ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Large language models-in-the-loop: Leveraging expert small artificial intelligence models for multilingual anonymization and de-identification of protected health information

Murat Gunay[1]*, Bunyamin Keles[2†], and Raife Hizlan[1†]

[1]Department of Research and Development, AI Handed LLC, Lewes, Delaware, United States of America

[2]Department of Health Management, Hacettepe University Institute of Social Sciences, Ankara, Turkey

Abstract

The rise of chronic diseases and pandemics, such as COVID-19 has emphasized the need for effective patient data processing while ensuring privacy through anonymization and de-identification of protected health information. Anonymized data facilitates research without compromising patient confidentiality. This paper introduces expert small artificial intelligence (AI) models developed using the large language model (LLM)-in-the-loop methodology to meet the demand for domain-specific de-identification of named entity recognition (NER) models. These models overcome the privacy risks associated with LLMs used through application programming interfaces by eliminating the need to transmit or store sensitive data. More importantly, they consistently outperform LLMs in de-identification tasks, offering superior performance and reliability. Our de-identification NER models, developed in eight languages—English, German, Italian, French, Romanian, Turkish, Spanish, and Arabic—achieved F1-macro score averages of 0.931, 0.960, 0.955, 0.937, 0.930, 0.963, 0.957, and 0.922, respectively. These results establish our de-identification NER models as the most accurate healthcare anonymization solutions, surpassing existing small models and even general-purpose LLMs, such as GPT-4o. While Part I of this series introduced the LLM-in-the-loop methodology for biomedical document translation, this second paper showcases its success in developing cost-effective expert small NER models in de-identification tasks. Our findings lay the groundwork for future healthcare AI innovations, including biomedical entity and relation extraction, demonstrating the value of specialized models for domain-specific challenges.

Keywords: De-identification; Health Insurance Portability and Accountability Act; Protected health information; Patient safety; Large language models-in-the-loop; Anonymization

†*These authors contributed equally to this work.*

***Corresponding author:**
Murat Gunay
(murat.esra.gunay@gmail.com)

Citation: Gunay M, Keles B, Hizlan R. Large language models-in-the-loop: Leveraging expert small artificial intelligence models for multilingual anonymization and de-identification of protected health information. *Artif Intell Health*. 2026;3(1):138-151. doi: 10.36922/AIH025120021

Received: March 19, 2025

1st revised: June 19, 2025

2nd revised: July 2, 2025

3rd revised: July 22, 2025

4th revised: August 21, 2025

Accepted: August 26, 2025

Published online: September 19, 2025

1. Introduction

Patient data are essential for improving public health, expanding preventive health services, preventing diseases, and formulating necessary health policies. Recent studies show that almost all (99%) hospitals in the United States[1] use electronic health records

(EHR). Similarly, Wales, Scotland, Denmark, and Sweden have implemented EHRs in the past few years. In contrast, the United Kingdom still lacks a nationally accessible health data system, despite the COVID-19 pandemic underscoring the critical importance of EHR data.[2] EHRs enable the examination of disease trends, support predictive modeling, and inform the development of health policies.

Technology, which has become more complex and has developed with medical practices, necessitates the development of methods that will protect patient privacy.[3] With information security and information leakage recently gaining more importance, patient safety may have significant consequences beyond ethical violations in fundamental and health law.[4] Personal data are sensitive information associated with an individual and protected by various laws.[5] Personal privacy data in healthcare, known as protected health information (PHI), includes private information, such as a patient's health history, treatments received, and more.[6]

EHRs contain both valuable clinical information and PHI. While EHRs are a rich data source for research, their usability is restricted due to the confidentiality of PHI.[7-10] For example, the Health Insurance Portability and Accountability Act (HIPAA) regulates the use of 18 types of PHI, such as name, phone number, dates, and more (Table 1).[11,12] Therefore, PHI must be extracted from the text before EHR data can be used. Automating de-identification systems is needed since manually extracting PHI is time-consuming and costly. In addition, coordination between annotators is also an important consideration.[13,14] While early approaches to de-identification relied on complex rules to detect PHI, recent developments use machine learning (ML) methods and train on expert-annotated records. Hybrid systems integrate practices as features into statistical models, such as conditional random fields (CRFs).[15]

The de-identification method makes it possible to use EHRs in research by removing confidential information.[15] Basic de-identification rules include removing direct identifying statements, such as name, date, and more. Advanced statistical methods anonymize the data, reducing de-identification risk.[16] However, new techniques may also introduce unknown privacy risks. Therefore, continuous evaluation and improvement efforts are necessary.[17] Advanced methods can enable extensive collections of EHRs to be used efficiently and securely in research.

