*0.298*	*1.05*
Lehman et al. (2021)	5.54%	**0%**	0.251	0.116	1.39
$p = 0.99$	1.91e-3%	0.12%	**0.433**	0.253	**1.22**
$p = 0.95$	1.91e-3%	0.12%	0.485	**0.306**	1.26
$k = 1000$	5.75e-3%	0.11%	0.435	0.246	1.23

Table 3: Text quality metrics for each corpus of text. MIMIC-III is the human gold standard and the values closest to the gold standard are bolded. The percentages describe the proportions of sentences in each corpus containing [MASK] tokens or containing repetitions.

The clinical BERT model used in this article is trained on clinical data, but uses a BERT model pre-trained on non-sensitive data as its basis. This is good from a privacy perspective, as it means that names that are emitted when sampling from the model are of uncertain origin. Detecting a name in the output is thus a weaker signal, as the name might simply be memorized from the first phase of training on non-sensitive data. However, Gu et al. (2021) show that pre-training with only medical data can yield stronger results, suggesting that this approach may become more prevalent in the future.

Further research into extracting training data for BERT models trained solely on sensitive data would shed light on the potential risks of this approach. The model in this article is also uncased, meaning that it is only trained on lowercase tokens. This means that it has a harder time distinguishing entities that are normally capitalized, like names, from other words. Investigating the impact of not lowercasing the data would be interesting since this is a design choice that may not be suitable for languages where the casing is important.

More robust metrics for measuring privacy leakage from training data extraction attacks would also be of use. The metrics used in this article and by Lehman et al. (2021) strongly suggest that detecting a link between a patient's name and a condition is very difficult. A very small number of samples contain any such possible associations, and many of these are likely to be false positives. This is both due to the ambiguity of many of the detected names and being slightly more likely to find a condition not associated with the named patient.

It is also unclear what risks are acceptable from a legal perspective. Regulations such as the GDPR have strict requirements to avoid risk for identification. At the same time, the GDPR also contains language stating that "the costs of and the amount of time required for identification" (European Commission 2018) should be taken into consideration when making risk assessments. Clarifications from legal scholars are necessary for these and other results in the privacy domain to be contextualized and applicable to real applications.

7 Conclusions

The sampling methods used in this article show a significant improvement regarding the linguistic quality of the samples, as shown in Table 3. At the same time, Table 1 shows that the prevalence of patients and their conditions within the generated samples is stable. This suggests that privacy leakage is not strongly correlated with the quality of the sampling techniques.

Nucleus sampling, first described as a technique for sampling from the autoregressive model GPT-2 (Holtzman et al. 2020), is also shown to be an effective technique for sampling from the masked language model BERT. Further research into how to sample quality text from masked language models is an interesting topic, but our research indicates that advances in that direction do not have significant privacy implications.

It cannot be ruled out that other sampling techniques, regardless of their linguistic quality, may be able to extract training data more effectively. Carlini et al. (2020) showed that the risk of an adversary successfully extracting training data from GPT-2 is significant. Our results, together with those of Lehman et al. (2021), strongly suggest that the risk of successfully sampling sensitive data from a BERT-based model is much smaller when compared to GPT-2.

Acknowledgments

A special thanks to Sarthak Jain and Eric Lehman for their patient assistance with reproducing their experiments from Lehman et al. (2021) and for making their data available to us. We are also grateful to the DataLEASH project for funding this research work.

References

Al Badawi, A.; Hoang, L.; Mun, C. F.; Laine, K.; and Aung, K. M. M. 2020. Privft: Private and fast text classification with homomorphic encryption. *IEEE Access* 8: 226544–226556.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Berg, H.; Chomutare, T.; and Dalianis, H. 2019. Building a De-identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 118–125.

Berg, H.; Henriksson, A.; and Dalianis, H. 2020. The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, 1–11.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2020. Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805* .

Dalianis, H.; and Berg, H. 2021. HB Deid-HB De-identification tool demonstrator. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 467–471.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)(2019)*.

European Commission. 2018. Recital 26 - Not applicable to anonymous data. URL https://gdpr.eu/recital-26-not-applicable-to-anonymous-data/.

Gatt, A.; and Krahmer, E. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61: 65–170. ISSN 1076-9757. doi:10.1613/jair.5477. URL https://www.jair.org/index.php/jair/article/view/11173.

Geman, S.; and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6): 721–741.

Ghazvininejad, M.; Levy, O.; Liu, Y.; and Zettlemoyer, L. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324* .

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv:2007.15779 [cs]* URL http://arxiv.org/abs/2007.15779. ArXiv: 2007.15779.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. *arXiv:1706.04599 [cs]* URL http://arxiv.org/abs/1706.04599. ArXiv: 1706.04599.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=rygGQyrFvH.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python. doi:10.5281/zenodo.1212303. URL https://doi.org/10.5281/zenodo.1212303.

Huang, K.; Altosaar, J.; and Ranganath, R. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342 [cs]* URL http://arxiv.org/abs/1904.05342. ArXiv: 1904.05342.

Johansen, A. 2010. Markov Chain Monte Carlo. In Peterson, P.; Baker, E.; and McGaw, B., eds., *International Encyclopedia of Education (Third Edition)*, 245–252. Oxford: Elsevier, third edition edition. ISBN 978-0-08-044894-7. doi:https://doi.org/10.1016/B978-0-08-044894-7.01347-6. URL https://www.sciencedirect.com/science/article/pii/B9780080448947013476.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3(1): 160035. ISSN 2052-4463. doi:10.1038/sdata.2016.35. URL https://www.nature.com/articles/sdata201635. Number: 1 Publisher: Nature Publishing Group.

Kraljevic, Z.; Searle, T.; Shek, A.; Roguski, L.; Noor, K.; Bean, D.; Mascio, A.; Zhu, L.; Folarin, A. A.; Roberts, A.; Bendayan, R.; Richardson, M. P.; Stewart, R.; Shah, A. D.; Wong, W. K.; Ibrahim, Z.; Teo, J. T.; and Dobson, R. J. B. 2021. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine* 117: 102083. ISSN 0933-3657. doi:10.1016/j.artmed.2021.102083. URL https://www.sciencedirect.com/science/article/pii/S0933365721000762.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* btz682. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btz682. URL http://arxiv.org/abs/1901.08746. ArXiv: 1901.08746.

Lehman, E.; Jain, S.; Pichotta, K.; Goldberg, Y.; and Wallace, B. C. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? In *Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*.

Nakamura, Y.; Hanaoka, S.; Nomura, Y.; Hayashi, N.; Abe, O.; Yada, S.; Wakamiya, S.; and Aramaki, E. 2020. KART: Privacy Leakage Framework of Language Models Pre-trained with Clinical Records. *arXiv:2101.00036 [cs]* URL http://arxiv.org/abs/2101.00036. ArXiv: 2101.00036.

Nasr, M.; Shokri, R.; and Houmansadr, A. 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)*, 739–753. IEEE.

Parmar, P. V.; Padhar, S. B.; Patel, S. N.; Bhatt, N. I.; and Jhaveri, R. H. 2014. Survey of various homomorphic encryption algorithms and schemes. *International Journal of Computer Applications* 91(8).

Piantadosi, S. T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21(5): 1112–1130. ISSN 1531-5320. doi:10.3758/s13423-014-0585-6. URL https://doi.org/10.3758/s13423-014-0585-6.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8): 9.

Salazar, J.; Liang, D.; Nguyen, T. Q.; and Kirchhoff, K. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659* .

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. IEEE.

Wang, A.; and Cho, K. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094* .

Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; and Yu, Y. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, 1097–1100. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5657-2. doi:10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

PAPER II

USING MEMBERSHIP INFERENCE ATTACKS TO EVALUATE PRIVACY-PRESERVING LANGUAGE MODELING FAILS FOR PSEUDONYMIZING DATA

Thomas Vakili and Hercules Dalianis. 2023. Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data. In *Proceedings of the 24th Nordic Conference on Computational Linguistics*, Tórshavn, Faroe Islands.

Author Contributions Thomas Vakili was responsible for designing and carrying out the experiments. The manuscript was written by Thomas Vakili, with comments from Hercules Dalianis.

Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data

Thomas Vakili and Hercules Dalianis
Department of Computer and Systems Sciences (DSV)
Stockholm University, Kista, Sweden
{thomas.vakili,hercules}@dsv.su.se

Abstract

Large pre-trained language models dominate the current state-of-the-art for many natural language processing applications, including the field of clinical NLP. Several studies have found that these can be susceptible to privacy attacks that are unacceptable in the clinical domain, where personally identifiable information (PII) must not be exposed.

However, there is no consensus regarding how to quantify the privacy risks of different models. One prominent suggestion is to quantify these risks using membership inference attacks. In this study, we show that a state-of-the-art membership inference attack on a clinical BERT model fails to detect the privacy benefits of pseudonymizing data. This suggests that such attacks may be inadequate for evaluating token-level privacy preservation of PIIs.

1 Introduction

State-of-the-art results in natural language processing typically rely on large pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) or models in the GPT family (Radford et al., 2019). Multiple studies have found that their large number of parameters can cause PLMs to unintentionally memorize information in their training data, making them vulnerable to privacy attacks (Carlini et al., 2019, 2021). At the same time, other studies have shown that training PLMs using domain-specific data yields better results on domain-specific tasks (Lee et al., 2020; Lamproudis et al., 2021). In the clinical domain, these combined findings pose a significant challenge: training PLMs with clinical data is necessary to achieve state-of-the-art results. However, PLMs can be vulnerable to privacy attacks that are especially dangerous when training with clinical data. Broadly speaking, these attacks can be divided into two classes: training data extraction attacks and membership inference attacks.

1.1 Privacy Attacks

Training data extraction attacks are the more severe class of attacks. An adversary who successfully mounts such an attack can extract details about training data that were used to train a PLM. Carlini et al. (2021) show that GPT-2 is vulnerable to such attacks. Several studies (Nakamura et al., 2020; Lehman et al., 2021; Vakili and Dalianis, 2021) have tried to mount similar attacks on BERT models. To this date, there are no examples of successful training data extraction attacks targeting BERT models.

Membership inference attacks (MIAs) do not aim to *extract* training data from models. Instead, these attacks try to discern whether or not a datapoint was present in a model's training data. Inferring that a datapoint has been present in the training data is less severe than extracting it but could, for example, reveal if a patient has visited a set of clinics.

MIAs have been proposed as a proxy for measuring the degree of memorization in machine learning models (Shokri et al., 2017; Murakonda and Shokri, 2020; Mireshghallah et al., 2022). Both training data extraction attacks and MIAs rely on some degree of memorization in the model. However, MIAs do not require any algorithms that generate the memorized data. By focusing solely on detecting memorization, MIAs are used to estimate a worst-case degree of privacy leakage. Indeed, MIAs are the basis for the ML Privacy Meter developed by Murakonda and Shokri (2020).

Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 318–323
May 22-24, 2023

1.2 Protecting Datapoints or Tokens?

One special property of natural language data is that many words in a sentence can be replaced with synonyms without changing the overall semantics of the sentence. This feature is interesting from a privacy perspective and is the basis for *pseudonymization*.

Pseudonymization is the process of replacing sensitive information with realistic surrogate values. For example, names are replaced with other names or with placeholders. These kinds of sensitive words or phrases are rarely important for the utility of the data, neither for fine-tuning models (Berg et al., 2021; Vakili and Dalianis, 2022), pretraining models (Verkijk and Vossen, 2022; Vakili et al., 2022), nor for general research purposes (Meystre et al., 2014a,b). One important example of this is MIMIC-III (Johnson et al., 2016), which contains a large number of electronic health records in which sensitive words or phrases have been manually replaced with placeholders. This dataset is widely employed in clinical machine learning and is considered to be relatively safe.

One fundamental assumption of pseudonymization is that the higher-level semantics of a text are not important from a privacy perspective. For example, an electronic health record describing a patient visiting a hospital is not sensitive if we cannot infer *who* the patient is, *when* the visit took place, and so on. One way of viewing this is that the data are not primarily sensitive on the datapoint level, but on the token level.

1.3 Membership Inference Attacks and Pseudonymization

Manual pseudonymization is a time-consuming process. Many institutions lack the resources to manually pseudonymize data on the scale required for modern machine learning models or even for less data-intensive qualitative clinical research. An alternative is to use *automatic pseudonymization*. Automatic pseudonymizers typically rely on named entity recognition (NER) to detect sensitive information. The detected entities are then either replaced with realistic surrogates or with placeholders. However, NER systems are rarely perfectly accurate. Imperfect recall leads to some sensitive entities remaining after processing the data, which is undesirable from a privacy perspective.

Because systems performing automatic pseudonymization fail to detect some sensitive entities, it is important to measure the privacy implications of this. A straightforward approach is to consider the recall of the NER model that powers the system. This metric can be used to estimate the number of sensitive entities that remain in the data. Such estimates are useful for determining the sensitivity of an automatically pseudonymized dataset. However, they are less ideal for judging the privacy risks of a machine learning model trained using the dataset. Assuming that the trained model has memorized every single sensitive entity is overly pessimistic.

Estimating the privacy risks of models using MIA, as suggested by Mireshghallah et al. (2022), is an attractive alternative that would allow pseudonymization to be compared to other privacy-preserving techniques. However, MIAs are designed to measure the memorization of entire datapoints rather than the memorization of sensitive tokens. This poses a challenge to the paradigm of using MIAs to estimate the privacy risks of machine learning models trained using pseudonymized data.

In this study, we show that the state-of-the-art MIA described by Mireshghallah et al. (2022) cannot distinguish between a model trained using real or pseudonymized data. These results suggest that using this attack to quantify privacy risks fails to capture privacy gains from pseudonymizing training data.

2 Methods and Data

This study closely mirrors the experimental setup used by Mireshghallah et al. (2022) in order to minimize discrepancies stemming from differences in implementation details. The datasets and models are based on resources introduced by Lehman et al. (2021). The experiments aim to examine whether or not membership inference attacks can distinguish between a model trained using real or pseudonymized data.

2.1 Data

This study uses the ClinicalBERT-1a model trained by Lehman et al. (2021). They train a model using pseudonymized clinical notes from a subset of MIMIC-III. This specific model is of the same size as BERT-base (Devlin et al., 2019) and uses this model's parameters as a starting point for continued pre-training to adapt the model to the clinical domain. The corpus used to train

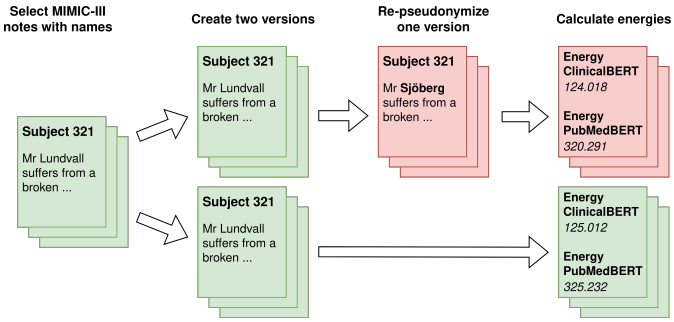

Figure 1: Our experiments use a filtered subset of MIMIC-III that only contains records with named (but pseudonymized) patients. One subset, the *Pseudo* subset, has been used to create the ClinicalBERT model used as the target for the attack. Another version, referred to as the *Real* dataset, is re-pseudonymized and acts as a stand-in for the original sensitive raw data.

the model is also available. Mireshghallah et al. (2022) perform their membership inference experiments using the training data for the BERT model and MIMIC-III data that was not used for training the model. The method also needs a reference model, and this study follows their example by also using PubMed-BERT (Gu et al., 2021) for this purpose.

This study focuses specifically on MIAs' ability to discern whether or not a model has been trained using pseudonymized data. A filtered version of MIMIC-III containing only sentences with names is created to ensure that the results reflect this distinction. This dataset contains a total of 236,114 datapoints. A pseudonymized version of the dataset is created in which all names have been replaced with other names.

After replacing all the names, we have two datasets where each sentence differs solely in what names are used. The dataset used to train the model will be referred to as the *Pseudo* dataset, and the re-pseudonymized dataset will be referred to as the *Real* dataset. This mimics the situation where we have a model trained on *perfectly pseudonymized* training data. Figure 2 illustrates the scenario that is simulated. Ideally, the membership inference attack should indicate that replacing all names with pseudonyms has made the model much safer.

2.2 Predicting Membership

This study uses the same procedures as Mireshghallah et al. (2022) since their method is the current state-of-the-art membership inference attack targeting masked language models like BERT. The method works by analyzing how the target model reacts to a datapoint as compared to a reference model. The target and reference models, in our case ClinicalBERT and PubMed-BERT, differ in that the target model has been trained using sensitive data that the reference model has not been exposed to.

A variety of different measurements can represent the reaction of the model. Following the example of Mireshghallah et al. (2022), we use the *normalized energy values* calculated for every datapoint. These values $E_\theta(S)$ are calculated by estimating the probability of a sequence of tokens S given a set of masking patterns M for a model with the parameters θ:

$$E_\theta(S) = \frac{1}{|M|} \sum_{m \in M} e_\theta(S, m)$$

$$e_\theta(S, m) = \sum_{i \in m} \log\Big[p_\theta(S_i \mid S_m)\Big]$$

S_i is the token at index i and S_m is the altered sequence S to which the masking pattern m has been applied. These normalized energy values are calculated for three datasets, for both the target model and the reference model:

In-data Parts of the dataset used to train the target model. In this study, the two datasets described in Section 2.1 fill this function, as illustrated in Figure 1.

Out-data A second dataset known *not* to belong to the target models training data. This subset

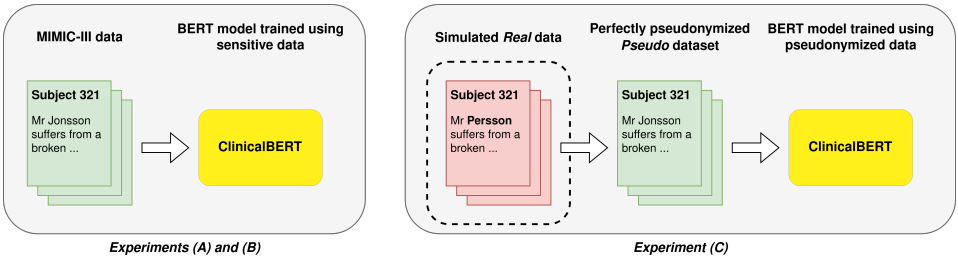

Figure 2: This study simulates the scenario in which a perfectly pseudonymized dataset has been used for continued pre-training of a BERT model. The version of MIMIC-III used to create the ClinicalBERT model from Lehman et al. (2021) is re-identified with pseudonyms and is used in experiments (A) and (B). We refer to this dataset as the *Pseudo* dataset. In experiment (C), we simulate the original, pre-pseudonymized MIMIC-III by populating the data with other names and call this version the *Real* dataset.

of MIMIC-III is also used in Mireshghallah et al. (2022).

Threshold data A third dataset disjoint from the *Out-data* and known not to belong to the target models training data. A subset of i2b2 (Stubbs and Uzuner, 2015) is used, as in Mireshghallah et al. (2022).

The normalized energy values of the target and reference models are compared for the threshold data, resulting in a threshold. This threshold is used to classify if a datapoints belongs to the *In-data* or the *Out-data* based on the difference between the energy values of the datapoint obtained from the models. The intuition behind this method is that if the target model has memorized a datapoint, then its energy value will be noticeably higher relative to the reference model's energy value. The threshold is set so that 90% of the datapoints in the *threshold dataset* are correctly classified as non-members (Mireshghallah et al., 2022). We also calculate the AUC to provide a threshold-independent assessment of the privacy risks.

This study examines the claim that membership inference attacks can be used to quantify privacy gains from using various privacy-preserving techniques. The scenario modeled in these experiments simulates the situation where the privacy-preserving technique is perfect pseudonymization. Every datapoint with a named patient in the training data for ClinicalBERT has a corresponding datapoint in the *Real* dataset where the name is different. In such a scenario, no real names are left in the training data to memorize. Thus, the risk of leaking any name of a patient is zero, representing a substantial increase in privacy. If the attack accurately quantifies these privacy gains, then we would expect it to perform worse when the data has been pseudonymized.

3 Results

Three different attacks are performed using three different datasets as the in-data. The accuracy, precision, and recall values of each attack are listed in Table 1. Experiment (A) mirrors the setup used by Mireshghallah et al. (2022). Experiments (B) and (C) use the subsets of MIMIC-III that only contain names. There are only very small differences in the correctness of the classifications, regardless of the configuration used.

Table 1 also lists the AUC, which represents a threshold-independent evaluation of the MIAs. The AUC varies more than the other three metrics. However, the difference between experiments (A) and (B) is larger than that between experiments (B) and (C). This is despite the fact that the *In-data* for experiments (A) and (B) come from the same population. The difference in AUC between experiments (B) and (C) is 0.017.

Experiments (A) and (B) represent cases where we have not performed any pseudonymization of the training data. That is, the *In-data* are used to train the BERT model without employing any privacy-preserving techniques. Experiment (C) is the result of the simulated scenario where perfect pseudonymization is employed to preserve the privacy of the data. In other words, the model is not exposed to any real names during training. The privacy gains from using this technique are not reflected by the metrics in Table 1.

	In-data	Out-data	Threshold	Accuracy	Precision	Recall	AUC
(A)	*Pseudo, random sample*	*Held-out*	*i2b2*	*0.771*	*0.990*	*0.548*	*0.916*
(B)	Pseudo, names only	Held-out	i2b2	0.780	0.990	0.566	0.882
(C)	Real, names only	Held-out	i2b2	0.770	0.990	0.548	0.865

Table 1: The membership inference attack is run with three different configurations. Experiment (A) uses a random sample of MIMIC-III used in Mireshghallah et al. (2022) as in-data, and all experiments use the same out-data as they do. Experiments (B) and (C) use the datasets described in Section 2.1 for the in-data. The accuracy of each attack is displayed alongside the recall and the precision values. The threshold-independent AUC value is also listed.

4 Discussion and Conclusions

This study focuses specifically on protecting names. Future research would benefit from analyzing additional categories of PII. However, the data and models created by Lehman et al. (2021) focus specifically on names. This class of PII is used in this study to facilitate comparisons with earlier studies.

The results from the three experiments in Table 1 are very similar to each other. At the same time, experiment (C) represents a scenario in which a very strong privacy-preserving measure has been employed to increase the privacy of the target model. If the studied MIA is an accurate way of quantifying the privacy benefits of using pseudonymization, then we would expect the MIA to be much less accurate in experiment (C). The fact that the MIA works nearly as well for experiments (A) and (B) as for (C) indicates that using this attack to quantify memorization does so on a datapoint level. This may be useful for evaluating techniques such as differentially private pretraining (Li et al., 2022), which operate on entire datapoints.

