Saarland University

Department of Language Science and Technology

Master Thesis

Submitted as part of the degree requirements of the MSc in Language Science and Technology, Saarland University

Instruction-Tuning LLaMA for Synthetic Medical Note Generation: Bridging Data Privacy and Utility in Downstream Tasks

Author:
Lotta Kiefer
Matriculation: 7039244

Supervisors:
Prof. Dr. Dietrich Klakow
Jesujoba Alabi

Additional Advisor:
Prof. Dr. Hercules Dalianis

28.11.2024

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged. I assure that the electronic version is identical in content to the printed version of the Master's thesis.

Saarbrücken, 28.11.2024

Place, Date

Signature

Acknowledgement

I would like to thank Hercules Dalianis for enabling my research internship at the Natural Language Processing Research Group at DSV, Stockholm University, and for his advice, guidance, and friendly support throughout my time in Stockholm. Furthermore, I want to thank Thomas Vakili who offered me valuable insights and technical support during the experiments. I thank Aron Henriksson, Martin Hansson, Tyr Hullmann, Mohamad Homam Mawaldi, and Martin Mladenov for sharing their expertise with me and helping me build upon their studies.

I also want to thank my supervisors Dietrich Klakow and Jesujoba Alabi for helping me shape this research with their support, valuable feedback, and great advice throughout numerous discussions during the process of my Master's Thesis.

Finally, I am very grateful to my family and friends for always supporting me throughout my studies.

Abstract

Recent advancements in Natural Language Processing (NLP) have unlocked transformative potential for medical applications. The discharge summaries contained in Electronic Health Records (EHRs) could serve as an ideal source for medical research and the development of such applications. However, this progress is constrained by the private nature of these records, which strongly limits the availability of high-quality training data. To address this, we propose a novel framework employing LLaMA-3.1-8B for generating synthetic English and Swedish medical notes. Instruction-tuning with ICD-10 descriptions aims at producing data that balances privacy and utility while overcoming challenges such as diversity reduction and medical incoherence observed in prior approaches. Comprehensive evaluation reveals that the synthetic data exhibits a broad vocabulary, strong privacy protections, and high utility for tasks like Named Entity Recognition (NER) and medical coding for both English and Swedish. The synthetic notes demonstrate on-par performance with real data in NER tasks and show potential for state-of-the-art results in medical coding with increased dataset size. Differences in utility were found to be most likely attributable to some widespread noise contained in the synthetic dataset. While some artifacts remain, user studies involving medical professionals found no significant differences in readability or medical coherence compared to real data. Privacy evaluations confirmed low proximity to real data, mitigating risks of sensitive information leakage. This study establishes a robust foundation for synthetic medical note generation, addressing privacy and data sparsity challenges in clinical NLP. The results highlight synthetic data as a promising alternative for training high-performance medical systems, paving the way for privacy-preserving, scalable, and effective solutions in healthcare.

Contents

List of Figures

List of Tables

List of Abbreviations

AE	Autoencoders
AI	Artificial Intelligence
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representations from Transformers
BART	Bidirectional Auto-Regressive Transformer
CITI	Collaborative Institutional Training Initiative
DL	Deep Learning
DP	Differential Privacy
DSV	Department of Computer and System Sciences
ED	Emergency Department
EHR	Electronic Health Record
EMR	Exact Match Ratio
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GPT	Generative Pretrained Transformer
GPU	Graphics Processing Unit
HIPAA	Health Insurance Portability and Accountability Act
ICD	International Classification of Diseases
ICL	In-Context Learning
ICU	Intensive Care Unit
ISO	International Organization for Standardization
LLaMA	Large Language Model Meta AI
LLM	Large Language Model
LSTM	Long Short-Term Memory
MAP	Mean Average Precision
MIMIC	Medical Information Mart for Intensive Care
MIT	Massachusetts Institute of Technology
NER	Named Entity Recognition
NLP	Natural Language Processing
OOF	Out-of-Family
PEFT	Parameter-Efficient Fine-Tuning
PHI	Protected Health Information
QLoRA	Quantization Low-Rank Adaptation
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SEPR	Stockholm Electronic Patient Record Corpus
SOTA	State-of-the-Art
WF	Within-Family
WHO	World Health Organization

1 Introduction

Recent advancements in Deep Learning (DL) and Natural Language Processing (NLP) have unlocked significant potential for developing medical assistant tools, such as automatic diagnosis coding for discharge summaries. However, as in any DL domain, the availability and quality of training data are critical determinants of system performance. Electronic Health Records (EHR) provide a comprehensive digital repository of patients' healthcare information, including demographics, medical history, diagnoses, and medications. These records, which encompass both structured tabular data and unstructured free-text notes, could serve as an ideal source for medical research and tool development. In particular, free-text notes, with their detailed and nuanced descriptions of a patient's health, hold great potential for advancing biomedical research and creating powerful medical applications with the help of NLP methods (Wu et al., 2022; Yogarajan et al., 2020).

Despite this potential, the sensitive nature of EHR data imposes strict limitations on its accessibility, resulting in a shortage of high-quality training data for robust medical artificial intelligence (AI) systems (Murtaza et al., 2023). Synthetic data generation offers a promising solution to this challenge, provided the generated data preserves the task-critical properties of real data while safeguarding patient privacy by avoiding the leakage of Protected Health Information (PHI). Recent years have seen an increasing interest in generating synthetic free-text medical notes, especially with the advent of large language models (LLMs). However, many existing approaches struggle with a trade-off between privacy and utility: while synthetic notes may enhance privacy protection, they often lose substantial utility for training downstream models (Baumel et al., 2024; Melamud & Shivade, 2019). Additionally, LLM-generated synthetic data frequently suffer from reduced diversity, shown in a small vocabulary, and may exhibit medical incoherence (Falis et al., 2024; Hullmann & Hansson, 2024; Libbi et al., 2021; Mawaldi & Mladenov, 2024).

This work aims to address these challenges by leveraging LLaMA-3.1-8B (Dubey et al., 2024), a state-of-the-art (SOTA) LLM, for generating synthetic English and Swedish medical notes that strike a balance between privacy and utility. Our proposed approach involves instruction-tuning LLaMA using textual descriptions of ICD-10 diagnosis and procedure codes, enabling content control and fostering greater diversity in the generated data. To evaluate our approach, we employ a comprehensive set of evaluation methodologies focused on fidelity, privacy, utility, and medical coherence. The utility evaluation encompasses critical tasks such as medical coding and Named Entity Recognition (NER), assessing the generalizability of the synthetic data across diverse applications. Further, a user study involving medical professionals provides insights into the medical coherence and readability of the synthetic documents.

Through this diverse assessment, we explore whether instruction-tuned LLaMA-3.1-8B can generate synthetic data that serves as a viable substitute for real-world data in terms of utility while preserving privacy. Our results demonstrate that the proposed framework

successfully narrows the gap between privacy and utility. By utilizing pseudonymized data during generation, the risk of data leakage is minimized. Our synthetic data achieves on-par performance with real data on certain tasks and demonstrates competitive, albeit slightly lower, utility compared to real data on others. Additionally, it overcomes diversity reduction by exhibiting a broad vocabulary and great variety that even surpasses the real data. Insights from the user study reveal no significant differences in readability or medical coherence between real and synthetic documents.

Overall, our results highlight the efficacy of this approach in generating high-quality English and Swedish medical notes, addressing privacy and data sparsity concerns in clinical NLP. This framework lays a solid foundation for future research and offers a path toward creating robust medical NLP tools suitable for real-world applications without compromising patient privacy.

2 Background and Related Research

This chapter introduces the foundational concepts necessary for the present work and provides a comprehensive overview of the current state of research on the generation of synthetic medical notes.

2.1 Electronic Health Records

An EHR is a digital version of a patient's health data, essentially replacing the traditional patient's paper chart. The International Organization for Standardization (ISO) defines EHRs generically as "a digital repository of a patient's medical information that documents their entire healthcare journey in real time" (International Organization for Standardization, 2004). The primary purpose of EHRs is to securely store and exchange health information across multiple healthcare providers and institutions, enabling continuous and efficient healthcare delivery and supporting clinical decision-making (Häyrinen et al., 2008). EHRs typically include structured data such as demographic information, vital signs, laboratory test results, medication prescriptions, and diagnosis and procedure codes. In addition to this structured information, the patient record usually contains a discharge summary, i.e. a free-text medical note written by healthcare staff, such as doctors or nurses, detailing the patient's health status upon release from the hospital. For this study, discharge summaries, along with diagnosis and procedure codes, are the key focus and are described in more detail below. Other structured data contained in EHRs is not relevant to this study and is therefore not examined in detail.

2.1.1 Discharge Summaries

A discharge summary provides comprehensive documentation of a patient's hospital stay, from admission to discharge. It facilitates communication between hospitals and primary care providers by detailing symptoms, treatments, diagnoses, and other relevant information. Additionally, discharge summaries offer patients and their families clear explanations of diseases and follow-up plans (Wimsett et al., 2014). However, the specific components included in discharge summaries can vary, and many studies have found that important information is often missing (Wimsett et al., 2014). In the U.S., the Joint Commission has established standards for discharge summaries, which should include the following components (The Joint Commission, 2024):

(i) Reason for hospitalization

(ii) Significant findings

(iii) Procedures and treatment provided

(iv) Patient's discharge diagnosis

(v) Patient and family instructions (as appropriate)

(vi) Attending physician's signature

However, Kind and Smith (2008) note that there are no clear definitions for the contents of these components, and it remains unclear how consistently these standards are followed in hospitals. Thus, discharge summaries are relatively free in their form and can vary substantially in structure and content depending on their author. They capture important health information in the form of free text enabling a more detailed storage of information than within the structured components of EHRs. For the purposes of this work, we use the terms "discharge summaries" and "medical notes" interchangeably, as making precise distinctions between different types of clinical free-text reports is not relevant to the scope of this study.

2.1.2 ICD-10 Codes

ICD-10 stands for the 10$^{\text{th}}$ version of the *International Classification of Diseases* (World Health Organization, 2016) and is a system that standardizes diagnosis and procedure codes to ensure that patients receive the correct level of care and that healthcare providers are accurately compensated for their services (Edin et al., 2023). ICD was first introduced by the World Health Organization (WHO) in 1948 and has since been periodically revised. ICD-10 was officially implemented in 2015 after several decades of development (Hirsch et al., 2016). A newer version, ICD-11, came into force in 2022 but is currently in a five-year transition phase before full implementation (Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM), 2024). To date, ICD-10 remains the most widely used medical coding system, encompassing approximately 155,000 codes and being employed in over 100 countries (Hirsch et al., 2016). The codes, which correspond to either diagnoses or procedures, consist of four to seven characters. The first character is always a letter, followed by a digit, and up to five additional characters that can be either alphabetical or numerical. The coding scheme follows a hierarchical classification structure, as illustrated in an example in Figure 2.1.

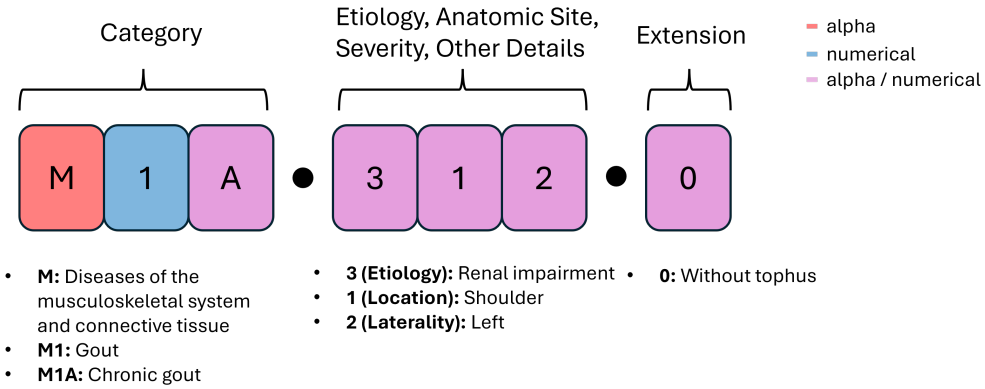


Figure 2.1: Example of ICD-10 structure: Chronic gout due to renal impairment, left shoulder, without tophus

2.2 Clinical NLP

Healthcare is a top priority in every country, yet it is a complex system constantly facing new challenges. One significant issue is the shortage of healthcare professionals, such as doctors and nurses, which many countries struggle with (Kempe, 2024; Murray, 2002). This shortage can have far-reaching consequences, including an overload of existing health workers. AI applications have the potential to support healthcare professionals and address other critical issues, such as medical errors and disparities in access to care (Goldberg et al., 2024). To build such tools, large-scale medical datasets are essential for computational processing. A substantial portion of medical information is stored in free-text form, such as in scientific publications and discharge summaries. Free-text notes, accounting for approximately 80% of the data within EHRs, are often more comprehensive than their structured counterparts and contain valuable detailed information (Wu et al., 2022; Xiao et al., 2018). As a result, NLP methods have increasingly influenced biomedical research. In particular, the advancements in NLP and LLMs offer great potential for future research and the development of powerful medical tools that can revolutionize healthcare (Demner-Fushman et al., 2009; Garner, 2004; Yogarajan et al., 2020).

A wide range of NLP methods has been utilized in prior research to develop various clinical applications. These include information retrieval, relation extraction, and text classification among others, which are commonly used to build clinical decision support systems, allocate medical resources, or create personal health assistants (Zhou et al., 2022). For a comprehensive listing of all prevalent NLP methods and applications in clinical NLP, we refer to Névéol et al. (2018), Wu et al. (2022), and Zhou et al. (2022). Aligned with this work, we focus on the two key applications of de-identification and medical coding and explore how NER is typically employed in areas of clinical NLP.

2.2.1 De-Identification

EHRs contain sensitive and private information about a patient in the form of personally identifiable information such as full name, birth date, or social security number and confidential information such as health status, sexual orientation, or ethnicity (Fang & Li, 2024). This information, which can be used to identify or potentially identify a patient, is referred to as PHI. In addition to patient-specific data, EHRs may also contain sensitive information related to the healthcare organization, such as details about critical technologies or financial transactions (Fang & Li, 2024). The unregulated and unfiltered distribution of such data poses a serious privacy risk and is a primary reason for the lack of publicly available health data that can be used for effective research.

De-identification is a process used to remove or obscure PHI from a dataset, ensuring that patient records can be used for research or as training data for medical models without compromising patient privacy (Vakili et al., 2022). Various methods of de-identification are employed, including substituting PHI with suitable surrogates (e.g., replacing “John” with “Peter”), masking the PHI (e.g., replacing "John" with underscores ___), using class labels (e.g., replacing “John” with “<First Name>”), or removing all sentences containing PHI (Berg et al., 2020). Research has shown that the way PHI is handled can significantly impact the utility of the data, with substitution using surrogates generally preserving the data’s utility most effectively (Berg et al., 2020).

Manually de-identifying EHR data is a time-consuming and costly process. However, advancements in NLP have made it possible to automate this task, thus being fast and cost-effective (Johnson et al., 2023). Various approaches have been proposed in the literature to automate the de-identification process, ranging from simple rule-based systems to more sophisticated neural methods (Kovačević et al., 2024). One commonly used and effective approach involves leveraging pretrained language models, such as BERT (or domain-specific variants fine-tuned for the clinical context), and adapting them to perform NER on PHI tags (Vakili et al., 2022). The identified entities can then be processed as needed, such as by substituting them with surrogates to protect patient privacy.

While some studies demonstrate how de-identification can reduce privacy risks in medical data while maintaining its usability for downstream tasks (Vakili et al., 2022), other research highlights potential risks associated with the process. Yogarajan et al. (2020) categorize these risks into two main areas: re-identification and the loss of utility, medical accuracy, and consistency across the data. Several studies have shown that re-identification poses a real risk, particularly due to quasi-identifiers, i.e., identifiers that are not explicitly identifiable but may allow for re-identification of individuals when combined with external data (Emam et al., 2011; Yogarajan et al., 2020). Additionally, research suggests that de-identification can compromise medical accuracy, potentially leading to a decrease in downstream task performance (Pantazos et al., 2017; Yogarajan et al., 2020), with varying impact based on the de-identification method used (Berg et al., 2020). Moreover, de-identification requires access to the original, real corpus, which limits the ability to create synthetic data beyond this source. As a result, de-identification alone has to date not enabled the widespread public sharing of clinical data.

In this work, we address these concerns by using de-identified real data as input to train an LLM for synthetic data generation. The generated synthetic data can then be uti-

lized for further tasks, such as developing clinical downstream applications. By employing pseudonymized data in the synthetic data generation process, we add an additional layer of safety, ensuring a high level of privacy protection.

2.2.2 Medical Coding

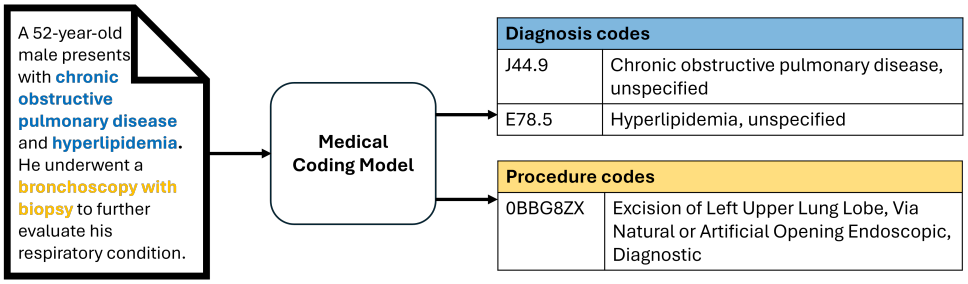


Figure 2.2: Illustrated process of automatic medical coding

Diagnoses and procedures are typically recorded using ICD-10 codes (see Chapter 2.1.2). This standardized system serves administrative, financial, and statistical purposes, making it a crucial component of EHRs. With around 155,000 codes, manual assignment by healthcare professionals is not only time-consuming but also error-prone (Burns et al., 2012; Stausberg et al., 2008; Tseng et al., 2018). Automated medical coding, which involves the automatic identification of ICD-10 codes in the discharge summary of an EHR, offers a promising solution to streamline this process and is a popular application in clinical NLP. Figure 2.2 illustrates this process with the help of an exemplary medical note. Medical coding automation can significantly reduce the administrative burden on healthcare professionals, saving time and potentially minimizing errors during manual coding. Given that properly coded health records are critical for clinical decision-making, public health surveillance, research, and reimbursement, automated clinical coding has the potential to enhance healthcare efficiency greatly and is an extensively researched area (S. Ji et al., 2024).

The task is typically posed as a multi-label multi-class classification task, where each discharge summary $|d|$ with tokens $d = \{t_1, t_2, \ldots, t_{|d|}\}$, has the goal of predicting $y \subseteq Y$, where y is a subset of all possible ICD-10 codes Y (Dong et al., 2022; Huang et al., 2022). The large number of ICD-10 codes makes this task, even with SOTA NLP methods, extremely challenging. Further obstacles arise from the scarcity of clinical training data, the need to handle long documents, and the strong imbalance of codes within the dataset (Dong et al., 2022; Edin et al., 2023). While earlier attempts at this task relied on rule-based and symbolic approaches, the expansion of DL methods and the introduction of various (large) language models has resulted in a vast amount of new methodologies studied to build high-performing medical coding systems (Dong et al., 2022; S. Ji et al., 2024). S. Ji et al. (2024) provide a comprehensive review of current DL approaches for automatic medical coding, which can generally be classified into four categories: recurrent neural networks, convolutional neural networks, neural attention mechanisms, and graph neural networks. Since ICD codes are arranged in hierarchical order, a lot of research focuses on the development

of hierarchical decoders (e.g., Dong et al., 2021; Xie et al., 2019). Additionally, the use of LLMs with few-shot prompting has demonstrated promising results (Yang et al., 2022, 2023). While research on English datasets predominates in this area, systems have also been developed for other languages, including Spanish (Miranda-Escalada et al., 2020), French (Tchouka et al., 2023), and Swedish (Lamproudis et al., 2024), among others.

Rather than discussing all mentioned methodologies, we will focus on the one most relevant to this work: PLM-ICD, introduced by Huang et al. (2022).

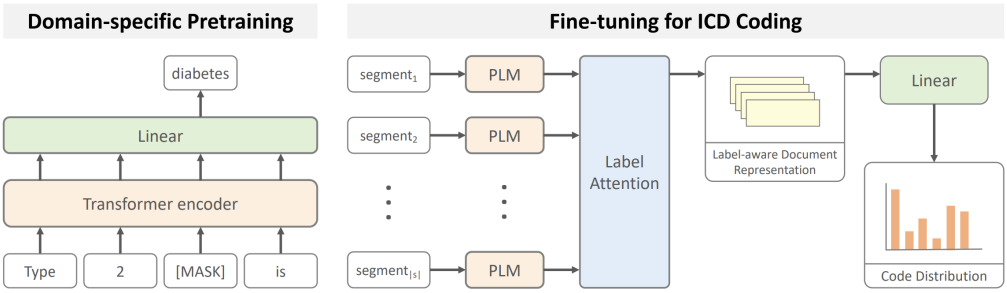


Figure 2.3: Illustration of PLM-ICD framework, Figure from Huang et al. (2022)

Figure 2.3 provides an overview of the PLM-ICD framework, reproduced from the original paper by Huang et al. (2022). This framework leverages a language model pretrained on clinical data. The authors propose a bidirectional encoder architecture, based on BERT (Devlin et al., 2019), that is further pretrained on medical texts, such as PubMedBERT (Gu et al., 2021) and RoBERTa-PM (Lewis et al., 2020). To address the issue of input length limitations, PLM-ICD employs *segment pooling*. This technique splits the document into chunks that are shorter than the model's maximum input length, encodes them, and then aggregates these chunk representations to form a representation of the entire document. This enables PLM-ICD to handle documents that exceed the model's maximum input length.

The framework also incorporates the *label-aware attention* mechanism, initially proposed by Vu et al. (2021). This mechanism enhances the pretrained model by learning label-specific representations that focus on key text fragments relevant to each label. After obtaining token-level hidden representations (H), the attention mechanism computes label-specific attention weights (A) via linear transformations. A weighted sum of H is then calculated to generate label-specific document representations (D), which are used to predict label probabilities through a sigmoid function. The model is trained by minimizing binary cross-entropy loss.

Huang et al. (2022) demonstrate that PLM-ICD achieves SOTA performance compared to previous approaches, with an ablation study validating the effectiveness of its three key components: domain-specific pretraining, segment pooling, and label-aware attention. This framework has since been adopted and further developed by subsequent research, including work by Edin et al. (2023). In their study, Edin et al. (2023) introduced a reproducibility framework, proposing replicable splits and preprocessing methods for the MIMIC-III and MIMIC-IV datasets and comparing the performance of several existing medical coding systems. Their results show that PLM-ICD outperforms the other systems, achieving a Micro F1 score of 58.5% on MIMIC-IV. We use medical coding as primary

downstream utility evaluation and follow the PLM-ICD implementation and dataset splits of MIMIC-IV from Edin et al. (2023) in this work and use their reported results as a reference for comparisons.

2.2.3 NER

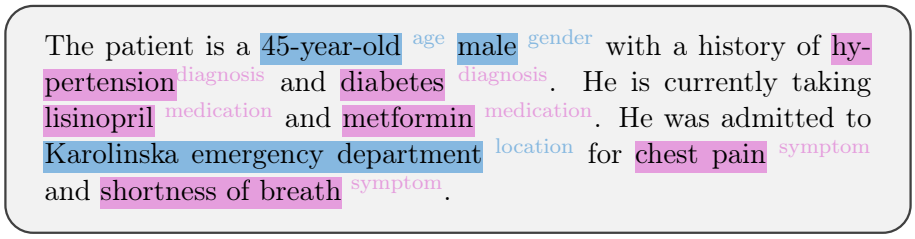


Figure 2.4: Example of NER with biomedical NER in purple and PHI NER in turquoise.

NER is the task of identifying, extracting, and classifying key entities mentioned in a text into predefined categories. This process helps to convert unstructured text into structured information, making it easier to analyze. NER is widely used across various applications and is one of the most common methodologies in clinical NLP (Wu et al., 2022). A popular approach for tackling this task is to use a pretrained language model, such as BERT, which can optionally be further pretrained on clinical domain-specific data. The model is then fine-tuned in a supervised manner for NER, typically framed as a multiclass classification task, where each token is assigned a label corresponding to one of the target entities or a zero label (Bose et al., 2021).

In clinical NLP, NER is applied in two main areas: (i) identifying PHI tags, such as names, dates, or locations, for de-identification (e.g., Kovačević et al., 2024; Libbi et al., 2021; Vakili et al., 2022), and (ii) identifying clinical and biomedical terms, such as diseases, symptoms, medications, or treatments (e.g., Durango et al., 2023; Hiebel et al., 2023). Figure 2.4 illustrates examples of NER tags for both applications. Given that vast amounts of unstructured text are difficult to process and the recognition of clinical concepts is crucial for clinical decision-making, the ability to automatically extract clinical concepts can significantly assist health professionals (Bose et al., 2021). Moreover, de-identification is a critical requirement for sharing clinical data, which is highly relevant for clinical research, as outlined in Chapter 2.2.1. This highlights the importance and widespread use of NER in clinical NLP.