According to HIPAA, there are two possible methods of identity masking. The "Expert Determination" method, which requires employing an expert in the field, involves a small risk in identifying the individual whose information is used and is performed using different statistical methods. In this method, the expert must have sufficient experience and knowledge. The other method is the "Safe Harbor" method, which involves de-identifying 18 pre-determined relevant identifiers that must be removed and/or modified from the corpus.[11,18] In studies using deep learning (DL) models, the Safe Harbor method is used, and the relevant PHI is de-identified. The de-identification process was primarily guided by the HIPAA framework for protecting patient health information. For non-English languages, entity definitions were carefully adapted to reflect HIPAA requirements and global privacy standards, such as the General Data Protection Regulation for European languages. Where relevant, local legal and regulatory frameworks were considered to ensure our approach aligns with both United States and international best practices for data privacy.

Table 1. Types of protected health information

No	Type of protected health information
1	Names
2	All geographic subdivisions smaller than a state
3	Dates
4	Telephone numbers
5	Vehicle identifiers
6	Fax numbers
7	Device identifiers and serial numbers
8	Emails
9	Uniform resource locators
10	Social security numbers
11	Medical record numbers
12	Internet protocol addresses
13	Biometric identifiers
14	Health plan beneficiary numbers
15	Full-face photographic images and any comparable images
16	Account numbers
17	Certificate/license numbers
18	Any other unique identifying number, characteristic, or code

The lack of comprehensive data privacy frameworks can lead to vulnerabilities, leaving sensitive patient information susceptible to breaches and misuse. Despite efforts to anonymize this data, reidentification is still feasible through just a few spatiotemporal data points.[19] Recent advancements in privacy-preserving technologies have increased adoption,[20] particularly in artificial intelligence (AI) and big data analytics. These technologies are vital in addressing major global health challenges by enhancing access to healthcare, promoting health, preventing diseases, and improving the overall experience for healthcare professionals and patients. AI, coupled with

big data analytics, is the backbone for many innovations in digital health, driving improvements in care delivery and decision-making processes. These domains are supported by additional technologies, such as the Internet of Things, next-generation networks (e.g., 5G), and privacy-preserving platforms, such as blockchain.[21,22]

However, questions remain regarding the accountability for AI and large language models (LLMs). Given that AI lacks autonomy and sentience, it cannot hold moral responsibility, leaving uncertainty about who should be accountable for its decisions and actions.[23]

LLMs, particularly GPT-4o, have demonstrated remarkable success in zero-shot de-identification tasks. However, significant challenges remain in utilizing proprietary paid application programming interfaces (APIs) and open-source small LLMs. Paid APIs raise privacy concerns, as hospitals are often unwilling to transmit sensitive patient data to cloud-based services. Conversely, locally deployed small LLMs present difficulties in production due to their limited accuracy, resource-heavy requirements, and complex deployment processes. In response to these challenges, we propose expert small AI models—lightweight, plug-and-play solutions that offer superior performance and accuracy compared to LLMs, while being more practical and efficient for deployment in secure, on-premises environments. Hence, we did not directly implement LLMs to address these challenges; instead, we developed expert small named entity recognition (NER) models using the LLM-in-the-loop methodology.

2. What is the LLM-in-the-loop?

From our perspective, the "LLM-in-the-loop" methodology is an integral part of the development process for expert small models, without relying on LLMs as the final solution. Instead of directly using LLMs for tasks, we utilize them selectively at various stages, such as synthetic data generation, rigorous evaluation, and agent orchestration, to improve the performance of smaller, domain-specific models. This approach allows us to benefit from the capabilities of LLMs while keeping the models efficient, focused, and specialized for specific tasks.