It remains to be shown which of the datapoint's characteristics are used to separate members from non-members. The results of our experiments suggest that using this MIA does not accurately quantify the privacy gains from using pseudonymization, which instead operates on the token level. While the scope of this short paper was limited to evaluating a state-of-the-art MIA for BERT models, future research should also evaluate other MIAs and a wider range of privacy-preserving techniques.

Acknowledgements

We want to thank Fatemehsadat Mireshghallah for sharing the data and code used in Mireshghallah et al. (2022). We are also grateful to the DataLEASH project for funding the research presented in this paper.

References

Hanna Berg, Aron Henriksson, Uno Fors, and Hercules Dalianis. 2021. De-identification of Clinical Text for Secondary Use : Research Issues. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 5: HEALTHINF*, pages 592–599. SciTePress.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *Proceedings of the 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 *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghas-

semi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035. Number: 1 Publisher: Nature Publishing Group.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2021. Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 790–797, Held Online. INCOMA Ltd.

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.

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, Online. Association for Computational Linguistics.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. Large Language Models Can Be Strong Differentially Private Learners. In *Proceedings of the Tenth International Conference on Learning Representations*.

Stéphane Meystre, Shuying Shen, Deborah Hofmann, and Adi Gundlapalli. 2014a. Can physicians recognize their own patients in de-identified notes? *Studies in Health Technology and Informatics*, 205:778–782.

Stéphane M. Meystre, Óscar Ferrández, F. Jeffrey Friedlin, Brett R. South, Shuying Shen, and Matthew H. Samore. 2014b. Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics*, 50:142–150.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347.

Sasi Kumar Murakonda and Reza Shokri. 2020. ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning. ArXiv:2007.09339 [cs].

Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. KART: Privacy Leakage Framework of Language Models Pre-trained with Clinical Records. *arXiv:2101.00036 [cs]*. ArXiv: 2101.00036.

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.

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. ISSN: 2375-1207.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29.

Thomas Vakili and Hercules Dalianis. 2021. Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In *Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021)*.

Thomas Vakili and Hercules Dalianis. 2022. Utility Preservation of Clinical Text After De-Identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing at ACL 2022*, pages 383–388, Dublin, Ireland. Association for Computational Linguistics.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Stella Verkijk and Piek Vossen. 2022. Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1098–1103, Marseille, France. European Language Resources Association.

PAPER III

SWECLINEVAL: A BENCHMARK FOR SWEDISH CLINICAL NATURAL LANGUAGE PROCESSING

Thomas Vakili, Martin Hansson, and Aron Henriksson. 2025. SweClinEval: A Benchmark for Swedish Clinical Natural Language Processing. In *Proceedings of The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*.

Author Contributions Thomas Vakili was responsible for coordinating the study. The experiments were carried out by Thomas Vakili and Martin Hansson. The manuscript was written primarily by Thomas Vakili and Martin Hansson, with contributions from all authors. Aron Henriksson provided extensive support in reviewing and improving the manuscript.

SweClinEval: A Benchmark for Swedish Clinical Natural Language Processing

Thomas Vakili, Martin Hansson, and Aron Henriksson
Department of Computer and Systems Sciences
Stockholm University, Kista, Sweden
{thomas.vakili, martin.hansson, aronhen}@dsv.su.se

Abstract

The lack of benchmarks in certain domains and for certain languages makes it difficult to track progress regarding the state-of-the-art of NLP in those areas, potentially impeding progress in important, specialized domains. Here, we introduce the first Swedish benchmark for clinical NLP: *SweClinEval*. The first iteration of the benchmark consists of six clinical NLP tasks, encompassing both document-level classification and named entity recognition tasks, with real clinical data. We evaluate nine different encoder models, both Swedish and multilingual. The results show that domain-adapted models outperform generic models on sequence-level classification tasks, while certain larger generic models outperform the clinical models on named entity recognition tasks. We describe how the benchmark can be managed despite limited possibilities to share sensitive clinical data, and discuss plans for extending the benchmark in future iterations.

1 Introduction

The field of natural language processing (NLP) has seen several important breakthroughs in the past decade. Currently, the field is dominated by pre-trained transformers models (Vaswani et al., 2017) that can be used to solve a wide and – ideally – diverse set of tasks. The capabilities of these models have to a large degree been tracked through the use of *benchmarks*, significantly helping to drive progress in the area. These evaluation suites test how the models perform on different pre-defined tasks and allow for comparisons between models and approaches.

While there are many benchmarks available, there are also many potential uses for NLP that they do not cover. Frequently, evaluations rely on English data (Joshi et al., 2020; Søgaard, 2022). However, a model performing well on an English benchmark in no way guarantees similar performance if the language changes. Additionally, benchmarks such as GLUE (Wang et al., 2018) tend to focus on tasks formulated for general-domain data. With increasing calls for NLP to be applied to specific domains, such as the clinical domain, there is a pressing need for benchmarks that address these areas.

The clinical domain, in particular, suffers from a lack of datasets for evaluating NLP systems. One critical reason for this is the inherently sensitive nature of clinical data. There are multiple studies (Carlini et al., 2021; Nasr et al., 2023) demonstrating the potential risks of using sensitive data for machine learning – let alone sharing data in their raw form. That said, there are some widely used resources for clinical NLP. Prominent examples include the various versions of MIMIC (Johnson et al., 2022) and the i2b2 datasets (Murphy et al., 2010). Crucially, these datasets predominantly evaluate NLP systems on data in English or other higher-resourced languages.

In this paper, we introduce the first Swedish benchmark based on real clinical NLP data: *SweClinEval*. This benchmark consists of datasets built from electronic health records from the Health Bank (Dalianis et al., 2015) and includes a wide range of clinical tasks. These tasks include three different document-level sequence classification tasks and three token-level named entity recognition (NER) tasks. This introduction of SweClinEval includes nine different models, and future additions will be added to the benchmarks online leaderboard[1].

The evaluations presented in this paper show that many models targeting Swedish data per-

[1]The leaderboard of SweClinEval is available at: https://sweclineval.dsv.su.se

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 767–775
March 3-4, 2025

form strongly on our benchmark. However, the performances vary, and several interesting trends emerge from our results. These results highlight the importance of continuing to focus on domain-specific evaluations for languages other than English. Our results demonstrate the current state of Swedish clinical NLP, and the benchmark serves as an important tool for monitoring progress in this important NLP domain.

2 Related Research

The NLP community has seen impressive advances in the past few years with the advent of LLMs. Several new model architectures have been proposed since Vaswani et al. (2017) described the transformer, and new models are released at a rapid pace. These LLMs aim to be general-purpose models, with task-specific applications requiring only smaller adjustments in the form of fine-tuning or prompt engineering. In response to this new paradigm, there has been an increasing focus on creating benchmarks that capture the nuanced difference in performance in the growing plethora of models.

2.1 General-Domain Benchmarks

Benchmarks come with different objectives and designs. A prominent example is the GLUE (Wang et al., 2018) family of benchmarks. The original *General Language Understanding Evaluation* (GLUE) benchmark aimed to, as the name suggests, capture a wide range of capabilities that act as proxies for natural language understanding. As models have become more powerful, the NLP community has responded with more varied and difficult benchmarks. These include the SuperGLUE (Wang et al., 2019) benchmark that introduces more difficult tasks, and the XGLUE benchmark (Liang et al., 2020) that also examines the multilingual capabilities of models.

2.2 Swedish Benchmarks

The vast majority of papers at NLP conferences focus on English data (Søgaard, 2022), to the detriment of smaller and less well-resourced languages. The introduction of multilingual benchmarks such as XGLUE is in part a response to this dominance of English-only datasets.

Another development is the creation of language-specific benchmarks. For Swedish, this trend has materialized in the form of benchmarks such as the Superlim[2] (Berdicevskis et al., 2023) and OverLim[3] benchmarks. These benchmarks mirror the structure of the GLUE family of benchmarks, but use datasets that specifically use Swedish data.

An important benchmark, especially for the purposes of this paper, is the ScandEval (Nielsen, 2023) benchmark. This benchmark is multilingual but focuses mainly on the Scandinavian language family. LLMs for these languages have been found to benefit from training on shared datasets. The ScandEval benchmark was also used to determine which models to benchmark, as detailed in Section 3.2.

2.3 Clinical Benchmarks

The most commonly used benchmarks aim to measure general-purpose capabilities in a general-domain setting. However, many important applications of NLP are domain-specific. In this paper, we focus on NLP for clinical data, which has several domain-specific features. Due to the setting in which they are produced, clinical data are often riddled with domain-specific acronyms and terminology that can be harder for general-domain models to process (Dalianis, 2018). Furthermore, clinical datasets are difficult to share due to the inherently sensitive nature of the data.

Nevertheless, there have been efforts to create benchmarks that measure the clinical or biomedical capabilities of LLMs. BLURB (Gu et al., 2021) is a benchmark in the vein of GLUE and includes a wide range of clinical tasks. This benchmark highlighted the shortcomings of general-domain models and the benefits of using LLMs specific to the clinical domain. In contrast, the later Dr. Bench (Gao et al., 2023) benchmark shows that general-domain models can indeed outcompete domain-specific models on certain tasks. These diverging conclusions exemplify the need for diverse domain-specific benchmarks to monitor the progress of LLMs in the clinical domain.

A recent benchmark highly relevant for Swedish biomedical NLP is the *Swedish Medical Benchmark* introduced by Moëll and Farestam (2024). This benchmark is comprised of a selection of four datasets with multiple-choice questions. These datasets were collected from public

[2]Superlim is Swedish for super glue, a reference to the SuperGLUE benchmark.

[3]https://huggingface.co/datasets/KBLab/overlim

sources and probe LLMs for biomedical knowledge. A benefit of using publicly available data is that the data can be shared. On the other hand, such data are not representative of the types of clinical data and tasks encountered when creating, for example, a system interfacing with patient records.

The main contribution of this paper is the introduction of the SweClinEval benchmark. This benchmark is not only focused on the clinical domain, but is the first benchmark that monitors the state of Swedish clinical NLP using real electronic patient records for realistic clinical tasks.

3 Methods and Materials

Creating this first rendition of SweClinEval involved collecting resources for evaluation and deciding how to conduct the evaluations. This section describes the datasets used for the benchmark and the models that were tested, and how they were chosen. The design of the evaluations and the metrics used for comparing models are also described.

3.1 Datasets

The benchmark consists of six datasets that are part of the Health Bank (Dalianis et al., 2015) infrastructure[4] . The Health Bank consists of over 2 million Swedish electronic health records written between 2006 and 2014 from a range of different clinical units in Sweden. The datasets have been collected for more than a decade, either through manual annotation or by mining information from the Health Bank data. Three of the datasets are document-level classification tasks, and the other three are token-level NER tasks.

ICD-10 The Stockholm EPR Gastro ICD-10 Corpus (Remmer et al., 2021) is a document-level classification task where discharge summaries related to gastrointestinal patients are assigned high-level diagnosis code blocks. These 10 different code blocks encode information about what type of diagnosis was assigned to the patient. The task is a multi-label classification task, meaning that each document can be associated with more than one code block.

ADE The Stockholm EPR ADE ICD-10 Corpus (Vakili et al., 2022) is another document-level classification task that determines whether or not a discharge summary describes a patient suffering from an adverse drug event. This is a binary classification problem.

Factuality The Stockholm EPR Diagnosis Factuality Corpus (Velupillai, 2011; Velupillai et al., 2011) is the third document-level classification task. This manually annotated corpus assigns a *factuality* level to the diagnoses of each clinical note. These different levels describe the confidence with which a diagnosis was decided. The six different classes are: *Certainly Negative, Probably Negative, Possibly Negative, Possibly Positive, Probably Positive,* and *Certainly Positive*.

Factuality NER This version of the Stockholm EPR Diagnosis Factuality Corpus is a token-level NER task. The task involves assigning the same six labels to tokens in each document that indicate a diagnosis. The task is to both detect these diagnoses and assign them a factuality level. This version also includes an *Other* tag for clinically relevant information that is not indicating factuality.

Clinical Entity NER The Stockholm EPR Clinical Entity Corpus (Skeppstedt et al., 2014) is a manually annotated NER corpus that describes a task in which the model needs to identify clinically relevant terms. These are divided into four classes: *Diagnosis, Findings, Body Parts*, and *Drugs*. The model needs to detect tokens associated with these classes and assign them the correct labels.

PHI NER The final corpus used in the benchmark is the Stockholm EPR PHI Corpus (Dalianis and Velupillai, 2010). This corpus consists of patient records and has been manually annotated for named entities describing personally identifiable protected health information (PHI). Each instance of PHI is assigned one of nine classes: *First Name, Last Name, Age, Phone Number, Partial Date, Full Date, Location, Health Care Unit,* and *Organization*.

Additional statistics about the six datasets are listed in Table 1. None of the datasets have been

[4]This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

adapted for use with prompt-style autoregressive language models. This limitation is reflected in the model selection for this paper and adapting the datasets for broader use is left to future iterations of SweClinEval.

3.2 Models

Nine different models were included for the experiments in this paper and are listed in Table 2. Two of these – SweDeClin-BERT and SweClin-BERT – were specifically created for use in Swedish clinical NLP and have previously shown strong performance on the datasets in SweClinEval (Vakili et al., 2024). Additionally, seven general-domain models known to perform well for Swedish data were included. These seven models were selected based on their performance in the ScandEval (Nielsen, 2023) benchmark.

The majority of the models are based on the BERT/RoBERTa architecture (Devlin et al., 2019; Liu et al., 2019). The RemBERT (Chung et al., 2020) and Multilingual E5 Large (Wang et al., 2024) models are based on their own transformer architectures. These two models also exhibit the greatest language diversity in their training data. The training data for the *RoBERTa Large* and *BERT Large* models from AI Sweden are also multilingual. These were trained using *The Nordic Pile* corpus (Öhman et al., 2023) which consists mainly of Scandinavian and English data.

Crucially, all nine models are encoder models. This is a limitation imposed by the nature of the datasets, as described in the previous section. It is possible to restructure datasets so that they can be used autoregressively. However, such a conversion would be non-trivial and is left for future research.

3.3 Evaluation Procedure

All nine models were trained and evaluated using the six datasets. To ensure a fair estimate of each model's performance, the evaluations were done using 10-fold cross-validation. This allowed us to calculate the average performance alongside the standard deviation, enabling a more fair comparison. The comparisons were based on the F_1 scores of each cross-validation.

For each fold in the cross-validation, models were trained for a maximum of three epochs. Early stopping was enabled, and the best-performing checkpoint was used to predict the test set in each fold. The F_1 scores used for the comparisons were based on the average score from each fold and the standard deviation. For the NER tasks, these were the token-level micro F_1 scores. The *PHI NER* task uses the IOB scheme to mark where an entity begins and ends, and this distinction was included in the evaluation. The document-level sequence classification tasks instead rely on F_1 scores weighted for the support of each class in the test set.

4 Results

Nine models were evaluated using 10-fold cross-validation for six different datasets, resulting in 540 evaluations. The average F_1 scores and their deviations are listed in Table 3.

For the sequence-level classification tasks, the highest average F_1 scores are consistently obtained using the domain-adapted models. The same is not true for the token-level NER tasks. For these tasks, the highest F_1 scores were obtained by the general-domain *RoBERTa Large* model from AI Sweden. However, the domain-adapted *SweDeClin-BERT* model has the second-highest average F_1 scores for the *Factuality NER* and *Clinical Entity NER* tasks.

The different average F_1 scores vary substantially between the best- and worst-performing models. Nevertheless, the standard deviations are large. This means that many of the averages are within a standard deviation of a competing model. This necessarily limits the analysis into which models are *best*, since randomness has a strong influence on the variability in the F_1 scores.

In addition to the predictive performance, Table 4 also lists the processing time of each model when performing inference. Unsurprisingly, the smaller models are faster to run. These figures are based on the HuggingFace implementations of each model running on an *Nvidia RTX A5000* GPU. Although the exact inference time will depend on the hardware available, the number indicate the relative cost of running these model in a production environment.

5 Discussion

A few trends emerge from the results in the previous section. There are also some limitations and pointers to future work that are important to discuss. However, we begin by discussing the findings from our results.

As previously mentioned, the highest average F_1 scores in the sequence classification tasks are

Task	Type	Classes	Documents	Tokens
ICD-10	Classification	10	6,062	930,550
ADE	Classification	2	21,725	931,778
Factuality	Classification	6	3,710	102,223
Factuality NER	NER	7	3,822	286,205
Clinical Entity NER	NER	4	3,120	178,672
PHI NER	NER	9	29,560	282,820

Table 1: Six different datasets were used in the benchmark evaluation. Three of these are NER tasks and three are sequence classification tasks. This table lists the datasets alongside their size, the number and classes, and the types of classification they target.

Model	Parameters	Paper
SweDeClin-BERT	125 M	(Vakili et al., 2022)
SweClin-BERT	125 M	(Lamproudis et al., 2021)
KB-BERT Base	125 M	(Malmsten et al., 2020)
AI Nordics BERT Large	335 M	N/A[5]
AI Sweden RoBERTa Large	355 M	N/A[6]
AI Sweden BERT Large	369 M	N/A[7]
KB-BERT Large	370 M	N/A[8]
Multilingual E5 Large	560 M	(Wang et al., 2024)
RemBERT	576 M	(Chung et al., 2020)

Table 2: In this initial edition of the SweClinEval benchmark, nine different models were evaluated. All models are encoder models, and they are listed here in order of parameter count. When available, the paper that introduced the model is listed. SweDeClin-BERT and SweClin-BERT are the only models created specifically for Swedish clinical NLP.

achieved by the domain-adapted models. This indicates that, at least for these tasks, domain adaptation results in better performance on clinical NLP tasks. On the other hand, this finding is not as clear when examining the NER tasks. While the domain-adapted models perform competitively, the best-performing model on all three NER tasks is AI Sweden's *RoBERTa Large* model.

Crucially, the models differ greatly in size. The smaller models are around three times smaller than the medium-sized models, and more than four times smaller than the largest models. The comparatively strong performance of the domain-adapted models, which are both small, is more impressive when seen from this perspective. Domain adaptation seems to allow smaller models to compete with larger counterparts. Naturally, this leads to the question of whether this finding holds true for larger models, too. The two clinical models are initialized from *KB-BERT Base*, and an interesting direction for future work could be examining if initializing from larger models produces analogous results. The *RoBERTa Large* model from AI Sweden would be an interesting candidate, given its strong performance on the NER tasks. In any case, the benefits from domain adaptation align with many previous studies (Gu et al., 2021; Lamproudis et al., 2021).

Perhaps somewhat surprisingly, parameter count itself does not seem to be a determining factor in what models are the strongest. This is not only the case when comparing domain-adapted and general-domain models. For example, *KB-BERT Base* and *KB-BERT Large* were both trained by the same organization, and are from the same model family. The main difference between the

[5] https://huggingface.co/AI-Nordics/bert-large-swedish-cased

[6] https://huggingface.co/AI-Sweden-Models/roberta-large-1160k

[7] https://huggingface.co/AI-Sweden-Models/bert-large-nordic-pile-1M-steps

[8] https://huggingface.co/KBLab/megatron-bert-large-swedish-cased-165k

Model	Size	ICD-10	Factuality	ADE
		Classification	Classification	Classification
SweDeClin-BERT	S	0.832±0.011	**0.735±0.018**	**0.203±0.022**
SweClin-BERT	S	**0.836±0.014**	0.731±0.021	0.196±0.014
KB-BERT Base	S	0.801±0.015	0.671±0.017	0.185±0.012
AI Nordics BERT Large	M	0.811±0.012	0.657±0.025	0.192±0.013
AI Sweden RoBERTa Large	M	0.816±0.018	0.594±0.126	0.159±0.028
AI Sweden BERT Large	M	0.816±0.012	0.654±0.032	0.167±0.057
KB-BERT Large	M	0.801±0.013	0.683±0.019	0.190±0.011
Multilingual E5 Large	L	0.824±0.013	0.525±0.074	0.192±0.015
RemBERT	L	0.823±0.010	0.379±0.059	0.149±0.050
Model	**Size**	**Factuality**	**Clinical Entity**	**PHI**
		NER	**NER**	**NER**
SweDeClin-BERT	S	0.623±0.024	0.766±0.034	0.945±0.012
SweClin-BERT	S	0.610±0.018	0.754±0.038	0.938±0.014
KB-BERT Base	S	0.600±0.025	0.743±0.039	0.941±0.025
AI Nordics BERT Large	M	0.612±0.026	0.721±0.039	0.948±0.010
AI Sweden RoBERTa Large	M	**0.641±0.011**	**0.779±0.036**	**0.965±0.009**
AI Sweden BERT Large	M	0.513±0.185	0.738±0.038	0.854±0.285
KB-BERT Large	M	0.552±0.025	0.697±0.046	0.936±0.012
Multilingual E5 Large	L	0.603±0.019	0.511±0.339	0.608±0.037
RemBERT	L	0.417±0.026	0.600±0.075	0.947±0.011

Table 3: Nine encoder models were evaluated for sequence classification using six different clinical tasks. Three of the tasks were sequence classification tasks, and three were token-level NER tasks. The performance is summarized using F_1 with standard deviations. The highest F_1 of each task is bolded, and the second highest is underlined. Models are ordered according to ascending parameter count as listed in Table 2 and categorized as ***S**mall, **M**edium,* or ***L**arge* models.

Model	Sequence	NER
SweDeClin-BERT	2.86 ms	2.85 ms
SweClin-BERT	2.86 ms	2.84 ms
KB-BERT Base	2.88 ms	2.87 ms
AI Nordics BERT Large	5.60 ms	5.56 ms
AI Sweden RoBERTa Large	6.91 ms	6.05 ms
AI Sweden BERT Large	5.60 ms	5.56 ms
KB-BERT Large	8.76 ms	8.67 ms
Multilingual E5 Large	6.08 ms	6.03 ms
RemBERT	9.38 ms	9.36 ms

Table 4: The different models used in the benchmark use different architectures and are of different sizes. This table lists the time of each model for inference on one sample, both for sequence classification and NER.

models is that the larger model consists of more parameters and was trained using a much larger corpus. Nevertheless, *KB-BERT Base* actually outperforms its larger counterpart in some cases. While the large standard deviations call for cautious interpretations of the results, it is at least clear the larger model is not outperforming its smaller competitor.

On the other hand, parameter count clearly influences the inference speed of the models, as indicated in Table 4. While this is not surprising, it is worth mentioning. Other benchmarks, such as the GLUE benchmark, do not always present this information. However, inference speed can be important in practice, especially when differences in performance are small. Smaller and faster models require less expensive hardware, which can be important in cases where it is not possible to use cloud providers to run the models. This is frequently the case for clinical uses, due to the sensitivity of clinical data.

6 Conclusions

In this paper, we present SweClinEval – the first Swedish benchmark for clinical NLP. We evaluate

a wide range of encoder-style LLMs for six different Swedish clinical NLP tasks. This effort represents the first such evaluation to be conducted, and forms a basis for future monitoring of the advances in Swedish clinical NLP.