The versatility and importance of NER also contribute to its popularity as a downstream task used to evaluate the utility of synthetically generated medical notes (e.g., Hiebel et al., 2023; Libbi et al., 2021). An additional advantage of NER for this evaluation is that the annotations required for training NER models can be synthetically generated along with the notes themselves, saving time and cost associated with manual annotation (Libbi et al., 2021). However, research has shown that synthetic data can be useful for training a model on a downstream NER task without being necessarily linguistically coherent (Libbi et al., 2021). This suggests that synthetic data, while effective for NER tasks, may not be suitable for other downstream tasks, particularly if it lacks medical coherence. Additional

evaluation methods might be needed to determine whether synthetically generated data can serve as a comprehensive substitute for real data or if distinct synthetic datasets must be generated for different downstream tasks.

We choose NER alongside medical coding as a downstream utility task to evaluate our generated synthetic data. This combination allows for robust utility assessments and facilitates direct comparison with prior work.

2.3 Synthetic Data Generation

All data generated by computational models is considered synthetic data. The rise of generative models has enabled the creation of vast amounts of high-quality synthetic data, such as synthetic text produced by LLMs. Synthetic data serves many purposes, with one popular research application being its use as training data for other models. This approach is largely motivated by three factors (Jordon et al., 2022):

(i) **Data sparsity**: High-performing deep learning models typically require large amounts of diverse data. Often, high-quality, task-specific data is limited, for example when dealing with low-resource languages (Feng et al., 2021). Additionally, labeled data, which is essential for supervised model training, can be costly and time-consuming to annotate. Synthetic (annotated) data has shown to be an effective substitute or supplement, improving model performance on real-world data (Puri et al., 2020).

(ii) **Data privacy**: Research and model training require data distribution, but privacy concerns in certain domains can restrict data sharing and processing, as seen in healthcare and finance (Assefa et al., 2021; Murtaza et al., 2023). Synthetic data allows model training while preserving privacy (Assefa et al., 2021; Libbi et al., 2021; Tang et al., 2023).

(iii) **De-biasing**: Machine learning models are known to inherit historical biases from their training data, such as biases related to gender or race, resulting in unfair or inaccurate performance across different populations (Gallegos et al., 2024). Synthetic data can help address these biases, promoting the development of fairer models (Tiwald et al., 2021).

In the healthcare sector, synthetic data is used primarily due to a combination of data privacy and data sparsity concerns. The sensitive nature of healthcare data limits its availability for research, and the data that is accessible is often insufficient for training high-performing models, particularly for specialized tasks (Hiebel et al., 2023; Tang et al., 2023).

The evaluation of synthetic data typically falls into three categories, each crucial for determining the success of the generation process (Budu et al., 2024):

(i) **Fidelity**: Fidelity refers to how closely the synthetic data resembles the real data, capturing the variable dependencies and statistical properties of the original dataset (Budu et al., 2024). It is commonly measured by comparing data statistics, variable distributions, pairwise correlations, or using distance metrics.

(ii) **Utiliy**: Utility assesses whether the synthetic data can be used in place of real data for similar tasks (Budu et al., 2024). A standard approach for measuring utility involves predictive modeling, i.e., training ML models on both real and synthetic data for downstream tasks and comparing their performance results when tested on real-world data.

(iii) **Privacy**: Privacy examines the risk of information leakage from real to synthetic data, determining whether the synthetic dataset reveals any sensitive information from the original data used for its creation (Budu et al., 2024). Privacy is typically evaluated using distance or disclosure metrics.

Since EHRs contain extensive tabular data, including explicit identifiers (such as names and IDs), demographic information, and medical events recorded over time (e.g., lab results, prescription data, or diagnosis codes), a significant amount of research has focused on synthetic data generation for structured EHR data. These variables represent diverse data types, including categorical, ordinal, numerical, and date formats (Hernandez et al., 2022). Modeling such mixed-type time-series data within synthetic datasets poses a considerable challenge, with approaches ranging from classical statistical or regression methods to DL approaches like Autoencoders (AE) and Generative Adversarial Networks (GAN) (Choi et al., 2017; Dash et al., 2020; Rankin et al., 2020). Hernandez et al. (2022) offer a comprehensive systematic review of these methods, noting the growing popularity of DL approaches, particularly GAN-based models, which have demonstrated promising performance in recent years. Nevertheless, as this work focuses on generating synthetic unstructured free text rather than tabular data, and the methodologies for achieving these goals differ significantly, we will not explore tabular EHR synthesis further and instead focus on the current state of research for generating synthetic free-text medical notes.

To provide an overview of the current research landscape on synthetic medical note generation, a bibliographic search was conducted. Google Scholar was utilized to identify relevant publications from the past ten years (01.01.2014 to 21.10.2024), using the keywords *synthetic medical notes*, *synthetic clinical notes*, and *synthetic discharge summaries*. Additional studies were retrieved from the reference sections of included articles. To qualify for inclusion in this overview, publications had to meet the following criteria:

(i) Present an approach for generating synthetic free-text medical notes.

(ii) Evaluate the synthetic data, with a minimum assessment for utility.

(iii) Be written in English.

Recent publications that are not (yet) peer-reviewed were deliberately included due to their high relevance to this work.

Table 2.1: Research Overview of Methodologies for Generating and Evaluating Synthetic Medical Notes

Authors	Language	Model	Method	Utility Evaluation	Privacy Evaluation
Guan et al. (2018)	English	GAN	Training + REINFORCE	Diagnosis Classification	-
Lee (2018)	English	LSTM	Training	Diagnosis Classification	-
Melamud and Shivade (2019)	English	LSTM	Training	Lexical-semantic association, Natural Language Inference, Letter Case Information	Sequential Pointwise Differential Training Privacy
Amin-Nejad et al. (2020)	English	Vanilla Transformer + GPT2	Training + Fine-tuning	Readmission prediction + Phenotype Classification	-
Kasthurirathne et al. (2021)	English	SeqGAN	Training	NER	Presence Disclosure
Li et al. (2021)	English	CharRNN + SegGAN + GPT-2 + CTRL	Training + Fine-tuning	Clinical NER + Event Sequence Similarity)	-
Libbi et al. (2021)	Dutch	LSTM + GPT-2	Training + Fine-tuning	PHI NER	r-BM25, ROUGE-3, ROUGE-5 and User study
Hiebel et al. (2023)	French	BLOOM + GPT-2	Fine-tuning	Clinical NER	8-gram overlap
Tang et al. (2023)	English	ChatGPT	Prompting	Clinical NER + Relation Extraction	Embedding illustration
Baumel et al. (2024)	English	GPT-3 + PHI-2	Fine-tuning (with DP)	Clinical NER	Robust Membership Inference Attack scores
Belkadi et al. (2024)	English	Bio_ClinicalBERT	Masked Language Modelling	Clinical NER	De-Identification Performance + Presence Disclosure
Falis et al. (2024)	English	GPT-3.5	Prompting	Medical Coding	-
Hullmann and Hansson (2024)	Swedish	KB-Bert	Fine-tuning	PHI NER	8-gram overlap
Kumichev et al. (2024)	Russian	GPT-4 + LLaMA-2-7B	Prompting + Fine-tuning with Knowledge Graph	Medical Coding	-
Litake et al. (2024)	English	ChatGPT + LLaMA-2-70b	Prompting	Classifying Acute Renal Failure	-
Mawaldi and Mladenov (2024)	English	LLaMA-2-7B	Fine-tuning	Clinical NER	-
Micheletti et al. (2024)	English	BERT, RoBERTa, BiomedNLP-PubMedBERT, T5, BART, SciFive-large-Pubmed PMC	Masked + Causal Training	Clinical NER	-
Ren et al. (2024)	English	Bio Clinical BERT + ClinicalBERT + RoBERTa-base + Clinical-Longformer	Masked Training	Clinical NER	-
Singh et al. (2024)	Indian	LLaMA-3-8B Instruct + Gemma + Mistral-7B-Instruct-v0.1	ICL	PHI NER	-

Table 2.1 presents the results of the bibliographic search, which identified 19 relevant articles published between 2018 and 2024. This overview does not aim to exhaustively cover all existing literature, nor does it serve as a systematic review. Instead, its purpose is to provide an overview of the current research landscape, focusing on the methodologies employed for generating and evaluating synthetic medical notes. Given recent advances in language modeling and the resulting methodological diversity, this overview is particularly timely and relevant. We propose that this table could serve as a starting point for a future systematic review. Key findings and conclusions from this search will be discussed, with a more detailed examination of selected publications.

The generation of synthetic free-text EHRs remains in its early stages, though it became increasingly popular with advances in language modeling (Rankin et al., 2020). This trend is evident from the publication dates in Table 2.1, where over half of the included studies were published in the current year, with several appearing only after the experimental phase of this work concluded. While the majority of research focuses on English data, a subset also investigates synthetic generation of discharge summaries in languages such as French (Hiebel et al., 2023), Dutch (Libbi et al., 2021), Swedish (Hullmann & Hansson, 2024), Indian (Singh et al., 2024), and Russian (Kumichev et al., 2024).

A wide range of models and methodologies have been tested for the generation process. Early research focused on training various model architectures from scratch, but there is a clear trend toward leveraging pretrained LLMs through transfer learning. The methodologies employed include fine-tuning, prompting, and in-context learning (ICL), reflecting the evolving nature of the field.

One of the earliest works was conducted by Guan et al. (2018), who trained a GAN framework using the REINFORCE algorithm (Williams, 1992) with disease features as input to generate corresponding notes. Their results showed that the synthetic data performed similarly to real data in a simple disease classification downstream task. Several other studies have utilized Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997), a popular architecture for language processing due to its ability to capture long-range dependencies. For instance, Melamud and Shivade (2019) trained a 2-layer LSTM on English discharge summaries for text generation. Although qualitative observations revealed clear differences from real data, their findings indicated that the generated notes retained genuine properties of the real data. Amin-Nejad et al. (2020) implemented a Vanilla Transformer architecture (Vaswani et al., 2017) for generating clinical notes. Their results demonstrated its potential for data augmentation. However, in low-resource scenarios, the synthetic data was of insufficient quality to enhance performance when used as augmentation.

In addition to the Vanilla Transformer architecture, Amin-Nejad et al. (2020) also explored the potential of pretrained LLMs. They fine-tuned GPT-2 (Radford et al., 2019) for the task, which showed promising performance in low-resource scenarios. However, GPT-2 encountered difficulties handling long sequence-to-sequence downstream tasks, where its generated data demonstrated lower utility compared to the data generated by the Vanilla Transformer. Similarly, Libbi et al. (2021) fine-tuned GPT-2 to generate Dutch discharge summaries and compared the pretrained model to the LSTM architecture used in Melamud and Shivade (2019). While GPT-2 generated more coherent text, the LSTM-generated training data exhibited superior downstream performance on an NER task for

de-identification. This suggests that models trained from scratch can generate synthetic notes with higher utility than smaller pretrained LLMs. However, fine-tuning more recent and powerful pre-trained models, such as GPT-3 (T. Brown et al., 2020) and LLaMA-2-7B (Touvron et al., 2023), has shown promising results for various downstream tasks, including medical coding (Kumichev et al., 2024) and NER (Baumel et al., 2024; Mawaldi & Mladenov, 2024).

In addition to fine-tuning, ICL and prompting have been employed to guide LLMs in generating the desired synthetic notes. These methods offer significant advantages in terms of saving computational resources and time compared to fine-tuning or training models from scratch. While ICL requires only a few examples in the prompt, which is particularly beneficial in data-scarce scenarios, the prompt examples must be carefully selected to ensure diversity and coverage. This approach carries a higher risk of generating a biased dataset. Tang et al. (2023) and Litake et al. (2024) reported successful results when using synthetic data generated through prompting with ChatGPT (OpenAI, 2024) and LLaMA-2-70B (Touvron et al., 2023) as training data for clinical NER and clinical classification models.

However, other studies have highlighted limitations of employing LLMs without prior fine-tuning. Falis et al. (2024) employed GPT-3.5 in a zero-shot manner to generate English discharge summaries and tested their usefulness for augmenting training data for medical coding systems. The authors reported unnatural text with spurious information and a lack of diversity. While synthetic data augmentation slightly hindered overall performance, it reduced out-of-family (OOF) errors. Even though these results are rather unsatisfactory, it is important to consider that medical coding is a complex multi-label classification task. It is more challenging than NER or binary classification tasks, which may explain the differences in utility observed across studies, as medical coding might require data of higher quality and coherence than simpler tasks.

In contrast, Kumichev et al. (2024) addressed medical coding using synthetic Russian medical notes to upsample rare and challenging classes. The authors fine-tuned LLaMA-2-7B (Touvron et al., 2023) and used GPT-4 (Touvron et al., 2023) in a zero-shot manner with a medical knowledge graph integrated into the framework to sample relevant medical information in the prompts. They observed performance improvements in the medical models when upsampling with synthetic data generated by both approaches. However, their findings are not directly comparable to those of Falis et al. (2024), as their evaluation focused on a limited subset of rare codes and involved a substantially smaller number of overall codes in the Russian dataset. Notably, their results suggest that prompting GPT-4 can achieve synthetic data quality comparable to fine-tuning the substantially smaller LLaMA-2-7B model.

While there is a visible trend toward using pretrained autoregressive models like GPT or LLaMA, some studies opt for masked language modeling (Belkadi et al., 2024; Ren et al., 2024). This preference can likely be attributed to a study conducted by Micheletti et al. (2024), where both causal and masked language models were used to generate synthetic text. Their findings showed that masked language models consistently outperformed causal language models in the evaluation of the generated texts. Additionally, Belkadi et al. (2024) emphasize the privacy guarantees and controlled content advantages when using masked language models. However, it is important to note that the comparison in Micheletti et al.

(2024) only involved smaller causal models (with hundreds of millions of parameters), which limits the generalizability of their findings. In contrast, SOTA causal language models typically contain billions of parameters, making the comparison with smaller models less meaningful.

When examining the *Utility* and *Privacy* evaluation columns of Table 2.1, one key observation becomes apparent: there is no consensus on evaluation practices. The majority of studies focus on NER as a downstream utility task, but the implementation of NER varies across studies, with no standardized benchmarks, base models, or training procedures. Other studies focus on building downstream classification models, ranging from simple binary classification tasks (e.g., Litake et al., 2024) to more complex multilabel classification tasks (e.g., Falis et al., 2024). A significant portion of the reported studies does not perform explicit privacy evaluations, either assuming privacy is ensured in their frameworks or leaving privacy evaluation to future work. This may be due to the inherent difficulty of operationalizing privacy measurements. For the studies that do include privacy evaluation, some rely on disclosure metrics (e.g., Belkadi et al., 2024; Kasthurirathne et al., 2021), while others use distance metrics (e.g., Hiebel et al., 2023; Libbi et al., 2021). Since it is challenging to draw definitive conclusions from these metrics alone, some complement these quantitative evaluations with manual assessments and user studies for a more thorough privacy evaluation (e.g., Libbi et al., 2021).

While generation and evaluation approaches differ significantly, some key findings are reported consistently across various studies. In particular, three main issues commonly arise in the generation of synthetic medical notes:

(i) **Privacy Utility Trade-Off:** A widely reported challenge is the trade-off between privacy and utility. Studies have found that utility tends to decrease as privacy protection increases, which is consistent with the finding that downstream models trained on synthetic data often perform less effectively than those trained on real data (Baumel et al., 2024; Melamud & Shivade, 2019).

(ii) **Diversity Reduction:** Several papers report a decrease in variety within synthetic medical notes (Hullmann & Hansson, 2024; Libbi et al., 2021; Mawaldi & Mladenov, 2024). This is typically evidenced by a reduced vocabulary, which may negatively impact downstream performance.

(iii) **Incoherence:** Many studies identify issues with medical incoherence in synthetic notes based on manual inspections, reflecting inconsistencies or inaccuracies in clinical logic and sequence (Falis et al., 2024; Hiebel et al., 2023; Libbi et al., 2021).

The privacy utility trade-off is a central dilemma in synthetic data generation. The less synthetic data resembles real data, the greater the privacy protection. However, this reduced resemblance also causes synthetic data to lose key properties needed for training effective models, thereby impairing model performance on real-world data. Integrating Differential Privacy (DP) into the fine-tuning process can provide mathematically grounded privacy protection by adding controlled random noise to mask individual data points (Baumel et al., 2024). Yet, due to the inherent privacy utility trade-off, this increase in privacy typically results in decreased data utility (e.g., Baumel et al., 2024), which has discouraged some authors from incorporating DP into their frameworks (Libbi et al., 2021). DP

also introduces added computational complexity, and achieving reliable privacy guarantees with DP can be challenging (H. Brown et al., 2022; Igamberdiev et al., 2024; Libbi et al., 2021). For these reasons, DP's use is controversial, not only due to its tendency to degrade utility and increase computational demands but also because some researchers question whether it can truly provide reliable assurance about privacy (H. Brown et al., 2022).

Diversity reduction presents a two-fold problem. First, a decrease in diversity can lead to significant differences in resemblance between synthetic and real data, as the synthetic data fails to represent the full distribution of real-world data. Second, diversity is crucial for high model performance across many tasks. Consequently, diversity reduction can lead to a significant decrease in synthetic data utility (Hullmann & Hansson, 2024; Libbi et al., 2021).

Previous research suggests that training data can be effective even when it lacks grammatical accuracy or topical coherence (Hiebel et al., 2023; Libbi et al., 2021). For example, Libbi et al. (2021) found that their LSTM model achieved higher utility in a NER task compared to their GPT-2 model, despite the latter demonstrating greater coherence. This finding indicates that medical coherence may not be a strict requirement for training data suitability in certain downstream tasks. Libbi et al. (2021) argue that syntactic accuracy could be more crucial than semantic coherence for NER applications. However, while NER tasks may not require coherent narratives, other applications, such as medical chatbots, depend on accurate medical information to function effectively. For synthetic data to serve as a viable substitute for real data across diverse applications, it must be medically coherent to ensure general utility, especially for models intended for real-world use and decision support.

The huge diversity in evaluation metrics found in this search for both utility and privacy makes it extremely challenging to compare the effectiveness of different approaches used for medical note generation and impedes future research from building upon existing work. This overview provides a foundation for a systematic review of evaluation metrics, intending to establish robust baselines and benchmarks to enable meaningful and comparable assessments of different generation methods.

We propose a novel generation framework leveraging a SOTA LLM through instruction-tuning. This approach aims to induce variety in the synthetic data while maintaining utility, ensuring medical coherence, and providing strong privacy protection. This aims at addressing consistently reported issues in the generation of synthetic medical notes in a unified manner. To evaluate this framework, we introduce a diverse set of methodologies that assess fidelity, privacy, generalizable utility, and medical coherence. This comprehensive evaluation is crucial to ensure that synthetic notes can reliably substitute real medical records. We recommend that future research adopt this diversity in evaluation methods and work towards establishing a broadly applicable evaluation benchmark. The proposed framework is detailed in the following chapter.

3 Methodology

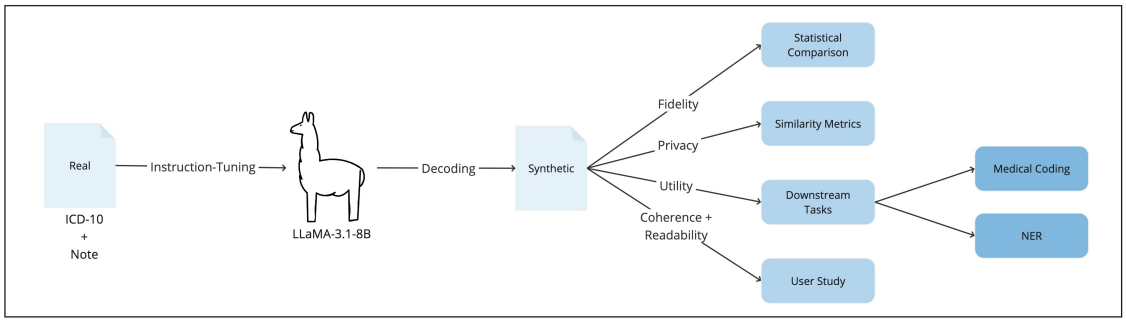


Figure 3.1: Overall approach employed in this work

The primary objective of this work is to generate synthetic medical notes in both English and Swedish that maintain high utility while ensuring robust privacy protection. The synthetic notes are intended to fully replace real data, rather than serve as augmentation, to effectively address privacy concerns. To achieve this, we introduce a novel framework, as depicted in Figure 3.1, designed for generating synthetic medical notes. This framework leverages the corresponding ICD-10 codes of the notes for instruction-tuning the LLaMA-3.1-8B model, enabling the generation of a versatile synthetic dataset in a simple and adaptable manner. The evaluation is divided into four main components: assessing the overall similarity between synthetic and real notes, evaluating the privacy preservation within the notes, examining their utility in downstream tasks with a focus on medical coding, and analyzing their readability and medical coherence through a user study. The following sections will provide a detailed description of each step involved in the process.

3.1 Dataset Creation

This chapter introduces the English and Swedish datasets that form the foundation of this study and outlines the process by which the relevant data was extracted and preprocessed.

3.1.1 MIMIC-IV

The English dataset used in this study was extracted from the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset (Johnson et al., 2023), which contains EHRs sourced from the Beth Israel Deaconess Medical Center obtained between 2008 and 2019.

While MIMIC-IV is publicly available, users must complete a training program provided by the Collaborative Institutional Training Initiative (CITI) (CITI Program, 2024) through the Massachusetts Institute of Technology (MIT) to obtain credentialed access. The dataset includes 524,000 admission records from over 257,000 distinct patients, with EHRs from patients admitted to either the Emergency Department (ED) or the Intensive Care Unit (ICU), resulting in a rich and diverse clinical dataset that includes data from closely monitored patients (Johnson et al., 2023).

The EHRs consist of both structured data, such as demographics, patient measurements, and ICD codes, and unstructured data in the form of free-text clinical notes, including discharge summaries and radiology reports. These clinical notes have been de-identified by removing all PHI tags and replacing them with underscores. For this study, only discharge summaries that include corresponding ICD-10 diagnosis and procedure codes are considered relevant. The discharge summaries are organized into sections, including *Chief Complaint, History of Present Illness, Past Medical History, Brief Hospital Course, Physical Exams, and Discharge Diagnoses*. However, not every discharge summary contains all of these sections, and the scope and detail of each component vary significantly across reports.

Following the methodology outlined by Edin et al. (2023), relevant discharge summaries were filtered, and ICD-10 codes that occurred fewer than 10 times were excluded. This resulted in a dataset of 122,279 documents, containing a total of 7,942 unique ICD-10 codes. Figure 3.2 displays the code distribution across all ICD-10 chapters within the full MIMIC datasets. Leveraging the hierarchical structure of ICD-10, which categorizes diseases into 22 distinct chapters (based on the first three characters of each code), these chapters encompass a wide range of medical fields. The full descriptions of each chapter, along with their corresponding character codes, are listed in Table A.1 in the Appendix. As shown in Figure 3.2, MIMIC includes codes from 15 chapters representing over 2% of the codes of the total dataset. The three most prominent chapters are XXI, IX, and IV, which correspond to *Factors influencing health status and contact with health services*, *Diseases of the circulatory system* and *Endocrine, nutritional and metabolic diseases*.

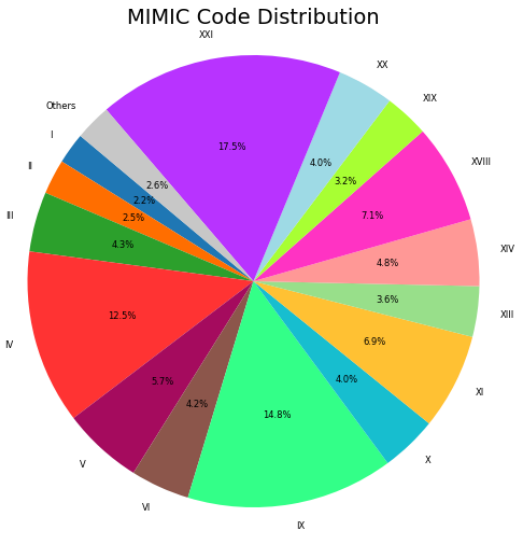


Figure 3.2: MIMIC-IV Chapter Distribution

The dataset was divided into three splits, following the approach outlined by Edin et al. (2023). These splits contain 72.9%, 16.2%, and 10.9% of the documents, respectively, and were created to ensure that the majority of ICD-10 codes are represented in all three sets. For this study, the splits are referred to as MIMIC-L, MIMIC-M, and MIMIC-S, with MIMIC-L being the largest split, MIMIC-M the medium-sized split, and MIMIC-S the smallest.

Initial experiments using LLMs on the MIMIC discharge summaries revealed that the models struggled to process the documents effectively due to certain sections, such as lab results or medication lists. These sections are highly structured, often containing numerous digits, and lack integration into the free-text narrative of the summaries. To address this, a preprocessing approach was developed to remove such structured information, ensuring that the remaining text contained all relevant details in free-text form. Additionally, we identified sections that provided less critical information, did not convey information due to pseudonymization, or repeated content from other parts of the summary. These sections were also removed to make the summaries more concise and easier for the LLMs to process. Since each discharge summary has a slightly different structure and variations in spelling, this task was handled using regular expressions, which carries the risk of not capturing all unwanted sections uniformly across documents. Table 3.1 provides an overview of the sections that were removed as part of the preprocessing. This process resulted in cleaner, more focused notes that were more manageable for the LLMs.