Recently, there has been a growing emphasis on the work done within the scope of LLM-in-loops. Studies have shown that LLMs perform better on tasks traditionally completed by humans,[24-26] and the potential for effective utilization of LLMs is emphasized. It is seen that the innovative approach of "LLM-in-the-loop" is used in different fields today. In a study conducted to analyze social media content and reveal hidden themes,[27] the advanced capabilities of LLMs were leveraged to gain a deeper understanding of social media content by analyzing social media messages, discovering thematic structures and nuances in texts, and effectively matching texts to themes. Another study using the LLM-in-loops technique to improve the performance of LLMs was aimed at continuously improving the model outputs through iterative feedback loops, and this was applied in a study in the medical field. The aim was to increase the accuracy and reliability of the model and reduce hallucinations. The LLM-in-loops study, which integrated human expert evaluation of model outputs, feedback provision, and subsequent retraining, focused on reducing model errors and obtaining more reliable results in medical question-answering and summarization tasks.[28]

Another study, which examined the potential of LLMs to recognize and examine intertextual relationships in biblical and Koine Greek texts, highlighted how LLMs evaluate different intertextual scenarios and how these models can detect direct quotations, allusions, and echoes between texts. The study also mentioned the ability of LLMs to generate intertextual observations and connections and the potential of these models to reveal new insights. However, it is noted that the model had difficulties with long query texts and can create incorrect intertextual connections, which reveals the importance of expert evaluation.[29]

We first used the "LLMs-in-the-loop" method in the context of biomedical document translation.[30] In the previous work, we demonstrated its success in developing cost-effective expert small NER models for de-identification tasks. Our findings laid the groundwork for future healthcare AI innovations, including biomedical entity and relation extraction, and demonstrated the value of specialized models for domain-specific challenges.

As we navigate the evolving landscape of AI in healthcare, the LLM-in-the-loop methodology stands out as a transformative approach. Recent studies have highlighted its capacity to enhance the performance of the models by leveraging human expertise to refine outputs continuously. This innovative strategy addresses the traditional challenges faced in biomedical text processing, such as accuracy and reliability, and mitigates issues, such as hallucinations that commonly occur in AI-generated content. By fostering a symbiotic relationship between human input and ML, we pave the way for advanced applications, including more effective biomedical entity extraction and improved medical summarization techniques. Ultimately, this research underscores the significance of integrating human intelligence with AI capabilities, setting the stage for more robust and trustworthy healthcare solutions.

3. Background

The de-identification model, called the NER classification model, can be considered under four headings:[31] (i) Rule-based models, (ii) ML models, (iii) hybrid models, and (iv) DL models.

Techniques, such as rule-based models and dictionaries can be easily implemented without labels but are vulnerable to input errors.[31-34] ML methods, such as support vector machines and CRFs can recognize complex patterns but require large amounts of labeled data and feature engineering, and are poor at generalization.[35-37] Hybrid systems combine rule-based and ML models, providing high accuracy but requiring intensive feature engineering.[38,39]

Considering the disadvantages of the past three approaches in creating the de-identification systems, the latest state-of-the-art systems employ DL techniques to achieve better results than the best hybrid systems without requiring a time-consuming feature engineering process. DL is an ML subset using multilayered artificial neural networks and is very successful in most natural language processing (NLP) tasks. Recent advances in DL and NLP (especially in the field of NER) enable the systems to outperform the winning hybrid system proposed by Yang and Garibaldi[39] on the 2014 i2b2 de-identification challenge dataset.[31,35]

De-identifying unstructured data is a widely recognized challenge[40] in NLP, involving two key tasks: Identifying PHI and replacing it through masking or obfuscation. Research has primarily focused on PHI identification. Early de-identification approaches,[41,42] especially in healthcare, were rule-based, using regular expressions, syntactic rules, and specialized dictionaries to detect PHI, such as phone numbers and emails. However, they struggled with identifying more complex entities, such as names and professions and required significant adjustments to function in different datasets, limiting their flexibility. The 2014 i2b2 project[34] introduced automatic de-identification, fueling the advancement of ML and DL models for more accurate PHI detection. Early ML methods, such as CRF,[43] used hand-crafted features and lexical rules,[44] signaling a shift to more adaptive and scalable approaches.

Work in the de-identification context has achieved human-level accuracy in de-identifying clinical notes from research datasets. Still, challenges remain in scaling this success to large, real-world environments. The hybrid context-based model outperformed traditional NER models by 10% in the i2b2-2014 benchmark. It also has significantly fewer errors (93% accuracy) compared to ChatGPT (60% accuracy).[45]

Large language-based methods have been used in the development of de-identification models. However, these are still in the early stages, and further development is still needed to protect the privacy and security of health data.[46] The continued need to use APIs in LLM models and the challenge of storing patient data reveal that expert models are still needed.