The results of this first evaluation indicate several interesting trends. The benchmark results suggest that domain adaptation is an effective strategy for improving the performance of LLMs in the clinical domain, at least for small LLMs. Future research should examine whether this also holds for larger models. Furthermore, the evaluations also show that parameter count alone is not enough to perform strongly in the tasks included in our benchmark.

The aim of this paper is to enable monitoring of the progress within Swedish clinical NLP. Due to privacy constraints, the data cannot be shared. We strongly encourage others interested in Swedish clinical NLP to contact us for inclusion in the benchmark. This pragmatic approach to benchmarking enables us to monitor the progress that is being made, which SweClinEval makes possible.

6.1 Limitations

A limitation of the current version of the benchmark is that it only supports encoder models. This is unfortunate, as there is a strong trend towards using autoregressive models both in fine-tuning and few-shot settings. Future versions of the benchmark would benefit from including versions of the datasets that allow non-encoder models to be evaluated. This is not trivial but, as demonstrated by the ScandEval benchmark, it is possible and is an aim for future iterations of the benchmark. Furthermore, we aim to extend the benchmark with more datasets for tasks such as summarization and question-answering.

A more significant limitation of SweClinEval is that currently, only parts of the data can be shared. This restriction is due to privacy regulations surrounding the inherently sensitive clinical data from which the datasets were created. However, two of the datasets – the *Stockholm EPR PHI Corpus* and the *Stockholm EPR ICD-10 Corpus* – are available in automatically de-identified form for academic users. As the regulatory environment around secondary use of private information changes, it may be possible to share the data more freely in the future. For now, our view is that SweClinEval is a pragmatic solution that allows the Swedish NLP community to monitor the progress in Swedish clinical NLP.

References

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. https://doi.org/10.18653/v1/2023.emnlp-main.506 Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

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 *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. http://arxiv.org/abs/2010.12821 Rethinking embedding coupling in pre-trained language models.

H. Dalianis, A. Henriksson, M. Kvist, S. Velupillai, and R. Weegar. 2015. HEALTH BANK - A workbench for data science applications in healthcare. In *CEUR Workshop Proceedings*. CEUR-WS.

Hercules Dalianis. 2018. https://doi.org/10.1007/978-3-319-78503-5 *Clinical Text Mining*. Springer International Publishing, Cham.

Hercules Dalianis and Sumithra Velupillai. 2010. https://doi.org/10.1186/2041-1480-1-6 De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. *Journal of Biomedical Semantics*, 1(1):6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. https://doi.org/10.18653/v1/N19-1423 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. 2023. https://doi.org/10.1016/j.jbi.2023.104286 Dr.bench: Diagnostic reasoning benchmark for clinical natural

language processing. *J. of Biomedical Informatics*, 138(C).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. https://doi.org/10.1145/3458754 Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2022. https://doi.org/10.13026/7VCR-E114 MIMIC-IV.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2021. Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 790–797, Held Online. INCOMA Ltd.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. https://doi.org/10.18653/v1/2020.emnlp-main.484 XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv:2007.01658 [cs]*. ArXiv: 2007.01658.

Birger Moëll and Fabian Farestam. 2024. https://sltc2024.github.io/abstracts/moell_farestam.pdf Swedish Medical Benchmark, an evaluation framework for LLMs in the Swedish medical domain.

Shawn N. Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C. Chueh, Susanne Churchill, and Isaac Kohane. 2010. https://doi.org/10.1136/jamia.2009.000893 Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association: JAMIA*, 17(2):124–130.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. https://doi.org/10.48550/arXiv.2311.17035 Scalable Extraction of Training Data from (Production) Language Models. ArXiv:2311.17035 [cs].

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Sonja Remmer, Anastasios Lamproudis, and Hercules Dalianis. 2021. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In *Proceedings of RANLP 2021: Recent Advances in Natural Language Processing, RANLP 2021, 1-3 Sept 2021, Varna, Bulgaria*, pages 1158–1166.

Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158. Publisher: Elsevier.

Anders Søgaard. 2022. Should We Ban English NLP for a Year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2024. End-to-end pseudonymization of fine-tuned clinical BERT models. *BMC Medical Informatics and Decision Making*, 24:162.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 4245–4252. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sumithra Velupillai. 2011. Automatic classification of factuality levels: A case study on Swedish diagnoses and the impact of local context. In *Fourth International Symposium on Languages in Biology and Medicine, LBM 2011*.

Sumithra Velupillai, Hercules Dalianis, and Maria Kvist. 2011. Factuality levels of diagnoses in Swedish clinical text. *Studies in Health Technology and Informatics*, 169:559–563.

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*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. https://doi.org/10.18653/v1/W18-5446 GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. http://arxiv.org/abs/2402.05672 Multilingual e5 text embeddings: A technical report.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. https://doi.org/10.48550/arXiv.2303.17183 The Nordic Pile: A 1.2TB Nordic Dataset for Language Modeling. ArXiv:2303.17183.

PAPER IV

DOWNSTREAM TASK PERFORMANCE OF BERT MODELS PRE-TRAINED USING AUTOMATICALLY DE-IDENTIFIED CLINICAL DATA

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Author Contributions Thomas Vakili was responsible for coordinating the study. The method was designed by Thomas Vakili, with input from the other authors. The experiments were carried out by Thomas Vakili and Anastasios Lamproudis. The manuscript was written primarily by Thomas Vakili, with contributions from all authors. Hercules Dalianis and Aron Henriksson provided extensive support in reviewing and improving the manuscript.

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4245–4252
Marseille, 20-25 June 2022

Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, Hercules Dalianis
Department of Computer and Systems Sciences (DSV), Stockholm University, Kista, Sweden
{thomas.vakili, anastasios, aronhen, hercules}@dsv.su.se

Abstract

Automatic de-identification is a cost-effective and straightforward way of removing large amounts of personally identifiable information from large and sensitive corpora. However, these systems also introduce errors into datasets due to their imperfect precision. These corruptions of the data may negatively impact the utility of the de-identified dataset. This paper de-identifies a very large clinical corpus in Swedish either by removing entire sentences containing sensitive data or by replacing sensitive words with realistic surrogates. These two datasets are used to perform domain adaptation of a general Swedish BERT model. The impact of the de-identification techniques is assessed by training and evaluating the models using six clinical downstream tasks. The results are then compared to a similar BERT model domain-adapted using an unaltered version of the clinical corpus. The results show that using an automatically de-identified corpus for domain adaptation does not negatively impact downstream performance. We argue that automatic de-identification is an efficient way of reducing the privacy risks of domain-adapted models and that the models created in this paper should be safe to distribute to other academic researchers.

Keywords: Privacy-preserving machine learning, pseudonymization, de-identification, Swedish clinical text, pre-trained language models, BERT, downstream tasks, NER, multi-label classification, domain adaptation

1. Introduction

Natural Language Processing (NLP) research is currently dominated by so-called pre-trained language models based on transformers (Vaswani et al., 2017), which were popularized by the introduction of the BERT model by Devlin et al. (2019). These language models typically consist of millions – even billions – of parameters that are learned from enormous corpora. The success of pre-trained language models in general-domain tasks has prompted research into whether these models also succeed in medical-domain tasks.

Language models are taught to model language by learning the statistical distributions of the words in their training data. However, words often have different meanings depending on in which domain they are used. The word *chest* has a dual meaning in everyday language – something used for storage or a region of the body – but only one of these is relevant in a medical context. A language model which has learned the word *chest* from a general-domain corpus may have a representation of the word that is sub-optimal in the medical domain.

Indeed, many researchers have found that performance on domain-specific tasks is helped by adapting existing language models or pre-training new models using in-domain data (Lee et al., 2019; Beltagy et al., 2019; Lamproudis et al., 2021; Lamproudis et al., 2022b; Lamproudis et al., 2022a). Better performance means that the models will be more useful in helping medical professionals improve patient outcomes.

However, the scale of the data used to train these models means that researchers cannot know what sensitive information the corpora contain. In the medical domain, we can be certain that the texts contain sensitive information. This is cause for concern since pre-trained language models are susceptible to privacy attacks (Bender et al., 2021).

This paper examines one way of reducing the privacy risks: automatic de-identification. Two different approaches are studied: pseudonymization (Sweeney, 1996; Dalianis, 2019) and removal of sensitive data. Two different clinical BERT models are created by applying these techniques to the pre-training data. The impact of automatic de-identification on the performance of the models is then evaluated on downstream tasks.

2. Related Research

The two main topics of this paper are automatic de-identification and the privacy risks of large language models. This section introduces these concepts by providing a brief summary of results related to the topic of this paper.

2.1. Privacy Attacks on Language Models

Large pre-trained language models are susceptible to a wide range of attacks on privacy. One reason for this is due to their size, which gives them a tendency to unintentionally memorize parts of their training data. The attacks can generally be separated into two main categories:

Training data extraction An attacker that successfully mounts a model inversion attack is able to extract details about the training data. One example of a training data extraction attack was mounted by Carlini et al. (2020). They managed to extract entire passages from IRC logs from the model GPT-2 (Radford et al., 2019).

Membership inference If an attacker is able to discern whether or not a datapoint was part of the training data, they have successfully mounted a membership inference attack (Shokri et al., 2017). Although these attacks are typically less severe than training data extraction attacks, they can also expose sensitive data.

To the best of our knowledge, there are no examples of successful training data extraction attacks on BERT models. Lehman et al. (2021) and Vakili and Dalianis (2021) found that BERT models are at least less susceptible to such attacks than GPT-2. Both studies attempted to extract training data from a BERT model trained on a version of MIMIC-III (Johnson et al., 2016) which had its masked entities populated with realistic but fake values.

Nakamura et al. (2020) performed a related attack that attempted to re-predict pseudonymized information. They trained a BERT model on a version of the MIMIC-III (with inserted surrogate values) and then re-masked the surrogate entities in this dataset. They then attempted to reconstruct the surrogate names but did not succeed, concluding that this does not seem to be a viable attack.

Lehman et al. (2021) also performed membership inference attacks on their BERT model. Their results indicated a small risk of memorizing patients' names. At the same time, they were not able to link a patient's name to any of their conditions. Jagannatha et al. (2021) also performed membership inference attacks on BERT and found that there is a risk of privacy leakage from BERT models. However, this risk is significantly smaller than for models like GPT-2.

2.2. De-Identification of Clinical Text Data

The electronic health records (EHRs) used in clinical NLP are inherently sensitive. For example, the data used in this study was found to have an estimated protected health information (PHI) density[1] of 1.57% (Henriksson et al., 2017). However, the PHI density varied considerably across medical specialties and classes of clinical notes. For example, almost 20% of the sentences in discharge summaries contained at least one PHI. The prevalence of PHI has caused many researchers to explore ways of reducing the risks to patient privacy that comes with using their health data. One active area of research is automatic de-identification.

Automatic de-identifiers typically rely on named entity recognition (NER) models to detect sensitive data in datasets. Thus, the recall of the model needs to be balanced against its precision. In this context, the classic precision-recall trade-off translates to one between utility and privacy. Low recall means that a lot of sensitive data will be undetected, but a low precision results in a dataset where a lot of non-sensitive data is corrupted.

[1]PHI density was defined as the number of PHI mentions divided by the number of tokens.

Berg et al. (2020) used various high recall models to de-identify several Swedish clinical datasets. This did not seem to lower the utility of the datasets, as training with the datasets did not significantly decrease downstream performance. The authors tried out four strategies for the de-identification: pseudonymization (replacing sensitive data with surrogates), masking the sensitive data, replacing a sensitive word with its class name (e.g., replacing "John" with "First Name"), and removing the sensitive data along with the sentence in which it appeared. All of the downstream tasks were NER tasks and were approached using a machine learning algorithm based on conditional random fields (CRFs). The tasks were clinical entity identification, adverse drug effect identification, and cervical cancer symptom detection. Pseudonymization resulted in the smallest negative impact on the downstream tasks, while the sentence removal strategy resulted in a greater deterioration of the performance.

Vakili and Dalianis (2022) automatically de-identified three Swedish clinical datasets using pseudonymization. Each dataset was associated with a task: two sequence classification tasks (ICD-10 classification and factuality classification) and one NER task (clinical entity recognition). Different BERT models were trained using unaltered and pseudonymized data, and the performances on all tasks were compared. There was no significant difference in the performance of the models trained on unaltered data and the models trained on pseudonymized data.

Obeid et al. (2019) de-identified clinical data and evaluated the impact of this by building detectors of altered mental status (AMS) using a variety of machine learning models. These included Naïve Bayes Classifiers, Single Decision Trees, Random Forests, and Multilayer Perceptrons. The deep learning models performed the strongest, but no model showed any significant deterioration in performance when trained using de-identified text instead of the original text.

No automatic de-identification system has perfect recall, and some sensitive data will remain in a processed corpus. However, pseudonymizing the data makes it difficult to determine which data are real and which data are pseudonymized. Carrell et al. (2019) explored the concept of *hiding in plain sight* (HIPS). They were able to train a tagger to distinguish between pseudonymized data and data that were HIPS in a pseudonymized dataset. The tagger performed significantly better than random guessing but had a high rate of false positives and false negatives. Thus, the authors concluded that HIPS is still helpful for protecting privacy.

This study applies two of the de-identification approaches outlined in Berg et al. (2021) to a clinical corpus data. However, the data used in this paper is much larger in scale and is used to pre-train language models rather than to build task-specific classifiers.

3. Data

The clinical data used to train and evaluate the BERT models originate from the Karolinska University Hospital. The data are stored in the research infrastructure Health Bank – The Swedish Health Record Research Bank[2] (Dalianis et al., 2015) at DSV/Stockholm University.

3.1. EHRs from the Health Bank

The BERT models in this paper were pre-trained using a 17.9 GB subset of the Health Bank. The clinical texts come from a large number of clinical units and encompass over 2 million EHRs[3]. This dataset is comparable in size to the general domain Swedish corpus of newspapers, Swedish Wikipedia, and government documents that was used to pre-train *KB-BERT* (Malmsten et al., 2020).

These EHRs were de-identified according to the process outlined in Section 4.1, and the resulting dataset was used to train two BERT models, as will be described in Section 4.2. Lamproudis et al. (2021) also use this dataset in its unaltered form to train the baseline model used for evaluating the impact of de-identifying the pre-training data.

3.2. Datasets for Downstream Tasks

Five manually annotated datasets, all created from the Health Bank, were used to evaluate the downstream performance of the models. All of the downstream tasks concern clinical NLP tasks and make it possible to compare the BERT models to each other.

Stockholm EPR Gastro ICD-10 Corpus A Gastro ICD-10 data set consisting of 6,062 gastro-related discharge summaries and their assigned ICD-10 diagnosis codes. The data set encompasses 4,985 unique patients and 795,839 tokens. The data are divided into 10 groups that correspond to different body parts; the ICD-10 codes range from K00 to K99. Each group contains several codes (Remmer et al., 2021).

Stockholm EPR PHI Corpus A PHI data set of 4,480 annotated entities and 380,000 tokens. The PHIs correspond to nine PHI classes: *First Name*, *Last Name*, *Age*, *Phone Number*, *Location*, *Health Care Unit*, *Organization*, *Full Date*, and *Date Part* (Dalianis and Velupillai, 2010).

Stockholm EPR Clinical Entity Corpus A clinical entity data set comprising 70,852 tokens and 7,946 annotated entities corresponding to four clinical entity classes *Diagnosis*, *Findings*, *Body parts*, and *Drugs* (Skeppstedt et al., 2014).

Stockholm EPR Diagnosis Factuality Corpus A factuality diagnosis data set encompassing six levels of annotations regarding the factuality of a diagnosis. The data set consists of 3,710 samples with 7,066 annotated entities *Certainly Positive*, *Probably Positive*, *Possibly Positive*, *Possibly Negative*, *Probably Negative*, and *Certainly Negative* encompassing 240,000 tokens (Velupillai et al., 2011; Velupillai, 2011). The dataset is used for two tasks. One is a NER task where the goal is to identify tokens specifying diagnoses and assigning them a factuality label. The second task treats the sample as a single datapoint and performs a multi-label classification of the entire sample to predict its factuality.

Stockholm EPR ADE ICD-10 Corpus A newly introduced ADE corpus containing 16,858 samples encompassing 634,000 tokens. The samples are distributed over 12 different ICD-10 codes describing adverse drug events. The task is treated as a binary classification task where positive samples have been assigned a specific ICD-10 code that denotes an adverse drug event. Negative samples in each group have been assigned a code describing a similar condition that was not drug-induced. The goal of the task is to determine whether or not the condition defined by the ICD-10 code was induced by an ADE.

4. Experiments

The study encompasses three steps. First, the Health Bank corpus is processed to detect and deal with sensitive data. This leads to two different clinical corpora that are then used for domain-adaptive pre-training. The resulting models are evaluated on downstream tasks, and the results are compared to other models trained on the Health Bank data. This section gives a detailed account of the experiments and their results.

4.1. De-Identifying the Health Bank

A NER model was built based on a clinical BERT model trained by Lamproudis et al. (2021) using the *Stockholm EPR PHI Corpus*. The model was used to detect the nine PHI classes described in Section 3.2 and by Dalianis and Velupillai (2010). This model was then applied to the 17.9 GBs of EHRs extracted from the Health Bank. This processing uncovered a large amount of possibly sensitive data. The number of detected instances for each PHI type is listed in Table 1.

Two approaches to de-identification were taken, as illustrated in Figure 1. In the first approach, which we refer to as *pseudonymization*, each detected entity was replaced by a realistic surrogate value of the same class. For example, a detected name will be replaced with another (generated but realistic) name. Pseudonymization preserves the semantics of the text as long as the entity has been correctly classified and allows the model to

[2]Health Bank: `http://dsv.su.se/healthbank`

[3]This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

PHI Type	# Predicted Instances	NER Recall	NER Precision
Health Care Unit	19,659,127	80%	87%
Partial Date	19,374,711	83%	94%
Last Name	14,332,309	97%	96%
First Name	12,525,688	97%	98%
Full Date	10,459,935	55%	77%
Location	3,158,031	89%	85%
Age	2,064,111	35%	47%
Organisation	1,078,115	36%	71%
Phone Number	1,262,313	40%	63%

Table 1: The PHI types in order of frequency as classified by the de-identification system. The per-class recall and precision for the NER model are also displayed and were calculated on the test data from Dalianis and Velupillai (2010). In total, 83,914,340 sensitive entities are found in 49,715,558 sentences.

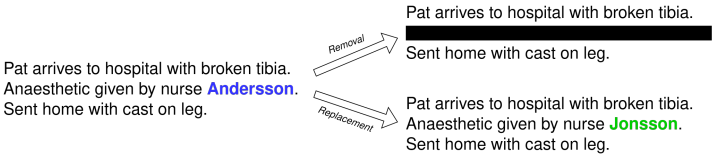

Figure 1: This hypothetical example illustrates the two approaches taken to de-identify the data. One approach *replaces* the sensitive data with realistic surrogates and is used to train the model *KB-BERT + Pseudo*. The other approach instead *removes* the entire sentence from the dataset and this filtered dataset is used to train the model *KB-BERT + Filtered*.

learn essentially the same information without exposing any sensitive information.

The second and more aggressive approach is to remove all sentences that contain sensitive entities. This approach removes 49,715,558 out of 364,385,114 sentences in the original dataset. In other words, 13.65% of all sentences were identified as containing sensitive entities. The removal of these sentences reduced the size of the dataset by approximately 19%.

Combined with the total number of entities shown in Table 1, these statistics indicate a slight tendency for sensitive entities to cluster in the same sentences, with around 1.69 entities per sensitive sentence. If this tendency holds for the entire dataset, then removing entire sentences should help remove some additional sensitive entities that the de-identifier has missed.

4.2. Training the BERT Models

The models in this paper are trained using a setup similar to Lamproudis et al. (2021), whose model is used for comparison in this study. Their model was trained using unaltered sensitive EHR data and is referred to as *KB-BERT + Real* in this paper. The two new models are built using the datasets described in Section 4.1:

KB-BERT + Pseudo The data used to train this model has had all sensitive entities (as listed in Table 1) replaced with realistic surrogates of the same entity class.

KB-BERT + Filtered This model is built using the dataset where all sentences found to contain sensitive data have been *removed*. This filtered version of the dataset is 19% smaller than the version used to train *KB-BERT + Pseudo*.

Both models were trained using *KB-BERT* (Malmsten et al., 2020) as the starting point and are the same size as $BERT_{BASE}$ (Devlin et al., 2019). As in Lamproudis et al. (2021), the vocabularies of both models are identical to that of *KB-BERT*. Pre-training was resumed for three epochs of the datasets using hyperparameters shown in Table 2.

One way in which the training of these two models differs from *KB-BERT + Real* is that our training data does not contain any document boundaries. This means that some datapoints in the training data contain two sentences from different clinical notes. In theory, this can harm the training process. As will be shown in Section 4.3, it does not seem to matter very much in practice.

4.3. Evaluating on Downstream Tasks

After training each model for three epochs, the resulting models were fine-tuned and evaluated on each of the six downstream tasks described in Section 3.2.

Table 3 displays the results of the downstream evaluation. Each model, except for *KB-BERT*, is evaluated on all three epochs, and we report the best result of the three evaluations. The best result is selected as the

Hyperparameter	Value
Max epochs	3
Batch size	256
Training sequence length	512
Mask probability	15%
Optimizer	Adam
Learning decay rate	Linear
Learning rate	1e-4
Dropout	0.1
Warm-up steps	10,000

Table 2: The hyperparameters used for continuing the pre-training with *KB-BERT* as a starting point. These hyper-parameters were used to train *KB-BERT + Real* (Lamproudis et al., 2021), *KB-BERT + Filtered*, and *KB-BERT + Pseudo*.

aim of this study is not to determine the optimal number of epochs which could vary depending on the de-identification approach.

All three models outperform the non-clinical baseline *KB-BERT* on every clinical downstream task. This is expected and indicates that the models have adapted to the language of the domain. More surprisingly, de-identification does not lead to any discernible drop in performance. In fact, *KB-BERT + Pseudo* even outperforms *KB-BERT + Real* on some tasks.

5. Discussion & Conclusions

The results in Section 3.2 show that performance on downstream tasks is not harmed by de-identifying the data used for domain adaptation of language models. This section contextualizes these findings and provides suggestions for future research.

5.1. Absence of Performance Drops

Automatic de-identification leads to a certain degree of corruption of the training data. The models used in this paper have a strong level of precision for many entity classes, as shown in Table 1. On the other hand, the evaluation indicates that around 15% of all detected locations are actually something else. The de-identification system will then either erroneously replace the word with a location name – corrupting the data – or unnecessarily discard the sentence.

Surprisingly, Table 3 indicates that this does not adversely affect the usefulness of the resulting models on the downstream tasks. *KB-BERT + Pseudo* is trained on data that is possibly corrupted due to precision issues but still performs similarly to *KB-BERT + Real*.