Table 3.1: Different sections removed during preprocessing with justifications for removal.

Section	Justification
Name / Unit	Irrelevant, Pseudonomized
Admission Date / Discharge Date / Date of Birth	Irrelevant, Pseudonomized
Medications	Structured, numerical data
Lab results	Structured, numerical data
Vitals	Structured, numerical data
Facility	Irrelevant, Pseudonomized
Discharge Instructions	Irrelevant, Repetition

3.1.2 SEPR Corpus

The Swedish data used in this study is based on the Stockholm EPR Gastro ICD-10 Pseudo Corpus II (referred to as SEPR II), created as part of the work by Lamproudis et al. (2024). This corpus was extracted from the Health Bank Infrastructure (Dalianis et al., 2015), which includes the Stockholm Electronic Patient Record Corpus (SEPR). SEPR contains Swedish EHRs from over 2 million patients across more than 512 units at Karolinska University Hospital, collected between 2006 and 2014. The data is available for academic use through Stockholm University [1].

The SEPR II subset consists of 317,971 patient records, including 81,089 discharge summaries for 113,174 patients, with corresponding ICD-10 codes. All records within this

[1] Contact the Health Bank, https://www.dsv.su.se/healthbank, at Stockholm University for access.

dataset pertain to gastrointestinal conditions. The patient records have been automatically de-identified by replacing all PHI tags with realistic surrogates. The dataset contains a total of 415 unique ICD-10 codes.

Figure 3.3 illustrates the code distribution across all chapters contained in the full SEPR II dataset. Although all codes in SEPR II originate from ICD-10 chapter XI, which focuses on *Diseases of the digestive system*, they can be further subdivided into ten more specific subcategories. Detailed descriptions of these subcategories are provided in Table A.2 in the Appendix. SEPR II includes codes from eight of these ten subcategories representing over 2% of the codes contained in the full dataset. The three most prominent subcategories are VI, V, and II, corresponding to *Other diseases of intestines*, *Noninfective enteritis and colitis*, and *Diseases of oesophagus, stomach and duodenum*.

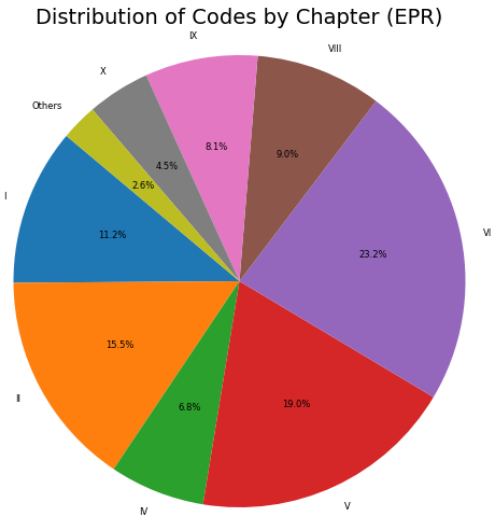


Figure 3.3: SEPR II Chapter Distribution

For this study, the SEPR II corpus was split into three sets containing 75%, 15%, and 10% of the data, referred to as SEPR-L, SEPR-M, and SEPR-S, respectively, in alignment with the naming conventions used for the English dataset. Since the Swedish discharge summaries are significantly more concise and consist only of free text, without the structured sections present in the MIMIC dataset, no additional preprocessing was applied to them.

3.1.3 Dataset Comparison

The English and Swedish corpora differ in several important properties:

(i) **Scope**: The MIMIC-IV dataset covers a wide range of medical domains from the ED and ICU, while the SEPR II corpus is restricted to gastrointestinal conditions.

(ii) **Unique ICD-10 Codes**: Reflecting the difference in scope, the Swedish corpus contains 415 unique ICD-10 codes, significantly fewer than the 7,942 unique codes in the English corpus.

(iii) **Discharge Summary Length**: On average, the Swedish discharge summaries contain only 12.5% of the tokens found in the English discharge summaries. Additionally, the Swedish notes do not have the structured subsections present in the English notes and instead serve as a single, shorter summary.

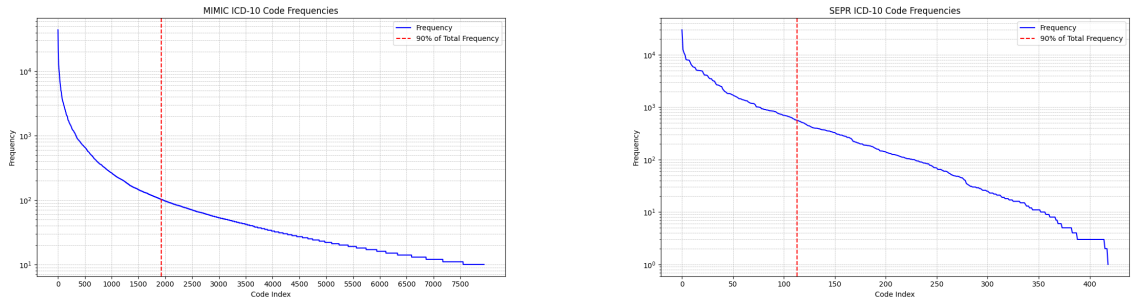
(iv) **Number ICD-10 Codes**: While the Swedish records have 1.09 ICD-10 codes assigned per note on average, the English records represent 15.65 codes on average.

Table 3.2 presents the statistical properties of the complete MIMIC-IV and SEPR II datasets as well as their respective splits, underscoring the differences between the two corpora. Although SEPR II contains approximately two and a half times as many documents as MIMIC-IV, MIMIC-IV includes more than three times the number of tokens, with significantly longer individual documents. This distinction is further emphasized in the code counts, where MIMIC-IV contains over five times the total codes found in SEPR II.

Table 3.2: Statistical comparison of the full English and Swedish datasets and their corresponding splits.

	Total Docs	Total Tokens	AVG Token/Doc	Vocabulary	Total Codes	AVG Code/Doc	Unique Codes
MIMIC-IV Total	**122,278**	**195,297,607**	**1,597**	**186,546**	**1,914,237**	**15.65**	**7,942**
MIMIC-L	89,098	141,991,892	1594	158,308	1,389,330	15.59	7,939
MIMIC-M	19,802	31,880.837	1610	63,750	210,771	15.76	7,935
MIMIC-S	13,378	21,424,878	1602	63,750	210,771	15.76	7,906
SEPR II Total	**317,750**	**63,401,085**	**200**	**532,954**	**347,188**	**1.09**	**419**
SEPR-L	237,968	47,496,536	200	455,728	259,778	1.09	419
SEPR-M	47,783	9,543,881	200	191,595	52,419	1.10	413
SEPR-S	32,027	6,365,636	199	153,931	35,031	1.10	411

Figure 3.4 illustrates the frequency distribution of ICD-10 codes within the MIMIC-IV (Figure 3.4a) and SEPR II (Figure 3.4b) datasets. It shows that for both datasets, a small subset of codes accounts for a large proportion of the overall code frequencies. The red line indicates how many codes contribute to 90% of all code appearances. For both datasets, this corresponds to only around one-fourth of all codes with the other three-fourths of codes distributed across the remaining 10%. Thus, we observe highly imbalanced frequencies for both the English and Swedish datasets.



(a) MIMIC-IV Code Frequencies

(b) SEPR II Code Frequencies

Figure 3.4: Distribution of codes frequencies for MIMIC-IV (left) and SEPR II(right). The logarithmic frequencies are shown on the y-axis with the code indices on the x-axis.

It is important to consider the differences between the corpora when analyzing and interpreting the results. A direct comparison of real and synthetic data is meaningful for each language individually, while any comparison between the two languages should be made with caution, taking the discussed differences into account.

3.2 Synthesizing Medical Notes

LLaMA-3.1-8B serves as the base model for generating synthetic datasets in this study. Meta's LLaMA 3.1 is recognized as a leading model family in SOTA autoregressive LLMs, demonstrating high performance across diverse tasks and competing with top models like GPT-4 (Vavekanand & Sam, 2024). LLaMA-3.1 is available in three configurations of 8 billion, 70 billion, and 405 billion parameters, trained on a multilingual dataset of over 15 trillion tokens, providing robust support for eight languages and context lengths up to 128,000 tokens. Being open source, the LLaMA models are broadly accessible for research applications. For this study, we selected the smallest version, LLaMA-3.1-8B, which offers a balance of computational efficiency and strong benchmark performance, making it suitable for environments with limited computational resources.

To adapt LLaMA-3.1-8B for generating synthetic medical notes, transfer learning was applied. We propose instruction-tuning the model using prompts containing textual descriptions of ICD-10 codes. Instruction-tuning involves fine-tuning the model to follow specific instructions, which we expect to yield several benefits for synthetic data generation and address previously reported challenges. Specifically, our approach aims to achieve the following:

(i) **Content control**: Prompting with ICD-10 codes allows direct control over the content of the medical notes, enabling generation within specific medical domains by guiding note content through targeted codes.

(ii) **Variety**: Synthetic datasets often face issues with diversity reduction, leading to limited utility of generated data. By prompting with varied ICD-10 codes, we can encourage greater variety in the synthetic notes.

(iii) **Automatic Labeling**: Including ICD-10 codes in the prompt means that the synthetic notes are automatically annotated, providing ready-made labels for subsequent supervised fine-tuning of downstream models.

The instruction template used in this study follows the well-known Alpaca format, originally introduced by Taori et al. (2023), which has demonstrated effectiveness for instruction-tuning in prior research. The template has the following structure:

```
Below is an instruction that describes a task, paired with an input that
provides further context.  Write a response that appropriately completes
the request.

### Instruction:
{instruction}

### Input:
{input}

### Response:
```

In this setup, the `Input` field contained textual descriptions of ICD-10 codes, while the `Response` field entailed the corresponding synthetic medical note. To identify an effective instruction for high-quality data generation, three different prompts were manually created aiming at a clear and precise formulation of the requirement. These prompts were then presented to ChatGPT (OpenAI, 2024), asking to select the best option and make necessary refinements. This approach was chosen based on research indicating that LLMs can generate effective instructions for specific tasks (Honovich et al., 2023). The final instruction used in the framework is shown in Figure 3.5. Future work could explore the potential of experimenting with different instruction prompts to improve data generation quality.

> Given a list of textual descriptions of procedure and diagnosis codes, generate a corresponding clinical discharge summary that provides comprehensive and relevant details about the patient's medical history, current condition and treatment received at the hospital.

Figure 3.5: English instruction used for instruction-tuning LLaMA-3.1-8B with the Alpaca template.

The fine-tuning process for this study was implemented using the Axolotl project (OpenAccess-AI-Collective, 2024), an open-source tool tailored for fine-tuning LLMs. Axolotl supports a variety of model architectures and configurations, along with easy integration of parameter-efficient, performance-enhancing techniques. Given limited computational resources, we applied 4-bit quantization with Quantization Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023). Unlike full model fine-tuning, QLoRA trains only small, low-rank adapter layers inserted into the model, using 4-bit quantization to compress model weights. QLoRA has been shown to effectively reduce memory usage while maintaining comparable model performance and is a popular method used for parameter-efficient fine-tuning (PEFT) (Xu et al., 2023). To further optimize instruction-tuning, DeepSpeed ZeRO Stage 3 (Rajbhandari et al., 2020; Rasley et al., 2020) was integrated, enabling efficient multi-GPU training by sharding optimizer states, gradients, and model parameters across parallel workers. Detailed training configurations are provided in Table C.1 in the Appendix.

For fine-tuning on the Swedish dataset, the same Axolotl framework was employed, with minor modifications in the source code to accommodate Swedish-specific prompts. The Swedish version of the adapted Alpaca template and associated instructions are included in the Appendix, Chapter B. The medical notes and ICD-10 codes from MIMIC-L and SEPR-L were utilized to transform the ICD-10 codes into descriptive text and integrate codes as input and medical notes as response into the Alpaca template. These prompts were then used for instruction-tuning LLaMA.

After instruction-tuning, the trained LoRA adapters were merged back into the base model for inference. For this stage, we used the open-source vLLM library (vllm-project, 2024), which optimizes attention memory handling via Paged Attention, significantly boosting inference speed (Kwon et al., 2023). This increase in efficiency is especially beneficial given the need to generate a large number of lengthy documents in this study. Configuration details for vLLM, including decoding parameters, can be found in Table C.2 in the Appendix.

To generate synthetic English datasets, we used textual ICD-10 code descriptions from MIMIC-S and MIMIC-L to create prompts for decoding. Following the methodology of Edin et al. (2023), MIMIC-M was designated as the test set for the utility evaluation, where only real data is required. To promote diversity and facilitate later filtering, five unique responses were generated for each prompt using random sampling. Similarly, for the Swedish datasets, two synthetic datasets were generated from prompts extracted from the ICD-10 code sequences of SEPR-M and SEPR-L, with the real SEPR-S dataset set aside for testing.

3.3 Assessment Methods

The following chapter introduces the various assessment methods used to evaluate the generated synthetic notes, focusing on fidelity, privacy, utility, and medical coherence.

3.3.1 Fidelity Evaluation

To evaluate the resemblance of the synthetic data to the real data, we compare key statistical features of both datasets, including token counts, sentence counts, and vocabulary size. This comparison is conducted using the synthetic and real MIMIC-S datasets for English and the synthetic and real SEPR-M datasets for Swedish. These statistical analyses provide initial insights into whether the synthetic data aligns with the distribution of the real data. Additionally, the unique token ratio offers an indication of the dataset's variety and helps assess whether it suffers from diversity reduction.

In addition, we perform a manual investigation of the synthetic MIMIC-S dataset to evaluate whether the synthetic data accurately captures the structure of real discharge summaries, including the representation of individual subsections. This manual review also aims to identify linguistic and contextual similarities and differences between the synthetic and real documents. It is important to note that this investigation was conducted without

the assistance of medical professionals and is intended to provide initial descriptive insights into the data rather than a detailed content-focused evaluation.

3.3.2 Privacy Evaluation

Operationalizing privacy in datasets poses significant challenges, with no established consensus on evaluation metrics in the literature (Kaabachi et al., 2023). Key difficulties include identifying dataset features that contribute to privacy risks, measuring the impact of these features on privacy, and defining thresholds that establish a dataset as private. DP is one prominent approach, designed to limit the influence of any single data point in a dataset to reduce re-identification risks. However, as discussed in Chapter 2.3, DP often comes at the cost of utility loss and additional computational demands.

Both the English and Swedish datasets used for this work underwent pseudonymization prior to synthetic generation, adding an extra layer of protection to the framework and minimizing the risk of exposing sensitive PHI. Consequently, the risk of privacy violations is naturally low, and no additional privacy-preserving measures were incorporated into the process. With this, we prioritize simplicity in our framework and allow the synthetic data generation process to retain higher utility without incurring the performance costs associated with DP.

Since privacy preservation is a core goal of generating synthetic medical notes, it is nonetheless essential to assess the extent of privacy protection achieved in the synthetic data. The use of distance metrics is a common approach to measure the similarity between synthetic and real data, which serves as a proxy for re-identification risk. Following the methodology in Libbi et al. (2021), we employ ROUGE-5 (Lin, 2004) for MIMIC, to calculate the 5-gram overlap between synthetic and real training data. As a baseline comparison, we also calculate the 5-gram overlap between the real MIMIC-S and MIMIC-L datasets. We compare our results to Libbi et al. (2021) by reporting average, median, minimum, and maximum recall scores for the whole datasets and the 122 document pairs with the highest proximity. To gain further insights, the medical notes with the 20 highest recall scores are then manually reviewed to determine if these similarities pose privacy concerns or if they are due to general, non-identifying content.

Additionally, ROUGE-5 will be used to assess whether using the same code sequences from the model's fine-tuning data for generation is justifiable. In most cases, data used in testing (in this case, synthetic data generation) should not overlap with training data to ensure reliable performance metrics and avoid data repetition, or in our case the risk that the model is merely copying sequences from its training data. By comparing ROUGE-5 scores between a synthetic hold-out set (MIMIC-S) and a subset of the synthetic MIMIC-L documents, we can assess whether the reuse of code sequences from the training data in inference leads to unacceptable levels of similarity or if it can be employed without raising additional privacy concerns over the use of new code sequences.

For the Swedish data, we follow the methodology of Hiebel et al. (2023) by calculating 8-gram overlaps between synthetic and real data to assess the risk of the model reproducing long sequences from its training data. Specifically, we compare the synthetic SEPR-M dataset to the real SEPR-L dataset. As a baseline, we also calculate 8-gram overlaps

within the real SEPR-M and SEPR-L datasets. Additionally, we compare our findings to the 8-gram overlap results reported by Hullmann and Hansson (2024), who used the same metric to evaluate privacy in their synthetic Swedish medical notes. While these similarity scores provide an initial indication of privacy risk, they are difficult to interpret without manual review to confirm if overlapping sequences pose privacy concerns. Due to limited proficiency in Swedish, no manual investigation was conducted, leaving it as a direction for future research.

3.3.3 Utility Evaluation

As discussed in Chapter 2.3, various downstream tasks are used to evaluate synthetic data. Automatic medical coding, in particular, is a highly complex task due to the large number of ICD-10 codes involved. To accurately capture the contents of a medical note and map it to a list of ICD-10 codes, the note must exhibit a certain level of medical coherence, particularly in terms of combining related diagnoses and procedures. Previous research has shown that medical coherence is not as critical for other tasks, such as NER (Hullmann & Hansson, 2024; Libbi et al., 2021). Therefore, medical coding is likely a more suitable task for assessing the overall utility of synthetic notes, as it reflects the quality of the notes in capturing relevant medical information.

Furthermore, our proposed methodology provides a synthetic dataset annotated with ICD-10 codes, making it directly applicable for the supervised training of medical coding systems without the need for further adaptation. Since two previous studies employed similar methodologies and datasets for synthetic medical note generation (Hullmann & Hansson, 2024; Mawaldi & Mladenov, 2024), we also include downstream utility tasks of clinical NER for the English dataset and PHI NER for the Swedish dataset, as used in those works. This allows for a comparison with prior results and an evaluation of whether the proposed method, which focuses on medical coding, can also support utility for other downstream tasks, thereby demonstrating its generalizability in substituting real data.

Medical Coding
For training the medical coding systems, the PLM-ICD framework introduced by Huang et al. (2022), as detailed in Chapter 2.2.2 was used. We followed the adapted implementation of Edin et al. (2023), who provide a reproducible medical coding framework compatible with different model architectures trained on MIMIC-IV and the earlier MIMIC-III dataset. The base model for the English dataset is RoBERTa-PM, a pretrained model that achieved the best results in a comparison by Huang et al. (2022).

For the Swedish data, we used SweDeClin-BERT (Vakili et al., 2022) as the Swedish medical pretrained base model. Slight modifications were made to adapt the preprocessing pipeline, enable training on Swedish data, and address occasional model collapse issues reported by the original authors. All other settings and configurations were adapted from Edin et al. (2023), including truncating longer documents to 4000 tokens and tuning thresholds, both of which have been shown to enhance performance. In line with Edin et al. (2023), we report nine evaluation metrics, listed and described in Table 3.3.

Table 3.3: Description of the different performance metrics reported for the medical coding models.

Metric	Description
Micro & Macro AUC-ROC	Area under the receiver operating characteristic curve: Probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one.
Micro F1	Harmonic mean of Micro Precision and Micro Recall.
Macro F1	Arithmetic mean of Macro Precision and Macro Recall.
EMR: Exact match ratio	Percentage of instances where all codes were predicted correctly.
Precision@k: Precision@8 and Precision@15	Micro Precision among top 8 and 15 predicted codes.
Precision@R / R-Precision	Micro Precision among top k codes, where k equals the true number of relevant codes.
MAP: Mean average precision	Micro Precision considering the exact rank of all relevant codes in the document.

Edin et al. (2023) run every model ten times to report mean performance and standard deviation. To save time and computational resources the runs were limited to three times in this work. To identify significant differences in model performance, two-sample t-tests were conducted for all metrics with a determined significance level of $\alpha = 0.05$.

To ensure that the results reported in Edin et al. (2023) can be reproduced and to validate the implementation, the model was initially trained on MIMIC-L and evaluated on MIMIC-M using the original preprocessing methods. Next, the model was trained on a version of MIMIC-L that had been preprocessed using the method described in Chapter 3.1.1 (i.e., MIMIC-L Short), to confirm that shortening the documents does not adversely affect performance. For all subsequent experiments, either the shortened real data or synthetic data generated from the shortened data were used. MIMIC-M was consistently employed as test set to ensure comparability across experiments.

To save computational resources and allow for further experiments, the model was also built upon MIMIC-S. Both synthetic MIMIC-L and MIMIC-S datasets were utilized as training data to compare performance. Additionally, the following sub-experiments were conducted to gain further insights into the utility of synthetic discharge summaries in training medical coding systems:

(i) **Filtering**: As described in Chapter 3.2, five different medical notes were generated for each prompt during decoding. Filtering training datasets has been proven to enhance model performance, especially in medical applications where high-quality data is essential for accurate results (e.g., Moore & Lewis, 2010). Since the medical notes used as training data for medical coding must contain sufficient information about each code, applying an existing medical coding system may be useful to assess whether the generated notes adequately cover the required codes. To achieve this, we used the PLM-ICD model provided by Edin et al. (2023) on each generated note from the synthetic MIMIC-S set. We then filtered out the notes with the highest and lowest Micro F1 scores from the five generated responses for each prompt, creating two new filtered training datasets.

(ii) **Balancing**: The imbalance of codes in medical datasets is a common issue when training medical coding systems, often resulting in suboptimal performance for less frequent codes (Edin et al., 2023; Lamproudis et al., 2024). To address this issue, two new training datasets were created. The first dataset ensures completely balanced code frequencies across all ICD-10 codes, while the second dataset adjusts the frequencies by increasing the least frequent codes to a minimum of 10 occurrences

and proportionally decreasing the more frequent codes. Both datasets maintain the original number of codes per note in the MIMIC-S set but with balanced or adjusted frequencies. Note, that the codes with adjusted frequencies were randomly re-ordered without taking into account co-occurrences of single codes.

(iii) **Increasing training size**: To evaluate the effect of training data size on model performance, two approaches were applied to increase the training set based on MIMIC-S. In the first approach, the training set size was doubled and tripled by using the first two and three generated outputs for each prompt. In the second approach, MIMIC-S was expanded by incorporating random subsets of the synthetic MIMIC-L dataset to double and triple the training set size. This not only evaluates the impact of training set size but also investigates whether the increased variety in the generated code sequences influences the model's performance.

The Swedish models were trained on the real and synthetic SEPR-L and SEPR-M datasets and tested on the real SEPR-S dataset. Again, performance differences were compared between the real-data and synthetic-data models calculating two-sample t-tests for each metric.

In a final experiment on medical coding, we evaluated the suitability of LLaMA-3.1-8B for our framework by comparing model performance on training data generated by the instruction-tuned base version of LLaMA to data generated through the integration of two specialized pretrained language models:

(i) **OpenBioLLM-8B:** Developed by Saama, OpenBioLLM is an open-source model built on the foundation of LLaMA-3-8B, specifically adapted for biomedical applications. This adaptation was achieved through extensive fine-tuning on a large biomedical corpus, along with Direct Preference Optimization and medical instruction-tuning to enhance its relevance to the biomedical field (Ankit Pal, 2024).

(ii) **AI-Sweden Models/LLaMA-3-8B** This model, released by AI Sweden, is a version of LLaMA-3-8B, specialized through continuous pretraining on a subset of the Nordic Pile, a dataset with over 227 billion tokens in Swedish, Danish, and Norwegian. The goal of this adaptation is to optimize the model for Nordic languages (AI Sweden Models, 2024).

The integration of OpenBioLLM aims to assess whether the framework, fine-tuning the base LLaMA model on medical notes, is sufficient for medical domain adaptation or if additional pretraining on biomedical data further enhances the quality of generated synthetic medical records. Similarly, AI Sweden's adapted LLaMA model helps determining if the base model, fine-tuned within our framework, can effectively handle Swedish medical texts, or if additional pretraining on Swedish is required. Since Meta has not specified the model's proficiency in Swedish, it is unclear whether Swedish was included in the pretraining data for the base model, or to what extent.

To better understand the differences in model performance between real and synthetic data, we conducted an error analysis focusing on several aspects that provide valuable insights. First, we analyzed the predicted codes from a general statistical perspective, such as calculating the total number of unique predictions, the average number of predictions per

document, and the distribution of predictions across ICD-10 chapters. Second, we assessed the impact of code frequency in the training data and document length in the test data on the F1 scores, as both factors are commonly reported to influence model performance. Third, we examined within-family (WF) and out-of-family (OOF) errors, i.e., the number of incorrect predictions that still fall within the same ICD-10 chapter, to investigate the nature of these errors and determine whether the models have potentially learned more than is apparent from the other reported metrics. Finally, we trained medical coding models on real data with artificially inserted noise and compared them to the synthetic-data model to better understand where differences in the utility between real and synthetic data might stem from.