4. Methodology

This section details the purpose of the research, the datasets employed, the methods for training and testing, the data preparation process, and the modeling and evaluation phases. The protection of personal data, compliance with legal regulations, and mitigation of risks associated with processing sensitive patient information are central to this study.

Our LLM-in-the-loop methodology leveraged LLMs at key stages, such as synthetic data generation, labeling, and evaluation, focusing on developing high-performance, expert small models. To this end, we used a combination of proprietary closed-source data, open-source datasets, and synthetic data, all annotated by our labeling team in accordance with i2b2 labeling logic. Incorporating synthetic data and LLM-assisted labeling further enhanced the scope and quality of our training datasets.

For English-language de-identification NER models, we utilized the entire dataset for training. The i2b2 test dataset served as the exclusive test set for evaluation purposes, allowing us to benchmark performance with high precision. For non-English languages, we applied an 80–20 split for training and testing. In addition, our medical translation models[30] were used to translate the English datasets into non-English languages, generating high-quality parallel datasets across multiple languages.

In the data pre-processing phase, we employed language-specific tools to ensure accurate de-identification across different languages. The "Stanza" library was utilized for Romanian-language tasks, while the Natural Language Toolkit library was used for other languages. Word tokenization for all datasets was performed using the "word-punct tokenizer" from the Natural Language Toolkit library.

For evaluation, we adopted the strict evaluation method, where both the chunk and the label had to match to be considered a correct prediction. This rigorous approach ensured the accuracy and reliability of our models, particularly in handling PHI.

By integrating proprietary, open-source, and LLM-synthesized datasets and utilizing real and translated data, this methodology demonstrates the capability of

expert small models to provide accurate, domain-specific de-identification solutions. Our approach minimizes reliance on large LLMs while ensuring privacy and top-tier performance in medical data anonymization.

The results in Tables 2 and 3 were achieved using a structured and detailed prompt to extract PHI from clinical notes. The prompt provided a comprehensive list of entity definitions, such as "AGE," "CITY," "DEVICE," and "ORGANIZATION," along with examples for clarity. It instructed GPT-4o to identify and mark entities using a consistent tagging format (e.g., BEGINER LABEL CHUNK ENDNER) while preserving the original text. Specific guidelines were included for nuanced cases, such as excluding titles (e.g., "Dr.") from names and marking only actual dates for the "DATE" label. This rigorous approach ensured precision in high-performing categories and highlighted areas for improvement in more challenging entities. The prompt used in the study is presented in Appendix A.

4.1. Datasets

"i2b2-2014" is a research project (https://portal.dbmi.hms.harvard.edu) on de-identification and heart disease in clinical texts, and its labeling logic was used in our study. For English-language de-identification NER models, we used a training dataset composed of approximately 78% synthetic, AI-generated data and 22% proprietary, closed-source data. The i2b2 training dataset was not used at any stage. The synthetic data were created by generating artificial EHR records and annotating PHI entities using our automated

Table 2. Categories and comparison of the de-identification model in English i2b2-protected health information

Protected health information/model owners	Our scores	Khin *et al.*[31]	Yang and Garibaldi[39]	Kocaman *et al.*[45]	GPT-4o
AGE	0.981	0.973	0.948	0.964	0.781
CITY	0.944	0.909	0.776	0.949	0.917
COUNTRY	0.881	0.805	0.303	0.920	0.802
DATE	0.978	0.987	0.976	0.996	0.494
DEVICE	0.762	-	-	0.286	0.217
DOCTOR	0.966	0.962	0.945	0.980	0.743
HOSPITAL	0.920	0.928	0.864	0.972	0.575
IDNUM	0.867	0.756	0.838	0.909	0.288
LOCATION-OTHER	1	-	-	0.722	-
MEDICALRECORD	0.942	0.979	0.971	0.980	0.716
ORGANIZATION	0.876	0.719	0.427	0.874	0.400
PATIENT	0.967	0.961	0.933	0.967	0.535
PHONE	0.868	0.970	0.952	0.978	0.456
PROFESSION	0.900	0.899	0.688	0.925	0.583
STATE	0.961	0.932	0.863	0.969	0.932
STREET	0.985	0.989	0.978	0.996	0.900
USERNAME	0.962	0.957	0.978	0.954	0.635
ZIP	0.989	0.982	0.986	0.982	0.975
Macro score average	0.931	0.919	0.840	0.863	0.548