KB-BERT + Filtered also performs comparably to *KB-BERT + Real* even though the data is reduced to a non-trivial degree. It does, however, perform noticeably worse on the PHI NER task. This is expected since the de-identification approach aims to remove all such entities from the continued pre-training.

5.2. Reliability of the De-Identification

The NER model used in this paper is evaluated on in-domain clinical NER data. This strongly suggests that the recall and precision estimates are accurate. Nevertheless, the efficacy of the de-identification can only be assessed using the testing data. Due to the very nature of the problem, this means that the amount of sensitive information remaining in the training data can only be estimated.

However, not all entity classes are equally sensitive. Table 3 shows that our system detects and de-identifies 97% of all first and last names which are arguably the most sensitive classes. Furthermore, an attacker cannot target a specific person as they do not know if their names are among the 3% retained in the dataset.

5.3. Releasing the Models?

As explained in Section 2.1, there has been a growing interest in evaluating how susceptible pre-trained language models are to privacy attacks. While GPT-2 has been found to be very susceptible to attacks, BERT seems to be more resilient.

The performance of the de-identification system suggests that the overwhelming majority of sensitive data are removed from the training data of our models. If only 3% of all names in the data used for domain adaptation are sensitive, and the risk of exposing *any* name is less than 10% (Jagannatha et al., 2021), then the risk of exposing a *real* name is very small.

Another feature of the approach taken in this paper is that the models use a pre-trained model as their starting point. This means that any memorized names can come both from the Health Bank or the data used to train *KB-BERT*. This can be viewed as a form of *hiding in plain sight* (HIPS). Thus, an attacker who has extracted a name not only needs to determine whether or not it is a surrogate but also whether it came from a sensitive or non-sensitive data source.

BERT models have been shown to be quite resistant to training data extraction attacks (Nakamura et al., 2020; Lehman et al., 2021; Vakili and Dalianis, 2021). Furthermore, the limited susceptibility to membership inference attacks (Lehman et al., 2021; Jagannatha et al., 2021) is likely negligible when most of the data memorized by the model has been made non-sensitive through de-identification. Based on this, as well as other points made in this paper, we believe that the models can safely be shared among academic researchers. The model *KB-BERT + Pseudo* will be distributed under the name *SweDeClin-BERT*[4] once we have obtained the necessary permissions from the Swedish Ethical Review Authority.

5.4. Future Research

As noted in Section 2.2, previous research has shown that training on pseudonymized data can adversely impact model performance. In this paper, we show that

[4]This is short for **Swe**dish **De**-identified **Clin**ical **BERT**.

Model	ICD-10	PHI	Clinical Entity	Factuality	Factuality	ADE
	Classification	NER	NER	Classification	NER	Classification
KB-BERT	0.799	0.91	0.803	0.635	0.630	0.183
KB-BERT + Real	**0.833**	**0.941**	0.858	0.732	0.682	**0.199**
KB-BERT + Filtered	**0.833**	0.929	0.854	0.731	0.672	**0.199**
KB-BERT + Pseudo	0.832	**0.941**	**0.861**	**0.736**	**0.684**	0.191

Table 3: The table compares the downstream performances of each BERT model. *KB-BERT* and *KB-BERT + Real* are used as baselines. *KB-BERT* is also the starting point for the continued pre-training of all three models, as described in Section 4.2. All values are F_1-scores and the best results are bolded.

this does not seem to be a problem when pre-training for domain adaptation. However, the data used for the downstream tasks is unaltered sensitive data, and further research into the impacts of pseudonymization on task-specific training data is needed.

It could also be interesting to perform a similar experiment on English data. A natural candidate would be to use the freely available and anonymized MIMIC-III dataset (Johnson et al., 2016), though this would require replacing all the masked PHIs with realistic surrogates. This has been done by Lehman et al. (2021). On the other hand, using a non-anonymized dataset – as done in this paper – helps ensure that the results are realistic and not contingent on the quality of the surrogate selection.

Another way to avoid leaking private information is to use synthetic data. This can be generated using generative models. Generative models such as GANs[5] (Goodfellow et al., 2014) have successfully been applied to generate very realistic image data, targeting many different domains (Jetchev and Bergmann, 2017; Han et al., 2018; Brock et al., 2018).

Choi et al. (2017; Guan et al. (2018) use GANs to generate EHR data, and a more recent paper by Al Aziz et al. (2021) use generative transformer-based models to generate synthetic EHRs. None of these papers use the synthetic data to pre-train a new language model. Performance limitations are likely a barrier to generating a dataset of the scale needed for domain adaptation of a pre-trained language model.

5.5. Conclusions

This paper compares the impact of automatically de-identifying a large corpus which is used to domain-adapt Swedish BERT models. The consequences for the utility of the de-identified corpus are determined by comparing the downstream performance of the resulting BERT models with a model domain-adapted using an unaltered version of the corpus.

The results from six clinical downstream tasks show that there is no negative impact from using an automatically de-identified clinical corpus. Indeed, the results show a slight increase in performance for some tasks.

We suggest that practitioners who use clinical data for domain adaptation incorporate automatic de-identification into their workflow to decrease the risk of privacy leaks. Automatic de-identification is an easily implemented measure that reduces the risks of unintentionally memorizing sensitive information without harming utility.

[5]GAN stands for **G**enerative **A**dversarial **N**etwork.

Acknowledgements

We would like to thank Sonja Remmer for creating the *Stockholm EPR ADE ICD-10 Corpus*.

This work was partially funded by the *DataLEASH* project and by Region Stockholm through the project *Improving Prediction Models for Diagnosis and Prognosis of COVID-19 and Sepsis with NLP*.

References

Al Aziz, M. M., Ahmed, T., Faequa, T., Jiang, X., Yao, Y., and Mohammed, N. (2021). Differentially Private Medical Texts Generation Using Generative Neural Networks. *ACM Transactions on Computing for Healthcare*, 3(1):5:1–5:27, October.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Berg, H., Henriksson, A., and Dalianis, H. (2020). The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Louhi 2020, in conjunction with EMNLP 2020*, pages 1–11.

Berg, H., Henriksson, A., Fors, U., and Dalianis, H. (2021). De-identification of Clinical Text for Secondary Use : Research Issues. pages 592–599. SciTePress.

Brock, A., Donahue, J., and Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2020). Extracting Training Data from Large Language Models. *arXiv:2012.07805 [cs]*, December. arXiv: 2012.07805.

Carrell, D. S., Cronkite, D. J., Li, M. R., Nyemba, S., Malin, B. A., Aberdeen, J. S., and Hirschman, L. (2019). The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hiding in plain sight. *Journal of the American Medical Informatics Association*, 26(12):1536–1544, December.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, pages 286–305. PMLR, November. ISSN: 2640-3498.

Dalianis, H. and Velupillai, S. (2010). De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. *Journal of Biomedical Semantics*, 1(1):6, April.

Dalianis, H., Henriksson, A., Kvist, M., Velupillai, S., and Weegar, R. (2015). HEALTH BANK- A Workbench for Data Science Applications in Healthcare. *CEUR Workshop Proceedings Industry Track Workshop*, pages 1–18, 1.

Dalianis, H. (2019). Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation, September 30, 2019, Turku, Finland*, number 166, pages 16–23. Linköping University Electronic Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, December. MIT Press.

Guan, J., Li, R., Yu, S., and Zhang, X. (2018). Generation of Synthetic Electronic Medical Record Text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380, December.

Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., and Nakayama, H. (2018). GAN-based synthetic brain MR image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 734–738, April. ISSN: 1945-8452.

Henriksson, A., Kvist, M., and Dalianis, H. (2017). Prevalence estimation of protected health information in swedish clinical text. In *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, pages 216–220. IOS Press.

Jagannatha, A., Rawat, B. P. S., and Yu, H. (2021). Membership Inference Attack Susceptibility of Clinical Language Models. *arXiv:2104.08305 [cs]*, April. arXiv: 2104.08305.

Jetchev, N. and Bergmann, U. (2017). The Conditional Analogy GAN: Swapping Fashion Articles on People Images. pages 2287–2292.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May. Number: 1 Publisher: Nature Publishing Group.

Lamproudis, A., Henriksson, A., and Dalianis, H. (2021). Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 790–797, Held Online, September. INCOMA Ltd.

Lamproudis, A., Henriksson, A., and Dalianis, H. (2022a). Evaluating Pretraining Strategies for Clinical BERT Models. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*.

Lamproudis, A., Henriksson, A., and Dalianis, H. (2022b). Vocabulary modifications for domain-adaptive pretraining of clinical language models. In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies – HEALTHINF*, volume 5, pages 180–188.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682, September. arXiv: 1901.08746.

Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., and Wallace, B. (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, Online, June. Association for Computational Linguistics.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with Words at the National Library of

Sweden–Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Nakamura, Y., Hanaoka, S., Nomura, Y., Hayashi, N., Abe, O., Yada, S., Wakamiya, S., and Aramaki, E. (2020). KART: Privacy Leakage Framework of Language Models Pre-trained with Clinical Records. *arXiv:2101.00036 [cs]*, December. arXiv: 2101.00036.

Obeid, J. S., Heider, P. M., Weeda, E. R., Matuskowitz, A. J., Carr, C. M., Gagnon, K., Crawford, T., and Meystre, S. M. (2019). Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in Health technology and Informatics*, 264:283.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Remmer, S., Lamproudis, A., and Dalianis, H. (2021). Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In *Proceedings of RANLP 2021: Recent Advances in Natural Language Processing, RANLP 2021, 1-3 Sept 2021, Varna, Bulgaria*, pages 1158–1166.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, May. ISSN: 2375-1207.

Skeppstedt, M., Kvist, M., Nilsson, G. H., and Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158.

Sweeney, L. (1996). (replacing personally-identifying information in medical records, the scrub system). In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association.

Vakili, T. and Dalianis, H. (2021). Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In *Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021)*, volume 3068. CEUR Workshop Proceedings.

Vakili, T. and Dalianis, H. (2022). Utility Preservation of Clinical Text After De-Identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing, BioNLP@ACL 2022*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008.

Velupillai, S., Dalianis, H., and Kvist, M. (2011). Factuality levels of diagnoses in Swedish clinical text. In *User Centred Networked Health Care*, pages 559–563. IOS Press.

Velupillai, S. (2011). Automatic classification of factuality levels: A case study on Swedish diagnoses and the impact of local context. In *Fourth International Symposium on Languages in Biology and Medicine, LBM 2011*.

PAPER V

END-TO-END PSEUDONYMIZATION OF FINE-TUNED CLINICAL BERT MODELS

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2024. End-to-end pseudonymization of fine-tuned clinical BERT models. *BMC Medical Informatics and Decision Making*, 24(1):162.

Author Contributions Thomas Vakili performed all the experiments and wrote the main parts of the article. Aron Henriksson gave suggestions on the experiment setup, contributed to the article, and commented on all parts. Hercules Dalianis gave suggestions on the experiment setup, contributed to the article, and commented on all parts.

Vakili *et al.*
BMC Medical Informatics and Decision Making (2024) 24:162
https://doi.org/10.1186/s12911-024-02546-8

BMC Medical Informatics and Decision Making

RESEARCH Open Access

End-to-end pseudonymization of fine-tuned clinical BERT models

Privacy preservation with maintained data utility

Thomas Vakili[1*], Aron Henriksson[1] and Hercules Dalianis[1]

Abstract

Many state-of-the-art results in natural language processing (NLP) rely on large pre-trained language models (PLMs). These models consist of large amounts of parameters that are tuned using vast amounts of training data. These factors cause the models to memorize parts of their training data, making them vulnerable to various privacy attacks. This is cause for concern, especially when these models are applied in the clinical domain, where data are very sensitive.

Training data pseudonymization is a privacy-preserving technique that aims to mitigate these problems. This technique automatically identifies and replaces sensitive entities with realistic but non-sensitive surrogates. Pseudonymization has yielded promising results in previous studies. However, no previous study has applied pseudonymization to both the pre-training data of PLMs and the fine-tuning data used to solve clinical NLP tasks.

This study evaluates the effects on the predictive performance of end-to-end pseudonymization of Swedish clinical BERT models fine-tuned for five clinical NLP tasks. A large number of statistical tests are performed, revealing minimal harm to performance when using pseudonymized fine-tuning data. The results also find no deterioration from end-to-end pseudonymization of pre-training and fine-tuning data. These results demonstrate that pseudonymizing training data to reduce privacy risks can be done without harming data utility for training PLMs.

Keywords Natural language processing, Language models, BERT, Electronic health records, Clinical text, De-identification, Pseudonymization, Privacy preservation, Swedish

Introduction

The popularization of the transformer architecture [1] in the past few years has led to rapid advances in natural language processing (NLP). Many benchmarks are now dominated by pre-trained language models (PLMs) that learn to model language using unlabeled corpora. There are many PLM architectures, and this article focuses on the BERT architecture [2], which is widely used and competitive in many NLP benchmarks. PLMs typically consist of hundreds of millions, even billions, of parameters which are trained on enormous amounts of unlabeled training data. The sizes of the corpora used to pre-train these models are typically in the range of tens of gigabytes or even terabytes of data. The BERT models used in this study consist of over 100 million parameters and are pre-trained on around 6 billion tokens [2, 3]. On the other end of the scale, the largest publicly available version of Llama 2 consists of 70 billion parameters tuned using a corpus spanning 2 trillion tokens [4].

PLMs have shown great promise in several NLP domains, and the clinical domain is no exception. State-of-the-art results in clinical NLP tend to rely on PLMs,

*Correspondence:
Thomas Vakili
thomas.vakili@dsv.su.se
[1] Department of Computer and Systems Sciences, Stockholm University, P.O. Box 7003, 164 07 Kista, Stockholm, Sweden

e.g., for temporal relation extraction [5], text similarity [6], concept normalization [7], adverse drug event extraction [8], medication event extraction [9] and information extraction [10]. However, while PLMs are generally pre-trained using readily available corpora in the general domain – e.g., Wikipedia and other data sources on the Internet – research suggests that using generic PLMs in highly specialized domains such as healthcare may be suboptimal due to significant domain differences [11, 12], even in the presence of large language models like T5-XL and GPT-3 [13]. This has motivated efforts to develop domain-specific clinical PLMs. There are different approaches to developing domain-specific PLMs [14], including pre-training a new language model from scratch with in-domain data, e.g., in the form of clinical text from electronic health records (EHRs). An alternative approach is to adapt an existing, generic PLM to the target domain by continuing to pre-train it with in-domain data. The vocabulary of the model can be retained or adapted to account for domain differences. This continued pre-training is known as *domain-adaptive pre-training* [15–17].

While PLMs have shown great promise in solving important NLP problems, their reliance on increasingly large numbers of parameters and vast corpora causes models to memorize parts of their training data [18–20]. This tendency is undesirable in many use cases but also has important implications for privacy. When models are domain-adapted using clinical data, these privacy risks must be mitigated. Clinical data often describes sensitive information that must be protected, not just for ethical reasons but also due to current regulations.

One way to reduce the privacy risks of using clinical data is to remove sensitive information. An important technique for doing so is called *pseudonymization*, which involves locating sensitive passages using named entity recognition (NER) and substituting them with realistic surrogate data. This technique has been applied to data for pre-training language models [21, 22] and for fine-tuning models [20, 23], with successful results. However, previous research has only studied these two training steps in isolation.

In this paper, we demonstrate the first example of a clinical language model that has been *fully pseudonymized* in both the domain-adaptive pre-training and fine-tuning steps. This is done by:

- Pseudonymizing datasets for five clinical downstream tasks.
- Fine-tuning and evaluating a total of 300 models through 10-fold cross-validation of 30 different combinations of pseudonymized data and models.
- Comparing all models in terms of F_1 to determine if any statistically significant differences in predictive performance exist.

The results show that end-to-end pseudonymization can be successfully applied to the pre-training and fine-tuning of language models. We find that end-to-end pseudonymization preserves privacy and simultaneously retains the utility of the data for domain-adaptive pre-training and fine-tuning of PLMs.

Background

This study focuses on mitigating the privacy issues of modern transformer models in NLP using pseudonymization. This section gives a more detailed motivation for how these models are vulnerable to privacy attacks and why pseudonymization is a good privacy-preserving technique. Other privacy-preserving techniques are discussed, and previous works on pseudonymization are presented to provide the context in which this study is situated.

Privacy attacks

As mentioned in the introduction, large language models can be susceptible to privacy attacks. This susceptibility is partially due to the self-supervised pre-training objectives that tend to involve reconstructing a noisy or obscured version of the training data. For example, BERT models are pre-trained using *masked language modeling* [2], which involves reconstructing a version of the training data in which some tokens have been replaced with `[MASK]` tokens. The pre-training is then performed using large text corpora with unknown quantities of sensitive information, and the learned features are encoded in millions or billions of parameters.

Privacy attacks targeting PLMs aim to extract information about their training corpora. The attacks do so by targeting the information encoded in the parameters of the models. Depending on the objective, these attacks can be categorized into two main classes. *Training data extraction attacks* aim to reconstruct data used to train a model. This is a severe form of attack since it can result in the disclosure of sensitive information about persons described in the training data of a model. In the clinical domain, this could mean exposing the details of a patient's medical history. Training data extraction attacks require an effective algorithm for sampling information from a model; however, such algorithms are not (yet) described for all models [24–27]. Nevertheless, there are examples of successful training data extraction attacks targeting generative systems such as GPT-2 and ChatGPT [19, 28].

Membership inference attacks aim to discern whether or not a particular datapoint has been used to train a target model [29]. In a clinical context, this information could reveal if a patient associated with an EHR has visited a set of clinical units associated with particular health problems. This category of attacks typically involves measuring the *reaction* of a model to a set of datapoints and using this information to distinguish between members and non-members of the training data [30, 31]. Successful membership inference attacks may pose a privacy threat in themselves, but are also often used as a building block in training data extraction attacks when determining whether the algorithm has extracted a real or spurious datapoint.

Privacy-preserving techniques

Several privacy-preserving techniques have been developed to mitigate the privacy threats described in the previous section. In this section, a non-exhaustive list of techniques will be described to provide context for why this study focuses on the pseudonymization of training data. Other promising and oft-mentioned techniques include differential privacy, homomorphic encryption, and synthetic training data.

Differential privacy is a notion of privacy that was originally designed for database records. The idea is that, given a datapoint d and two datasets D and D' differing only in that $d \in D$ while $d \notin D'$, the output of any aggregation of these datasets should be close to indistinguishable [32]. As it is typically formulated, we have (ϵ, δ) -differential privacy [33] for an aggregation M with range R if

$$P[M(D) \in R] \leq e^{\epsilon} P[M(D') \in R] + \delta.$$

Differential privacy has also been adapted for deep learning. The DP-SGD algorithm [34] is a differentially private version of the stochastic gradient descent algorithm commonly used to train neural networks. While differentially private learning has the advantage of having a formal mathematical definition, the ϵ and δ parameters can be difficult to choose and interpret. This issue is compounded by the fact that effective differential privacy typically works by adding noise to the aggregation (e.g., the training algorithm), which may hinder efficient training [35]. Furthermore, differential privacy was originally designed for database records, and some have argued that it is ill-suited to the unstructured nature of natural language [36].

In contrast to differential privacy, homomorphic encryption aims to protect the result of an input X and its output $M(X \mid D)$ rather than D (e.g., the data used to train a machine learning model M) itself. This is achieved by implementing M using operations that handle encrypted data, meaning that both X and $M(X \mid D)$ are knowable only to the person *using* the model [29]. Homomorphic encryption allows users to use a model owned by another party safely. The technique enables private inferences that do not disclose any information about the data to the owner of the model nor to any potential eavesdropper. However, it does not protect the *owner* of the model from attacks such as membership inference attacks or training data extraction attacks since the output of the inference is made available to the user initiating the inference.

With the growing availability of models capable of high-quality natural language generation, some have considered creating synthetic training data. This data, being synthetic, is assumed to be non-sensitive. By synthesizing data, the use of sensitive clinical data can be reduced [37] or done away with entirely [38, 39]. Synthetic data has been used in several studies to train well-performing fine-tuned clinical NLP models while limiting the risk of exposing private information from the original data [37–39]. There are fewer examples of models *pre-trained* using synthetic data. This is likely due to, at least in part, the computational costs of operating the large language models required to produce enough high-quality synthetic text. However, the example of GatorTronS [40] shows that this approach is indeed possible and that models pre-trained on synthetic text can perform well. On the other hand, the extent to which a synthetic text itself may contain sensitive data is poorly understood. The risk that the synthesizing language model accidentally generates parts of its own training data cannot be ruled out.

Automatic de-identification and pseudonymization

Many of the aforementioned privacy-preserving techniques are not specific to natural language data. Differential privacy, for example, was originally designed for database-style structured data where each row is to be protected. Unstructured natural language data stands out as a particularly high-dimensional data form. In contrast to structured database rows, it can be difficult to exhaustively specify all of the information contained in an EHR. On the other hand, another feature of textual data is that many words or phrases can be replaced with similar information without changing the overarching meaning of a text. Examples of this phenomenon are synonyms which, broadly speaking, are interchangeable words that have the same meaning.

Automatic de-identification typically relies on NER to remove sensitive entities, such as data constituting personally identifiable information (PII). These entities usually cover direct identifiers such as names, but also cover *quasi-identifiers* such as locations, ages, and dates. Quasi-identifiers are PII that do not directly identify a

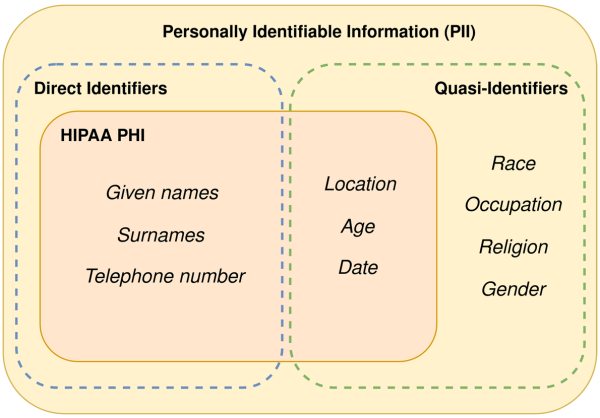

Fig. 1 The HIPAA regulation in the United States lists 18 types of PII, called Protected Health Information (PHI), that should be removed for privacy reasons. These cover most of the PII types that are typically considered to be direct identifiers. However, as the figure illustrates, there are many quasi-identifiers that are not covered by this PHI definition

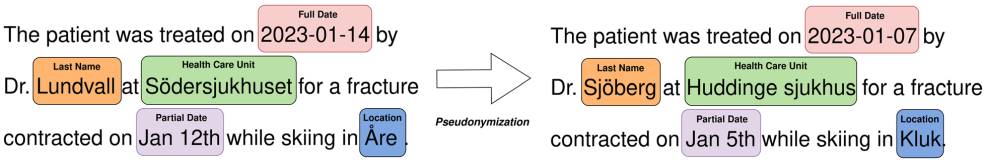

Fig. 2 The pseudonymizers used in this study replace detected sensitive entities with realistic surrogates. The figure illustrates some of the entities considered by the system. The surrogate values are selected to preserve as much information as possible. However, an adversary with knowledge of Swedish geography would realize that, in this example, Kluk is an unlikely place to go skiing

person, but that may do so when combined with other quasi-identifiers or with auxiliary information. A commonly used set of PII is the collection of entities designated as personal health information (PHI) by the HIPAA regulation [41] in the United States. Examples of PII, PHI, and how they relate to different types of identifiers can be found in Fig. 1. In this article, we use the broader term PII. However, the set of PII covered by the de-identifiers is based on the PHI described by the HIPAA regulation [42].