NER

In addition to medical coding, NER was used as a downstream task to assess the utility of the synthetic data. This choice was motivated by two main reasons: First, evaluating synthetic data on multiple downstream tasks provides a more comprehensive assessment of its general utility. Second, NER is widely used for utility evaluation in synthetic data research, enabling direct comparison with previous studies. Specifically, two studies similar to ours were selected as baselines, and their methodologies were replicated to facilitate result comparison.

For English, we used the framework from Mawaldi and Mladenov (2024) as a baseline. In their study, synthetic medical notes were created by fine-tuning LLaMA-2-7B to generate the *History of Present Illness* section from the *Chief Complaint* using data from the MIMIC III dataset. While their approach shares similarities with ours in using earlier versions of both the model and dataset, their method of transfer learning differs, as they incorporated part of the medical note in the prompts, whereas we used ICD-10 code transcriptions and generated only a subsection of the full note, whereas we generate the complete discharge summary. To evaluate the utility of their synthetic data, they annotated both real and synthetic subsets of 5,000 documents each using Med7 (Kormilitzin et al., 2021) for drug name identification and Stanza (Qi et al., 2020) for disease identification. These labeled datasets were then used to fine-tune BERT base (Devlin et al., 2019) for clinical NER with drug and disease labels. Their evaluation metrics included accuracy, F1-score, precision, and recall. Following this approach, we extracted subsets of 5,000 documents from both the real and synthetic MIMIC-S datasets to implement clinical NER as a downstream task for utility evaluation.

For Swedish, the implementation of Hullmann and Hansson (2024) was followed. They employed SEPR II, the same corpus used in this work, to generate synthetic data by fine-tuning KB-BART (KB Lab, 2023), using the first sentence of a medical note as a prompt to generate the remainder of the note. Their approach differs from ours in that they used an encoder-decoder model, whereas we used a decoder-only model and prompted with the beginning of the note where we used ICD-10 codes. However, their objective of generating full synthetic medical notes resembling those in SEPR II makes their results highly comparable. For utility evaluation, they conducted PHI NER as a downstream task, using SweDeClin-BERT (Bridal et al., 2022) to annotate real and synthetic datasets of 26,023 notes each with nine PHI tags. These annotated datasets were then used for supervised fine-tuning of SweDeClin-BERT, with the SEPR PHI Pseudo Manual Corpus (Velupillai et al., 2009), containing a manually annotated subset of 300 documents, serving as the evaluation gold standard. Evaluation metrics included accuracy, precision, recall,

and F1-score. In our work, we followed their methodology by extracting a subset from the synthetic SEPR-M dataset, matching their token count of 1,187,380 tokens, annotating them with SweDeClin-BERT NER, and using this data to fine-tune SweDeClin-BERT.

In summary, this work focuses on utility evaluation primarily through medical coding, implemented based on the approach from Edin et al. (2023). Additionally, by training NER models as a secondary downstream task, we aim to provide broader insights into the generalizable utility of the synthetic data. Moreover, the NER evaluation enables direct comparison with previous research on synthetic medical note generation, specifically with the clinical NER models of Mawaldi and Mladenov (2024) for English data and the PHI NER models of Hullmann and Hansson (2024) for Swedish data.

3.3.4 User Study

As discussed in Chapter 2.3, previous research suggests that training data can be effective for training downstream systems even when it lacks grammatical accuracy or topical coherence (Hiebel et al., 2023; Libbi et al., 2021). However, while some tasks may not require coherent narratives, others might depend on accurate medical information to function effectively. We argue that medical coherence is essential for synthetic data to serve as a viable substitute for real data across diverse applications.

The importance of grammatical correctness is more complex. Real EHRs often contain grammar errors, such as spelling mistakes, incomplete sentences, and incorrect punctuation (Lai et al., 2015). While complete grammatical accuracy in synthetic data could aid comprehension and processing, it could also create an unnatural distinction from real data. If desired to accurately mirror the stylistic nuances of real medical records, such imperfections should be therefore preserved in the synthetic data as well.

Quantitatively measuring medical coherence is challenging, as it requires profound medical expertise. To address this, a user study was conducted to evaluate the readability and medical coherence of a sample of synthetic and real document pairs corresponding to the same code sequences. Medical professionals were recruited as participants and asked to rate both synthetic and real notes on two criteria: readability and medical coherence. The ratings were on a bipolar scale from 1 to 5, with 1 indicating the lowest level of quality and 5 indicating the highest. Specifically, for readability, 1 meant *not natural at all* and 5 meant *completely natural: could be written by a doctor*. For medical coherence, 1 indicated *not coherent at all* and 5 meant *Perfectly coherent: Symptoms, diagnosis, procedures, etc. fit together perfectly*. The samples were presented in random order to eliminate any sequence-based biases.

The participants were unaware that they were being presented with partially synthetic medical notes. After each rating, and at the end of each note, they had the opportunity to provide justifications for their ratings and comment on any unusualities they observed during their review of the notes.

This study allows for an evaluation of the readability and medical coherence of the synthetic discharge summaries from a professional perspective, offering valuable insights into how easily synthetic notes can be distinguished from real notes in their quality. It also

provides indirect information about the potential for using these synthetic notes in real-world applications. If participants identified medical incoherencies or unusual elements not found in the real data, it would suggest that the quality of the synthetic data may not yet be sufficient for use in building clinical decision support systems that can be deployed in practical scenarios.

In addition to rating readability and medical coherence, participants were asked to perform manual diagnosis coding on the medical notes. This task served two purposes: first, to assess whether the presented notes accurately reflect the diagnoses provided as prompts in the text, and second, to examine the level of inter-annotator agreement between participants and their performance in comparison to medical coding models. This task was only presented to participants who confirmed to be experienced in medical coding, stating to perform this task *often* or *very often*.

The study samples were randomly selected from the dataset, with the MIMIC samples ranging from 300 to 600 words in length and the SEPR samples ranging from 100 to 200 words. This length range was chosen to balance the participants' reading effort while ensuring the documents were long enough to provide enough context and pose as a representative sample. Following pretests with five independent participants, whose feedback addressed issues like task clarity, potential ambiguities, and time requirements, the English questionnaire was reduced to three real-synthetic document pairs, while the Swedish questionnaire was shortened to four pairs. This adjustment ensured that study participation would take approximately 30 minutes. Since medical expertise was a key requirement for participation, all participants had to confirm that they possessed such expertise by being engaged in a profession in the medical field before starting the study. The questionnaire following the presentation of each medical note is included in the Appendix (see Chapter D).

3.4 Experimental Setup

The Department of Computer and System Sciences (DSV) at Stockholm University provided the study setting for this work. All experiments were conducted on a DSV-provided server equipped with four NVIDIA RTX A5000 graphical processing units (GPUs), each offering 24GB of RAM (NVIDIA, 2024). The server did not have internet access to ensure data security and control over the experimental environment.

3.5 Ethical Considerations

Both the MIMIC and SEPR corpora contain sensitive clinical data that require careful ethical consideration in their use. Relevant privacy agreements were signed prior to gaining access to the datasets. Additionally, an ethics training program provided by CITI, including the courses *Data or Specimens Only Research* and *Conflicts of Interest*, was completed before access to the MIMIC dataset was granted. All data protection regulations, including the General Data Protection Regulation (GDPR) (European Parliament & Council of the European Union, 2016) and the Health Insurance Portability and Accountability

Act (HIPAA) (Centers for Medicare & Medicaid Services, 1996), were strictly adhered to throughout the research. The data used in this study was pseudonymized minimizing the risk of data leakage, and only the data relevant to our experiments was processed. All experiments were conducted on a server provided by DSV, which was isolated from the internet. All models used in this study were loaded, trained, and saved locally on this server, ensuring no data was transmitted to online APIs. This research has been approved by the Regional Ethical Review Board in Stockholm under permission no. 2007/1625-31/5. While we do our best to protect patient privacy and maintain medical correctness in our synthetic notes, we do not recommend the distribution or use of any generated data or models trained on this data in real-world settings without further privacy evaluations.

Recognizing the environmental impact of using computationally intensive LLMs and energy-demanding GPUs, efforts were made to minimize the carbon footprint of this study. This was achieved by incorporating PEFT methods and limiting computational resources to the minimum required for the experiments.

4 Results

This chapter presents the results from the various assessment methods applied to the synthetic data.

4.1 Fidelity: Statistical Comparison and Manual Investigation

To assess the fidelity of the synthetic data, i.e., how closely it resembles the real data, a range of statistical features were compared between the real and synthetic corpora. For the English data, MIMIC-S was used as the baseline for comparison, while SEPR-M served as the reference dataset for the Swedish data. In the Appendix, examples of synthetic medical notes are provided for both English (see Figure E.1) and Swedish (see Figure E.2).

Table 4.1: Statistical comparison of the real and synthetic MIMIC-S datasets.

Statistical Comparison real vs. synthetic MIMIC-S								
	Total Docs	AVG Sent/Doc	AVG Token/Doc	AVG Token/Sent	Total Sent	Total Tokens	Unique Tokens	Unique Ratio
Real	13,378	72.4	1,285.5	17.7	969,065	17,197,361	96,380	0.006
Synth	13,378	79.34	1,638.8	20.66	1,061,409	21,923,740	233,845	0.011

Table 4.1 compares key statistical features between the real and synthetic MIMIC-S datasets. Notably, the synthetic data exhibits a greater total count of tokens and sentences, with sentences being, on average, three tokens longer in the synthetic dataset. Furthermore, synthetic documents contain, on average, seven more sentences per document, making them significantly longer than the real documents. The synthetic dataset also has a notably larger vocabulary, with over 59,000 more unique tokens than the real data. While the overall higher token count in the synthetic dataset accounts for the increase, the ratio of unique words is also higher in the synthetic data at 1.1%, compared to 0.6% in the real data.

This suggests that the fine-tuned LLaMA model can generate content in the synthetic discharge summaries that it did not encounter during fine-tuning, resulting in a synthetic dataset with greater variety than the original. This finding is particularly noteworthy since previous research has often reported reduced variety and a smaller vocabulary in synthetic datasets (Hullmann & Hansson, 2024; Libbi et al., 2021; Mawaldi & Mladenov, 2024). These limitations in variety can hinder downstream performance, so the increased vocabulary size observed in this study may help mitigate this issue.

Table 4.2: Statistical comparison of the real and synthetic SEPR-M datasets.

Statistical Comparison real vs. synthetic SEPR-M								
	Total Docs	AVG Sent/Doc	AVG Token/Doc	AVG Token/Sent	Total Sent	Total Tokens	Unique Tokens	Unique Ratio
Real	47,783	13.96	199.93	14.33	666,826	9,553,204	191,578	0.02
Synth	47,783	13.88	252.32	18.18	663,164	12,056,615	391,152	0.03

The statistical comparison of the real and synthetic SEPR-M datasets for the Swedish data is presented in Table 4.2. Overall, the comparison aligns closely with that of the MIMIC data. Unlike the synthetic MIMIC dataset, however, the synthetic SEPR dataset mirrors the real SEPR dataset in terms of the average number of sentences per document and contains a slightly lower total sentence count. Nonetheless, the synthetic SEPR dataset features longer sentences, with an average of four additional tokens per sentence, resulting in a higher total token count and a significantly larger vocabulary. Specifically, the synthetic data includes 1% more unique tokens than the real dataset. This suggests that our proposed method consistently introduces more variety into the synthetic data than prior approaches and even surpasses the variety present in the real data.

In addition to the statistical analysis, a manual review of the synthetic dataset was conducted to examine key characteristics and assess its resemblance to the real data. Due to language constraints, this investigation was limited to the MIMIC dataset. Table 4.3 summarizes the notable similarities and differences identified, along with illustrative examples.

Table 4.3: Some examples of key findings of the manual investigations in the synthetic MIMIC-S data.

Similiarities	
Abbreviations	[...] **PMH** past **MI**, **HTN**, **Afib** on **ASA** verapamil **CCB** digoxin [...]
Spelling Mistakes	[...] that was **suspcious** for [...] , he was **suppose** to undergo [...]
Pseudonymization	Mr. ___ is a ___ yo M admitted to the Acute Care Surgery Service on ___ with abdominal pain.
Differences	
Repetitions	[...] but no pain or **nausea nausea** [...]
Inconsistencies	[...]Patient is a ___ yo **M. She** denies any dizziness [...]
Hallucinations	Given this and the fact that patient is a **poor historian**, a follow up MRI ordered for ___.

The investigation found that the synthetic data successfully captures several core attributes of the real dataset. Structurally, the synthetic documents replicate the general format of discharge summaries, including distinct sections described in Chapter 3.1.1. The sentence style mirrors that of the original data, characterized by concise, sometimes incomplete sentences focused on delivering precise information without digression. Disease mentions appear naturally embedded in the notes, mostly without explicit repetition of their full names as provided in the prompt.

A closer examination revealed that the synthetic documents contain many abbreviations, along with occasional grammatical errors and spelling mistakes, reflecting patterns also

present in the training data. Given that discharge summaries typically include numerous abbreviations and errors (Dalianis, 2018; Lai et al., 2015), it is expected that these characteristics would persist in the synthetic data unless addressed during preprocessing. Additionally, the synthetic data seems to replicate the masked pseudonymization of MIMIC, as shown in the example provided in Table 4.3.

Despite these similarities, the synthetic data also displays certain characteristics that are more indicative of typical LLM-generated text. For instance, repetition is occasionally observed, either in isolated words (as shown in Table 4.3) or across longer passages spanning several sentences, which can create an unnatural flow. LLMs are also prone to generating hallucinations, where content may be factually incorrect, irrelevant, or fabricated (Z. Ji et al., 2023). While coherence issues will be further explored in the user study, some irrelevant sections have already been identified. For example, one document mentions an MRI ordered because the patient is described as a "poor historian", which seems contextually odd and irrelevant. Thus, this example likely illustrates an LLM hallucination. Additionally, while a detailed assessment of medical inconsistencies will be covered in the user study, some minor inconsistencies, such as mismatched gender references, were noted during this review.

Overall, the fidelity assessment shows that, while the synthetic data captures many characteristics of the real datasets, there are also notable distinctions. The statistical comparison reveals that synthetic documents generally contain more tokens. This feature could be adjusted during decoding if necessary. A positive differentiation is the increased variety, reflected in the larger vocabulary size, which appears to address limitations observed in prior approaches. Although preliminary findings indicate that LLM-related artifacts, such as repetitions and hallucinations, are relatively uncommon, their presence could still influence the synthetic data's utility. Issues related to medical coherence and readability will be further examined in Chapter 4.5.

4.2 Privacy: Similarity

To evaluate whether the generated synthetic datasets preserve privacy, two different similarity metrics for the English and Swedish datasets were employed. These metrics were used to compare the synthetic data to the training data used in the instruction-tuning process. The goal was to assess the risk of the synthetic data potentially leaking sensitive information from the training data.

4.2.1 MIMIC: ROUGE-5 Recall

To assess the similarity between the synthetic and real MIMIC datasets, ROUGE-5 scores were computed following the methodology of Libbi et al. (2021). Specifically, we compared the synthetic MIMIC-S dataset to the real MIMIC-L dataset and a subset of the synthetic MIMIC-L dataset (equal in size to the MIMIC-S dataset) to the real MIMIC-L dataset. To have a baseline of the proximity within real datasets, ROUGE-5 was additionally calculated between the real MIMIC-S and real MIMIC-L data, The recall scores for the entire dataset

as well as the 122 documents with the highest recall values are presented and compared to the results reported by Libbi et al. (2021) in Table 4.4.

Table 4.4: Comparison of ROUGE-5 Recall Scores: Real MIMIC-S, Synthetic MIMIC-S and a Subset of Synthetic MIMIC-L vs. Real MIMIC-L, compared to Libbi et al. (2021). The results are reported for the full dataset and the 122 document pairs with the highest scores.

	All Real/Synthetic Pairs				Highest 122 Real/Synthetic Pairs			
	AVG	Median	Min	Max	AVG	Median	Min	Max
Libbi et al. (2021)	0.031	0.026	0.000	1.000	0.207	0.143	0.025	1.000
MIMIC-S real	0.138	0.088	0.006	1.000	0.794	0.779	0.727	1.000
MIMIC-S synth	0.097	0.052	0.001	1.000	0.760	0.748	0.690	1.000
MIMIC-L synth	0.068	0.036	0.000	0.900	0.668	0.648	0.587	0.900

The baseline comparison of real MIMIC-S to real MIMIC-L demonstrates significantly higher ROUGE-5 scores than those achieved with both synthetic datasets. The average recall score for the real MIMIC-S dataset stands at 0.138, which is approximately double that of the synthetic MIMIC-S data. Moreover, among the 122 closest document pairs, the real-real comparison yielded the highest scores, though the difference here is less strong. From a privacy perspective, these findings are particularly encouraging. The low similarity of synthetic datasets to the training data, which is even lower than the similarity between real datasets, suggests that the fine-tuned model successfully generates diverse and novel discharge summaries rather than replicating segments from the training data, thus, ensuring high privacy protection.

The synthetic MIMIC-L subset shows overall lower ROUGE-5 scores compared to the MIMIC-S data, indicating that these synthetic documents are, on average, more distant from the real documents. This suggests a lower risk of re-identification, implying that the synthetic data generated from MIMIC-L prompts poses a reduced privacy risk. Thus, this result supports the approach of generating synthetic datasets from MIMIC-L code sequences, i.e. from the same code sequences the LLaMA model was trained on without presenting a higher security risk than when applied to new code sequences.

When compared to the results of Libbi et al. (2021), both synthetic datasets of this work exhibit higher ROUGE-5 scores. However, a direct comparison is not very meaningful, as Libbi et al. (2021) used a different dataset. As discussed in Chapter 3.1.1, the MIMIC documents follow a highly structured format, where certain 5-gram overlaps, such as *Major Surgical or Invasive Procedure*, are expected and even desired to produce realistic data. This is also highlighted in the higher recall scores of the baseline. Therefore, a more detailed investigation is needed to assess privacy concerns more accurately. To further explore this, we conduct a more in-depth analysis of the 20 highest-risk documents from the synthetic MIMIC-S dataset, based on their ROUGE-5 scores.

Table 4.5 demonstrates why a high ROUGE-5 score does not necessarily indicate a privacy concern by showing the total count of occurrences of the overlapping 5-grams found in

the 20 most similar documents in MIMIC-L. None of the identified 5-grams are unique, and most appear frequently, with a median occurrence of 1005 times. This suggests that the presence of overlapping 5-grams does not directly map to specific real documents in the training data. However, Table 4.6 reveals that within these 20 documents, there are long overlapping sequences of up to 216 words that significantly increase the risk of re-identification. A manual investigation of the document pairs yielded two key findings:

(i) No leakage of PHI could be found.

(ii) Long overlapping passages enable mapping to original documents.

Finding (i) was expected, as the original data is already pseudonymized, meaning PHI was already filtered out in the real data. Finding (ii), however, suggests that applying this method to non-pseudonymized data could pose a privacy risk, and adjustments, such as the integration of DP, would be necessary to mitigate this risk. To further reduce the likelihood of re-identification, ROUGE-5 recall scores could be used to filter out documents that are too similar to the original dataset.

However, the average and median ROUGE-5 recall scores for the entire dataset show that most synthetic documents differ substantially from the original data, contributing to the creation of a new and varied synthetic dataset. Given that no identical documents were found between the synthetic and real data, and in light of the significant differences observed, we have opted not to apply additional filtering in this work.

Table 4.5: Total counts of overlapping 5-grams in MIMIC-L, extracted from the 20 most similar document pairs between synthetic MIMIC-S and real MIMIC-L.

AVG	Median	Min	Max
3249	1005	6	88910

Table 4.6: Word lengths of the longest overlapping word sequences in the 20 most similar document pairs between synthetic MIMIC-S and real MIMIC-L

AVG	Median	Min	Max
93	79	19	216

4.2.2 SEPR: 8-Gram Overlap

For the Swedish privacy evaluation, we followed the approach outlined by Hiebel et al. (2023), calculating 8-gram overlap between the real and synthetic datasets. We then compared our results to those of Hullmann and Hansson (2024), who reported 8-gram overlap results for their Swedish synthetic dataset, which was also generated from the SEPR corpus. Table 4.7 presents the 8-gram overlap between the synthetic SEPR-M and real SEPR-L datasets, with the overlap between real SEPR-M and SEPR-L serving as a baseline for comparison. The results are also compared to those reported by Hullmann and Hansson (2024).

Table 4.7: 8-gram overlap between synthetic SEPR-M and real SEPR-L in comparison to baseline and Hullmann and Hansson (2024)

	8-Gram Overlap
Hullmann and Hansson (2024)	0.02442
Synthetic SEPR-M	0.00179
Baseline	0.00488

As shown in Table 4.7, the 8-gram overlap between the synthetic SEPR-M dataset and the real SEPR-L dataset is smaller than the overlap between the real SEPR-M and SEPR-L datasets, as well as the results reported by Hullmann and Hansson (2024). This suggests that the fine-tuned LLaMA model did not simply replicate parts of the training data but generated a new and varied dataset. The lower overlap compared to the real SEPR-M dataset may be attributed to the higher vocabulary and unique token ratio discussed in Chapter 4.1, which could also explain the substantially lower overlap compared to the results of Hullmann and Hansson (2024), who reported a smaller vocabulary in their synthetic dataset.

However, it is important to note that the overlap alone does not allow for conclusions regarding the risk of privacy leakage. These results only indicate that the model successfully generated a dataset that differs from the training data. To assess the privacy risk more thoroughly, further evaluations, such as a manual investigation, would be required to ensure that documents do not undesirably leak private information or allow for direct mappings to real documents. Due to language constraints, we refrain from such an investigation for this work.

Overall, the privacy assessment indicates that the proposed framework can effectively generate new medical notes rather than merely replicating sequences encountered during instruction-tuning. However, longer duplicated sequences were identified in the most similar document pairs in the real-synthetic MIMIC comparison. Additionally, the similarity metrics used in this assessment alone cannot offer a fully reliable privacy evaluation without further investigation. Therefore, integrating additional assessment methods or privacy-preserving techniques into the fine-tuning process would be essential when working with data that has not been pseudonymized in advance.

4.3 Utility: Medical Coding

This chapter presents the results of the utility evaluation for medical coding models trained on synthetic data. We first present the results of models trained on the MIMIC datasets, followed by those trained on SEPR data and data generated using the domain-specific versions of LLaMA.

4.3.1 English Models

The results of the medical coding models trained on the synthetic datasets are presented in Table 4.8, alongside comparisons to models trained on real MIMIC datasets. The real MIMIC-M set serves as test data for all reported experiments. We reproduced the implementation by Edin et al. (2023) using MIMIC-L as training data, and our model results fall within the reported standard deviation for all metrics except Exact Match Ratio (EMR), where we achieved a 0.2% higher score. To confirm that the adapted preprocessing method described in 3.1.1 does not impact model performance, we also trained a model on the preprocessed MIMIC-L Short dataset and compared its results to Edin et al. (2023). Similarly, the model's performance on all metrics except EMR fell within the reported standard deviation, with our preprocessing method yielding a 0.1% higher EMR.

Overall, no significant differences were observed between the performance of our reproduced model, the model trained with our adapted preprocessing method, and the results reported by Edin et al. (2023). This finding supports the assumption that much of the information in discharge summaries is irrelevant to the task of medical coding, justifying the document shortening step during preprocessing and before being used to fine-tune LLaMA. Accordingly, the synthetic training datasets can be considered short versions, since LLaMA was trained on the preprocessed, shortened MIMIC-L dataset before the generation process.

To establish a real baseline using a smaller dataset, enabling more experiments while saving time and computational resources, we trained a model on the preprocessed real MIMIC-S dataset. As expected, this model is outperformed by the model trained on the larger MIMIC-L dataset, aligning with previous findings that the amount of training data plays a crucial role in model performance (Edin et al., 2023). Despite this, the model trained on MIMIC-S achieved a mean micro F1 score of 48.2%, representing a non-trivial performance that serves as a useful comparison point for models trained on synthetic data.

The models trained on the synthetic MIMIC-L and MIMIC-S datasets were compared to their real data counterparts. As shown in Table 4.8, both synthetic models were significantly outperformed by the real-data models across most metrics except EMR for both models and Micro AUC and Macro F1 for the synthetic MIMIC-S model. However, the synthetic models still demonstrate competitive performance. Notably, the model trained on synthetic MIMIC-L outperforms the real-data model trained on MIMIC-S, suggesting that synthetic data has the potential to achieve superior results as the training set size increases. Given that one advantage of the proposed data-generation framework is its ability to produce arbitrary amounts of synthetic data, this finding indicates a promising direction for training models on synthetic data that could potentially surpass SOTA models trained on real datasets.

In the following, the results of the sub-experiments addressing filtering, balancing, and increased training data will be presented.

(i) **Filtering:** The five outputs generated for MIMIC-S were filtered by using them as input to the medical coding model from Edin et al. (2023). Using the outputs with the highest and lowest micro F1 scores, two new training datasets were created to train medical coding models. It was expected that these models would show significantly higher and lower performance, respectively, compared to the synthetic MIMIC-S

model trained on the first output of the decoding stream. However, as shown in Table 4.8, no significant differences were found between any of the three models. Thus, the proposed filtering method cannot be used to improve model performance or to create a dataset suitable for preference tuning. However, the results suggest that the generation model is capable of producing consistent outputs, where each of the first five outputs obtained during decoding represents the prompted ICD-10 codes similarly well.