Table 3. i2b2 test set scores using GPT-4o

Entity	Precision	Recall	F1-score
B-AGE	0.688	0.937	0.791
B-CITY	0.948	0.904	0.925
B-COUNTRY	0.908	0.718	0.832
B-DATE	0.808	0.834	0.821
B-DEVICE	0.132	0.625	0.217
B-DOCTOR	0.956	0.810	0.877
B-HOSPITAL	0.916	0.675	0.775
B-IDNUM	0.340	0.672	0.531
B-MEDICALRECORD	0.960	0.794	0.869
B-ORGANIZATION	0.303	0.695	0.422
B-PATIENT	0.852	0.779	0.814
B-PHONE	0.757	0.726	0.741
B-PROFESSION	0.695	0.637	0.665
B-STATE	0.902	0.974	0.937
B-STREET	0.933	0.927	0.930
B-USERNAME	0.563	0.728	0.635
B-ZIP	1.000	0.993	0.997
I-AGE	0.175	0.453	0.253
I-CITY	0.872	0.852	0.862
I-COUNTRY	0.800	0.615	0.696
I-DATE	0.755	0.755	0.755
I-DEVICE	0.133	1.000	0.235
I-DOCTOR	0.490	0.767	0.605
I-HOSPITAL	0.891	0.715	0.793
I-IDNUM	0.392	0.550	0.458
I-LOCATION	0.114	0.121	0.118
I-MEDICALRECORD	0.763	0.457	0.571
I-ORGANIZATION	0.246	0.750	0.370
I-PATIENT	0.535	0.652	0.587
I-PHONE	0.749	0.755	0.752
I-PROFESSION	0.628	0.693	0.659
I-STATE	0.917	0.688	0.786
I-STREET	0.839	0.964	0.897
I-ZIP	0.714	0.625	0.667
O	0.986	0.984	0.985
Macro average	0.5819	0.6247	0.5775
Weighted average	0.970	0.967	0.968

Note: Data using the "beginning, inside, outside" (BIO) format.

labeling pipeline. The i2b2 test dataset served as the exclusive evaluation benchmark set, per standard practices. For non-English models, the entire training data were derived by translating[30] the English dataset into the target language, followed by an 80:20 split for training and testing. No open-source test sets were used for non-English languages; the i2b2 test set was used exclusively for English evaluation.

In addition, we utilized several NLP techniques and open-source third-party tools (LangTest by John Snow Labs: https://langtest.org/) to enhance and augment the training datasets. Although the i2b2 2014 dataset was not utilized for training purposes, we provide relevant information and statistics to offer a more comprehensive understanding of its role in our evaluation process. i2b2/UTHealth is a dataset focused on identifying medical risk factors for coronary artery disease in the medical records of diabetic patients, where risk factors include hypertension, hyperlipidemia, obesity, smoking status, and family history, as well as diabetes, coronary artery disease, and indicators suggestive of the presence of these diseases.[47] The i2b2 dataset consists of 1,304 progress notes of 296 diabetic patients. All PHIs in the i2b2-2014 dataset were already de-identified by the dataset's creators before our study, using automated replacement of real entities with synthetic, randomly generated identifiers (e.g., fictitious names or dates). We did not apply any further de-identification to this dataset. Our study used the i2b2-2014 test set exclusively for evaluation and benchmarking. The term "randomly" refers to the replacement strategy used by the original dataset providers to ensure that PHI tokens were substituted with realistic but non-identifying values. The PHIs in this dataset were first categorized into HIPAA categories and then into i2b2-PHI categories, as shown in Table 4. Overall, the i2b2 dataset contains 56,348 sentences with 984,723 individual tokens, of which 41,355 are individual PHI tokens representing 28,867 particular PHI instances.[31]

Table 4. Protected health information categories of the HIPAA and our study's entities

HIPAA	i2b2 dataset	Our dataset
Name	Patient, doctor, username	Patient, doctor profession
Profession	Profession	Profession
Location	Street, city, state, country, zip, hospital, organization	Street, city, country, zip, hospital, location, organization
Age	Age	Age
Date	Date	Date
Contact	Phone, fax, email, URL, IP address	Phone, fax, email
ID	Medical record, ID no, SSN, license no	Medical record, ID, ID no, SSN, sex, family

Abbreviations: HIPAA: Health Insurance Portability and Accountability Act; ID: Identification; IP: Internet protocol; SSN: Social security number; URL: Uniform resource locator.