De-identification is typically done in two main steps. First, the NER model of the de-identifier is used to detect entities that are PII. Next, these are *sanitized* in some way. Examples of sanitization techniques include replacing entities with their class name, masking them with a nondescript placeholder, and replacing them with surrogate values. This study focuses on the last strategy—*pseudonymization*—which replaces sensitive entities with realistic replacement values of the same entity type. These should preferably be chosen cleverly to preserve as much semantic information as possible without harming privacy. An example of how this process can work is illustrated in Fig. 2.

The goal of pseudonymization is to remove the PII most likely to be used to re-identify individuals. However, it is important to recognize that pseudonymizers are never perfect. The NER models that power them often have imperfect recall and precision. Imperfect recall is a privacy issue since low recall implies that the model will miss sensitive entities that should be sanitized. On the other hand, low precision will result in many non-sensitive entities being replaced with inappropriate values. In the worst case, poor precision can lead to task-relevant words being replaced with irrelevant information, corrupting the datapoint and potentially having a negative impact on data utility. Both the low-recall and low-precision scenarios are illustrated in Fig. 3.

Pseudonymization is related to but different from *anonymization*. Although the terms are sometimes used interchangeably in the literature, anonymization

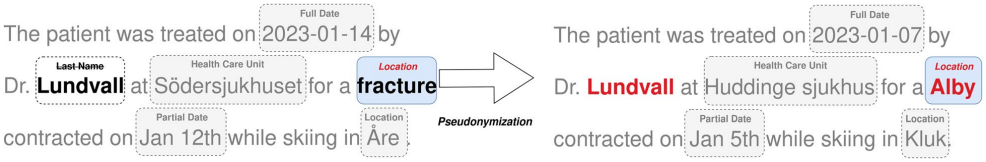

Fig. 3 The NER models that power pseudonymizers are never perfect. When recall is insufficient, they will miss names such as *Lundvall*, which will remain exposed in the text. When there are problems with precision, non-sensitive words will be changed to irrelevant replacement values. In the worst case, a clinically relevant term like *fracture* may be replaced with a surrogate PII entity that harms data utility

is typically associated with stronger privacy guarantees. For example, when the term is used in the GDPR[1] it is often understood as implying complete and irreversible removal of any information that can be used to partially or fully identify an individual [44]. Pseudonymization, as understood in this study, does not fulfill this stricter requirement. Rather, it is a process that *enhances* the privacy of data.

In contrast to other techniques within the field of privacy-preserving machine learning, pseudonymization is a text-specific technique for privacy preservation that harnesses the particular characteristics of natural language. When successfully applied, pseudonymization preserves the overall semantics of a datapoint while removing sensitive information. This scenario increases the privacy of a dataset while preserving its utility. However, when precision is not high enough, erroneous classifications and subsequent replacements will lead to a corruption of the data. The aim of this study is to demonstrate that, with a reasonably strong NER model, this does not happen often enough to harm the utility of the data for pre-training or fine-tuning clinical BERT models.

Utility for machine learning using sanitized text

An early study on using pseudonymized EHRs is described by Yeniterzi et al. [45]. The authors trained NER models for detecting PII using both the pseudonymized and the original data. They found that the results deteriorated significantly when training on pseudonymized data and evaluating on unaltered text, with the F_1 score falling from 0.862 to 0.728.

Lothritz et al. [23] study the impact of de-identification on a wide range of general-domain datasets. They employ a variety of sanitization strategies, including two pseudonymization strategies of different sophistication. They evaluate these strategies using ERNIE [46] and BERT models on eight different downstream tasks. Their results show that de-identification harms the utility of their datasets, but that this harm was small. The results also show that pseudonymization yields the strongest performance among the considered sanitization strategies.

Another study using sanitized text for machine learning is described by Berg et al. [47]. The authors pseudonymized Swedish clinical texts and then used them to train two different machine learning algorithms to detect PII. These algorithms were then evaluated on real Swedish clinical text data. The study aimed to enable sanitized training data to be transferred between hospitals for performing de-identification tasks. The authors tried two machine learning algorithms: conditional random fields (CRF) and long short-term memory (LSTM) networks. CRF gave the best results on training on sanitized text and de-identifying real clinical text; however, the performance on identifying several PII classes deteriorated, with the overall recall decreasing from 85% to 50%. This effect was primarily observed for the PII classes *Location*, *Health Care Units* and *Full Date*.

Berg et al. conducted another study [48] using four different strategies to sanitize the training data for downstream tasks, where models with different levels of recall were used to sanitize a set of Swedish datasets for clinical NER. Using a model with high recall is a good strategy in terms of privacy since it will identify more sensitive entities. However, these benefits may come at the expense of lower precision and more false positives. The study evaluated four different strategies for sanitizing the datasets: pseudonymization, masking the sensitive entities, replacing them with their class name, and removing the entire sentences in which sensitive entities were detected. The impact of sanitizing the data was evaluated by training CRF models for three clinical NER tasks using different sanitized datasets. Overall, the pseudonymization strategy had the smallest negative impact on the downstream tasks, while the sentence removal strategy resulted in a larger performance deterioration.

The overlap between PII and clinical entities is a source of potential harm to utility and has been thoroughly investigated by Berg et al. [48]. It was found that only one percent of clinical entities were affected by the

[1] The GDPR is the General Data Protection Regulation of that is applied throughout the European Union [43].

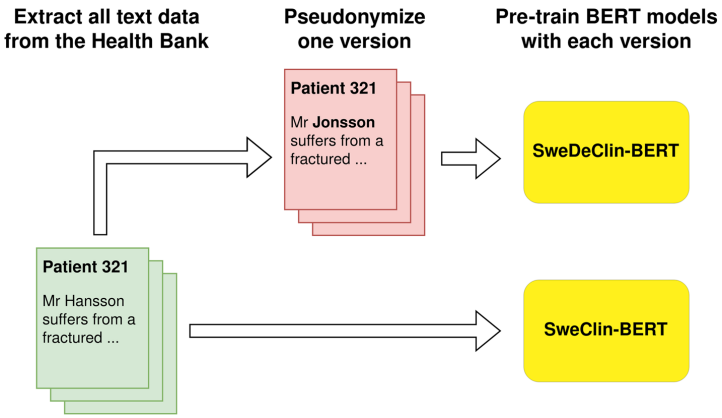

Fig. 4 This study uses two different clinical BERT models created in earlier studies. SweClin-BERT is trained on a sensitive version of the Health Bank corpus [51], whereas SweDeClin-BERT is trained on a version that has been automatically pseudonymized [22]. Both models are initialized with the weights of KB-BERT [52]

de-identification process. The worst affected PII classes were *Health Care Unit* and *Person* (first and last name), which tended to overlap with the clinical entities *drug, body part, disorder* and *finding*. A later study [49] indicated that the risk of misclassifying eponyms (e.g., diseases like *Alzheimer disease* that are named after medical doctors) is lower when using BERT-based PII classifiers compared to earlier approaches. However, clinical entities are diverse, and there are other cases where misclassifications could be an issue.

Vakili et al. [22] evaluated the impact of pre-training BERT models using de-identified and unaltered data. Two sanitizing strategies were used: pseudonymization and sentence removal. Two models were adapted to the clinical domain by pre-training using clinical data sanitized with each strategy. The resulting models were then evaluated on six downstream tasks. The results showed no negative impact from pre-training using de-identified data compared to using unaltered data. Similarly, Vakili & Dalianis [20] evaluated the impact of fine-tuning a clinical BERT model using pseudonymized or unaltered datasets. They evaluated their approach using three downstream tasks, again finding no significant difference between training on unaltered or pseudonymized data. This study further builds upon the previous studies and provides deeper examinations of the interactions between pseudonymization and data utility. Furthermore, we demonstrate that pseudonymization can be applied both to the pre-training *and* fine-tuning data without harming the performance on clinical NLP tasks.

Methods and materials

This study relies on a large number of datasets and models, mainly created using data from the Swedish Health Record Research Bank (Health Bank)[2]. The original data were collected from the Karolinska University Hospital [50] and consist of a large number of Swedish EHRs[3]. This section describes the data and models used in the experiments, and how these experiments were carried out.

Clinical BERT models

This study examines the impact of pseudonymization applied to data for domain-adaptive pre-training and fine-tuning BERT models. As illustrated in Fig. 4, two different PLMs are used. One—*SweDeClin-BERT*—that has been trained using pseudonymized pre-training data [22], and another model—*SweClin-BERT*—that was trained on the unaltered version of the same dataset [51]. Both models were initialized using weights from the Swedish general-domain KB-BERT model [52] and were adapted to the clinical domain by pre-training for three epochs over the Health Bank corpus. Figure B1 in Appendix B contains a diagram showing how the models relate to other parts of the Health Bank.

The Health Bank corpus used for domain-adaptive pre-training consists of approximately 2.8 billion words

[2] http://www.dsv.su.se/healthbank

[3] This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

Table 1 The five tasks were based on four different clinical corpora from the Health Bank. This table lists the size of each corpus in terms of the number of documents and tokens. The table also specifies the number of possible classes and whether the tasks are document-level or token-level classification tasks

Corpus	Documents	Tokens	Classes	Level
ICD-10	6,062	930,550	10	Document
ADE	21,725	931,778	2	Document
Factuality	3,710	102,223	6	Document
Factuality NER	3,822	286,205	6	Token
Clinical NER	3,120	178,672	4	Token

which is comparable to the 3.3 billion words used to train KB-BERT [3]. We did not pre-train for more than three epochs for reasons of resource efficiency. This is justified by prior work using the same data [53] showing that longer pre-training was unnecessary when starting from a general-domain model.

Five clinical downstream tasks

The utility of the models and datasets after and before pseudonymization was assessed using five clinical NLP tasks. The five tasks are based on corpora from the Health Bank infrastructure and are summarized[4] in Table 1 and described in this section. The utility of each pseudonymization configuration was examined by measuring the performance of models fine-tuned on these tasks. Below is a list of the datasets as well as the abbreviated names used in Table 1 and other tables in the paper.

Stockholm EPR Gastro ICD-10 Corpus I (ICD-10)

The Gastro ICD-10 dataset consists of gastro-related discharge summaries and their assigned ICD-10 diagnosis codes. The discharge summaries relate to 4,985 unique patients. The ICD-10 codes are divided into 10 groups corresponding to different body parts; the ICD-10 codes range from K00 to K99. Each group contains several codes [55].

Stockholm EPR Clinical Entity Corpus (Clinical NER)

A clinical entity dataset encompassing 157,123 tokens and 20,675 annotated entities assigned to four clinical entity classes *Diagnosis*, *Findings*, *Body parts*, and *Drugs* [56]. The goal of the task is to identify and correctly label the clinical entities.

Stockholm EPR Diagnosis Factuality Corpus (Factuality NER)

A factuality diagnosis dataset specifying six levels of confidence regarding the factuality of a diagnosis. The dataset encompasses 6,865 annotated entities[5] labeled as *Certainly Positive*, *Probably Positive*, *Possibly Positive*, *Possibly Negative*, *Probably Negative*, or *Certainly Negative* [57, 58]. The task consists of identifying tokens in the corpus specifying diagnoses and assigning them a factuality label.

Stockholm EPR Diagnosis Factuality Corpus (Factuality)

A dataset which is a variation of the *Stockholm EPR Diagnosis Factuality NER Corpus* that instead assigns a factuality level to the entire document. The classification task is a multi-label classification problem where the model needs to predict the factuality of each document. The labels are the same as in the NER version of the task.

Stockholm EPR ADE ICD-10 Corpus (ADE)

The ADE corpus contains 21,725 discharge summaries describing adverse drug events (ADEs). The task is a binary classification task, where positive samples have been assigned an ICD-10 code that denotes an ADE. Negative text samples in each group have been assigned an ICD-10 code describing a diagnosis that is not drug-induced. The task is to determine whether the diagnosis defined by the ICD-10 code was induced by an ADE or not [22].

Pseudonymization

The pseudonymization performed in this study relies on NER to locate sensitive entities that should be replaced. Two such NER models are used. Both are based on BERT and are fine-tuned on the Stockholm EPR PHI Corpus [42]. This corpus contains 380,000 tokens and 4,480 manually annotated entities in nine classes based on the American HIPAA regulation. One model *pseudo+* uses a non-pseudonymized Swedish clinical BERT model [59] and another, slightly weaker model called *pseudo* is based on SweDeClin-BERT [22]. Tables 2 and 3 list the per-class performance of both NER models as measured using the test splits of their training data. Figure B2 in Appendix B shows how these models relate to other parts of the Health Bank.

Two pseudonymized versions of each dataset described in the previous section were created, one for each NER model. Sensitive entities were detected and then replaced with realistic surrogate values based on the method

[4] The number of tokens was calculated using the Punkt tokenizer for Swedish in NLTK [54].

[5] The dataset also contains 199 entities annotated for purposes irrelevant to our experiments. These annotations were ignored.

Table 2 The recall and precision of the *pseudo+* model for each PII type are displayed. The model is a clinical BERT model [59] that has been fine-tuned and evaluated using the *Stockholm EPR PHI Corpus* [42]

PII Class	Recall	Precision
Age	100%	100%
First Name	100%	100%
Last Name	98%	98%
Partial Date	99%	97%
Full Date	90%	91%
Phone Number	81%	68%
Health Care Unit	85%	94%
Location	100%	100%
Organization	71%	100%

Table 3 The recall and precision of the *pseudo* model for each PII type are displayed. The model is based on the pseudonymized SweDeClin-BERT model and has been fine-tuned and evaluated using the *Stockholm EPR PHI Corpus* [42]

PII Class	Recall	Precision
Age	100%	100%
First Name	97%	98%
Last Name	96%	97%
Partial Date	99%	98%
Full Date	87%	91%
Phone Number	93%	89%
Health Care Unit	89%	88%
Location	89%	81%
Organization	29%	80%

Table 4 Sensitive entities detected by the *pseudo* model

PII Class	Factuality NER	Clinical NER	ICD-10	Factuality	ADE
Age	1,392	1,149	3,060	1,353	2,995
First Name	528	274	1,185	510	3,965
Last Name	1,105	274	1,829	1,062	4,257
Partial Date	681	554	11,371	644	4,305
Full Date	128	137	18,875	125	22,296
Phone Number	148	45	141	142	460
Health Care Unit	3,554	2,005	3,365	3,406	7,635
Location	110	78	510	105	689
Organization	5	1	37	4	59
Total words	*253,124*	*191,202*	*798,120*	*239,722*	*788,930*

Table 5 Sensitive entities detected by the *pseudo+* model

PII Class	Factuality NER	Clinical NER	ICD-10	Factuality	ADE
Age	955	764	2,565	929	2,257
First Name	523	283	1,378	506	3,884
Last Name	1,055	707	1,904	1,016	4,064
Partial Date	369	316	5,740	355	2,995
Full Date	110	121	12,703	107	17,552
Phone Number	118	39	75	113	172
Health Care Unit	4,285	2,282	12,654	4,117	9,751
Location	182	102	1,217	172	985
Organization	4	12	6	1	66
Total words	*253,124*	*191,202*	*798,120*	*239,722*	*788,930*

described in this section. The number of sensitive entities detected by the pseudonymizers is displayed in Tables 4 and 5. These numbers include both false and true positives and indicate the degree to which the data were altered in the pseudonymization process.

An overview of the algorithm for surrogate selection is available in Dalianis [60], which describes the first version of the pseudonymizer. The system has been further refined since its initial conception. One adaption made from the original pseudonymizer is that the name lists used to replace first and last names have been expanded to include a wider range of names. The original system only considered the most common Swedish names, while the current system chooses from 244,000 first names and 34,000 surnames. However, a limitation of the pseudonymizer is that it lacks functionality for replacing organizations. As shown in Tables 4 and 5, organizations are very infrequent, meaning that the privacy and performance implications are limited.

The pseudonymizer created by Dalianis [60] replaces many entities using word lists. For example, a gendered name is replaced with another name typically associated with the same gender, and a gender-neutral name is replaced with a gender-neutral name. Streets and places in Stockholm randomly with other streets in Stockholm. Similarly, other locations in Sweden are replaced with locations in the same county, and similar logic exists to replace country names with names of countries in the same continent. Health care units are changed to other health care units using a list of known clinics. Other entities are changed using rules. Postal codes are replaced with more common postal codes with large

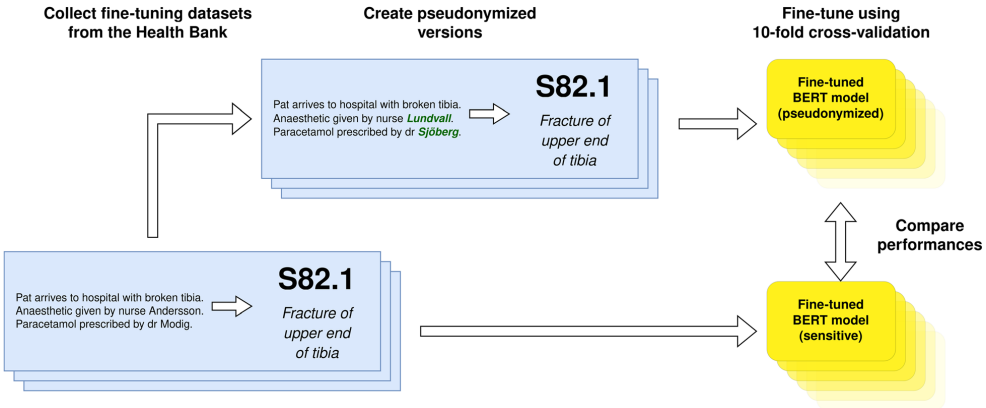

Fig. 5 Every dataset described in the "Five clinical downstream tasks" section was pseudonymized using both the *pseudo* and *pseudo+* de-identifiers. SweDeClin-BERT and SweClin-BERT were fine-tuned using the non-pseudonymized and the two pseudonymized versions of the datasets. All models were compared based on the F_1 scores aggregated from the 10-fold cross-validation of each model

populations. Dates are shifted one or two weeks earlier or later. Years and ages are handled similarly and are increased or decreased by a small and random number of years. Phone numbers are changed to other phone numbers according to the formatting rules for Swedish phone numbers.

Evaluating the impact of pseudonymization

As previously discussed, pseudonymization often entails a certain degree of data corruption. The main experiment in this study examines this effect on the downstream performance of clinical BERT models pre-trained and fine-tuned on pseudonymized clinical training data.

Once the datasets for the five clinical downstream tasks had been pseudonymized, a series of evaluations were carried out. Each version of every dataset was used to fine-tune and test both BERT models using 10-fold cross-validation [61], as illustrated in Fig. 5. Since the pseudonymization procedure is a deterministic pre-processing step, the pseudonymized models are tested on pseudonymized folds. The repeated training and evaluation using different splits resulted in a range of evaluation metrics used to estimate the mean and standard deviation of each configuration. The configurations were compared based on their F_1 scores[6] [63]. All fine-tuning configurations ran for a maximum of 10 epochs, with early stopping implemented to avoid overfitting and unnecessary computations.

In total, 30 different combinations of models and datasets were evaluated using 10-fold cross-validation. For every downstream task, we compare the difference in the performance of all combinations of models and pseudonymization approaches. The difference between each pair was tested for statistical significance using a *Mann-Whitney U test*[7] [64, 65] by comparing the F_1 scores of every fold in both models' 10-fold cross-validations.

Results

The 30 different model-dataset configurations combined with the 10-fold cross-validation resulted in 300 fine-tuned models. The evaluations of these models were used to produce F_1 for each configuration and downstream task. The means and standard deviations of each evaluation are listed in Table 6. From studying the columns of the table, it is apparent that most of the values are within a standard deviation of each other.

Comparing every configuration within each downstream task resulted in 150 Mann-Whitney U tests being performed. Out of these, 126 comparisons showed no statistically significant difference for $p < 0.05$. The remaining 24 comparisons showed varying degrees of statistical significance. To facilitate a focused analysis of the results, a curated sample of the significant results is listed in Table 7. These are limited to the cases where using real data outperformed using pseudonymized data, as these examples challenge the

[6] Metrics for the token classification tasks were calculated using the `seqeval` library [62] in `strict` mode.

[7] This test is also sometimes referred to as a Wilcoxon rank-sum test.

Table 6 The table compares the performance of each combination of models and datasets. The scores are the mean F_1 scores together with their standard deviation based on the results from the 10 folds. **P** stands for pre-training data and **F** for fine-tuning data. A ✗ denotes that no pseudonymization was done, a ✓ that it was done using the *pseudo* model and a + means that pseudonymization was performed using the *pseudo*+ model

Pseudonymized		Factuality	Clinical Entity	ICD-10	Factuality	ADE
P	F	NER	NER	Classification	Classification	Classification
✗	✗	*0.686±0.013*	*0.851±0.012*	*0.821±0.012*	*0.729±0.020*	*0.186±0.009*
✗	✓	0.639±0.038	0.843±0.011	0.810±0.011	0.725±0.021	0.190±0.017
✗	+	0.668±0.024	0.841±0.011	0.814±0.008	0.726±0.018	0.188±0.014
✓	✗	0.696±0.019	0.861±0.011	0.835±0.010	0.726±0.025	0.188±0.011
✓	✓	0.663±0.048	0.856±0.009	0.825±0.010	0.716±0.016	0.198±0.013
✓	+	0.695±0.013	0.853±0.011	0.832±0.007	0.733±0.022	0.205±0.018

Table 7 Out of 24 statistically significant results, 11 are cases where using non-pseudonymized data yields better results than using pseudonymized data. All of these find this effect with regard to the fine-tuning data. The *p-value* is the result of the Mann-Whitney U test for determining if the *Weaker model* performs worse than the *Stronger model*. For each model, **P** indicates whether the pre-training data was pseudonymized, and **F** indicates if the fine-tuning data was pseudonymized. Again, a ✗ denotes that no pseudonymization was done, a ✓ that it was done using the *pseudo* model and a + means that pseudonymization was performed using the *pseudo*+ model

Row	Task	Weaker model		Stronger model		*p*-value
		P	F	P	F	
(1)	ICD-10	✗	✓	✗	✗	0.0378
(2)	Factuality NER	✗	✓	✗	✗	0.0014
(3)	Clinical NER	✗	+	✗	✗	0.0269
(4)	ICD-10	✗	✓	✓	✗	0.0007
(5)	Clinical NER	✗	✓	✓	✗	0.0029
(6)	Factuality NER	✗	✓	✓	✗	0.0005
(7)	ICD-10	✗	+	✓	✗	0.0011
(8)	Clinical NER	✗	+	✓	✗	0.0022
(9)	Factuality NER	✗	+	✓	✗	0.0156
(10)	ICD-10	✗	✗	✓	✗	0.0086
(11)	Clinical NER	✗	✗	✓	✗	0.0226

main hypothesis of the study. The full list of the 24 statistically significant differences is provided in Table A1 of Appendix A.