(ii) **Balancing:** Given that highly unbalanced code frequencies are known to hinder model performance, we created two new synthetic datasets with a more balanced code frequency distribution. These datasets were designed to maintain the same number of training samples as MIMIC-S and match the number of codes per document. One dataset was fully balanced, while the other adjusted the lowest frequency codes to have a minimum occurrence of ten. Both datasets were used as training data. The performance metrics in Table 4.8 show that the model trained on the fully balanced dataset achieved near-zero scores for most metrics, failing to predict any code with confidence above the threshold. Similarly, the model trained on the dataset with a minimum frequency of ten exhibited very low performance, significantly worse than the model trained on synthetic MIMIC-S data. While this model managed to predict codes with confidence above the threshold, nearly all predictions were incorrect.

The complete balancing of codes is expected to underperform on an imbalanced test set, as high-frequency codes typically contribute significantly to the performance of classification systems. Another potential issue is the random distribution of codes used to create these datasets. Certain disease codes naturally co-occur, while others may result in rare or incompatible combinations. A potential solution could involve adding a step to the dataset creation process where medical expertise or statistical patterns are used to ensure that codes are recombined in a more meaningful manner. Overall, our approach to balancing code frequencies failed to improve model performance.

(iii) **Increasing training data:** To evaluate the effect of training data size, the MIMIC-S dataset was doubled and tripled in two different ways. First, by using the second and third outputs generated during inference for the same prompts, i.e., using the same code sequences with multiple outputs (referred to as "Dupli"). Second, by augmenting MIMIC-S with documents from the synthetic MIMIC-L dataset, selecting random subsets of the same document number as MIMIC-S (referred to as "New"). The results show a noticeable increase in performance metrics when doubling the training size using both approaches, with a higher performance increase when supplementing MIMIC-S with new data. When tripling the dataset size, the performance continues to improve with the inclusion of new data even though this improvement is very subtle. When using duplicated code sequences, the model performance does not further increase when tripling instead of doubling the data set size. For both dataset sizes, the model incorporating new code sequences significantly outperforms the model duplicating code sequences across almost all metrics. These results show on the one hand the importance of varied samples in the training data and on the other hand that increasing the size of training data has a large positive effect when working with smaller data sets, which is more difficult to achieve the larger the data set is.

The results of the synthetic and real MIMIC medical coding models overall demonstrate that synthetic documents generated from MIMIC and LLaMA-3.1-8B have the potential to serve as effective training data for creating high-performing models. However, the models trained on synthetic data still do not perform at the same level as those trained on real MIMIC data. A promising outcome is that model performance improves with increasing dataset size, which should be explored further in future work. Additionally, the issue of imbalance in the training dataset could theoretically be addressed using synthetic data. While our approach to balancing code frequencies did not yield successful results, we encourage future work to further explore this possibility.

Table 4.8: Results of medical coding model trained on different MIMIC datasets. Significance: *p<0.05, **p<0.01, ***p<0.001. The color of the asterisk indicates the reference model used to test for statistical differences and is additionally shown next to the model name..

	Classification					Ranking			
	AUC-ROC		F1		EMR	Precision@k		R-precision	MAP
Training Data	Micro	Macro	Micro	Macro		8	15		
Edin et al. (2023)	99.2±0.0	96.6±0.2	58.5±0.7	21.1±2.3	0.4±0.0	69.9±0.6	55.0±0.6	57.9±0.8	61.9±0.9
Real MIMIC-L	99.2	96.6	58.9	22.8	0.6	70.3	55.4	58.4	62.5
Real MIMIC-L Short	99.2±0.0	96.6±0.1	58.7±0.4	22.8±1.8	0.4±0.0	70.0±0.3	55.1±0.4	58.0±0.4	62.0±0.5
Real MIMIC-S Short	96.9±0.4	83.5±1.7	48.2±1.7	3.8±0.8	0.1±0.0	59.7±1.7	45.1±1.7	46.0±1.9	46.4±2.3
Synth MIMIC-L*	**98.9±0.0	**94.8±0.0	*54.8±0.8	*15.9±0.7	0.3±0.1	**65.5±1.0	*50.9±1.1	*53.9±0.8	*56.9±1.0
Synth MIMIC-S*	95.4±0.2	*77.0±0.9	*37.6±0.6	2.1±0.3	0.1±0.0	*48.3±0.6	*35.5±0.6	*35.8±0.7	*34.8±0.7
MIMIC-S Best F1*	95.5±0.4	77.3±1.7	36.6±1.9	2.1±0.4	0.0±0.0	46.6±2.6	34.5±1.8	34.7±2.0	33.5±2.3
MIMIC-S Worst F1*	95.7±0.3	78.4±1.3	37.8±1.5	2.2±0.3	0.0±0.1	48.4±1.8	35.8±1.3	36.1±1.4	35.1±1.7
MIMIC-S Balanced*	**48.4±3.0	***50.0±0.1	**0.4±0.0	**0.2±0.1	0.0±0.0	***0.2±0.2	***0.1±0.1	***0.1±0.1	***0.3±0.0
MIMIC-S Min10*	*89.1±1.5	***50.5±0.5	***17.7±0.1	***0.1±0.0	0.0±0.0	***24.5±0.0	***18.1±0.3	***18.4±0.0	***14.7±0.2
MIMIC-S 2x Dupli	97.8±0.1	88.3±0.2	50.1±0.7	8.1±0.7	0.2±0.1	61.7±0.8	47.2±0.7	48.7±0.8	50.0±0.9
MIMIC-S 3x Dupli	97.7±0.1	87.9±0.1	49.7±0.4	9.3±1.4	0.2±0.0	61.4±0.5	47.0±0.4	48.5±0.6	49.6±0.5
MIMIC-S 2x New*	***98.4±0.1	***91.8±0.4	*52.3±0.8	10.7±1.0	*0.2±0.0	*63.8±0.9	*49.3±0.8	*51.1±0.9	*53.1±1.1
MIMIC-S 3x New*	***98.8±0.1	***93.7±0.3	**53.5±0.9	*12.1±1.7	**0.2±0.0	**65.1±0.8	**50.5±0.8	**52.5±1.0	**55.1±1.2

4.3.2 Swedish Models

The results of the Swedish medical coding models are presented in Table 4.9. For comparison, we also report the results from Lamproudis et al. (2024), who trained a medical coding model on SEPR II. However, their results are only partially comparable to ours, as a different training algorithm was used, and the dataset splits are not defined in detail, meaning there is no identical test set. The PLM-ICD model trained on SEPR-L shows at 60.2% a similar Micro F1 score to that reported in Lamproudis et al. (2024). When using the significantly smaller SEPR-M as training data, the performance drops to a mean Micro F1 of 52.4%. The models trained on the respective synthetic datasets exhibit a performance decrease similar to what was observed with the English datasets. However, their performance can still be considered competitive, especially given the enhanced privacy in comparison to models trained on real data. Compared to the MIMIC models, it is notable that the SEPR models show much higher EMR and much lower Precision@8 and Precision@15 metrics, while R-precision remains comparable. This discrepancy is likely due to the significantly fewer codes per document in SEPR compared to MIMIC, enabling higher EMR scores and making Precision@8 and Precision@15 less suitable as evaluation metrics. As described in Chapter 3.1, there are several key differences between the English and Swedish datasets, which make a direct comparison less meaningful. However, the respective differences between real and synthetic data are comparable and, as discussed, quite

similar. This suggests that our proposed framework generates synthetic data of similar utility for training medical coding systems in English and Swedish.

Table 4.9: Results of medical coding model trained on different SEPR datasets. Significance: *p<0.05, **p<0.01, ***p<0.001

	Classification					Ranking			
	AUC-ROC		F1		EMR	Precision@k		R-precision	MAP
Training Data	Micro	Macro	Micro	Macro		8	15		
Lamproudis et al. (2024)	-	-	61.0	-	-	-	-	-	-
Real SEPR-L	99.3±0.1	97.0±0.2	60.2±0.9	23.4±1.1	48.8±0.7	12.3±0.1	6.9±0.0	60.4±0.7	71.9±0.7
Synth SEPR-L	98.8±0.2	**95.3±0.1	**54.7±0.6	***14.3±0.7	*44.1±0.8	**11.7±0.1	**6.6±0.0	**54.5±0.5	*66.3±0.9
Real SEPR-M	98.5±0.0	92.2±1.2	52.4±0.5	15.0±0.9	40.5±1.2	11.5±0.0	6.6±0.0	52.1±0.8	64.4±0.6
Synth SEPR-M	*98.1±0.1	89.7±0.8	**45.9±1.1	*8.2±1.3	*30.5±2.6	*10.9±0.1	**6.3±0.0	**45.3±1.4	**58.4±1.2

4.3.3 Effect of Domain Adaptation

To evaluate the suitability of the LLaMA-3.1-8B base model within our framework, we integrated two additional pretrained models into the data generation process. Synthetic data generated by these models was used to train medical coding models and then compared against models trained on the synthetic dataset generated by LLaMA-3.1-8B. To examine the impact of domain-specific pretraining, OpenBioLLM-8B, a model adapted to the medical domain, was employed to generate a synthetic dataset from MIMIC-S code sequences. Surprisingly, as shown in Table 4.10, the performance of the model trained on this dataset did not significantly differ from the MIMIC-S model trained on synthetic data generated by the fine-tuned base model, despite the common expectation that domain-specific pretraining enhances model performance (Gururangan et al., 2020). To further assess whether language-specific pretraining could improve Swedish synthetic records, a synthetic SEPR-M dataset was generated using a fine-tuned version of LLaMA-3-8B from AI Sweden. However, once again, no significant difference emerged between the models trained on this dataset and the synthetic SEPR-M dataset generated using the base version of LLaMA-3.1-8B.

Overall, these results suggest that fine-tuning LLaMA-3.1-8B on discharge summaries is sufficient to adapt the model to both the medical domain and the Swedish language. Thus, further adaptation through additional pre-training does not appear to influence the generation output. Whether this holds true for other languages remains to be investigated, as it is unclear how much Swedish data was included in the training set of the base model. Similarly, outcomes might vary with different medical models, as the effectiveness of domain adaptation heavily depends on the data and methods used (Lu et al., 2024). Furthermore, it should be noted that the adapted models were based on the older LLaMA-3-8B version. This difference in model versions could also explain why the domain-specific models did not generate data of higher utility. For the purposes of this work, however, the base version of LLaMA-3.1-8B seems to be a good fit for the synthetic generation framework.

Table 4.10: Results of medical coding model trained on data obtained from further domain-specific models compared to models trained on datasets generated using LLaMA-3.1-8B base. None of the differences is statistically significant with $p \geq 0.05$

	Classification					Ranking			
	AUC-ROC		F1		EMR	Precision@k		R-precision	MAP
Training Data	Micro	Macro	Micro	Macro		8	15		
Domain-Specific Pretraining: MIMIC-S									
Base	95.4±0.2	77.0±0.9	37.6±0.6	2.1±0.3	0.1±0.0	48.3±0.6	35.5±0.6	35.8±0.7	34.8±0.7
Pretrained	95.7±0.6	80.8±5.0	36.3±0.7	1.7±0.1	0.0±0.0	45.6±0.7	33.5±0.7	34.2±0.6	33.1±0.9
Language Pretraining: SEPR-M									
Base	98.1±0.1	89.7±0.8	45.9±1.1	8.2±1.3	30.5±2.6	10.9±0.1	6.3±0.0	45.3±1.4	58.4±1.2
Pretrained	98.3±0.0	92.0±0.3	48.8±0.1	10.8±0.2	36.4±0.2	11.2±0.0	6.4±0.0	48.2±0.2	61.0±0.2

4.3.4 Error Analysis

The error analysis is conducted on MIMIC data to help analyze and interpret performance differences. The analysis focuses on several aspects that aim to identify differences between the real and synthetic MIMIC-L and MIMIC-S models. This aims to better understand the differences between real and synthetic data and to evaluate which properties of the synthetic data reduce its utility for training medical coding models.

Predictions

Table 4.11: Some statistics about code predictions by real and synthetic MIMIC-L and MIMIC-S models compared to target codes from the test set.

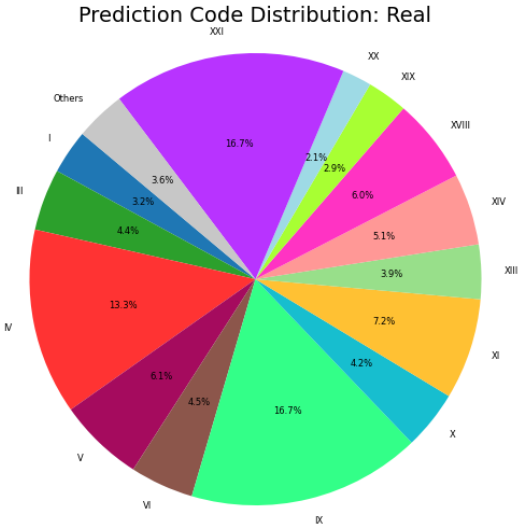
	Unique Codes	Correctly Predicted	AVG Codes/Doc	Total Codes	Exact Match	Subset Match	Threshold
Test Set	7,935	-	15.86	314,136	0.1%	1.5%	-
Real MIMIC-L	4,701	59.2%	14.3	283,784	0.8%	5.1%	0.38
Synth MIMIC-L	4,027	50.7%	14.1	279,171	0.8%	5.7%	0.29
Real MIMIC-S	1,053	13.3%	12.6	250,450	0.2%	11.5%	0.32
Synth MIMIC-S	1,125	14.2%	13.70	271,247	0.3%	17.4%	0.21

Table 4.11 highlights key prediction characteristics for the real and synthetic MIMIC-S and MIMIC-L models. Across all models, fewer unique codes are predicted than those present in the targets of the test set. The real MIMIC-L model correctly predicts approximately 59% of the codes at least once, while the synthetic MIMIC-L model correctly predicts around 51%. The MIMIC-S models show a considerable drop in correctly predicted codes, with only 13% for the real model and 14% for the synthetic model. On average, each model predicts 1 to 3 fewer codes than the targets, resulting in an overall lower code count across the test set. These findings suggest that the models might have difficulties predicting rare codes resulting in a substantially smaller amount of unique counts.

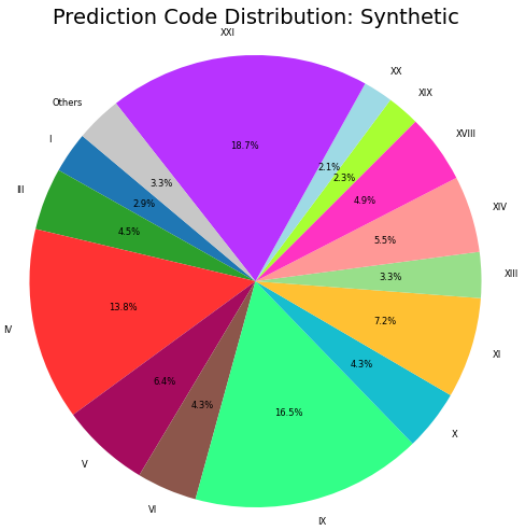
Table 4.11 also details the exact matches and subset matches between predicted code sequences and those in the training data, shedding light on the models' capacity to generate new code combinations. The real test set exhibits low overlap with training data sequences, showing an exact match ratio of 0.1% and a subset match ratio of 1.5%. While exact match

ratios increase across all models, the rise is more pronounced in the MIMIC-L models, reaching an exact match ratio of 0.8%. This may be due to the shorter average predicted code sequences in the MIMIC-S models, which reduces the likelihood of exact matches. However, subset matches increase more substantially in the MIMIC-S models, reaching 11.5% for real and 17.4% for synthetic MIMIC-S, suggesting difficulty in generating novel code combinations, which may contribute to reduced performance.

Thresholds tuned for optimal performance varied, ranging from 0.21 for the synthetic MIMIC-S model to 0.38 for the real MIMIC-L model. Consistent with findings from Edin et al. (2023), these variations underline the importance of model-specific threshold tuning over a fixed 0.5 threshold to prevent suboptimal or biased performance reporting.



(a) Chapter Distribution: Real



(b) Chapter Distribution: Synthetic

Figure 4.1: Code chapter distributions of predicted codes by the real-data MIMIC-L (top) and synthetic-data MIMIC-L (bottom) models.

Figure 4.1 illustrates the code distributions across different ICD chapters for the predictions of the real-data MIMIC-L model (top) and the synthetic-data MIMIC-L model (bottom). As shown, the chapter-wise distributions are nearly identical, covering the same 14 chapters, each with more than 2% of the codes. This distribution is also very similar to the real chapter distribution, as depicted in Figure 3.2 in Chapter 3.1.1, with one difference being the inclusion of a 15th chapter, representing over 2% of the codes in the real data. Notably, a large portion of the predictions falls within Chapters IV, IX, and XXI, which are also the most frequently represented chapters in the training data. In conclusion, the chapter distributions of the predictions reveal that both models successfully replicate the real data's distribution while also predicting codes from underrepresented chapters, demonstrating that both models have learned to predict disease codes from a wide range of medical fields.

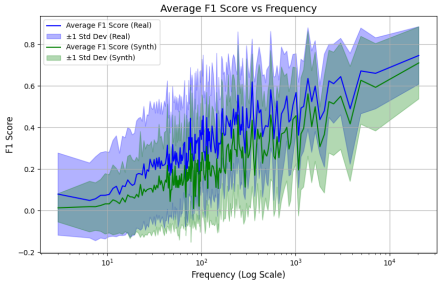
Code Frequency and Document Length

Figure 4.2 illustrates the impact of document length in the test set and code frequency in the training set on the F1 scores for models trained on real and synthetic MIMIC-L datasets. Both models show reduced performance on documents under 1000 words, an observation already reported by Edin et al. (2023), who found through manual inspection that shorter documents often lack essential information for accurate disease prediction. Table 4.12 reports the respective Pearson and Spearman correlation coefficients. The correlations for document lengths are split into ranges of 0-1000 and 1000-4000 words. This split reflects the observed trend: F1 scores increase with document length up to 1000 words, then decline as document length extends to the 4000-word truncation limit.

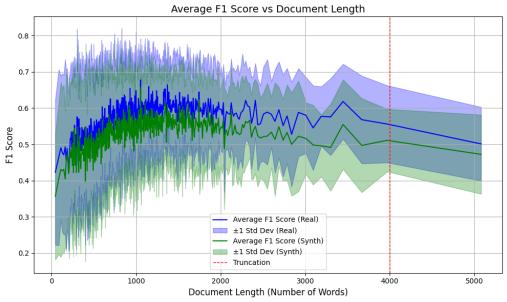
All correlations are statistically significant ($p<0.001$). Document length shows weaker correlations compared to code frequency, with positive correlations for documents up to 1000 words and negative correlations thereafter. The coefficients are slightly different than those reported in Edin et al. (2023). This may be due to the distinct preprocessing method applied here.

When comparing the real and synthetic models in Figure 4.1, similar trends are evident, with the synthetic model generally achieving lower F1 scores but following a comparable pattern in response to code frequency and document length. The synthetic model is slightly more sensitive to code frequency, while document length effects are marginally lower for short documents and slightly higher for longer ones compared to the real-data model.

Overall, the influence of document length on performance is minor, with the negative correlation being near zero, whereas code frequency has a substantial effect, aligning with findings from Edin et al. (2023) and Huang et al. (2022) and explaining the low count of unique codes discussed before. This highlights the need for additional training samples of low-frequency codes, as the models struggle to learn these codes effectively otherwise.



(a) F1 vs. Code Frequencies



(b) F1 vs. Document Length

Figure 4.2: Macro F1 vs. code frequencies in training data (left) and Micro F1 vs. document length in test data (right) of real (blue) and synthetic (green) MIMIC-L medical models

Table 4.12: Correlation between Macro F1 score and the logarithm of code frequency as well as Micro F1 score and document length split into length 0 to 1000 and 1000 to 4000. All correlations are significant with $p<0.001$.

	Code Frequency		Document Length			
			0-1000		1000-4000	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Real	0.48	0.51	0.19	0.19	-0.05	-0.05
Synthetic	0.50	0.52	0.17	0.16	-0.08	-0.06

OOF vs. WF errors

To better understand the incorrect predictions made by the models, we calculated WF and OOF errors. Table 4.13 presents the counts and percentages of correct and incorrect predictions as well as WF and OOF errors for models trained on real and synthetic MIMIC-L data. Overall, the synthetic-data model made slightly more errors than the real-data model even though exhibiting a lower count of altogether predictions. Both models exhibit a significantly higher proportion of WF errors than OOF errors, with a ratio of 79.4% for the real model and 82.5% for the synthetic model. This indicates that the models are often able to capture the general disease category, though they may miss the specific code, suggesting more learning occurred than what the reported performance metrics may reflect. Even though the real-data model made more errors than the synthetic-data model, a higher proportion of these errors are WF family, resulting in a slightly lower OOF error count for the synthetic model compared to the real model.

ICD-10 codes often vary by granularity, which can be difficult to capture in discharge summaries that may lack specific details. For instance, as Edin et al. (2023) illustrate, there are approximately ten different codes related to tobacco use and nicotine dependence; some codes achieve F1 scores over 50%, while others score 0%. They argue that class imbalances among highly infrequent codes result in the model strongly favoring more frequent ones. The high number of WF errors implies that while the model can identify the general disease area, code imbalance or insufficient detail in the summary likely leads to the prediction of an incorrect, more common code.

Table 4.13: Counts of overall correct and wrong predictions as well as WF and OOF family errors alongside percentages for the real-data and synthetic-data MIMIC-L models

Real MIMIC-L				Synth MIMIC-L			
Correct	Wrong	WF	OOF	Correct	Wrong	WF	OOF
176,270	107,514	85,356	22,158	162,939	116,232	94,733	21,499
62.1%	37.9%	79.4%	20.6%	58.4%	41.6%	81.5%	18.5%

Noise

The previous error analysis revealed similar error patterns between models trained on synthetic and real data, with the synthetic-data models generally performing slightly worse. This discrepancy may stem from noise present in the synthetic data, which could hinder its utility. The high vocabulary observed in the synthetic data suggests that while it adds diversity, it may also introduce noise. Two hypotheses were formulated to explain how this noise might be distributed across the synthetic medical notes:

- **H1:** All or most synthetic medical notes contain a certain percentage of noise.
- **H2:** Some synthetic medical notes contain a high percentage of noise, while others contain none or very little.

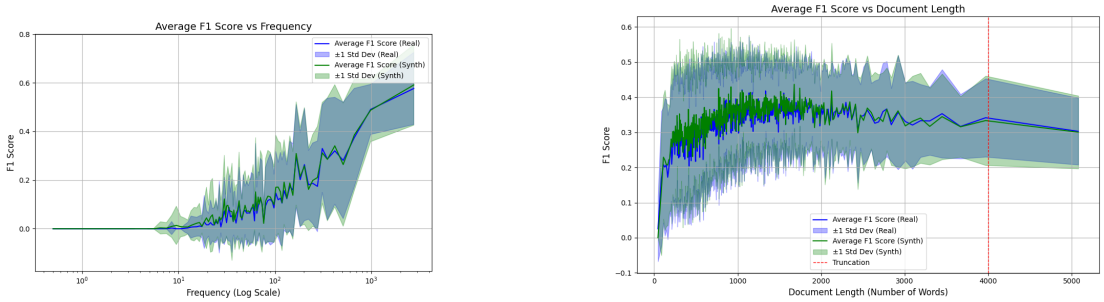
To test these hypotheses, noise was artificially introduced into the real MIMIC-S dataset by substituting words with random entries from the NLTK wordlist (Bird et al., 2009). For H1, a varying uniform percentage of words was replaced across all documents. For H2, 90% of words were replaced, but only in a random subset of varying size. These noisy datasets were used to train medical coding models, and their F1 scores were plotted against code frequency and document length to compare with the synthetic MIMIC-S model.

The results showed that under H1, substituting 20% of words across all documents produced F1-score curves very closely aligned with those of the synthetic data model (see Figure 4.3). Under H2, substituting 90% of words resulted in most similar curves to the synthetic MIMIC-S model when done for 15% of the documents (see Figure 4.4). Comparing Figures 4.2 and 4.4 demonstrates that the model trained on data with noise added as per H1 aligns more closely with the synthetic-data model than the model trained under H2.

This alignment strongly supports H1, indicating that noise in the synthetic data is distributed across all documents, reducing its utility for training medical coding models. This conclusion is further supported by the filtering experiments, which showed no substantial improvements in model performance. If the noise was localized to a subset, of documents, filtering would likely have led to significant enhancements. This reinforces the hypothesis that the noise is widespread and not concentrated in specific documents.