In the literature review, it is seen that there are relative limitations in terms of data sets in de-identification model studies other than English. For this reason, it can be stated that only a few de-identification models have been developed for different languages. In this respect, the de-identification models in different languages developed in this study will contribute to the literature and data scientists working on these models and the health institutions that will use them.

4.2. Experimental setup and metrics

4.2.1. Clinical English de-identification model

The corpus of clinical admission discharge and private clinical reports from private hospitals and healthcare organizations was used to develop the English de-identification model. Labeling was done according to the i2b2-2014 data principles as described previously. The labels employed in the model, which uses a fine-tuned version of the "microsoft/deberta-v3-small" model as an embedding, are shown in Table 5.

In the study, 10 labels were used for the rule-based method, and 18 labels were used for the DL methods. The training dataset was augmented for these labels since "ORGANIZATION," "PROFESSION," and "LOCATION-OTHER" entities gave low results due to the first training process with the DL method. The augmentation stages of the model were performed as follows. First, a fake chunk data frame was created for each label in various formats. Sentences with the labels "ORGANIZATION," "PROFESSION," and "LOCATION-OTHER" in the training dataset and CoNLL file were extracted. Each labeled chunk was removed and replaced with label abbreviations. The sentences were translated from English to the working language. For the translation, our medical translation models used the work by Keles *et al.*[30] The label abbreviations in the new sentences were replaced with new chunks of those labels from the fake data frame.

This new dataset was converted to "beginning, inside, outside" (BIO) format and added to the training dataset. The BIO tagging scheme is a widely adopted convention for NER tasks, where each token in a sentence is labeled as either the beginning (B) of an entity, inside (I) an entity, or outside (O) any entity. This format allows the model to accurately learn both the boundaries and the types of entities in text, and is particularly effective for training DL models for sequence labeling.

The model's performance, implemented with the DL method used in this study, was tested with the i2b2-2014

Table 5. English de-identification model labels

Method	Rule-based	Deep learning
Labels	ACCOUNT, DLN, EMAIL, FAX, IP, LICENSE, PLATE, SSN, URL, VIN	AGE, CITY, COUNTRY, DATE, DEVICE, DOCTOR, HOSPITAL, IDNUM, LOCATION-OTHER, MEDICAL RECORD, ORGANIZATION, PATIENT, PHONE, PROFESSION, STATE, STREET, USERNAME, ZIP

Abbreviations: ID: Identification; IP: Internet protocol; URL: Uniform resource locator; VIN: Vehicle identification number.

test set. It was observed that the retrained dataset with augmented labels showed better classification results when evaluated using the i2b2 2014 test set.[33] In the de-identification study conducted in English and with the DL method, a learning rate of 2×10^5, a max sentence length of 512, a batch size of two, and ten epochs of training were used. For the rule-based method, regexes suitable for each format were created for the selected labels.

4.2.2. Non-English de-identification models

To understand which labels could be used in de-identification models and which labels would be appropriate for which aggregates, and to determine the principles, the labeling team organized meetings with relevant hospital staff to develop the models in German, French, Italian, Romanian, Spanish, and Turkish. Data collected from clinical admission reports, discharge reports, and special clinic reports obtained from hospitals and health institutions were labeled according to these principles.

The training was conducted with the obtained data set. In the study, the 0.20 parts of the dataset determined during the division process were used as the test dataset. The dataset was pre-processed and converted into BIO format. For German: bert-base-german-cased, for Italian: bert-base-italian-cased, for French: camembert-bio-base, for Romanian: bert-base-ro-cased, for Turkish: bert-base-turkish-cased, and for Spanish: roberta-base-biomedical-clinical-es were used as embeddings.

The augmentation stages of the other language models were performed as follows. In the dataset used for the English de-identification model, a fake chunk data frame was created for each label in various formats. Each labeled chunk was removed and replaced with label abbreviations. The sentences were translated from English to the working language. For the translation, our medical translation models used the work by Keles *et al.*[30] The label abbreviations in the new sentences were replaced with new chunks of those labels from the fake data frame. This new data set was converted to BIO format and added to the train data set.