Notably, none of the statistically significant differences were cases where SweClin-BERT outperformed SweDeClin-BERT. This is apparent from the *P* column for the *Weaker model* only containing ✗'s. This implies that SweDeClin-BERT is a stronger model for the downstream tasks in this study. In that case, this difference in general model performance explains rows 4–11 of Table 7. Furthermore, there are no examples where training SweDeClin-BERT on different forms of fine-tuning data yielded statistically significant differences in performance.

The first three rows in Table 7 show that training SweClin-BERT using non-pseudonymized data sometimes yields statistically significant improvements compared to using pseudonymized data. This is found for one pair of configurations for three of the tasks. There are no examples where training on real data outperforms both forms of pseudonymized data. For example, the first row finds a statistically significant improvement from using real ICD-10 data rather than data pseudonymized using the *pseudo* model, but no significant difference is found if the *pseudo*+ model is used.

Discussion

The previous section presents several interesting findings. In this section, the results of the study are analyzed and contextualized. We also provide ideas for future work and discuss the limitations of our study.

Interpreting the significant results

The results of this study are based on a large number of Whitney-Mann U tests. When performing 150 statistical tests, there is a non-trivial risk of finding spurious statistical differences. The standard cut-off of $p < 0.05$ still risks finding differences by chance 1 out of 20 times. Nevertheless, there are some trends in Table 7 that are interesting to discuss.

First, it is notable that none of the statistically significant comparisons find that pre-training with real data outperforms pre-training with pseudonymized data. A similar result was indicated in a previous study by Vakili et al. [22]. However, it is important to note that only two pre-trained models were compared in this study. While the results strongly suggest that SweDeClin-BERT is better than SweClin-BERT, this does not mean that pre-training with pseudonymized data is better *in general*. Examining this would require pre-training many more BERT models with and without pseudonymizing the data. It would likely also require comparing pre-trained models initialized from random weights rather than the weights of a general-domain model. While this could be interesting to study, it is beyond the computational constraints imposed by the scope of this study.

Some of the statistically significant results in Table 7 do indicate that fine-tuning a non-pseudonymized model using unaltered data can yield stronger results than fine-tuning with pseudonymized data. However, this is only found for three of the five downstream tasks. Furthermore, none of these results hold for *both* of the pseudonymizers. The results in Table 6 also show that these examples are still within a standard deviation of each other. The results where fine-tuning on real data *does* outperform using both pseudonymized data (such as rows 4 and 7 of Table 7) are results where SweDeClin-BERT outperforms SweClin-BERT. Thus, these cases are better explained by the overall stronger results of SweDeClin-BERT. Crucially, for the purposes of this study, there are *no* examples of statistically significant differences where a model trained using end-to-end pseudonymization is outperformed by a non-pseudonymized version. The hypothesis of this study holds since we find no evidence of any significant deterioration from pre-training and fine-tuning using automatically pseudonymized data.

Quantifying privacy benefits

An important limitation of this study is that the privacy benefits of pseudonymization are only quantified in terms of the number of removed sensitive entities. This assumes that the sensitivity of the training data directly corresponds to the sensitivity of the model. This assumption may be pessimistic since it is unlikely that the trained model will memorize all remaining sensitive entities. On the other hand, relying on metrics such as recall and precision also obscures any particularities in the *specific* entities that are missed and if these could be more at risk of memorization.

Previous research has suggested that *membership inference attacks* can be used for estimating the degree of memorization in a model [30, 31, 66]. This approach can be effective for some privacy-preserving techniques, such as differentially private learning [34]. Unfortunately, this method has been shown to work poorly when applied to models trained using pseudonymized data [67].

The lack of robust methods for quantifying the privacy gains of pseudonymizing training data remains a significant drawback of the technique. For example, differentially private learning, as described in the background, gives rigorous mathematical privacy guarantees. In contrast, while the results in this article show that privacy can be gained without sacrificing data utility, the exact privacy gains remain unknown. However, the estimated amount of remaining PII in the training data provides an upper bound concerning the entities covered by the pseudonymizer. In any case, there is no consensus on how privacy *should* be measured from a regulatory standpoint. Indeed, according to some strict but prominent interpretations of the GDPR, legal use of data containing PII may be next to impossible [44]. The development of GDPR-compliant privacy metrics should preferably be conducted in communication with the legal community.

Domain-adaptive pre-training

Both pre-trained models—SweDeClin-BERT and SweClin-BERT—are initialized with the weights of the general-domain Swedish KB-BERT model. As shown by Lamproudis et al. [53], this allows them to converge faster when compared to pre-training from randomly initialized weights. This is beneficial from a resource perspective, as pre-training is both time and energy consuming. It can also have positive benefits for privacy, as the models have been trained using both sensitive and non-sensitive corpora.

While there are benefits to initializing the models from an already capable general-domain model, this decision is also a possible limitation of our methodology. While the previous study by Lamproudis et al. showed that

domain-adapted models and models pre-trained from scratch eventually converge, they did not look at whether pseudonymization may affect this result. Although PII constitutes a very small portion of the total data [68], it is plausible that pseudonymization introduces variability to the pre-training corpora. This added variability could make it easier or harder to learn. Whether pseudonymizing the pre-training corpora has any substantial impact on the rate of convergence or the final quality of a model trained from scratch is an interesting idea for future research.

Identifying PII in clinical text

The effectiveness of end-to-end pseudonymization as a privacy-preserving technique depends largely on the ability to accurately identify PII in the corpora used for pre-training and fine-tuning the clinical language models. In this study, a manually annotated PII corpus [42] was used to fine-tune clinical BERT models to identify PII. The performance of these models – estimated through evaluations on held-out test data from the *Stockholm EPR PHI Corpus* – is reported in Table 2 and 3. While both precision and recall are fairly high for most PII classes, we have not evaluated the performance of the model to identify PII in the downstream task corpora, nor in the pre-training corpus. A previous study showed that the performance of a CRF model trained on this PII corpus performed worse when applied to other types of clinical notes and that the performance varied quite considerably across different types of clinical notes, i.e. produced in different clinical specialties, written by persons in different professional roles, and under different headings [69]. In part, this may also be explained by the fact that the prevalence of PII varies across different types of clinical notes. While the overall PII density[8] was estimated to be around 1.57%, it was estimated to be as low as 0.97% for notes written by physiotherapists and as high as 2.19% in discharge notes [68].

The results of this study show that the utility of the models was not negatively affected by being trained on pseudonymized data compared to using the original sensitive data, allowing privacy risks to be reduced without sacrificing predictive performance. However, the utility would likely, at some point, be reduced if a pseudonymization system with poor precision substantially distorted the data. Here, two pseudonymizers with different performance levels were evaluated and the results did not indicate any significant difference in terms of their impact on data utility for fine-tuning clinical BERT models. However, previous work evaluating the impact of pseudonymization on the performance of clinical NER tasks showed that training pseudonymizers with higher recall at the expense of lower precision does eventually harm data utility [48]. In future work, it would be interesting to determine at what point – e.g., at a certain level of precision – that data utility starts to be significantly impacted. However, this tolerance threshold would likely need to be determined separately for different downstream tasks.

Sharing data and models

The clinical language model SweDeClin-BERT and the Stockholm EPR Gastro ICD-10 Pseudo Corpus are available for academic use worldwide[9]. Based on the results of this study, we plan to make the other pseudonymized corpora used in this study available as well. However, this requires supplementary ethical approval from the Swedish Ethical Review Authority. Moving forward, an interesting issue is whether it is also possible to make these pseudonymized clinical corpora and language models available to industry. This would enable commercial applications that could be used in real clinical settings. The benefits of sharing data and models must also be balanced against the privacy risks of doing so. From a legal standpoint, sharing data among academics can be justified due to the explicit provisions that the GDPR makes for research. These provisions do not apply to commercial use, making sharing data with commercial partners difficult.

As noted earlier in the discussion, there is no consensus regarding how privacy should be quantified when dealing with NLP models. The current flora of PLMs is heterogeneous, including both masked language models like BERT and generative models such as the GPT family. Risk assessments should likely be done on a per-model basis, given the vast differences between models in terms of architectures, the scale of their pre-training data, their number of parameters, and what privacy-preserving techniques have been applied. The models used in this study are based on the modestly-sized $\text{BERT}_{\text{BASE}}$ model, a non-generative model composed of approximately 110 million parameters. Although there have been several studies on the matter [24–27], there are no known examples of successful training data extraction attacks targeting BERT models.

It is important to note that the performance measures attained in this study do not necessarily hold for other sets of hospitals. All models and datasets use data from the Health Bank research infrastructure, which come

[8] Defined as the number of PII-labeled tokens divided by the total number of tokens.

[9] Contact the authors for details on how to gain access to the data and models.

from a specific set of medical clinical units. It is well-known that models trained on one set of data sources may perform worse when confronted with novel data [37]. Indeed, as noted in the previous section, performance can vary even within a set of data sources. Further complicating the situation, the clinical domain generally struggles with the many restrictions on sharing data. While understandable and justified from a privacy perspective, these restrictions make it difficult to evaluate models and datasets cross-institutionally. Nevertheless, two studies applying SweDeClin-BERT to new data have been carried out [70, 71], with encouraging results.

Conclusion

This study evaluates the impact of pre-training and fine-tuning using automatically pseudonymized training data. Two clinical BERT models, one trained on real data and one trained on pseudonymized data, are evaluated on five clinical downstream tasks. The datasets for these tasks are used both in unaltered form and in pseudonymized versions. The results from evaluating all different configurations of models and datasets are tested using Mann-Whitney U tests.

The analysis of the statistically significant tests finds limited evidence supporting that, in some cases, fine-tuning non-pseudonymized PLMs may work better if using non-pseudonymized data. Such an effect, if real, is small. Furthermore, we find no cases where pre-training and fine-tuning using pseudonymized data end-to-end harms utility. This demonstrates that pseudonymization can decrease the privacy risks of using clinical data for NLP without harming the utility of the machine learning models.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02546-8.

Supplementary Material 1.

Acknowledgements
We are grateful for the support for this study from the DataLEASH project and its funder, Digital Futures.

Authors' contributions
Thomas Vakili performed all the experiments and wrote the main parts of the article. Aron Henriksson gave suggestions on the experiment setup, contributed to the article, and commented on all parts. Hercules Dalianis gave suggestions on the experiment setup, contributed to the article, and commented on all parts.

Funding
Open access funding provided by Stockholm University. Thomas Vakili and Hercules Dalianis were funded by DataLEASH: LEarning And SHaring under Privacy Constraints through Digital Futures and the Strategic Research Area Information and Communication Technology the Next Generation (ICT TNG) of the Swedish government.

Availability of data and materials
The datasets generated and analyzed during the current study are not publicly available due to legal and ethical privacy concerns, as discussed in the section on "Sharing data and models". The pseudonymized versions of the ICD-10 corpus as well as the pre-trained SweDeClin-BERT model, are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679. Collecting informed consent was not possible due to the nature of the data. All methods were carried out in accordance with the Declaration of Helsinki [72]. In compliance with item 32 of the Declaration of Helsinki, the need for informed consent was waived by the Swedish Ethical Review Authority.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 August 2023 Accepted: 21 May 2024
Published online: 12 June 2024

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc.; 2017. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
2. Devlin J, Chang MW, Lee K, Toutanova K. 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). Minneapolis: Association for CompLinguistics; 2019. pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.
3. Vakili T. Attacking and Defending the Privacy of Clinical Language Models [Licentiate thesis]. Stockholm University. Kista: Department of Computer and Systems Sciences, Stockholm University; 2023. https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-216693.
4. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al.. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. https://doi.org/10.48550/arXiv.2307.09288. ArXiv:2307.09288.
5. Lin C, Miller T, Dligach D, Bethard S, Savova G. A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis: Association for Computational Linguistics; 2019. pp. 65–71.
6. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H, et al. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview. JMIR Med Inf. 2020;8(11):e23375.
7. Luo YF, Henry S, Wang Y, Shen F, Uzuner O, Rumshisky A. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. J Am Med Inform Assoc. 2020;27(10):1529–e1.
8. Mahendran D, McInnes BT. Extracting adverse drug events from clinical notes. In: AMIA Summits on Translational Science Proceedings, vol. 2021. 2021. pp. 420.
9. Mahajan D, Liang JJ, Tsou CH, Uzuner Ö. Overview of the. n2c2 shared task on Contextualized Medication event extraction in clinical notes. J Biomed Inform. 2022;2023:104432.
10. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics; 2022. pp. 1998–2022.

11. Lewis P, Ott M, Du J, Stoyanov V. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Virtual: Association for Computational Linguistics; 2020. pp. 146–157.
12. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc (HEALTH). 2021;3(1):1–23.
13. Lehman E, Hernandez E, Mahajan D, Wulff J, Smith MJ, Ziegler Z, et al. Do We Still Need Clinical Language Models? 2023. ArXiv:2302.08091.
14. Lamproudis A, Henriksson A, Dalianis H. Evaluating Pretraining Strategies for Clinical BERT Models. In: Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022). Marseille: ELRA Language Resources Association; 2022. pp. 410–416.
15. Tai W, Kung H, Dong XL, Comiter M, Kuo CF. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Virtual: Association for Computational Linguistics; 2020. pp. 1433–1439.
16. Koto F, Lau JH, Baldwin T. IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics; 2021. pp. 10660–10668.
17. Lamproudis A, Henriksson A, Dalianis H. Vocabulary modifications for domain-adaptive pretraining of clinical language models. In: Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies – HEALTHINF, vol. 5. Virtual: SciTePress; 2022. pp. 180–188.
18. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. New York: Association for Computing Machinery; 2021. pp. 610–623. https://doi.org/10.1145/3442188.3445922.
19. Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting Training Data from Large Language Models. In: Proceedings of the 30th USENIX Security Symposium. 2021. pp. 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.
20. Vakili T, Dalianis H. Utility Preservation of Clinical Text After De-Identification. In: Proceedings of the 21st Workshop on Biomedical Language Processing at ACL 2022. Dublin: Association for Computational Linguistics; 2022. pp. 383–388. https://doi.org/10.18653/v1/2022.bionlp-1.38. https://aclanthology.org/2022.bionlp-1.38.
21. Verkijk S, Vossen P. Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille: European Language Resources Association; 2022. pp. 1098–1103. https://aclanthology.org/2022.lrec-1.118.
22. Vakili T, Lamproudis A, Henriksson A, Dalianis H. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In: Proceedings of the 13th Language Resources and Evaluation Conference LREC 2022. Marseille; 2022. pp. 4245–4252. https://aclanthology.org/2022.lrec-1.451/.
23. Lothritz C, Lebichot B, Allix K, Ezzini S, Bissyandé TF, Klein J, et al. Evaluating the Impact of Text De-Identification on Downstream NLP Tasks. In: The 24rd Nordic Conference on Computational Linguistics, NoDaLiDa 2023. Tórshavn; 2023. https://aclanthology.org/2023.nodalida-1.2/.
24. Nakamura Y, Hanaoka S, Nomura Y, Hayashi N, Abe O, Yada S, et al. KART: Privacy Leakage Framework of Language Models Pre-trained with Clinical Records. 2020. ArXiv: 2101.00036.
25. Jagannatha A, Rawat BPS, Yu H. Membership Inference Attack Susceptibility of Clinical Language Models. 2021. ArXiv: 2104.08305.
26. Lehman E, Jain S, Pichotta K, Goldberg Y, Wallace B. Does BERT pretrained on clinical notes reveal sensitive data? In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics; 2021. pp. 946–959. https://doi.org/10.18653/v1/2021.naacl-main.73. https://aclanthology.org/2021.naacl-main.73.
27. Vakili T, Dalianis H. Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In: Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021). 2021. https://ceur-ws.org/Vol-3068.
28. Nasr M, Carlini N, Hayase J, Jagielski M, Cooper AF, Ippolito D, et al.. Scalable Extraction of Training Data from (Production) Language Models. 2023. https://doi.org/10.48550/arXiv.2311.17035. ArXiv:2311.17035.
29. Torra V. Guide to Data Privacy: Models, Technologies, Solutions. Undergraduate Topics in Computer Science. Cham: Springer International Publishing; 2022. https://doi.org/10.1007/978-3-031-12837-0. https://link.springer.com/10.1007/978-3-031-12837-0.
30. Murakonda SK, Shokri R. ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning. 2020. https://doi.org/10.48550/arXiv.2007.09339. ArXiv:2007.09339.
31. Mireshghallah F, Goyal K, Uniyal A, Berg-Kirkpatrick T, Shokri R. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics; 2022. pp. 8332–8347.
32. Dwork C, McSherry F, Nissim K, Smith A. Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S, Rabin T, editors. Theory of Cryptography. Lecture Notes in Computer Science. Berlin: Springer; 2006. pp. 265–284. https://doi.org/10.1007/11681878_14.
33. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In: Vaudenay S, editor. Advances in Cryptology - EUROCRYPT 2006. Lecture Notes in Computer Science. Berlin: Springer; 2006. pp. 486–503. https://doi.org/10.1007/11761679_29.
34. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep Learning with Differential Privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. CCS '16. New York: Association for Computing Machinery; 2016. pp. 308–318. https://doi.org/10.1145/2976749.2978318.
35. Anil R, Ghazi B, Gupta V, Kumar R, Manurangsi P. Large-Scale Differentially Private BERT. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi: Association for Computational Linguistics; 2022. pp. 6481–6491. https://aclanthology.org/2022.findings-emnlp.484.
36. Brown H, Lee K, Mireshghallah F, Shokri R, Tramèr F. What Does it Mean for a Language Model to Preserve Privacy? In: 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22. New York: Association for Computing Machinery; 2022. pp. 2280–2292. https://doi.org/10.1145/3531146.3534642.
37. Li J, Zhou Y, Jiang X, Natarajan K, Pakhomov SV, Liu H, et al. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. J Am Med Inform Assoc. 2021;28(10):2193–201. https://doi.org/10.1093/jamia/ocab112. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8449609/
38. Libbi CA, Trienes J, Trieschnigg D, Seifert C. Generating Synthetic Training Data for Supervised De-Identification of Electronic Health Records. Futur Internet. 2021;13(5):136. https://doi.org/10.3390/fi13050136. https://www.mdpi.com/1999-5903/13/5/136. Multidisciplinary Digital Publishing Institute
39. Hiebel N, Ferret O, Fort K, Névéol A. Can Synthetic Text Help Clinical Named Entity Recognition? A Study of Electronic Health Records in French. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik: Association for Computational Linguistics; 2023. pp. 2320–2338. https://aclanthology.org/2023.eacl-main.170.
40. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023;6(1):1–10. https://doi.org/10.1038/s41746-023-00958-w. https://www.nature.com/articles/s41746-023-00958-w. Nature Publishing Group
41. CMS. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). 1996. http://www.cms.hhs.gov/hipaa/. Accessed 28 Aug 2023.
42. Dalianis H, Velupillai S. De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. J Biomed Semant. 2010;1(1):6. https://doi.org/10.1186/2041-1480-1-6.
43. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data

Protection Regulation) (Text with EEA relevance). Legislative Body: EP, CONSIL; 2016. http://data.europa.eu/eli/reg/2016/679/oj/eng.
44. Weitzenboeck EM, Lison P, Cyndecka M, Langford M. The GDPR and unstructured data: is anonymization possible? Int Data Priv Law. 2022;12(3):184–206.
45. Yeniterzi R, Aberdeen J, Bayer S, Wellner B, Hirschman L, Malin B. Effects of personal identifier resynthesis on clinical text de-identification. J Am Med Inform Assoc. 2010;17(2):159–68.
46. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, et al. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. Proc AAAI Conf Artif Intell. 2020;34(05):8968–8975. https://doi.org/10.1609/aaai.v34i05.6428. https://ojs.aaai.org/index.php/AAAI/article/view/6428.
47. Berg H, Chomutare T, Dalianis H. Building a De-identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text. In: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Hong Kong; 2019. pp. 118–125. https://aclanthology.org/D19-6215/.
48. Berg H, Henriksson A, Dalianis H. The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Louhi 2020, in conjunction with EMNLP 2020. Virtual: Association for Computational Linguistics; 2020. pp. 1–11.
49. Vakili T, Hullmann T, Henriksson A, Dalianis H. When Is a Name Sensitive? Eponyms in Clinical Text and Implications for De-Identification. In: Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024). St. Julian's: Association for Computational Linguistics; 2024. pp. 76–80. https://aclanthology.org/2024.caldpseudo-1.9.
50. Dalianis H, Henriksson A, Kvist M, Velupillai S, Weegar R. HEALTH BANK - A workbench for data science applications in healthcare. In: CEUR Workshop Proceedings. CEUR-WS; 2015.
51. Lamproudis A, Henriksson A, Dalianis H. Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). Held Online: INCOMA Ltd.. 2021. pp. 790–797. https://aclanthology.org/2021.ranlp-main.90.
52. Malmsten M, Börjeson L, Haffenden C. Playing with Words at the National Library of Sweden – Making a Swedish BERT. 2020. arXiv:2007.01658.
53. Lamproudis A, Henriksson A, Dalianis H. Evaluating Pretraining Strategies for Clinical BERT Models. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, et al., editors. Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille: European Language Resources Association; 2022. pp. 410–416. https://aclanthology.org/2022.lrec-1.43.
54. Bird S, Loper E. NLTK: The Natural Language Toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions. Barcelona: Association for Computational Linguistics; 2004. pp. 214–217. https://aclanthology.org/P04-3031.
55. Remmer S, Lamproudis A, Dalianis H. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In: Proceedings of RANLP 2021: Recent Advances in Natural Language Processing, RANLP 2021, 1-3 Sept 2021, Varna, Bulgaria; 2021. pp. 1158–1166.
56. Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. J Biomed Inform. 2014;49:148–58.
57. Velupillai S, Dalianis H, Kvist M. Factuality levels of diagnoses in Swedish clinical text. In: User Centred Networked Health Care. IOS Press; 2011. pp. 559–563.
58. Velupillai S. Automatic classification of factuality levels: A case study on Swedish diagnoses and the impact of local context. In: Fourth International Symposium on Languages in Biology and Medicine, LBM 2011. Singapore; 2011.
59. Lamproudis A, Henriksson A, Dalianis H. Vocabulary Modifications for Domain-adaptive Pretraining of Clinical Language Models. 2022. pp. 180–188. https://www.scitepress.org/PublicationsDetail.aspx?ID=ILgTQ0V6lDU= &t=1.
60. Dalianis H. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In: Proceedings of the Workshop on NLP and Pseudonymisation, September 30, 2019, Turku, Finland. 166. Linköping University Electronic Press; 2019. pp. 16–23.
61. James G, Sohil F, Sohali MU, Shabbir J, Witten D, Hastie T, et al. An introduction to statistical learning with applications in R. New York: Springer Science and Business Media; 2013. https://www.tandfonline.com/doi/full/10.1080/24754269.2021.1980261.
62. Nakayama H. seqeval: A Python framework for sequence labeling evaluation. 2018. https://github.com/chakki-works/seqeval. Accessed 12 Apr 2024.
63. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge University Press; 2008.
64. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat. 1947;18(1):50–60. https://doi.org/10.1214/aoms/1177730491. https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-18/issue-1/On-a-Test-of-Whether-one-of-Two-Random-Variables/10.1214/aoms/1177730491.full. Institute of Mathematical Statistics.
65. Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets. J Mach Learn Res. 2006;7(1):1–30. http://jmlr.org/papers/v7/demsar06a.html.
66. Mireshghallah F, Uniyal A, Wang T, Evans DK, Berg-Kirkpatrick T. An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022. pp. 1816–1826. https://aclanthology.org/2022.emnlp-main.119.
67. Vakili T, Dalianis H. Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data. In: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). Tórshavn: NEALT Proceedings Series; 2023. pp. 318–323. https://aclanthology.org/2023.nodalida-1.33.
68. Henriksson A, Kvist M, Dalianis H. Prevalence estimation of protected health information in Swedish clinical text. In: Informatics for Health: Connected Citizen-Led Wellness and Population Health. IOS Press; 2017. pp. 216–220.
69. Henriksson A, Kvist M, Dalianis H. Detecting protected health information in heterogeneous clinical notes. Stud Health Technol Inform. 2017;245:393–7.
70. Jerdhaf O, Santini M, Lundberg P, Bjerner T, Al-Abasse Y, Jonsson A, et al. Evaluating Pre-Trained Language Models for Focused Terminology Extraction from Swedish Medical Records. In: Proceedings of the Workshop on Terminology in the 21st century: Many Faces, Many Places. Marseille: European Language Resources Association; 2022. pp. 30–32. https://aclanthology.org/2022.term-1.6.
71. Bridal O, Vakili T, Santini M. Cross-Clinic De-Identification of Swedish Electronic Health Records: Nuances and Caveats. In: Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference, LREC 2022. Marseille; 2022. pp. 49–52. https://aclanthology.org/2022.lrec-1.451/.
72. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191–4. https://doi.org/10.1001/jama.2013.281053.