The noise in synthetic data likely originates from LLM artifacts, such as hallucinations and repetitions, as well as differences in the way diseases and procedures are described compared to real data. These discrepancies may not be immediately noticeable but subtly

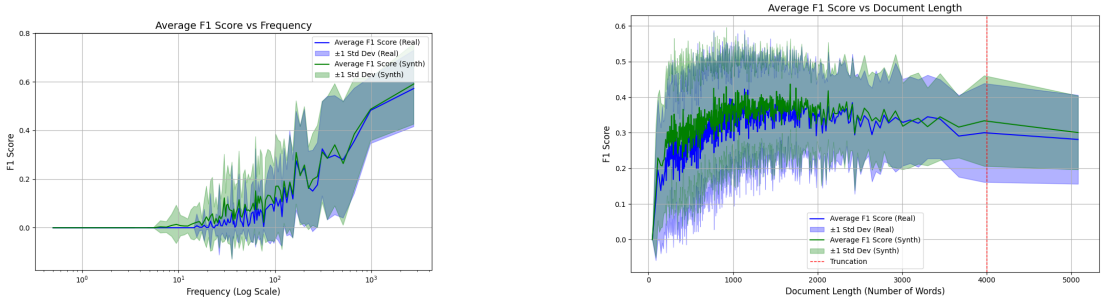
hinder performance when the models are evaluated on real-world data. To address this issue, future work should focus on a more detailed analysis of the noise in synthetic data and on refining the generation process to suppress unwanted artifacts without compromising diversity. Integrating explainability methods, such as feature attribution techniques in the medical models as described in Edin et al. (2024), could help identify dataset features that contribute to predictions and distinguish noise from valuable information. Such explainability methods not only offer insights into the quality of synthetic data but can also foster trust in clinical NLP systems by making model behavior more transparent. By mitigating the impact of noise and enhancing data quality, synthetic medical datasets can be further optimized for downstream tasks while maintaining their diversity and privacy-preserving properties.



(a) H1: F1 vs. Code Frequencies

(b) H1: F1 vs. Document Length

Figure 4.3: H1: Macro F1 vs. code frequencies in training data (left) and Micro F1 vs. document length in test data (right) of real MIMIC-S containing 20% noise in each document (blue) and synthetic MIMIC-S (green) medical models



(a) H2: F1 vs. Code Frequencies

(b) H2: F1 vs. Document Length

Figure 4.4: H2: Micro F1 vs. code frequencies in training data (left) and document length in test data (right) of real MIMIC-S containing 15% documents with 90% noise (blue) and synthetic MIMIC-S (green) medical models

4.4 Utility: NER

This Chapter presents the results of the second utility evaluation using the synthetic documents as training data for NER models. The results of training a clinical NER model on the MIMIC data are presented first, followed by the results of the PHI NER model trained on synthetic SEPR data.

4.4.1 Clinical NER

Following Mawaldi and Mladenov (2024), 5,000 random documents from both the synthetic and real MIMIC-S datasets were extracted and annotated using Med7 for drug names and Stanza for diseases. Table 4.14 shows the number of sentences, total words, and unique words for our real and synthetic subsets, compared to the datasets used by Mawaldi and Mladenov (2024). Since Mawaldi and Mladenov (2024) only generated the *History of Present Illness* section of the dataset, our datasets are substantially larger. As discussed in Chapter 4.1, the vocabulary and unique token ratio in our synthetic data are higher than in the real data. In contrast, the synthetic data in Mawaldi and Mladenov (2024) has a substantially smaller vocabulary compared to the real dataset, even though the synthetic dataset contains more tokens overall.

Table 4.14: Statistical properties of synthetic and real NER MIMIC training sets compared to Mawaldi and Mladenov (2024)

	Sentences	Total Words	Unique Words
Real Mawaldi and Mladenov (2024)	73,526	1,030,875	60,186
Real Our Work	377,161	6,440,092	63,211
Synth Mawaldi and Mladenov (2024)	112,322	1,411,143	26,019
Synth Our Work	428,119	8,275,997	123,905

Table 4.15 presents the label and unique label counts in our dataset, compared to those reported by Mawaldi and Mladenov (2024). Unsurprisingly, our datasets contain substantially more labels, reflecting their larger size. Similar to the findings regarding vocabulary, we observe more unique labels in our synthetic dataset compared to the real dataset, with the ratio of unique labels being slightly higher in the synthetic data. In contrast, Mawaldi and Mladenov (2024) report fewer unique labels in their synthetic dataset, despite it containing more labels overall. This further highlights the common issue of diversity reduction in synthetic datasets, which the generation approach of this work appears to have overcome.

Table 4.15: Disease and drug label counts of real and synthetic NER training sets from our work compared to Mawaldi and Mladenov (2024)

	Diseases	Unique Diseases	Drugs	Unique Drugs	Total Entities
Real Mawaldi and Mladenov (2024)	61,628	10,623	20,967	1,729	82,595
Real Our Work	267,549	49,827	104,933	6,213	372,482
Synth Mawaldi and Mladenov (2024)	123,476	50,90	28,092	652	151,568
Synth Our Work	309,661	88,528	136,636	13,435	446,297

Following Mawaldi and Mladenov (2024), we used the annotated datasets to fine-tune BERT base for clinical NER, evaluating performance on a test set of 1,000 real documents. Note, that the test sets used in Mawaldi and Mladenov (2024) and our work are not identical due to the lack of a reproducible split, making comparisons approximate. Table 4.16 presents the clinical NER model results from our study alongside those reported by Mawaldi and Mladenov (2024). Our models, trained on both real and synthetic data,

achieved a weighted F1 score of 0.87. All metrics of our real and synthetic models and the synthetic-data model by Mawaldi and Mladenov (2024) are nearly identical. The real-data model from Mawaldi and Mladenov (2024) performed slightly better, with a weighted F1 of 0.89.

These consistent performance metrics across models trained on varying token counts and label diversity suggest that automatic annotation in training data may achieve an upper limit of performance under the tested conditions. Errors in the automatic labeling process may propagate through successive models when being used for training. This might limit further gains in model accuracy even when increasing training data size or label counts. One potential improvement could involve integrating annotation directly into the synthetic generation process, as demonstrated by Libbi et al. (2021).

Nonetheless, the proposed method demonstrates strong clinical NER performance using synthetic data, comparable to real data while enhancing privacy. The lack of performance gains with additional training data suggests that smaller datasets might suffice, offering the potential for reduced computational costs. Our findings indicate that synthetic MIMIC data effectively trains clinical NER models without compromising performance, and the higher label variety within the synthetic data may benefit more complex NER models requiring diverse training labels. Future work should explore the minimum data requirements for optimal performance and the potential benefits of label variety observed in the synthetic data.

Table 4.16: Performance of clinical NER models trained on real and synthetic MIMIC data compared to Mawaldi and Mladenov (2024).

		Our models					Mawaldi and Mladenov (2024)				
		Precision	Recall	F1	Accuracy	Support	Precision	Recall	F1	Accuracy	Support
Real	Diseases	0.82	0.89	0.84	-	10868	0.85	0.90	0.87	-	13920
	Drugs	0.93	0.95	0.94	-	3965	0.94	0.95	0.95	-	4539
	Weighted	0.84	0.90	0.87	-	14833	0.87	0.92	0.89	-	18459
					0.98					0.98	
Synth	Diseases	0.82	0.88	0.85	-	10868	0.83	0.88	0.85	-	113920
	Drugs	0.93	0.95	0.94	-	3965	0.93	0.93	0.93	-	4539
	Weighted	0.85	0.90	0.87	-	14833	0.85	0.89	0.87	-	18459
					0.98					0.98	

4.4.2 PHI NER

Similar to our approach, Hullmann and Hansson (2024) developed a synthetic dataset based on SEPR II. To enable comparison, we followed their PHI NER evaluation methodology by extracting a random subset of our synthetic SEPR-M dataset that matches the synthetic token count of 1,187,380 reported by Hullmann and Hansson (2024). We used SweDeClin-BERT NER to annotate this data with nine distinct PHI tags, and evaluated model performance on a manually annotated test set of 300 patient notes that served as the gold standard in Hullmann and Hansson (2024). Table 4.17 displays the label counts for our synthetic subset compared to the real, synthetic, and test sets in Hullmann and Hansson (2024). Our synthetic dataset contains substantially more PHI labels, particularly in the categories of *First_ Name* and *Last_ Name* as well as *Full_ Date*. This difference may relate to the vast vocabulary observed in our synthetic data.

Table 4.17: Comparison of label count contained in Synthetic, Real, and Test datasets as reported in Hullmann and Hansson (2024) and a subset of synthetic SEPR-M with equivalent token count

	Hullmann and Hansson (2024)			Our Work
	Synthetic	Real	Test	Synthetic
Organisation	110	113	0	230
Age	802	933	59	1,344
Full_Date	3,892	2,659	532	7,557
Location	777	389	157	688
First_Name	2,185	2,443	917	5,626
Health_Care_Unit	4,777	1,489	1,170	3,496
Phone_Number	15	34	226	137
Last_Name	2,346	1,973	930	8,367
Date_Part	4,217	4,407	778	5,893
Total	19,121	14,440	4,769	33,338

The annotated datasets were used to fine-tune SweDeClin-BERT, a Swedish BERT model adapted to the medical domain. Table 4.18 presents the results. Hullmann and Hansson (2024) reported a significant performance gap between their real and synthetic data models, with the real model outperforming the synthetic by 0.10 in Macro F1 and 0.19 in Weighted F1 scores. In contrast, our synthetic-data model achieves substantially higher performance than the synthetic model from Hullmann and Hansson (2024), even exceeding their real-data model on average, reaching a Macro F1 score of 0.72 and a Weighted F1 score of 0.99. Note, that the *Organization* label is absent in the target set, leading to zero metrics for that label. For consistency and comparability, we include it in the evaluation despite this limitation.

The performance gain of our synthetic model is likely due to the larger label count, offering more training samples. However, the rather small performance difference between our synthetic model and the real-data model reported by Hullmann and Hansson (2024) despite the significantly larger label count, suggests a potential advantage in real data, such as reduced noise. It is important to note that, as with the clinical NER model, performance could be constrained by the limitations of automatic annotation, which may introduce errors that hinder further improvements.

In summary, both the clinical and PHI NER model results suggest that the generated synthetic data offers high utility, comparable to real data while providing increased privacy protection. The observed increase in unique labels and vocabulary within the synthetic data appears beneficial for training NER models. The combined findings from both the medical coding and the NER downstream tasks point to the general utility of synthetic data generated by the proposed framework. Future research should explore additional downstream tasks to validate the generalizability assumption. Additionally, issues arising from automatic annotations of training data should be investigated to better understand and interpret performance metrics.

Table 4.18: Performance of PHI NER models trained on real and synthetic SEPR data compared to Hullmann and Hansson (2024)

	Hullmann and Hansson (2024)						Our Work			
	Real			Synthetic			Synthetic			
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	
Organisation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Age	0.79	0.97	0.87	0.96	0.95	0.96	0.94	0.96	0.95	
Full_Date	0.90	0.97	0.93	0.73	0.98	0.84	0.93	0.95	0.94	
Location	0.90	0.25	0.38	0.75	0.18	0.29	0.70	0.32	0.44	
First_name	0.95	0.97	0.96	0.70	0.74	0.72	0.71	0.97	0.82	
Health_Care_Unit	0.35	0.58	0.43	0.22	0.46	0.30	0.51	0.67	0.58	
Phone_Number	0.88	0.86	0.87	0.80	0.45	0.58	0.78	0.93	0.85	
Last_Name	0.92	0.95	0.94	0.87	0.73	0.79	0.60	0.97	0.74	
Date_Part	0.96	0.97	0.96	0.82	0.97	0.89	0.91	0.97	0.94	
Accuracy			0.98			0.97				0.99
Macro AVG	0.74	0.72	0.70	0.65	0.61	0.60	0.71	0.77	0.72	
Weighted AVG	0.91	0.89	0.86	0.64	0.75	0.67	0.99	0.99	0.99	

The employed utility evaluation focused on using established clinical NLP approaches built upon real data to assess synthetic data utility rather than optimizing high-performing competitive models. Future work should aim to further adapt and refine models to leverage the properties of synthetic data, thereby achieving optimal performance.

4.5 User Study: Readability and Medical Coherence

This chapter presents the results of the user study investigating readability and medical coherence of the generated medical notes.

4.5.1 MIMIC Samples

Ten participants took part in the MIMIC study, all of whom confirmed having professional medical knowledge and working as doctors. The medical fields they specialize in are listed in Table 4.19, with some participants reporting expertise in more than one field. The years of experience range from a minimum of two years to a maximum of 38 years. Five out of the ten participants reported performing ICD-10 coding *often* or *very often*. These participants were asked to perform medical coding on the discharge summaries, while the other five participants, who indicated they *sometimes*, *never*, or *rarely* perform ICD-10 coding, were not presented with this task. The participants were recruited in Germany, and none of them were native English speakers. This should be taken into account when interpreting the results.

Table 4.19: Medical fields in which study participants work

Medical field	Count
Anesthesia	1
Dentistry	1
Emergency surgery	1
Gynecology	1
Neuropediatrics	1
Oncology	1
Orthopedics	1
Pediatrics	1
Radiology	1
Trauma surgery	1
Urology	1
Unspecified	2

The participants rated three real and three synthetic documents, each belonging to the same code sequences, on readability and medical coherence using a bipolar scale with 5 levels. The results of these ratings are presented below. Figure 4.5 shows that readability was slightly higher for the real documents. Table 4.20 reveals a mean of 3.7 and a median of 4 for the real documents, compared to the synthetic documents, which had a mean of 3.3 and a median of 3.0. A similar pattern emerged for the medical coherence ratings. The real documents were again rated higher, with a mean of 3.5 and a median of 4.0, while the synthetic documents slightly lagged behind, with a mean of 3.3 and a median of 3.0. However, a t-test revealed that neither of these differences was statistically significant $p \geq 0.05$.

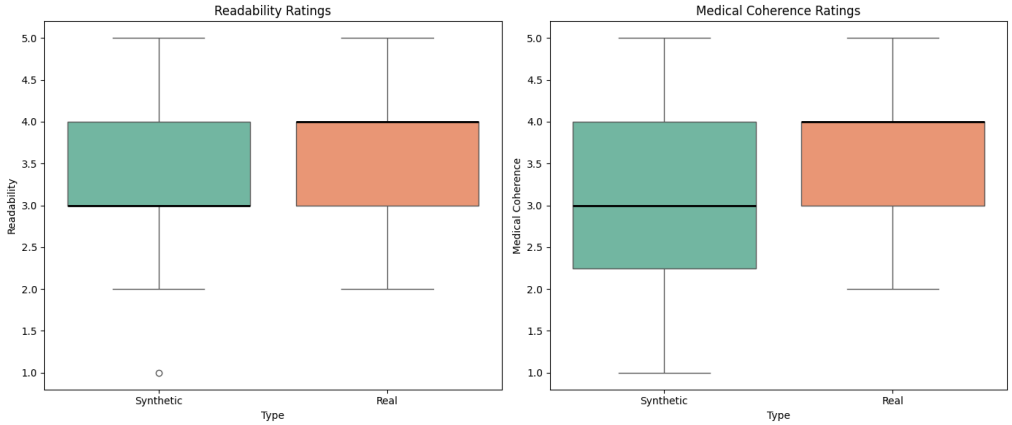


Figure 4.5: Comparison of readability and medical coherence averages across all documents and participants between real and synthetic documents in the MIMIC study. The boxes represent the interquartile range (IQR), with whiskers extending 1.5 times the IQR, and the median value indicated by a black line. The differences are not statistically significant with $p \geq 0.05$.

	Mean		Median	
	Real	Synthetic	Real	Synthetic
Readability	3.7	3.3	4.0	3.0
Medical Coherence	3.5	3.3	4.0	3.0

Table 4.20: Comparison of mean and median values of readability and medical coherence ratings between real and synthetic documents in the MIMIC study

Figure 4.6 illustrates the differences for each of the three real-synthetic document pairs. The real documents achieved higher or identical mean and median readability and medical coherence scores across all document pairs, except for document pair three, where the synthetic document received a slightly higher mean rating. Independent two-sample t-tests revealed none of the differences as statistically significant ($p \geq 0.05$).

Furthermore, the boxplots in 4.6 show a wide range of ratings for each document. Pairwise Cohen's Kappa calculations yielded a mean score of 0.03 for readability and 0.06 for medical coherence. These low scores indicate a low level of agreement between the participants. This variability may be explained by the different medical fields in which the participants work. Given that each participant has a different medical focus and varying years of experience, it is likely that their personal experiences and expertise influenced their ratings. Furthermore, since none of the participants are native English speakers, their proficiency in English may have influenced their ratings, particularly those related to readability.

Overall, all documents received a mean rating above 3 for both readability and medical coherence, indicating that none of the documents were perceived as entirely unrealistic or unpleasant to read. The non-significant difference between real and synthetic documents suggests that the synthetic documents are of a similar quality to the real ones. However, the wide range of ratings that are sometimes very low for both real and synthetic documents implies there may be structural or content issues within the real documents that have carried over to the synthetic ones.

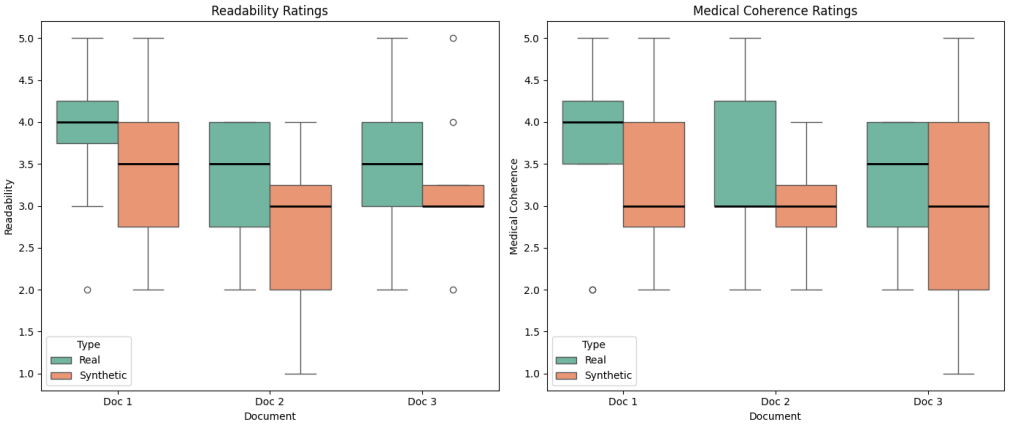


Figure 4.6: Comparison of readability and medical coherence scores averaged across participants for single document pairs in the MIMIC study. The boxes represent the IQR with whiskers 1.5 times the IQR and outliers as points outside this range. The median value is indicated by black lines. None of the differences are statistically significant with $p \geq 0.05$.

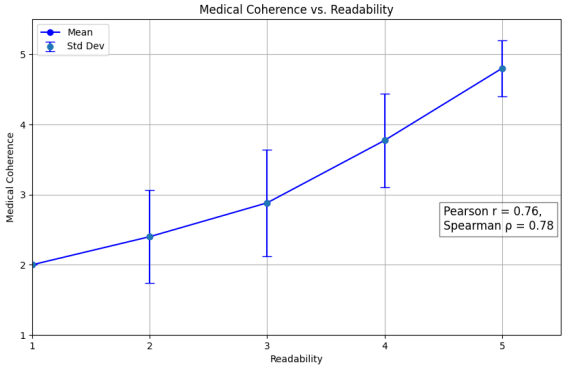


Figure 4.7: Correlation between Readability (x-axis) and Medical Coherence (y-axis) in the MIMIC study. There is a strong positive correlation ($p<0.001$) with a Pearson correlation coefficient of 0.76 and a Spearman correlation coefficient of 0.78.

A strong positive correlation between readability and medical coherence, with a Pearson correlation coefficient of 0.76 and Spearman correlation coefficient of 0.78, as shown in Figure 4.7, indicates that some documents are generally of higher quality than others. However, the low inter-annotator agreement complicates the interpretation of these results. To better understand potential issues within the documents, some of the participants' justifications for their answers are presented in the following.

Table 4.21: Participants' justifications for their ratings on readability and medical coherence, with counts for synthetic and real documents

	Synthetic	Real
Too many abbreviations	|||	
Hard to follow / unpleasant reading	|||	
Repetition			
Diagnosis missing			
Contradicting recommendations			
Insufficient information about treatment			
Insufficient information about patient's status	||		
Diagnosis does not fit MRI			

Table 4.21 presents some participants' justifications for their readability and medical coherence ratings. The first three justifications regard readability, while the last four concern medical coherence. For each justification, the number of references to real and synthetic documents is also noted.

For readability, several participants criticized the synthetic documents for containing excessive abbreviations and being difficult to follow which are issues that may be interconnected. Notably, this criticism was never applied to real documents. One participant mentioned that repetitions in a real document hindered readability. The justifications for medical coherence ratings fall into two main categories: missing or insufficient information, and contradictory information. These issues were noted for both real and synthetic documents.

Additionally, some participants criticized missing lab results; however, these omissions were intentional due to preprocessing and can therefore be neglected within the scope of this interpretation. It is important to note that providing justifications was optional, thus, the content and number of comments per document is less meaningful but still offers insight into potential issues within the discharge summaries.

One participant mentioned that their ratings were influenced by their medical specialty, making some documents easier for them to read than others. This observation supports the idea that low inter-annotator agreement may stem from the varied medical backgrounds of the participants.

In summary, the user study revealed no statistically significant differences between real and synthetic documents in terms of readability and medical coherence ratings, though real documents achieved slightly higher scores. It is important to note that this study evaluated a small sample of documents, so the interpretation of these results should be cautious. A strong correlation between readability and medical coherence suggests that certain documents are of generally higher quality than others. Low inter-annotator agreement may stem from participants' varied medical specialties. Overall, the synthetic documents appear to be of comparable quality to real ones, maintaining medical coherence and a similar language style. This result supports the broader and more generalizable applicability of the synthetic documents, even in contexts requiring medical coherence. However, some low ratings for both synthetic and real samples indicate existing issues that should be further investigated. Controlling for these issues in the generation process could help prevent the transfer of problematic elements from real to synthetic documents making them of even higher utility than real documents.

In addition to the ratings, five participants with experience in ICD-10 coding were asked to assign codes to the discharge summaries. Table 4.22 presents the mean, maximum, and minimum Micro F1 scores for each document, as well as the mean number of predicted codes and the maximum number of overlapping codes between two participants. The results reveal very low mean F1 scores, with an average of only 6.1% across all documents. Even the maximum F1 score of 40.0% is significantly lower than what this work's best medical coding models could achieve. When examining the number of predicted codes, it becomes clear that a key factor contributing to these low scores is the fact that, on average, less than half of the target codes were predicted.

While the real documents achieved a slightly higher average F1 score compared to the synthetic documents with a difference of 1.6%, this difference provides limited insight into how well the discharge summaries represent the correct codes, given the generally low scores. A manual review of the codes revealed that the granularity of the codes was a significant factor in the low F1 scores. Participants often did not provide the codes with sufficient granularity to match the targets, for example, by only specifying the first three digits of a code, or by deviating only in the last digits, reflecting fine-grained distinctions. Thus, many errors are WF errors, similar to what was observed in the error analysis of the medical coding models (see Chapter 4.3.4). Since participants were not instructed on the required granularity, more specific instructions would likely have led to higher F1 scores, making direct comparisons to model performance less meaningful.

The final column of Table 4.22 shows the maximum overlap of codes between any two

participants for each document. Notably, there were no identical sets of codes between any two participants for any of the documents, with a maximum overlap of two to three codes per document. This shows not only discrepancies between the provided codes and the target but also a low level of agreement among participants.

The purpose of presenting this task is not to directly compare the F1 scores to those of the previously reported models. A significantly larger sample size and more participants would be required for such comparisons. Additionally, as noted, the instructions given to participants should be adjusted to avoid biases in the results. Furthermore, in a real-world setting, coders typically specialize in a specific medical field, narrowing the range of codes they would encounter. Nonetheless, these results highlight the complexity of the ICD-10 coding task and why it is known to be error-prone, with little agreement even among experienced coders. This underscores the need for high-performance automatic medical coding systems to support and improve this process. The use of synthetic training data is one promising path for enabling the practical application of these systems. At the same time, these findings emphasize the importance of carefully selecting gold standard labels for training such systems, as these labels often come from manually labeled real data that may contain errors themselves.

Table 4.22: Micro F1 scores for ICD-10 predictions of participants in MIMIC study, alongside averaged number of predictions and maximum code overlap between any two participants

	Mean F1	Min. F1	Max. F1	Target Length	Mean Pred Length	Max. Code Overlap
Real 1	4.6	0.0	18.2	10	5.8	3
Synthetic 1	0.0	0.0	0.0	10	4.3	2
Real 2	0.0	0.0	0.0	14	3.7	3
Synthetic 2	3.3	0.0	13.3	14	3.5	2
Real 3	16.0	0.0	40.0	7	4.6	2
Synthetic 3	12.5	0.0	22.2	7	4.8	2
All synthetic	5.3	0.0	22.2	10.3	4.7	3
All real	6.9	0.0	40.0	10.3	4.2	2
Combined	6.1	0.0	40.0	10.3	4.5	3

4.5.2 SEPR Samples

The SEPR user study involved four participants: three nurses and one doctor, all native Swedish speakers, all of whom stated to perform ICD-10 coding *often* or *very often*, and were given the task of coding the medical notes. The participants rated four real and four synthetic notes on readability and medical coherence. Each pair of real and synthetic notes corresponded to the same ICD-10 code.

Given the small number of participants, the evaluation will be kept brief. However, the inclusion of this study is particularly important, as it provides valuable qualitative insights into the contents of the synthetic Swedish documents, especially considering the lack of further manual investigation due to language constraints.

Figure 4.8 presents an overall comparison of readability and medical coherence ratings. While the ratings were generally higher for the SEPR samples than for the MIMIC samples, the pattern remained similar, with synthetic notes generally receiving slightly lower ratings than real notes (see Table 4.23). However, the differences were not statistically significant $p \geq 0.05$.

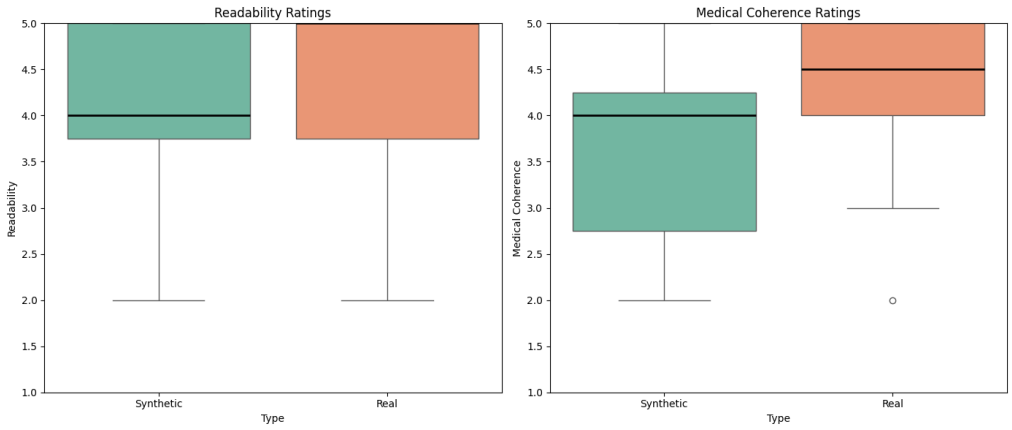


Figure 4.8: Comparison of readability and medical coherence averages across all documents and participants between real and synthetic documents in the SEPR study. The boxes represent the interquartile range (IQR), with whiskers extending 1.5 times the IQR, and the median value indicated by a black line. The differences are not statistically significant with $p \geq 0.05$.

	Mean		Median	
	Real	Synthetic	Real	Synthetic
Readability	4.3	3.9	5.0	4.0
Medical Coherence	4.3	3.6	4.5	4.0

Table 4.23: Comparison of mean and median values of readability and medical coherence ratings between real and synthetic documents in the SEPR study

Figure 4.9 compares the ratings for the four document pairs. The only statistically significant difference was found in the medical coherence ratings for document pair 3 ($p < 0.05$). This difference was driven by the very high ratings for the real document, rather than exceptionally low ratings for the synthetic document. Although the inter-annotator agreement was slightly higher than in the MIMIC study with a Mean Cohen's Kappa of 0.15 for readability and 0.16 for medical coherence, the agreement remains low, as evidenced by the range of ratings visible in Figure 4.9.

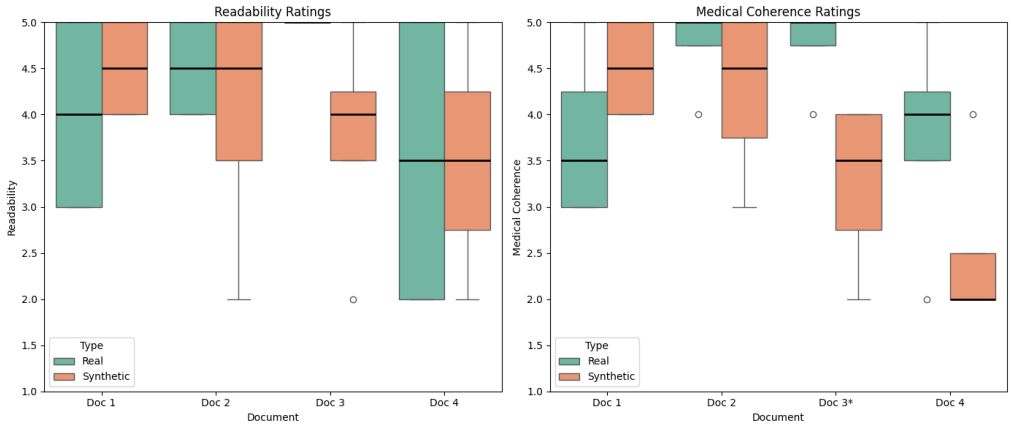


Figure 4.9: Comparison of readability and medical coherence scores averaged across participants for single document pairs in the SEPR study. The boxes represent the IQR with whiskers 1.5 times the IQR and outliers as points outside this range. The median value is indicated by black lines. The only statistical significant difference with $p \geq 0.05$ was found for the medical coherence ratings of document pair 3.

As shown in Figure 4.10, there was again a strong positive correlation between readability and medical coherence, suggesting that some medical notes are perceived as generally higher quality than others.

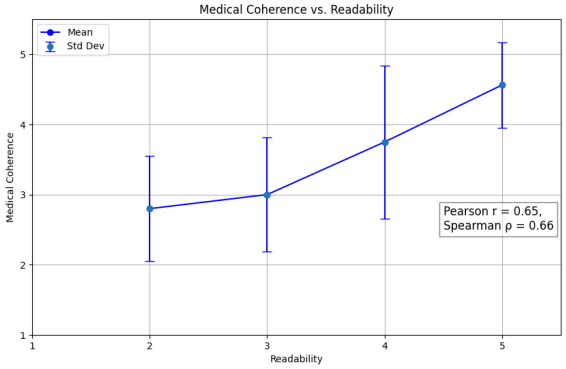


Figure 4.10: Correlation between Readability (x-axis) and Medical Coherence (y-axis) in the SEPR study. There is a strong positive correlation ($p<0.001$) with a Pearson correlation coefficient of 0.65 and a Spearman correlation coefficient of 0.66.

Table 4.24 presents the Micro F1 scores from the ICD-10 coding tasks. The mean F1 score for the SEPR study (18.35%) was higher than in the MIMIC study (6.1%), likely due to the shorter length of the SEPR notes and the fact that only one target code was required per document. As in the MIMIC study, low F1 scores were largely due to fine-grained discrepancies with the target codes. However, unlike the MIMIC study, additional performance drops in the SEPR study resulted from an excessive number of codes being assigned rather than too few. While the manual coding performed in this study was considerably worse than the best-performing models, these results are not directly comparable due to the aforementioned limitations, but they highlight the challenges in manual ICD-10 coding and the potential benefits of automation.

Table 4.24: Micro F1 scores for ICD-10 predictions of participants in SEPR study, alongside averaged number of predictions and maximum code overlap between any two participants

	Mean F1	Min. F1	Max. F1	Target Length	Mean Pred Length	Max. Code Overlap
Real 1	44.5	0.0	66.7	1	2.3	2
Synthetic 1	34.4	0.0	66.7	1	2.3	2
Real 2	17.9	0.0	100	1	2.0	1
Synthetic 2	0.0	0.0	0.0	1	1.3	0
Real 3	0.0	0.0	0.0	1	2.5	0
Synthetic 3	50.0	0.0	100	1	1.3	1
Real 4	0.0	0.0	0.0	1	3	2
Synthetic 4	0.0	0.0	0.0	1	1.8	1
All synthetic	21.1	0.0	100	1	1.7	2
All real	15,6	0.0	100	1	2.5	2
Combined	18.35	0.0	100	1	2.1	2

In conclusion, the SEPR study supports the findings of the MIMIC study, showing that the generated Swedish notes are comparable to the English notes in terms of readability and medical coherence. Since the differences between real and synthetic notes were not statistically significant, the results suggest that synthetic data could be a viable substitute for real data, even in contexts requiring medical coherence. The SEPR study's findings support the general applicability of synthetic documents across languages and settings, reinforcing the broader conclusions drawn from the MIMIC study.

5 Discussion

This study presents a framework for generating synthetic medical notes that effectively balance privacy and utility while maintaining medical coherence, diversity, and content richness. By leveraging instruction-tuning with ICD-10 textual descriptions, the framework addresses common challenges in synthetic data generation, such as diversity reduction, and demonstrates promising performance across multiple evaluations.

Comparing this framework comprehensively to prior research remains challenging due to the wide variety of evaluation methods employed in the field. Nonetheless, selective comparisons with existing studies and an analysis of reported difficulties suggest that the proposed approach addresses several common limitations in synthetic medical note generation.

The fidelity analysis revealed that synthetic notes exhibit key similarities with real notes while introducing some differences. For instance, synthetic notes tend to be longer on average, which is controllable during decoding if desired. More significantly, synthetic datasets feature a broader vocabulary and higher unique word ratio compared to real datasets, a notable improvement over prior work that often reported diversity reduction resulting in a strongly reduced vocabulary (Hullmann & Hansson, 2024; Libbi et al., 2021; Mawaldi & Mladenov, 2024). Diversity reduction restricts the utility and generalizability of synthetic data by narrowing its range of patterns. Our framework appears to mitigate this issue, likely due to the use of ICD-10 descriptions in the prompts. These descriptions generate diverse and variable inputs, encouraging broader content generation and richer vocabulary. Unlike methods that prompt using fragments of medical notes, which can lack variety, this approach also enables scalable generation of synthetic datasets across various medical domains, independent of the size of the original dataset.

Despite these positive findings, the synthetic notes exhibit some artifacts typical of LLM outputs, such as hallucinations, repetitions, and occasional incoherence. While these issues seem relatively rare, they introduce noise that could affect their utility for downstream applications and enable differentiation from real notes. Further refinements are needed to minimize such artifacts and enhance overall data quality.

The utility of the synthetic datasets was evaluated through two downstream tasks: NER and medical coding. In the NER tasks, the synthetic data performed comparably to real data and outperformed synthetic datasets generated by previous methods. For example, models trained on synthetic SEPR data performed better than models trained on synthetic data from earlier studies and even slightly better than models trained on real data of equivalent size when compared to Hullmann and Hansson (2024). This improvement is likely attributable to the higher variety in the synthetic data, which yielded richer training labels. However, the nearly identical performance across all models in the clinical NER task, when compared to (Mawaldi & Mladenov, 2024), suggests limitations in the implementation of

this task. This outcome may be attributed to its reliance on automatic annotation, which introduces errors during the labeling process and affects the quality of the training data.

In medical coding tasks, our models trained on synthetic data underperformed compared to those trained on real data of equivalent size but demonstrated significant improvement when the volume of synthetic training data was increased. This indicates that synthetic datasets could achieve SOTA performance with sufficient volume, albeit at the cost of increased computational resources. Given the substantial privacy advantages of synthetic data, this trade-off may be justified. The ability to generate arbitrary amounts of synthetic data makes these results particularly promising. While the filtering and balancing approaches employed in this study did not improve performance, future work should explore these strategies further. The flexibility to shape synthetic datasets to meet specific needs and optimize model training is a key advantage that should be fully leveraged to maximize utility.

Attributing the difference in performance to variations in the datasets is challenging. The error analysis revealed similar error patterns in both real and synthetic datasets, with the synthetic data generally performing slightly worse, similar to how models trained on reduced real data might behave. Inserting artificial noise into the real dataset revealed very similar performance patterns to the synthetic-data model when adding 20% of random words in every document and using it as training data. This suggests that the synthetic data contains widespread noise explaining its lower utility compared to the real data. Consequently, the observed variety in the synthetic data, while generally desirable, may also contribute to the noise that downgrades the utility of the dataset.

While medical models trained on synthetic data do not yet achieve parity with real-data models, they represent a substantial improvement over previous efforts in this domain. For instance, Falis (2024) generated synthetic datasets by employing GPT-3.5 in a zero-shot setting and reported a negative effect on overall performance when using the synthetic data as augmentation. This disparity likely stems from the data generation method itself, highlighting the importance of fine-tuning models for such tasks. Zero-shot employment, constrained by its reliance on the LLMs' pretaining knowledge and the given prompt, may be insufficient for teaching models to produce discharge summaries that accurately capture the content and context of the prompted codes.

The comparison of different downstream tasks also highlights the limitations of relying on simpler tasks like NER to assess the utility of synthetic data which is a common practice (see 2.1). As pointed out in earlier work, NER does not necessarily require medical coherence (Libbi et al., 2021), and as evidenced in the performance difference in this work, a strong utility for NER is not necessarily transferable to other tasks. More complex tasks, such as medical coding, can provide a more robust evaluation of utility.

The privacy evaluation, conducted using similarity metrics such as ROUGE-5 and 8-gram overlap, indicated that synthetic notes had lower proximity to training data than real notes, reflecting strong privacy protections. However, manual inspection revealed instances of longer sequence copying in the most similar English document pairs. While pseudonymization mitigated risks in this study, implementing additional privacy-enhancing methods, such as DP techniques, could strengthen these protections in future work. From a fidelity perspective, lower similarity scores between synthetic-real document pairs compared

to real-real pairs may be linked to the increased variability in synthetic data, which, as discussed, may also bear negative effects.

Medical coherence, a critical factor for synthetic notes, was assessed through a user study involving medical professionals to rate readability and medical coherence of English and Swedish samples. Although real documents tended to receive slightly higher ratings, the differences were not statistically significant, indicating that the framework can generate realistic and coherent medical content in both languages. While earlier work struggled to create medical coherent notes (Falis et al., 2024; Hiebel et al., 2023; Libbi et al., 2021), this achievement is likely due to instruction-tuning with ICD-10 descriptions, which guide the model's content generation and help to create contiguous notes. However, the user study also revealed that some issues inherent in real notes may have been transmitted to synthetic notes, highlighting the need to involve medical professionals in refining datasets to address these challenges. By identifying and specifying such issues, targeted controls can be implemented in the generation process to prevent their transmission to synthetic documents, thereby enhancing the overall quality and reliability of the synthetic data.

Manual ICD-10 coding of the user study highlighted the complexity and error-proneness of this task, as evidenced by low agreement among coders and generally low-performance scores. While these results are biased for various reasons and thus, not directly comparable to the coding models implemented in this work, these findings underscore the need for high-performing automated medical coding systems to enhance consistency, reduce human error, and alleviate administrative burdens in healthcare. Synthetic data offers a privacy-preserving solution for training such systems, overcoming the barriers posed by the sensitive nature of real patient data.

6 Limitations and Future Directions

The framework presented for generating synthetic English and Swedish medical notes offers promising insights; however, several limitations affect the interpretability, applicability, and generalizability of the results. Future work should address these limitations by building upon our framework, focusing on optimizing the generation process and evaluation methods. Below, these limitations are discussed in more detail, along with proposed future directions derived from the findings of this work.

Integration of Medical Expertise
Aside from the user study, no medical expertise was integrated into the development or evaluation of the proposed framework. As a result, manual assessments should be considered preliminary and limited in depth. Since the primary goal of synthetic note generation and clinical NLP systems, such as automated medical coding systems, is practical application, future work should incorporate appropriate medical expertise. Expert feedback could help address real-world requirements, enhance data quality, and improve system reliability. For example, medical experts could help design more realistic code sequences to address issues with low-frequency codes or expand datasets in meaningful ways. They could also identify shortcomings in real discharge summaries to avoid transferring these issues into synthetic datasets.

Language Constraints
Language constraints limited the depth of investigation and assessment for the Swedish-generated data. To gain a more comprehensive understanding of its quality and privacy-preserving properties beyond quantitative performance metrics, future work should include evaluations conducted by native Swedish speakers. Furthermore, testing the framework with other languages is recommended to evaluate its generalizability. The uncertainty about the extent of Swedish content in the pretraining data of LLaMA-3.1-8B highlights the need to test additional languages.

Dataset Limitations
The study was conducted using two distinct datasets: one focusing on gastrointestinal conditions in Swedish and the other on ED notes in English. While these datasets represent specific medical domains, the generalizability of the proposed approach to other medical domains remains unclear. Future research should evaluate synthetic note generation for additional medical specialties to ensure broader applicability.

Furthermore, previous studies have shown that medical notes often contain errors and inconsistencies, particularly in the assignment of ICD-10 codes. This work assumes that real notes are suitable for building clinical NLP systems and that the assigned ICD-10 codes serve as a reliable gold standard for medical coding systems. The proposed framework for medical note generation builds on the assumption that the ICD-10 codes adequately represent the content of the note. However, this assumption may inadvertently introduce

errors into the framework. Future research should investigate this aspect further, focusing on evaluating the accuracy and consistency of real medical notes and their ICD-10 annotations. Addressing these issues could lead to more robust synthetic data generation and improve the reliability of medical coding systems.

Model and Fine-Tuning Considerations
While LLaMA-3.1-8B demonstrated suitability for the proposed framework, future comparisons with alternative models could reveal valuable insights. This includes exploring larger models within the LLaMA family, different language model families such as Falcon (Almazrouei et al., 2023), or models building on different architectures. Additionally, the fine-tuning process could be further optimized by experimenting with different PEFT methods, hyperparameter configurations, or prompt variations. Exploring instruction-tuned base models such as LLaMA-3.1-8B Instruct could also yield improvements. These experiments, while outside the scope of this work due to time and computational constraints, present promising approaches for further optimization of the generation process.

Privacy Considerations
This work refrained from integrating privacy-enhancing methodologies to maintain simplicity and avoid the additional computational costs and utility trade-offs associated with techniques like DP. However, privacy evaluations revealed risks of longer sequences from the training data appearing in the synthetic dataset. Future research should explore privacy-enhancing methods to balance the privacy utility trade-off effectively. Moreover, expanding privacy evaluations, e.g., by assessing presence disclosure risks, would provide deeper insights into the risks associated with using and distributing synthetic discharge summaries, which is particularly essential if they are intended for real-world applications.

Evaluation Framework and Generalizability
Last, as discussed in Chapter 2, there is no consensus on standardized privacy and utility evaluations for synthetic medical notes. In this work, the synthetic data was tested as training data for two distinct downstream tasks, demonstrating its suitability and indicating, alongside strong performance in the user study, its generalizable utility. However, additional downstream tasks should be tested to further ensure the generalizability of its utility. The development of a unified evaluation benchmark for both utility and privacy is urgently needed to facilitate comparisons across different approaches and assess the performance of the proposed framework relative to other methods.

7 Conclusion

This work introduces a novel framework for generating synthetic English and Swedish medical notes, addressing the critical shortage of high-quality medical data needed to train robust clinical NLP applications. The framework successfully narrows the privacy-utility gap by generating synthetic datasets with high variety, robust privacy protections, and medical coherence. Leveraging ICD-10 codes for instruction-tuning resulted in discharge summaries with a broad vocabulary, that achieved strong utility for NER tasks and showed potential for SOTA performance in medical coding with further enhancements such as increased training data. The lower utility of the synthetic data for training medical coding models is most likely attributable to widespread noise contained in the dataset. LLaMA-3.1-8B model has proven suitable for this task, adapting to the medical domain and Swedish language without requiring prior domain-specific adaptation, all while maintaining computational efficiency due to its relatively small model size.

To further enhance the framework, we recommend addressing the limitations identified in this study and exploring additional refinements. The variety of applied evaluation tasks lays the groundwork for a unified benchmarking system, enabling consistent comparisons across different synthetic data generation approaches. The promising results of this work suggest that, with continued research, the privacy utility trade-off can be fully overcome. Synthetic clinical datasets could ultimately match or even exceed the performance of real data in training clinical applications. Achieving this goal would represent a major step forward in enabling privacy-preserving and effective clinical NLP solutions for healthcare.

Bibliography

AI Sweden Models. (2024). Llama-3-8b. https://huggingface.co/AI-Sweden-Models/Llama-3-8B

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., & Penedo, G. (2023). *The falcon series of open language models*. ArXiv. https://doi.org/10.48550/arXiv.2311.16867

Amin-Nejad, A., Ive, J., & Velupillai, S. (2020). Exploring transformer text generation for medical dataset augmentation. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 4699–4708). European Language Resources Association. Retrieved May 8, 2024, from https://aclanthology.org/2020.lrec-1.578

Ankit Pal, M. S. (2024). Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B

Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., & Veloso, M. (2021). Generating synthetic data in finance: Opportunities, challenges and pitfalls. *Proceedings of the First ACM International Conference on AI in Finance*. https://doi.org/10.1145/3383455.3422554

Baumel, T., Manoel, A., Jones, D., Su, S., Inan, H., Aaron, Bornstein, & Sim, R. (2024). *Controllable synthetic clinical note generation with privacy guarantees*. ArXiv. https://arxiv.org/abs/2409.07809

Belkadi, S., Ren, L., Micheletti, N., Han, L., & Nenadic, G. (2024). *Generating synthetic free-text medical records with low re-identification risk using masked language modeling*. ArXiv. https://arxiv.org/abs/2409.09831

Berg, H., Henriksson, A., & Dalianis, H. (2020). The impact of de-identification on downstream named entity recognition in clinical text. In E. Holderness, A. Jimeno Yepes, A. Lavelli, A.-L. Minard, J. Pustejovsky, & F. Rinaldi (Eds.), *Proceedings of the 11th international workshop on health text mining and information analysis* (pp. 1–11). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.louhi-1.1

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., & Ghosh, P. (2021). A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, *11*(18), 8319. https://doi.org/doi.org/10.3390/app11188319

Bridal, O., Vakili, T., & Santini, M. (2022). Cross-clinic de-identification of Swedish electronic health records: Nuances and caveats. In I. Siegert, M. Rigault, & V. Arranz (Eds.), *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* (pp. 49–52). European Language Resources Association. https://aclanthology.org/2022.legal-1.10

Brown, H., Lee, K., Mireshghallah, F., Shokri, R., & Tramèr, F. (2022). What does it mean for a language model to preserve privacy? *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2280–2292. https://doi.org/10.1145/3531146.3534642

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Budu, E., Etminani, K., Soliman, A., & Rögnvaldsson, T. (2024). Evaluation of synthetic electronic health records: A systematic review and experimental assessment. *Neurocomputing*, *603*, 128253. https://doi.org/https://doi.org/10.1016/j.neucom.2024.128253

Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). (2024). *Icd-11 übersetzung - bfarm* [Accessed 26 September 2024]. https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-11/uebersetzung/_node.html

Burns, E. M., Rigby, E., Mamidanna, R., Bottle, A., Aylin, P., Ziprin, P., & Faiz, O. D. (2012). Systematic review of discharge coding accuracy. *Journal of Public Health (Oxford)*, *34*(1), 138–148. https://doi.org/10.1093/pubmed/fdr054

Centers for Medicare & Medicaid Services. (1996). The health insurance portability and accountability act of 1996 (hipaa). http://www.cms.hhs.gov/hipaa/

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), *Proceedings of the 2nd machine learning for healthcare conference* (pp. 286–305, Vol. 68). PMLR. https://proceedings.mlr.press/v68/choi17a.html

CITI Program. (2024). Research, ethics, compliance, and safety training. https://about.citiprogram.org/

Dalianis, H., Henriksson, A., Kvist, M., Velupillai, S., & Weegar, R. (2015). Health bank - a workbench for data science applications in healthcare. *CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-1381/paper1.pdf

Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records*. Springer Nature. https://doi.org/10.1007/978-3-319-78503-5

Dash, S., Yale, A., Guyon, I., & Bennett, K. P. (2020). Medical time-series data generation using generative adversarial networks. In M. Michalowski & R. Moskovitch (Eds.), *Artificial intelligence in medicine* (pp. 382–391). Springer International Publishing. https://doi.org/10.1007/978-3-030-59137-3_34

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, *42*(5), 760–772. https://doi.org/10.1016/j.jbi.2009.08.007

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, *36*, 10088–10115. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. ArXiv. https://doi.org/10.48550/arXiv.1810.04805

Dong, H., Suárez-Paniagua, V., Whiteley, W., & Wu, H. (2021). Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, *116*, 103728. https://doi.org/10.1016/j.jbi.2021.103728

Dong, H., Falis, M., Whiteley, W., et al. (2022). Automated clinical coding: What, why, and where we are? *npj Digital Medicine*, *5*, 159. https://doi.org/10.1038/s41746-022-00705-7

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., ... Zhao, Z. (2024). *The llama 3 herd of models*. ArXiv. https://arxiv.org/abs/2407.21783

Durango, M. C., Torres-Silva, E. A., & Orozco-Duque, A. (2023). Named entity recognition in electronic health records: A methodological review. *Healthcare Informatics Research*, *29*(4), 286–300. https://doi.org/10.4258/hir.2023.29.4.286

Edin, J., Junge, A., Havtorn, J. D., Borgholt, L., Maistro, M., Ruotsalo, T., & Maaløe, L. (2023). Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2572–2582. https://doi.org/10.1145/3539618.3591918

Edin, J., Maistro, M., Maaløe, L., Borgholt, L., Havtorn, J. D., & Ruotsalo, T. (2024). An unsupervised approach to achieve supervised-level explainability in healthcare records. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 4869–4890). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.280

Emam, K. E., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLOS ONE*, *6*(12), e28071. https://doi.org/10.1371/journal.pone.0028071

European Parliament & Council of the European Union. (2016). *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)]. https://data.europa.eu/eli/reg/2016/679/oj