Publisher's Note

PAPER VI

DATA-CONSTRAINED SYNTHESIS OF TRAINING DATA FOR DE-IDENTIFICATION

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2025. Data-constrained synthesis of training data for de-identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27414–27427, Vienna, Austria. Association for Computational Linguistics.

Author Contributions Thomas Vakili designed and performed all the experiments and wrote the main parts of the paper. Aron Henriksson and Hercules Dalianis gave suggestions on the experimental setup, contributed to the paper, and commented on all parts.

Data-Constrained Synthesis of Training Data for De-Identification

Thomas Vakili, Aron Henriksson, and Hercules Dalianis
Department of Computer and Systems Sciences
Stockholm University, Kista, Sweden
{thomas.vakili, aronhen, hercules}@dsv.su.se

Abstract

Many sensitive domains – such as the clinical domain – lack widely available datasets due to privacy risks. The increasing generative capabilities of large language models (LLMs) have made *synthetic datasets* a viable path forward. In this study, we domain-adapt LLMs to the clinical domain and generate synthetic clinical texts that are machine-annotated with tags for personally identifiable information using capable encoder-based NER models. The synthetic corpora are then used to train synthetic NER models. The results show that training NER models using synthetic corpora incurs only a small drop in predictive performance. The limits of this process are investigated in a systematic ablation study – using both Swedish and Spanish data. Our analysis shows that smaller datasets can be sufficient for domain-adapting LLMs for data synthesis. Instead, the effectiveness of this process is almost entirely contingent on the performance of the machine-annotating NER models trained using the original data.

1 Introduction

Many useful applications in NLP involve domains where the data are sensitive. These privacy risks, and the accompanying limits to sharing data, have traditionally been solved through *de-identification*. This process involves finding parts of the text that can be used to identify an individual. Such information is typically referred to as personally identifiable information (PII). After locating PII, the passages need to be processed to remove or obscure the PII. Traditionally, this time-consuming work has been done manually. Automatic de-identification (Meystre et al., 2010) is a machine-driven approach that typically relies on named entity recognition (NER) to detect PII that needs to be removed.

Unfortunately, the PII datasets that exist to assist in privacy preservation are themselves sensitive and can usually not be shared. This circularity, together with the increasing generative capabilities of large language models (LLMs), has led to a growing interest in overcoming data limitations by eschewing the use of real data altogether. Instead, one can use generated *synthetic* corpora.

Previous studies have mainly been concerned with evaluating the privacy of the synthetic text (Yue et al., 2023; Miranda et al., 2024) or with creating the strongest-performing model possible using synthetic data (Libbi et al., 2021; Hiebel et al., 2023; Liu et al., 2025). Our study instead examines how synthetic data can be produced under constrained resources. This understudied problem is common in clinical institutions that lack resources, both in terms of data and computational hardware.

We carry out a systematic evaluation of key factors impacting the utility of LLM-generated synthetic data as training data for downstream tasks. Specifically, we study synthetic NER data for PII detection – an important task in the privacy-sensitive healthcare domain. Synthetic clinical data are generated using domain-adapted LLMs and machine-annotated using fine-tuned NER models. The evaluations focus primarily on the *utility* of synthetic data for training NER models.

Through extensive experimentation, we investigate the impact on utility of (i) the amount of data used for domain adaptation of the synthesizing LLM, (ii) the quality of the machine annotator, (iii) the amount of synthetic data generated, and (iv) model size. We also quantify the diversity and privacy of the generated data, and carry out experiments across two languages – Swedish and Spanish. Our main contributions are:

1. Demonstrating that moderately-sized LLMs can be adapted to the clinical domain to produce high-utility text with relatively small amounts of in-domain data.

2. Showing that using synthetic machine-annotated data allows for training NER mod-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27414–27427

els that perform only slightly worse compared to using real, sensitive data, while reducing the risk of exposing sensitive information in the original data.

3. Finding that, for the task of detecting PII, using larger generative LLMs for synthesis does not yield clear improvements in terms of utility. Rather, downstream performance relies on having a high-quality gold standard NER model for providing machine annotations.

2 Related Research

There have been several approaches to generating synthetic clinical data for a number of languages and for different purposes. Broadly speaking, most prior works have either focused on maximizing the utility of the synthetic data, or on studying the privacy characteristics of synthetic corpora.

While most papers studying data synthesis contain some form of privacy analysis, other papers have this as their main focus. Several papers study how differentially private learning impacts the utility (Yue et al., 2023; Igamberdiev et al., 2024) and the privacy of the data (Miranda et al., 2024). While privacy is an important justification for synthesizing data, it is not the main focus of our paper.

The second main current in the literature explores how to optimize synthesis to create the best possible synthetic corpora. These papers synthesize data using locally domain-adapted LLMs (Ive et al., 2020; Hiebel et al., 2023), or using instruction-tuned models (Kiefer, 2024; Liu et al., 2025). They show that high-utility data synthesis is possible. However, fewer papers systematically examine the conditions required for success.

In our literature review, two studies stand out as particularly relevant to this study. Libbi et al. (2021) synthesize a Dutch corpus for PII detection using a GPT-2 model (Radford et al., 2019) domain-adapted using 1 million documents and add rule-based machine annotations. Our study follows the same overall process for synthesis, but uses much less data and more modern NLP techniques. Xu et al. (2023) similarly create synthetic corpora and experiment with constraining the total amount of data used, but do so for the relation extraction task. In this paper, we focus on a different task – NER for PII detection. Furthermore, in contrast to both studies, we systematically evaluate the impact of constraining data alternately for *both* domain adaptation *and* machine annotation, try two different model sizes, synthesize corpora of different sizes, and validate our results across two languages.

3 Data and Methods

In this study, we investigate the impact of various factors related to generating synthetic data for fine-tuning encoder language models on downstream tasks in the healthcare domain. Specifically, we study the possibility of generating synthetic clinical text for training NER models for detecting PII. The synthetic text is created by a domain-adapted generative LLM and then machine-annotated for PII using a fine-tuned encoder model. This process follows the structure of previous works (Libbi et al., 2021) and is illustrated in Figure 1.

3.1 Generative Models

The aim of this study is to examine the feasibility of generating training data for NER models detecting PII. The foundation of the training data are synthetic texts, generated using autoregressive LLMs. Two model families are used as a base for domain adaptation to the clinical domain.

GPT-SW3 For Swedish, we use the *GPT-SW3* model (Ekgren et al., 2024). This autoregressive language model was trained using approximately 320 billion tokens. The data were mainly composed of Scandinavian texts and 35.3% of the data is Swedish.

FLOR The autoregressive model used to generate Spanish data is the *FLOR* model (Da Dalt et al., 2024). The model was initialized with the weights of the multi-lingual BLOOM model (Scao et al., 2023) and trained with continued pre-training. The data used spanned 140 billion tokens and was composed of equal parts English, Spanish, and Catalan data.

Both models are used autoregressively, without instruction tuning. The hypothetical – but often occurring – scenario motivating the study's design is where researchers have access to a small and sensitive NER dataset that cannot be shared outside of their organization. Zero-shot synthesis is an alternative strategy, but we leave it to future research to evaluate if this approach can yield clinical texts that are sufficiently similar to the real data. Instead, we perform different degrees of domain-adaptive fine-tuning to train the LLMs to produce such texts.

The primary experiments in Section 4.1 used the versions of FLOR and GPT-SW3 with 6.3

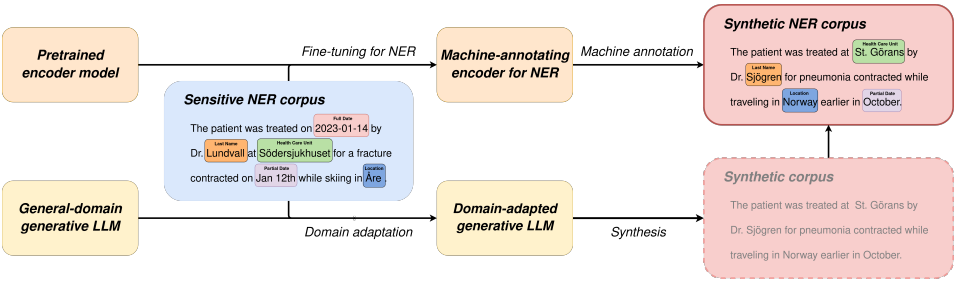

Figure 1: This figure illustrates the steps for creating the synthetic corpora. A gold standard corpus is used to train a NER model and to domain-adapt a general-domain LLM to produce synthetic clinical data. The LLM is used to generate synthetic data, and the NER model is used to add machine annotations. Later, these corpora are used to train synthetic NER models that are evaluated on gold-standard test data.

billion and 6.7 billion parameters, respectively. Smaller models with 1.3 billion parameters were used in Section 4.4 to investigate the impact of using smaller models for synthesis.

3.2 NER Datasets

This study focuses on a particular type of NER dataset – NER for detecting PII. Such datasets exist for several languages but are, as discussed in the introduction, very difficult to share and access. This is particularly true for datasets targeting the clinical domain. In this study, two datasets for detecting PII in clinical data are used.

SEPR PHI The *Stockholm EPR PHI*[1] *Pseudo Corpus (SEPR PHI)* is a Swedish dataset from five different healthcare units consisting of 100 patient records split into 21,553 sentences. The corpus spans 282,766 tokens, where 6,755 are manually annotated for nine different PII classes (Velupillai et al., 2009). The dataset has then been pseudonymized, meaning that all the annotated entities have been replaced with realistic pseudonyms (Dalianis, 2019).

MEDDOCAN The second dataset is *MEDDOCAN* – a Spanish dataset consisting of 1,000 medical texts. These are based on clinical cases augmented with PII from auxiliary sources (Marimon et al., 2019). The texts were then manually annotated for 19 different PII. Out of 504,569 tokens, 41,859 are tagged as PII. The documents were originally divided into train, test, and development sets. In this paper, these three subsets have been merged before resampling them into new subsets for five-fold cross-validation.

Each dataset was split into three subsets: one for training, another for validation, and a third held-out subset for testing. The training sets are used to domain-adapt the generative models and to train the machine-annotating NER models. The purpose of the validation subsets is twofold. First, they are used to monitor the training processes when fine-tuning the models for synthesis and NER. Once the models for synthesis have been trained, the validation data also serve as starting points when creating the synthetic corpora. Finally, the quality of the NER models trained using the synthetic corpora is evaluated using the held-out test sets, as these are not in any way part of the training or synthesis processes.

3.3 Encoder Models

This study trains NER models by fine-tuning pretrained encoder models. Two different models are used, one for each language.

SweDeClin-BERT We use *SweDeClin-BERT* (Vakili et al., 2022) for Swedish data as it has previously shown strong performance on the SEPR PHI corpus. This BERT-style model is based on the general-domain KB-BERT model (Devlin et al., 2019; Malmsten et al., 2020) and adapted to the clinical domain through continued pre-training on the Swedish Health Bank corpus (Dalianis et al., 2015). It consists of 125 million parameters.

roberta-base-bne For Spanish data, we use the *roberta-base-bne* model trained by Gutiérrez-

[1]Protected health information (PHI) is a term defined by the HIPAA regulation (HHS, 1996). The PHI term covers a subset of the types of information that constitute PII.

Fandiño et al. (2022). This RoBERTa-based model (Liu et al., 2019) consists of 125 million parameters. It was trained using a large Spanish corpus collected by the National Library of Spain and performs strongly on the MEDDOCAN task.

These models are used for two purposes. First, they are fine-tuned using the gold standard datasets. These gold models are used to add machine annotations to the synthesized corpora. They are also used as baselines for the later, synthetic NER models. The synthetic models are also initialized from the pre-trained encoder models, but are fine-tuned using the synthetic, machine-annotated corpora. The gold and synthetic NER models are then evaluated and compared to each other in order to measure the performance implications of using synthetic data.

3.4 Synthesizing and Evaluating the Corpora

The generative models were domain-adapted by fine-tuning them for autoregressive language modeling using QLoRA (Dettmers et al., 2023) as implemented in the Axolotl framework[2] with $r = 8$ and $\alpha = 32$. A comprehensive specification of all hyperparameters is available in Appendix A. GPT-SW3 was domain adapted using the Swedish SEPR PHI corpus and FLOR using the Spanish MEDDOCAN corpus. Domain adaptation was carried out with varying amounts of data to determine the impact on the utility of the synthetic data.

As mentioned in Section 3.2, the validation sets used for monitoring the fine-tuning process were also used for data synthesis. The 5% validation subsets were used to create starting points for generating text, as suggested by Libbi et al. (2021). The starting points were created by taking the first three words of each document in the validation sets.

Synthetic corpora have an intriguing advantage over real corpora: they can be arbitrarily large. Taking this feature into account, each three-word starting point was used to create 80 new samples. Consequently, the synthetic corpora are four times larger than the gold-standard corpora. In Section 4.5, the benefits of exploiting this feature are examined through experiments that use smaller amounts of synthetic data.

Synthesis was done using the vLLM library (Kwon et al., 2023). We use nucleus sampling (Holtzman et al., 2020) with $p = 0.95$, and the minimum token length is set to 10. The maximum token length is set to the length of the longest document in the validation set, or at least 50. The temperature was set to $t = 1.0$ after preliminary experiments showed that varying $0.8 \leq t \leq 1.2$ had very little impact on the results.

Finally, the synthetic texts were machine-annotated for PII entities. These were added using NER models fine-tuned on the gold-standard datasets. The gold-standard and synthetic NER models were trained for 6 epochs with a batch size of 16. The limited context window of the models was overcome, both during training and machine annotation, by splitting long documents into 128-word chunks. This chunking was done both during training and when using the gold-standard models for machine annotation.

The whole process, from domain adaptation to training and evaluating the synthetic NER models, was done through five-fold cross-validation. The utility of the synthetic corpora was measured using the token-level F_1 score. These evaluations rely on each fold's held-out gold-standard test data.

4 Experiments

The main contribution of this paper is a systematic investigation into how the different steps of synthetic corpora creation respond to data constraints. This section describes these experiments and their results. All experiments, except for those in Section 4.5, take advantage of the unbounded nature of synthetic corpora and allow them to be four times larger than the gold standard datasets. The performance of the NER models is summarized using token-level F_1 scores tested on gold-standard data. All values are averages and standard deviations that are calculated based on the results from the five-fold cross-validation.

4.1 Constraining the Total Amount of Data

The first experiment of this study investigated how much data is required to produce a well-performing NER model for detecting PII. The amount of data used for domain-adapting the generative model and fine-tuning the gold NER model is varied. This models the common situation where there is limited access to data, and demonstrates what performance can be expected for different data sizes.

Within each fold, this is scaled to between 5% and 95% of the training data in the fold. Four different amounts are used: 5%, 25%, 50%, and 95%. These subsets correspond to the *Sensitive*

[2] https://github.com/axolotl-ai-cloud/axolotl

% of fold	SEPR PHI			MEDDOCAN		
	Gold	Synthetic	Δ	Gold	Synthetic	Δ
5%	0.707 ± 0.037	0.724 ± 0.035	-0.017 ± 0.051	0.931 ± 0.012	0.309 ± 0.060	0.622 ± 0.061
25%	0.871 ± 0.010	0.847 ± 0.010	0.024 ± 0.014	0.967 ± 0.003	0.964 ± 0.005	0.003 ± 0.006
50%	0.908 ± 0.007	0.885 ± 0.010	0.023 ± 0.012	0.973 ± 0.004	0.970 ± 0.004	0.003 ± 0.006
95%	0.926 ± 0.005	0.896 ± 0.007	0.029 ± 0.009	0.978 ± 0.005	0.973 ± 0.003	0.005 ± 0.006

Table 1: Between 5% and 95% of the training data in each fold was used to domain-adapt the generative LLMs and the machine-annotating encoder models. In this table, the F_1 scores are listed for both the NER models trained on gold standard data and the synthetic data, as well as the difference. The F_1 scores are the average scores and their standard deviation over all five folds.

NER corpus in Figure 1. The final 5% are used for validation and for creating prompts for synthesis.

Synthetic corpora have an advantage over real corpora: their size is only constrained by the amount of computing power available for generation. As explained in Section 3.4, this feature is incorporated by letting the synthetic data be four times larger than the original datasets. In later experiments, we analyze the extent to which this advantage helps.

Table 1 lists the average F_1 scores and their standard deviation for each tested configuration. Unsurprisingly, the performance of the gold models – the models trained on the real data – increases as more and more data are available for training. The MEDDOCAN models are more resilient to shrinking the amount of training data, but both cases show clear improvements as more data are available. The F_1 scores of the models trained on synthetic corpora follow a similar pattern. Increasing the data allows for more data to be used to domain-adapt the generative model and for training a better machine annotator.

4.2 Scaling the Data for Domain Adaptation

The experiments in the previous section show that it is indeed feasible to create well-performing NER models trained on synthetic data using our method. On the other hand, the results depend on the amount of data used for domain-adapting the synthesizing generative LLM and for fine-tuning the machine-annotating encoder model. In the previous experiment, the data were fixed for both purposes.

This section describes an ablation study that measures the impact of varying the amounts of data used for domain adaptation. As before, the synthetic corpora are allowed to be four times larger than the original corpora. In these experiments, the amount of data used for fine-tuning the machine annotator is kept constant at 95%, while the amount used for domain adaptation is varied between 5% and 95%. Additionally, we also use synthetic corpora generated *without* domain adaptation.

% for d.a.	SEPR PHI	MEDDOCAN
0%	0.547 ± 0.178	0.295 ± 0.011
5%	0.873 ± 0.014	0.313 ± 0.032
25%	0.877 ± 0.010	0.970 ± 0.005
50%	0.896 ± 0.007	0.970 ± 0.005
95%	0.896 ± 0.007	0.973 ± 0.003
Gold	*0.926 ± 0.005*	*0.978 ± 0.005*

Table 2: The amount of data used for **d**omain **a**daptation (d.a.) of the synthesizing generative LLM was varied from 0% to 95% of the training data in each fold. The average F_1 scores of the synthetic NER models and the gold-standard models are listed.

The average F_1 scores of the models resulting from these experiments are listed in Table 2. Unsurprisingly, the worst-performing models are those that were trained using corpora synthesized without domain adaptation. These results show that domain adaptation does matter. However, there are clearly diminishing returns from increasing the amount of data for domain adaptation. Increasing the amount of data from 50% to 95% produces nearly identical results.

4.3 Varying the Data for Machine Annotation

The experiments in Section 4.2 indicated that the synthetic corpora improve when more data are available to domain adapt the model generating the text. However, the values in Table 2 and the values from the original experiments in Table 1 differ greatly. The results indicate that using a strong machine annotator – as in Section 4.2 – explains more of the performance. Another set of experiments was conducted to examine this effect.

In contrast to the previous experiments, these

% for m.a.	SEPR PHI		MEDDOCAN	
	Gold	Synthetic	Gold	Synthetic
5%	0.707 ± 0.037	0.725 ± 0.039	0.931 ± 0.012	0.942 ± 0.010
25%	0.871 ± 0.010	0.858 ± 0.012	0.967 ± 0.003	0.967 ± 0.004
50%	0.908 ± 0.007	0.889 ± 0.005	0.973 ± 0.004	0.965 ± 0.009
95%	0.926 ± 0.005	0.896 ± 0.007	0.978 ± 0.005	0.973 ± 0.003

Table 3: The amount of data used to create the **m**achine **a**nnotator (m.a.) varied between 5% and 95% of the training data in each fold. This table compares the downstream F_1 scores of the synthetic and gold-standard NER models. The values are the average F_1 scores and their standard deviation across all five folds.

experiments use 95% of the data for domain adaptation of the generative model that produces the synthetic text. This corpus is still allowed to be four times larger than the original training corpus. The data used to create the machine-annotating NER model is varied between 5% and 95%.

Models were trained and evaluated using five-fold cross-validation, and the resulting F_1 scores are listed in Table 3. The average F_1 scores adhere closely to those of the gold standard model that is trained on real data. This is especially clear when contrasting the scores with what was shown in Table 2. This strongly suggests that the performance of the synthetic models is mainly explained by the amount of data available when creating the machine annotators.

4.4 Using Smaller Generative Models

Model size	SEPR PHI	MEDDOCAN
Small	0.883 ± 0.006	0.973 ± 0.004
Larger	0.896 ± 0.007	0.973 ± 0.003
Gold	*0.926 ± 0.005*	*0.978 ± 0.005*

Table 4: GPT-SW3 and FLOR are available in smaller versions than those used in the other experiments. This table compares the average downstream F_1 scores obtained using the smaller and larger versions for domain adaptation.

The generative LLMs used for domain adaptation in this study – GPT-SW3 and FLOR – are available in different sizes. The previous experiments have used the 6.3 billion and 6.7 billion versions of the models. Although these models are not very large from a research perspective, domain-adapting them still requires expensive hardware. In this experiment, we try synthesizing data using the smaller versions of these LLMs.

Both smaller versions consist of approximately 1.3 billion parameters. Table 4 lists the F_1 scores obtained when using 95% of the data for domain adaptation and for creating the machine annotator. Despite being around five times smaller than their larger counterparts, the smaller models yielded very similar results to their larger counterparts. This suggests that smaller models are a viable alternative, at least for synthesizing data for PII identification.

4.5 How Much Synthesis is Enough?

In all previous experiments, we have exploited the fact that synthetic corpora can be generated indefinitely. This has been represented by letting the corpora be four times larger than the training data. In this experiment, we examine the effect of removing this advantage. In addition to training on the four times larger corpora, we also trained models using corpora of the same size as the training corpora. Finally, we trained models using just 5% of the synthetic corpora. The data used for domain adaptation and fine-tuning the machine annotator was kept at 95%.

Synthesized amount	SEPR PHI	MEDDOCAN
5%	0.814 ± 0.008	0.938 ± 0.006
100%	0.889 ± 0.009	0.968 ± 0.005
400%	0.896 ± 0.007	0.973 ± 0.003
Gold	*0.926 ± 0.005*	*0.978 ± 0.005*

Table 5: The synthetic corpora in the other experiments are four times larger than the original gold standards. This table lists the downstream F_1 scores of NER models trained on varying amounts of synthetic data.

Table 5 shows that, for these datasets, generating extra data has a small impact on the results. Generating a synthetic corpus that is the same size as the original corpus yields downstream results that are within one standard deviation of the results from

generating a four times larger corpus. This is true both for MEDDOCAN and for SEPR PHI.

4.6 Diversity of the Generated Data

Three different metrics were used to quantify the data themselves. These were lexical diversity, the length of the documents, and the number of entities in the documents. The metrics were calculated both for the generated corpora and for the gold standard corpora. The lexical diversity was estimated by stemming each token and then dividing the number of unique stems by the total number of tokens. Stems were obtained using the Swedish and Spanish Snowball stemmers implemented in NLTK (Bird and Loper, 2004).

Table 6 lists the three metrics for all of the considered corpora. For the synthetic corpora, the average number of entities was estimated using the strongest machine annotator for each dataset trained using 95% of the gold corpora. The diversity and average lengths of the synthetic corpora could be calculated before machine annotation.

The lexical diversity of the synthetic data is fairly consistent, regardless of the amount of data used for domain adaptation. It is also consistently lower than in the gold corpora. This is likely due to the temperature being fixed across the configurations. As explained in Section 3.4, varying the temperature had a negligible impact on the downstream performance of the synthetic models. However, it is likely that the lexical diversity of the corpora would increase with higher temperatures.

The largest adjustment from adding domain adaptation is that the synthetic corpora become closer to the gold corpora in terms of the number of entities per document and in length. However, the average number of entities per document tends to be noticeably higher in the synthetic corpora than in the gold corpora.

4.7 Estimating Privacy

Creating a synthetic variant of a sensitive dataset only protects the original data if the synthetic and sensitive datasets are sufficiently different. A common proxy for measuring these risks is to study the n-grams of the original and synthetic corpora (Ive et al., 2020; Hiebel et al., 2023). We calculated the *n-gram recall* of each generated dataset and the training data from which it is derived. This metric measures the proportion of unique n-grams in a reference document that is shared with a candidate document[3]. In this experiment, the reference documents are the real sensitive corpora used for domain adaptation, and the synthetic corpora are the candidates. In other words, given a real corpus with a set of n-grams R and a synthetic corpus with n-grams S:

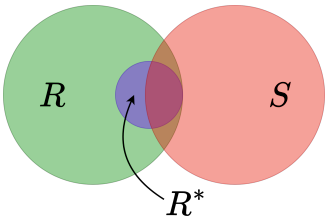

Figure 2: N-gram recall is calculated with all n-grams R in the gold standard corpora, as well as with those in the subset $R^* \subseteq R$ that overlap with sensitive entities. The recall is calculated by counting how many of these are shared with S.

$$\text{n-gram recall} = \frac{|R \cap S|}{|R|} \quad . \tag{1}$$

Since the data in this study are tagged for PII, a PII-sensitive n-gram recall is also used to estimate the degree of leakage of potentially sensitive information. Instead of considering all n-grams R in the reference document, as in Equation 1, this metric only considers the n-grams $R^* \subseteq R$ that overlap with sensitive entities in the gold standard corpus. The relation between R, R^*, and S is illustrated in Figure 2.

N-gram recall values were calculated for $n = \{3, 5, 10\}$. This was done for each of the five folds in the previous experiments. The n-grams were created by concatenating n tokens from the tokenizers of the domain-adapted generative models. The values for 5-grams are summarized in Table 7 and the other values in Tables 10 and 11 in Appendix B.

All three configurations produced similar patterns. The bottom row of Table 7 shows the average 5-gram recall scores of the synthetic corpora that were generated without any domain adaptation. These values were obtained by comparing these corpora to the 95% training corpora. These values serve as a useful baseline since any shared n-grams in these cases are purely incidental.

More interestingly, the n-gram recall values decrease as more data are used for domain-adapting

[3]It is similar to ROUGE (Lin, 2004) but is used on a corpus level rather than for comparing individual documents.

% for d.a.	SEPR PHI			MEDDOCAN		
	Diversity	**Doc. length**	**Doc. labels**	**Diversity**	**Doc. length**	**Doc. labels**
0%	4.28 ± 0.27	53.80 ± 25.21	1.62 ± 7.25	2.69 ± 0.10	542.42 ± 623.43	34.46 ± 111.86
5%	4.52 ± 0.30	18.28 ± 16.66	0.82 ± 4.04	2.58 ± 0.03	644.30 ± 1729.28	57.26 ± 181.33
25%	4.27 ± 0.22	16.73 ± 16.41	0.73 ± 3.95	2.44 ± 0.02	525.79 ± 367.61	56.34 ± 27.13
50%	4.38 ± 0.15	16.41 ± 16.32	0.81 ± 4.12	2.37 ± 0.05	519.07 ± 346.71	57.18 ± 25.18
95%	4.25 ± 0.29	16.69 ± 17.57	0.85 ± 4.36	2.40 ± 0.03	508.63 ± 347.05	57.91 ± 25.67
Gold	*6.26 ± 0.03*	*13.12 ± 18.52*	*0.31 ± 2.12*	*5.55 ± 0.03*	*510.28 ± 427.85*	*48.73 ± 19.97*

Table 6: The average lexical diversity, average length, and average number of annotated labels per document was calculated for the synthetic corpora and for the gold corpora. The synthetic corpora were machine annotated using models trained on 95% of the data. The **d**omain **a**daptation (d.a.) varied between using 95% of the data, to using none.

% for d.a.	SEPR PHI		MEDDOCAN	
	All 5-grams	**Sensitive 5-grams**	**All 5-grams**	**Sensitive 5-grams**
5%	0.328 ± 0.041	0.233 ± 0.066	0.005 ± 0.000	0.008 ± 0.001
10%	0.216 ± 0.002	0.154 ± 0.016	0.003 ± 0.000	0.006 ± 0.001
25%	0.183 ± 0.015	0.169 ± 0.021	0.003 ± 0.000	0.004 ± 0.000
50%	0.134 ± 0.021	0.141 ± 0.017	0.002 ± 0.000	0.003 ± 0.000
95%	0.122 ± 0.013	0.132 ± 0.010	0.002 ± 0.000	0.003 ± 0.000
0%	0.028 ± 0.002	0.047 ± 0.002	0.001 ± 0.000	0.001 ± 0.000

Table 7: 5-gram recall values were calculated for each synthetic corpus over five folds. We calculate both the general 5-gram recall and the recall for 5-grams overlapping with PII in the training corpora. The synthetic corpora varied in the amount of data used for **d**omain **a**daptation (d.a.) before generation. The bottom row shows the values for the synthetic corpora generated without domain adaptation when compared to the 95% gold corpora.

the generating model. Although this trend is not linear – there is a small increase between 10% and 25% for sensitive 5-grams in SEPR PHI – and the standard deviations are large, there is a clear decrease between using 5% and using 95% of the data. The most likely explanation is that the number of unique in-domain n-grams increases as the training data grows, meaning that each individual n-gram is less likely to be memorized. Conversely, using too little data for domain adaptation can cause the model to overfit.

While both general and sensitive n-gram recall decrease when more data are used, the sensitive n-gram recall is sometimes slightly higher. This could indicate that n-grams overlapping with sensitive entities may be at a higher risk of memorization, although this effect is very small. The n-gram recall values are also significantly higher in SEPR PHI than in MEDDOCAN. There is no clear explanation for this, other than that they differ substantially in their structure. MEDDOCAN contains fewer and longer documents, whereas SEPR PHI contains many but shorter documents. Inspecting the overlapping n-grams also revealed that many of them are related to dates or other vague but oft-occurring categories of PII.

5 Discussion

The results in Table 2 show that domain adaptation – to a point – was needed to generate synthetic data of sufficient quality. From a privacy perspective, there is, of course, a risk that domain adaptation causes sensitive data to be memorized and reproduced during data synthesis. The results in Table 7 indicate that a lower proportion of sensitive n-grams were reproduced when more data were used for domain adaptation. On the other hand, using fewer data minimizes the attack surface of the models. If the data requirements are lower, this may make it feasible to, for example, manually de-identify the data, or to audit an automatic de-identification of them.

An example of an application of our results is the cross-institutional validation of NLP systems. In sensitive domains such as the clinical domain, researchers are often barred from sharing their data and trained models due to privacy concerns. A

common situation is when research group (A) has created a system that works very well on their in-house data. Due to privacy regulations, another group (B) cannot share their data with (A), and this makes it difficult for (A) and (B) to validate if the system generalizes. An example of an attempt to work under these constraints is described by Bridal et al. (2022). They were able to make limited claims of generalization, but were limited due to the restrictions on sharing data. The method for synthesis explored in our paper would allow these research groups to share synthetic versions of their datasets or models, as these are much less sensitive than artifacts based on real data.

Training models using synthetic corpora is safer than using real data, but is no panacea. On the other hand, no currently existing technique for privacy preservation is sufficient when used in isolation. For example, NER-based automatic de-identification covers only a subset of PII (Pilán et al., 2022), and differentially private learning for NLP is difficult to implement properly (Brown et al., 2022; Miranda et al., 2024) and is often inefficient when done so (Igamberdiev et al., 2024). Synthetic data generation will likely be an important ingredient in many domains to overcome privacy issues.

6 Conclusions

Data synthesis is an attractive tool for dealing with data scarcity and privacy risks. However, synthesis itself can be challenging when access to data is constrained. Through extensive ablation studies – validated on models and data in two different languages – our experiments show that not all parts of the synthesis process are equally sensitive to these resource constraints. Domain adapting the models to create high-utility clinical text did not require using all of the data. The experiments show that using between 25% and 50% of the data can be enough for domain adaptation, at least for the studied datasets. Furthermore, the experiments also show that using smaller generative LLMs does not necessarily incur a big loss of utility. Instead, our experiments show that the most important factor is the data available to create machine annotations.

These findings are a valuable contribution to the clinical NLP and privacy communities, where sharing real data is often impractical or impossible due to legal constraints. These constraints, while necessary from a privacy perspective, also hinder collaboration. While our studies cannot fully ascertain the privacy of the synthetic corpora, the results indicate that they are less sensitive than the original data. When combined with other safety measures, such as de-identification and secure storage, synthetic data can serve as a basis for collaboration across institutional boundaries.

7 Limitations

The results demonstrated the importance of adapting the synthesizing LLM to the clinical domain in order to generate high-utility training data. For synthesizing corpora for PII detection, this was possible to do with very small amounts of data. However, it may be the case that PII detection is a task where the domain-specific details of the data are less important. Indeed, Libbi et al. (2021) argue that NER data, in general, retain their utility for machine learning even if, qualitatively, their contents lack coherence. In future work, it would be interesting to investigate if the same holds for clinical tasks that are more challenging and domain-specific tasks. Nevertheless, PII detection in the clinical domain is an important task in itself and any advances in this area will help to combat the issue of data scarcity.

Our process for synthesis also uses autoregressive language modeling without instruction-tuning. We opted for this design to make the task as simple as possible. As previously mentioned, this may not necessarily be a good design for other types of tasks where the document-level semantics matter more. For example, Kiefer (2024) synthesized data for the task of assigning diagnosis codes to discharge summaries by instruction-tuning models to create documents with different characteristics.

Another limitation of the study is the reliance on n-gram-based metrics for estimating privacy risks. This is a common practice (Ive et al., 2020; Hiebel et al., 2023) and can detect when data are being reproduced verbatim in the synthetic corpora. On the other hand, n-grams vary greatly in how sensitive they are. We try to address this with our n-gram recall metric that takes PII into account. However, we make limited claims about the privacy of the data and instead focus on their utility.

In Section 4.4, we find that smaller versions of the generative models could generate data of near-equal utility as their larger counterparts. This was especially clear when generating MEDDOCAN data. An interesting continuation would have been

to fine-tune an even smaller LLM using MEDDOCAN data. Unfortunately, the 1.3 billion version of FLOR that we use is the smallest one. GPT-SW3 is available in smaller versions, but proceeding to a monolingual analysis would lower the validity of the results and fell outside the scope of this study. Future work could try similar experiments with languages for which smaller models exist.

8 Ethical Statement

This work was conducted under ethical permission no. 2019-05679 granted by the Swedish Ethical Review Authority. MEDDOCAN is a publicly available dataset, where the PII in the documents are unrelated to the original patients. SEPR PHI is available on request and is a manually pseudonymized corpus where all identified PII have been replaced with surrogate values. Because of this, the privacy risks of the experiments in this paper are very small. Regardless, the experiments have been carried out in a computational environment in which only the authors and system administrators have had access to the data. Our experiments are also in accordance with the intended uses of the datasets.

The experiments conducted in this paper required considerable amounts of computational resources. We estimate that creating and evaluating the data and models for our experiments took approximately 130 GPU hours[4]. Luckily, the experiments were run in Sweden – where virtually all energy comes from sustainable sources. Nevertheless, the electricity expended when conducting these experiments could have been used for other purposes.

On the other hand, our results indicate that high-utility synthetic corpora can be created using small-scale data and without relying on the very largest LLMs. These results can be particularly helpful for researchers working in resource-constrained environments. This includes not only researchers in, e.g., the clinical domain, but also those working with low-resource languages. These parts of the NLP community are often under-served as increasing focus is placed on terabyte-scale datasets and LLMs with unwieldy amounts of parameters.

Even though synthetic data are safer than sensitive data, there is a risk that other researchers over-interpret our results and use them to justify irresponsible uses of synthetic data. Our focus on constraining the amounts of data used hopefully mediates some of these potential risks. Furthermore, we have thoroughly described the limitations of our results and the scope of our experiments.

[4]The calculations are available in Appendix C

Acknowledgments

We are grateful for the support for this study from the DataLEASH project and its funder, Digital Futures. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Olle Bridal, Thomas Vakili, and Marina Santini. 2022. Cross-clinic de-identification of Swedish electronic health records: Nuances and caveats. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 49–52, Marseille, France. European Language Resources Association.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What Does it Mean for a Language Model to Preserve Privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2280–2292, New York, NY, USA. Association for Computing Machinery.

Severino Da Dalt, Joan Llop, Irene Baucells, Marc Pamies, Yishi Xu, Aitor Gonzalez-Agirre, and Marta Villegas. 2024. FLOR: On the Effectiveness of Language Adaptation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7377–7388, Torino, Italia. ELRA and ICCL.

Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, volume 166, pages 16–23.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. HEALTH BANK - A workbench for data science applications in healthcare. In *Proceedings of the 27th Conference on Advanced Information Systems Engineering (CAiSE 2015)*, pages 1–18. CEUR-WS.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in neural information processing systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An Autoregressive Language Model for the Scandinavian Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68(0):39–60. Number: 0.

HHS. 1996. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Published: Online at https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *Proceedings of the Eighth International Conference on Learning Representations*.

Timour Igamberdiev, Doan Nam Long Vu, Felix Kuennecke, Zhuo Yu, Jannik Holmer, and Ivan Habernal. 2024. DP-NMT: Scalable Differentially Private Machine Translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 94–105, St. Julians, Malta. Association for Computational Linguistics.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, 3(1):69.

Lotta Kiefer. 2024. Instruction-Tuning LLaMA for Synthetic Medical Note Generation: Bridging Data Privacy and Utility in Downstream Tasks. Master's thesis, Saarland University.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Claudia Alessandra Libbi, Jan Trienes, Dolf Trieschnigg, and Christin Seifert. 2021. Generating synthetic training data for supervised de-identification of electronic health records. *Future Internet*, 13(5):136.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jinghui Liu, Bevan Koopman, Nathan J. Brown, Kevin Chu, and Anthony Nguyen. 2025. Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports. *Artificial Intelligence in Medicine*, 159:103027.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv:2007.01658 [cs]*. ArXiv: 2007.01658.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Jose Antonio Lopez Martin, and Marta Villegas. 2019. Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10:1–16.

Luis Miranda, Jocelyn Dunstan, Matías Toro, Federico Olmedo, and Felix Melo. 2024. Evaluating Privacy Risks in Synthetic Clinical Text Generation in Spanish. In *Latinx in AI @ NeurIPS 2024*.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, et al. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 4245–4252. European Language Resources Association.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International journal of medical informatics*, 78(12):e19–e26.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023. S2ynRE: Two-stage Self-training with Synthetic data for Low-resource Relation Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, Toronto, Canada. Association for Computational Linguistics.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

A Hyperparameters and Software Versions

Both GPT-SW3 and FLOR were fine-tuned using the hyperparameters in Table 8. These were selected based on examples from prior works and through small-scale experiments. SweDeClinBERT and roberta-base-bne were also trained using the same hyperparameters, listed in Table 9. These parameters were also selected based on examples from prior works. If the encoder models got stalled in the local minimum of only predicting the zero class, then training was restarted using the same data and hyperparameters. Due to the many different factors in our synthesis process, a full search of the hyperparameter space was not feasible. The hyperparameters used in the experiments proved sufficiently optimized to obtain high-utility models and clear results.

Parameter	Value
r	8
α	32
Dropout	0.05
Weight decay	0.1
Learning rate	0.0001
Batch size	16
Epochs	6

Table 8: The hyperparameters used for domain-adapting the generative LLMs using QLoRA.

Parameter	Value
Weight decay	0.00001
Learning rate	0.0001
Batch size	16
Epochs	6

Table 9: The hyperparameters used when fine-tuning the BERT/RoBERTa models for NER.

We also used several Python libraries to implement our experiments. The most important ones are NLTK (version 3.8.1) for stemming when computing diversity, vLLM (version 0.6.1) for synthesis, Axolotl (version 0.6.0) for domain adaptation (version 0.6.0), and Huggingface Transformers (version 4.44.0) for fine-tuning the NER models and for tokenizing the corpora before counting n-grams.

B 3-Grams and 10-Grams

As mentioned in Section 4.7, we calculated the n-gram overlaps for $n = 3, 5, 10$. Results for $n = 3$ are listed in Table 10 and $n = 10$ in Table 11. The results follow the same overall pattern as in Table 7 but are included for completeness. As expected, 3-grams are much more likely to occur in both corpora and 10-grams a lot *less* likely. Many 3-grams – due to subword tokenization – are not full words. In the main part of the paper, we chose

to present 5-grams because this struck a balance between what prior studies have used, and the fact that the models we study use sub-word tokenizers.

C Computational Requirements

The experiments in this paper consumed many GPU hours due to the large number of configurations required for the ablation study. The GPUs were provided by the National Academic Infrastructure for Supercomputing in Sweden. Unfortunately, the environment offers no straightforward way of computing the GPU hours. In this appendix, we estimate the GPU hours required to run the experiments based on data from the logs.

Domain-adapting the FLOR 6.3B model and synthesizing the corpora, for all amounts of domain-adaptation considered in this paper, took 7.5 hours. Domain-adapting GPT-SW3 took 10 hours. Both these processes used four *Nvidia A100* GPUs and a total of 70 GPU hours.

The encoder NER models were trained using single *Nvidia V100* GPUs. Training the SweDeClin-BERT models on one 95% portion of the data took approximately 6 minutes. Training roberta-base-bne on 95% of a MEDDOCAN fold took approximately 5 minutes. Based on other logs for other configurations, the time scales linearly with the amount of data. Based on this assumption, the models presented in this paper took an additional 60 GPU hours to train.

% for d.a.	SEPR PHI		MEDDOCAN	
	All 3-grams	**Sensitive 3-grams**	**All 3-grams**	**Sensitive 3-grams**
5%	0.583 ± 0.040	0.540 ± 0.061	0.066 ± 0.003	0.072 ± 0.006
25%	0.436 ± 0.019	0.426 ± 0.022	0.031 ± 0.000	0.039 ± 0.001
50%	0.365 ± 0.028	0.358 ± 0.033	0.026 ± 0.000	0.037 ± 0.001
95%	0.331 ± 0.018	0.321 ± 0.018	0.021 ± 0.000	0.034 ± 0.001
0%	0.180 ± 0.011	0.179 ± 0.007	0.019 ± 0.000	0.026 ± 0.001

Table 10: 3-gram recall values were calculated for each synthetic corpus. The values are averages and standard deviations over five folds.

% for d.a.	SEPR PHI		MEDDOCAN	
	All 10-grams	**Sensitive 10-grams**	**All 10-grams**	**Sensitive 10-grams**
5%	0.294 ± 0.019	0.040 ± 0.016	0.000 ± 0.000	0.000 ± 0.000
25%	0.137 ± 0.005	0.033 ± 0.008	0.000 ± 0.000	0.000 ± 0.000
50%	0.082 ± 0.008	0.022 ± 0.006	0.000 ± 0.000	0.000 ± 0.000
95%	0.052 ± 0.007	0.022 ± 0.004	0.000 ± 0.000	0.000 ± 0.000
0%	0.001 ± 0.000	0.002 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 11: 10-gram recall values were calculated for each synthetic corpus. The values are averages and standard deviations over five folds.