Falis, M., Gema, A. P., Dong, H., Daines, L., Basetti, S., Holder, M., Penfold, R. S., Birch, A., & Alex, B. (2024). Can gpt-3.5 generate and code discharge summaries? *Journal of the American Medical Informatics Association*, *31*(10), 2284–2293. https://doi.org/10.1093/jamia/ocae132

Fang, X., & Li, M. (2024). Privacy-preserving process mining: A blockchain-based privacy-aware reversible shared image approach. *Applied Artificial Intelligence*, *38*(1), 2321556. https://doi.org/10.1080/08839514.2024.2321556

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 968–988). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.84

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Garner, H. (2004). Engineering in genomics: The emerging in-silico scientist; how text-based bioinformatics is bridging biology and artificial intelligence. *IEEE Engineer-*

ing in Medicine and Biology, *23*(2), 87–93. https://doi.org/10.1109/memb.2004.1310989

Goldberg, C. B., Adams, L., Blumenthal, D., Brennan, P. F., Brown, N., Butte, A. J., Cheatham, M., DeBronkart, D., Dixon, J., Drazen, J., et al. (2024). To do no harm - and the most good - with ai in health care. *Nat Med*, *30*, 623–627. https://doi.org/10.1038/s41591-024-02853-7

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, *3*(1). https://doi.org/10.1145/3458754

Guan, J., Li, R., Yu, S., & Zhang, X. (2018). Generation of synthetic electronic medical record text. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 374–380. https://doi.org/10.1109/BIBM.2018.8621223

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8342–8360). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Häyrinen, K., Saranto, K., & Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International journal of medical informatics*, *77*(5), 291–304. https://doi.org/10.1016/j.ijmedinf.2007.09.001

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, *493*, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

Hiebel, N., Ferret, O., Fort, K., & Névéol, A. (2023). Can synthetic text help clinical named entity recognition? a study of electronic health records in french. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 2320–2338). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.170

Hirsch, J., Nicola, G., McGinty, G., Liu, R., Barr, R., Chittle, M., & Manchikanti, L. (2016). Icd-10: History and context. *American Journal of Neuroradiology*, *37*(4), 596–599. https://doi.org/10.3174/ajnr.A4696

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory [Conference Name: Neural Computation]. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Honovich, O., Scialom, T., Levy, O., & Schick, T. (2023). Unnatural instructions: Tuning language models with (almost) no human labor. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 14409–14428). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.806

Huang, C.-W., Tsai, S.-C., & Chen, Y.-N. (2022). PLM-ICD: Automatic ICD coding with pretrained language models. In T. Naumann, S. Bethard, K. Roberts, & A. Rumshisky (Eds.), *Proceedings of the 4th clinical natural language processing workshop* (pp. 10–20). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.clinicalnlp-1.2

Hullmann, T., & Hansson, M. (2024). *Generating synthetic training text from swedish electronic health records* [Master Thesis, Stockholm University]. DiVA. https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-231020

Igamberdiev, T., Vu, D. N. L., Kuennecke, F., Yu, Z., Holmer, J., & Habernal, I. (2024). DP-NMT: Scalable differentially private machine translation. In N. Aletras & O. De Clercq (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: System demonstrations* (pp. 94–105). Association for Computational Linguistics. https://aclanthology.org/2024.eacl-demo.11

International Organization for Standardization. (2004). *Health informatics - electronic health record - definition, scope, and context* [ISO Standard No. 20514:2005]. https://www.iso.org/standard/39525.html

Ji, S., Li, X., Sun, W., Dong, H., Taalas, A., Zhang, Y., Wu, H., Pitkänen, E., & Marttinen, P. (2024). A unified review of deep learning for automated medical coding. *ACM Computing Surveys*, *56*(12), 1–41. https://doi.org/10.1145/3664615

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, *10*(1), 1. https://doi.org/10.1038/s41597-022-01899-x

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic data – what, why and how?* ArXiv. https://arxiv.org/abs/2205.03257

Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Prasser, F., & Raisaro, J. L. (2023). Can we trust synthetic data in medicine? a scoping review of privacy and utility metrics. *medRxiv*, 2023–11. https://doi.org/10.1101/2023.11.28.23299124

Kasthurirathne, S. N., Dexter, G., & Grannis, S. J. (2021). Generative adversarial networks for creating synthetic free-text medical data: A proposal for collaborative research and re-use of machine learning models. *AMIA Summits on Translational Science Proceedings*, *2021*, 335–344.

KB Lab. (2023). Bart-base-swedish-cased. https://huggingface.co/KBLab/bart-base-swedish-cased

Kempe, S. (2024). Ärztemangel-kann die ki in der kardiologie unterstützen? *CardioVasc*, *24*(2), 14–15. https://doi.org/10.1007/s15027-024-3548-5

Kind, A. J., & Smith, M. A. (2008). Documentation of mandated discharge summary components in transitions from acute to subacute care. In K. Henriksen, J. B. Battles, M. A. Keyes, & M. L. Grady (Eds.), *Advances in patient safety: New directions and alternative approaches (vol. 2: Culture and redesign)*. Agency for Healthcare Research; Quality (US). https://www.ncbi.nlm.nih.gov/books/NBK43715/

Kormilitzin, A., Vaci, N., Liu, Q., & Nevado-Holgado, A. (2021). Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, *118*, 102086. https://doi.org/https://doi.org/10.1016/j.artmed.2021.102086

Kovačević, A., Bašaragin, B., Milošević, N., & Nenadić, G. (2024). De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artificial Intelligence in Medicine*, 102845. https://doi.org/10.1016/j.artmed.2024.102845

Kumichev, G., Blinov, P., Kuzkina, Y., Goncharov, V., Zubkova, G., Zenovkin, N., Goncharov, A., & Savchenko, A. (2024). Medsyn: Llm-based synthetic medical text generation framework. *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius,*

Lithuania, September 9-13, 2024, Proceedings, Part X, 215–230. https://doi.org/10.1007/978-3-031-70381-2_14

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626. https://doi.org/10.1145/3600006.3613165

Lai, K. H., Topaz, M., Goss, F. R., & Zhou, L. (2015). Automated misspelling detection and correction in clinical free-text records. *Journal of biomedical informatics*, *55*, 188–195. https://doi.org/10.1016/j.jbi.2015.04.008

Lamproudis, A., Svenning, T. O., Torsvik, T., Chomutare, T., Budrionis, A., Ngo, P. D., Vakili, T., & Dalianis, H. (2024). Using a large open clinical corpus for improved icd-10 diagnosis coding. *AMIA Annual Symposium Proceedings*, *2023*, 465–473.

Lee, S. H. (2018). Natural language generation for electronic health records. *NPJ digital medicine*, *1*(1), 63. https://doi.org/10.1038/s41746-018-0070-0

Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020). Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In A. Rumshisky, K. Roberts, S. Bethard, & T. Naumann (Eds.), *Proceedings of the 3rd clinical natural language processing workshop* (pp. 146–157). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.clinicalnlp-1.17

Li, J., Zhou, Y., Jiang, X., Natarajan, K., Pakhomov, S. V., Liu, H., & Xu, H. (2021). Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association: JAMIA*, *28*(10), 2193–2201. https://doi.org/10.1093/jamia/ocab112

Libbi, C. A., Trienes, J., Trieschnigg, D., & Seifert, C. (2021). Generating synthetic training data for supervised de-identification of electronic health records. *Future Internet*, *13*(5), 136. https://doi.org/10.3390/fi13050136

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013

Litake, O., Park, B. H., Tully, J. L., & Gabriel, R. A. (2024). Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *Journal of the American Medical Informatics Association*, *31*(6), 1404–1410. https://doi.org/10.1093/jamia/ocae081

Lu, W., Luu, R. K., & Buehler, M. J. (2024). *Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities*. ArXiv. https://arxiv.org/abs/2409.03444

Mawaldi, M. H., & Mladenov, M. (2024). *Synthetic data generation using large language models evaluating the utility of synthetic clinical text generated via fine-tuning llama-2 on mimic-iii data when used as training data for clinical named entity recognition* [Master Thesis, Stockholm University]. DiVA. https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-231326

Melamud, O., & Shivade, C. (2019). Towards automatic generation of shareable synthetic clinical notes using neural language models. In A. Rumshisky, K. Roberts, S. Bethard, & T. Naumann (Eds.), *Proceedings of the 2nd clinical natural language processing workshop* (pp. 35–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-1905

Micheletti, N., Belkadi, S., Han, L., & Nenadic, G. (2024). *Exploration of masked and causal language modelling for text generation*. ArXiv. https://arxiv.org/abs/2405.12630

Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., & Krallinger, M. (2020). Overview of automatic clinical coding: Annotations, guidelines, and solutions for

non-english clinical cases at codiesp track of clef ehealth 2020. *CLEF eHealth 2020*, *2696*. http://ceur-ws.org/Vol-2696/paper_263.pdf

Moore, R. C., & Lewis, W. (2010). Intelligent selection of language model training data. In J. Hajič, S. Carberry, S. Clark, & J. Nivre (Eds.), *Proceedings of the ACL 2010 conference short papers* (pp. 220–224). Association for Computational Linguistics. https://aclanthology.org/P10-2041

Murray, M. K. (2002). The nursing shortage: Past, present, and future. *JONA: The Journal of Nursing Administration*, *32*(2), 79–84. https://doi.org/10.1097/00005110-200202000-00005

Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review*, *48*, 100546. https://doi.org/10.1016/j.cosrev.2023.100546

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than english: Opportunities and challenges. *Journal of biomedical semantics*, *9*, 1–13. https://doi.org/10.1186/s13326-018-0179-8

NVIDIA. (2024). Nvidia. https://www.nvidia.com/

OpenAccess-AI-Collective. (2024). Axolotl. https://github.com/OpenAccess-AI-Collective/axolotl

OpenAI. (2024). Chatgpt (gpt-4). https://openai.com/chatgpt

Pantazos, K., Lauesen, S., & Lippert, S. (2017). Preserving medical correctness, readability and consistency in de-identified health records. *Health Informatics Journal*, *23*(4), 291–303. https://doi.org/10.1177/1460458216647760

Puri, R., Spring, R., Shoeybi, M., Patwary, M., & Catanzaro, B. (2020). Training question answering models from synthetic data. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 5811–5826). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.468

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

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

Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020). Zero: Memory optimizations toward training trillion parameter models. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16. https://doi.org/10.5555/3433701.3433727

Rankin, D., Black, M., Bond, R., Wallace, J., Mulvenna, M., & Epelde, G. (2020). Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Medical Informatics*, *8*(7), e18910. https://doi.org/10.2196/18910

Rasley, J., Rajbhandari, S., Ruwase, O., & He, Y. (2020). Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3505–3506. https://doi.org/10.1145/3394486.3406703

Ren, L., Belkadi, S., Han, L., Del-Pinto, W., & Nenadic, G. (2024). *Synthetic4health: Generating annotated synthetic clinical letters*. ArXiv. https://arxiv.org/abs/2409.09501

Singh, S., Gupta, S., Gupta, N., Sharma, N., Srivastava, L., Agarwal, V., & Modi, A. (2024). Generation and de-identification of Indian clinical discharge summaries using LLMs. In D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, & J. Tsujii (Eds.), *Proceedings of the 23rd workshop on biomedical natural language processing* (pp. 342–362). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.bionlp-1.26

Stausberg, J., Lehmann, N., Kaczmarek, D., & Stein, M. (2008). Reliability of diagnoses coding with ICD-10. *International Journal of Medical Informatics*, *77*(1), 50–57. https://doi.org/10.1016/j.ijmedinf.2006.11.005

Tang, R., Han, X., Jiang, X., & Hu, X. (2023). *Does synthetic data generation of LLMs help clinical text mining?* ArXiv. https://doi.org/10.48550/arXiv.2303.04360

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca

Tchouka, Y., Couchot, J.-F., Laiymani, D., Selles, P., & Rahmani, A. (2023). Automatic icd-10 code association: A challenging task on french clinical texts. *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, 91–96. https://doi.org/10.1109/CBMS58004.2023.00198

The Joint Commission. (2024). The joint commission. https://www.jointcommission.org/

Tiwald, P., Ebert, A., & Soukup, D. T. (2021). *Representative fair synthetic data*. ArXiv. https://arxiv.org/abs/2104.03007

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. ArXiv. https://arxiv.org/abs/2307.09288

Tseng, P., Kaplan, R. S., Richman, B. D., Shah, M. A., & Schulman, K. A. (2018). Administrative Costs Associated With Physician Billing and Insurance-Related Activities at an Academic Health Care System. *JAMA*, *319*(7), 691–697. https://doi.org/10.1001/jama.2017.19148

Vakili, T., Lamproudis, A., Henriksson, A., & Dalianis, H. (2022). Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 4245–4252). European Language Resources Association. https://aclanthology.org/2022.lrec-1.451

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vavekanand, R., & Sam, K. (2024). *Llama 3.1: An in-depth analysis of the next-generation large language model*. ResearchGate. https://doi.org/10.13140/RG.2.2.10628.74882

Velupillai, S., Dalianis, H., Hassel, M., & Nilsson, G. H. (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 [Mining of Clinical and Biomedical Text and Data Special Issue]. *International Journal of Medical Informatics*, *78*(12), e19–e26. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2009.04.005

vllm-project. (2024). Vllm. https://github.com/vllm-project/vllm

Vu, T., Nguyen, D. Q., & Nguyen, A. (2021). A label attention model for icd coding from clinical text. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. https://doi.org/10.5555/3491440.3491901

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*, 229–256. https://doi.org/10.1007/BF00992696

Wimsett, J., Harper, A., & Jones, P. (2014). Components of a good quality discharge summary: A systematic review. *Emergency Medicine Australasia*, *26*(5), 430–438. https://doi.org/10.1111/1742-6723.12285

World Health Organization. (2016). *International statistical classification of diseases and related health problems* (10th revision, Fifth edition, Vol. 1). https://icd.who.int/browse10/content/statichtml/icd10volume2_en_2016.pdf

Wu, H., Wang, M., Wu, J., Francis, F., Chang, Y.-H., Shavick, A., Dong, H., Poon, M. T., Fitzpatrick, N., Levine, A. P., et al. (2022). A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *NPJ digital medicine*, *5*(1), 186. https://doi.org/10.1038/s41746-022-00730-6

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association: JAMIA*, *25*(10), 1419–1428. https://doi.org/10.1093/jamia/ocy068

Xie, X., Xiong, Y., Yu, P. S., & Zhu, Y. (2019). Ehr coding with multi-scale feature attention and structured knowledge graph propagation. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 649–658. https://doi.org/10.1145/3357384.3357897

Xu, L., Xie, H., Qin, S.-Z. J., Tao, X., & Wang, F. L. (2023). *Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment.* ArXiv. https://arxiv.org/abs/2312.12148

Yang, Z., Kwon, S., Yao, Z., & Yu, H. (2023). Multi-label few-shot icd coding as autoregressive generation with prompt. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(4), 5366–5374. https://doi.org/10.1609/aaai.v37i4.2566

Yang, Z., Wang, S., Rawat, B. P. S., Mitra, A., & Yu, H. (2022). Knowledge injected prompt based fine-tuning for multi-label few-shot icd coding. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, *2022*, 1767–1781. https://api.semanticscholar.org/CorpusID:252762110

Yogarajan, V., Pfahringer, B., & Mayo, M. (2020). A review of automatic end-to-end de-identification: Is high accuracy the only metric? *Applied Artificial Intelligence*, *34*(3), 251–269. https://doi.org/10.1080/08839514.2020.1718343

Zhou, B., Yang, G., Shi, Z., & Ma, S. (2022). Natural language processing for smart healthcare. *IEEE Reviews in Biomedical Engineering*. https://doi.org/10.1109/RBME.2022.3210270

A ICD-10 Chapters

Table A.1: Description of MIMIC Chapters

Chapter	Blocks	Title
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D48	Neoplasms
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E90	Endocrine, nutritional and metabolic diseases
V	F00-F99	Mental and behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U99	Codes for special purposes

Table A.2: Description of SEPR Chapters

Chapter	Blocks	Title
I	K00-K14	Diseases of oral cavity, salivary glands and jaws
II	K20-K31	Diseases of oesophagus, stomach and duodenum
III	K35-K38	Diseases of appendix
IV	K40-K46	Hernia
V	K50-K52	Noninfective enteritis and colitis
VI	K55-K64	Other diseases of intestines
VII	K65-K67	Diseases of peritoneum
VIII	K70-K77	Diseases of liver
IX	K80-K87	Disorders of gallbladder, biliary tract and pancreas
X	K90-K93	Other diseases of the digestive system

B Swedish Prompt

The instructions translated into Swedish resulted in the following Alpaca template:

```
Nedan är en instruktion som beskriver en uppgift, parad med en ingång som
ger ytterligare sammanhang.  Skriv ett svar som pålämpligt sätt kompletterar
begäran.

### Instruktion:
{instruction}

### Ingång:
{input}

### Svar:
```

Figure B.1 illustrates the Swedish instruction filling the "Instruktion" field during instruction-tuning and generation of the synthetic SEPR datasets.

> Utifrån en lista med textuella beskrivningar av diagnoskoder, generera en motsvarande klinisk anteckning som ger omfattande och relevanta detaljer om patientens sjukdomshistoria, nuvarande tillstånd och behandling som mottagits påsjukhuset.

Figure B.1: Swedish instruction used for instruction-tuning LLaMA-3.1-8B with the Alpaca template.

C Configurations

Table C.1: Training configuration for instruction-tuning of LLaMA.

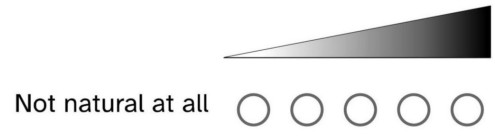
Parameter	**Value**
Load in 4-bit	true
Validation Set Size	0.05
Adapter Type	QLoRa
Sequence Length	6000 (Engl.), 4096 (Swed.)
Lora R	32
Lora Alpha	16
Lora Dropout	0.05
Gradient Accumulation Steps	2
Micro Batch Size	1
Number of Epochs	4
Optimizer	paged_adamw_32bit
Learning Rate	0.0002
LR Scheduler	cosine
Gradient Checkpointing	true
Flash Attention	true
Warmup Steps	10
Weight Decay	0.0
Deepseed	Deepspeed Zero3

Table C.2: Decoding configuration for synthetic note generation.

Parameter	Value
Decoding Strategy	Random sampling
Top-k	-1 (all)
Max tokens	6000 (Engl.), 4000 (Swed.)
Temperature	1.0
Repetition Penalty	1.5
N (Number of output sequences)	5

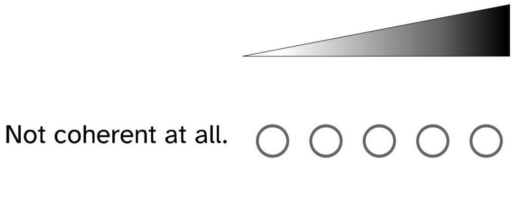
D User Study Questionnaire

20. On a scale of 1 to 5, how would you rate the readability and language style of this patient note? Disregard any contents of this note and focus only on linguistic aspects. Do not consider abnormalities that are due to pseudonimization.



21. If you can name any specific justification for your given answer, please state it here briefly. You're also encouraged to copy parts of the text that led to your previous answer and explain.

22. On a scale of 1 to 5, how would you rate the medical contents of this note? Disregard any linguistic unusualities and focus only on the medical information.



23. If you can name any specific justification for your given answer, please state it here briefly. You're also encouraged to copy parts of the text that led to your previous answer and explain.

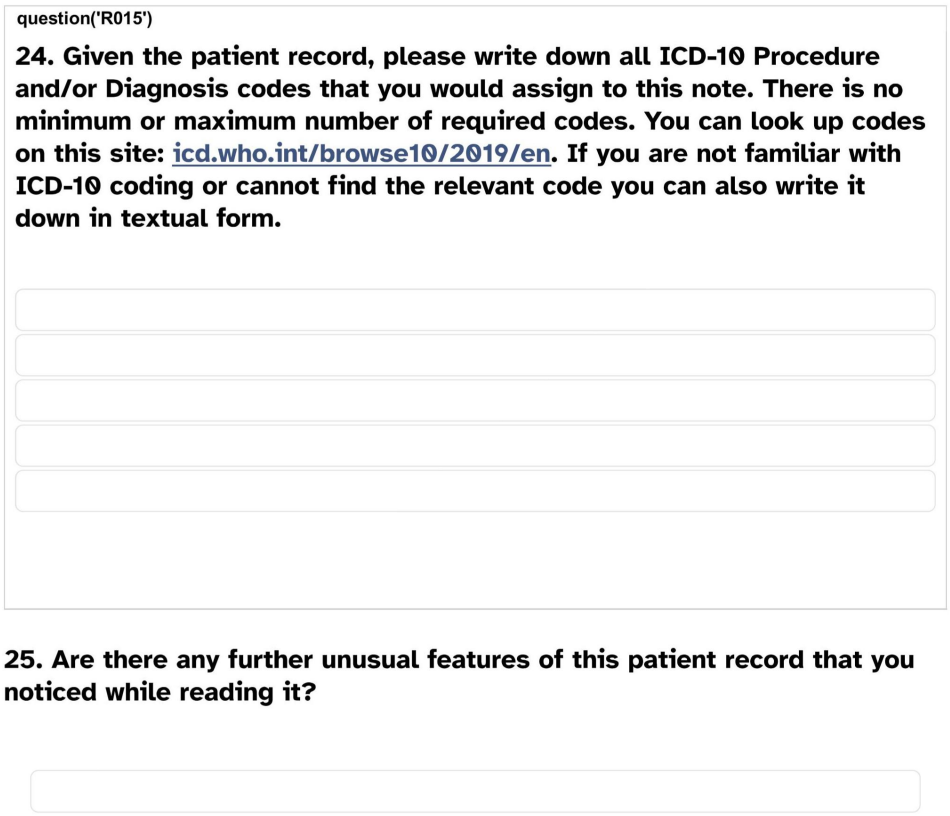


Figure D.1: Study Questionnaire

Figure D.1 depicts the user study questionnaire. Participants were presented with both synthetic and real medical notes. After reading each note, they were asked to respond to the presented questions. The ICD-10 coding task was assigned only to selected participants.

E Synthetic Medical Note Examples

Chief Complaint: left hip pain

Major Surgical or Invasive Procedure: ___: left THA by ___

History of Present Illness: ___ year old male with left hip osteoarthritis which has failed conservative management now presents for left hip replacement on ___

Past Medical History: Left hip pain s p cortisone injections last in ___, H o MI in ___, PCI s p stent ___ , ischemic , cardiomyopathy, H o afib in ___. off anti coagulation now , Stroke, HTN, osteoarthritis, hyperlipidemia, falls ,

Postoperative course was remarkable for the following: POD 1, calcium was 6.6 and magnesium was 1.9. This was repleted per recommendations. POD 2, phosphate was 1.9 and this was repleted. Patient's gabapentin was discontinued due to hypotension following 2 doses. Otherwise, pain was controlled with a combination of IV and oral pain medications. The patient received Aspirin 325 mg twice daily for DVT prophylaxis starting on the morning of POD 1. The surgical dressing was changed and the Silverlon dressing was removed 48 hours after surgery. A dry sterile dressing was reapplied. The patient was seen daily by physical therapy. Labs were checked throughout the hospital course and repleted accordingly. At the time of discharge the patient was tolerating a regular diet and feeling well. The patient was afebrile with stable vital signs. The patient's hematocrit was acceptable and pain was adequately controlled on an oral regimen. The operative extremity was neurovascularly intact and the dressing was intact. The patient's weight bearing status is weight bearing as tolerated on the operative extremity with no hip bridging or repetitive resistant hip flexion. Walker or two crutches. Wean as tolerated. Mr. ___ is discharged to home with services in stable condition. 13. Metoprolol Succinate XL 75 mg PO DAILY 14. Rosuvastatin Calcium 40 mg PO QPM

Discharge Disposition: Home With Service

Discharge Diagnosis: left hip osteoarthritis

Figure E.1: Example of synthetic MIMIC discharge summary extracted from the synthetic MIMIC-S dataset.

Inlagd för några veckor sedan pågrund av blod i avföringen. Kommer nu för endoskopisk undersökning (gastroskopi och koloskopi). Patienten mår utmärkt bra och har inte haft några symtom sedan han blev utskriven., Således tvåtill synes helt benigna polyper i kolon som idag avlägsnats med slyngan. Ingen blödning., Via inremitterande., Inremitterande., I distala sigmoideum påcirka 30 cm : s höjd från den ileocekala valveln noteras en cirka 4 mm - stor flack polyp med utseende som hyperplastisk polyp. Polypen avlägsnas med vanlig biopsitång. Inga omkringliggande förändringar. Inga tecken till blödning efteråt. I distala transversum samt proximala descendens ses ytterligare en cirka 5 mm stor flack polyp med utseende som hyperplastisk polyp. Denna avlägsnas med vanlig biopsitång. återigen inga omkringliggande förändringar. Inget blod i lumen., Dr Emelie Luusua ., Blod i avföring. Utredning.

Figure E.2: Example of synthetic SEPR discharge summary extracted from the synthetic SEPR-M dataset.