First, for each entity label (e.g., ORGANIZATION, PROFESSION, LOCATION-OTHER), we created a data frame of "fake chunks": Short text snippets (typically phrases or entity-level fragments) that could be substituted in place of real PHI. For example, for the label "ORGANIZATION," the fake chunk data frame included items, such as "Springfield General Hospital" or "Central Medical Associates." During augmentation, original EHR sentences with labeled PHI were modified by replacing real entities with randomly sampled fake chunks. For instance, the original EHR sentence reads, "The patient was transferred to Mercy Hospital under the care of Dr. Smith." After augmentation: "The patient was transferred to Springfield General Hospital under the care of Dr. Williams." This process was repeated for each target entity in the dataset. The modified sentences were then converted to the BIO format, where each token is labeled as "B-ENTITY," "I-ENTITY," or "O" (non-entity), as required for training sequence labeling models.

This approach aims to increase training diversity, prevent model memorization of real PHI, and ensure compliance with privacy standards. Detailed examples and code will be available in the project repository upon publication.

The de-identification research was performed with the DL method in seven languages other than English, a learning rate of 2×10^5, a max sentence length of 512, a batch size of 16 (batch size = 2 in Romanian), and ten epochs were trained.

5. Results

The results obtained from the de-identification NER models are shown in Table 2. In addition, the results obtained using GPT-4o and the comparison results of other studies utilizing the same dataset with the results obtained in this study are also included in the same table.

As seen in Table 2, the model realized in this study includes PHIs not used in other studies, and satisfactory results were obtained. When the performance results of the studies are compared with the results of this study, it is determined that new state-of-the-art values were obtained with this study. Although the train was performed with 18 PHI labels (DEVICE and LOCATION-OTHER labels were not used in other studies), and high scores of some labels were not obtained, the F1 macro score (0.931) obtained in this study was higher than the other models, and a new state-of-the-art value was achieved.

GPT-4o performs well in classes, such as "CITY," "COUNTRY," "ZIP," and "STATE," achieving high precision, recall, and F1-scores. However, it struggles significantly with "IDNUM," "LOCATION-OTHER," "ORGANIZATION," "EMAIL," "FAX," and "DEVICE," where the scores are notably low. The macro average (0.5757) indicates that the model's performance varies significantly across classes, with weaker performance in certain categories. In contrast, the micro average (0.5907) is slightly higher, reflecting the model's stronger performance in more frequent classes, but overall, the scores are low.

The results obtained for 13 labels in German, Italian, and French are shown in Table 6, while the results obtained for Turkish (13 labels), Spanish (14 labels), and Romanian (14 labels) are shown in Table 7.

The table presents F1-scores for de-identification tasks across German, Italian, and French datasets. Overall, the German model achieved the highest macro-average (0.960), followed by Italian (0.955) and French (0.937). "DATE" and "PHONE" categories exhibited consistently strong performance across all languages, achieving nearly perfect scores (0.995). In contrast, the "ORGANIZATION" category showed notable variability, with the French model scoring significantly lower (0.699). These results highlight the robustness of the models in categories, such as "AGE," "IDNUM," and "ZIP" while identifying areas for improvement in language-specific challenges, particularly for underperforming categories, such as "ORGANIZATION" in French (Table 6). However, since it was impossible to find any benchmark tests for these languages, comparing the scores obtained in this study was impossible.

Table 7 highlights strong performances for Turkish (0.963) and Spanish (0.957) models, followed by Romanian (0.930) and Arabic (0.922). Categories, such as "DATE," "PHONE," and "MEDICAL RECORD" achieved near-perfect scores across languages, demonstrating model robustness. Lower scores were observed for "CITY" and "ORGANIZATION" in Romanian and Arabic, indicating room for improvement. Missing or language-specific labels (e.g., EMAIL, SSN) show variability in evaluation, reflecting dataset differences. Turkish and Spanish excel in most categories, with consistent performance across diverse labels.

Table 3 presents the evaluation results for the B- and I- tags separately. The model achieved high overall accuracy (0.9672). Classes, such as "B-STATE," "I-CITY," and "I-COUNTRY" performed very well, while "B-EMAIL," "B-FAX," and "I-LOCATION" had lower precision and recall values, indicating challenges in identifying these entities. The macro average (0.5775) was lower than the weighted average (0.968), suggesting that less frequent or more difficult classes were pulling down the macro scores,

Table 6. German, Italian, and French de-identification model outputs

Language/labels	German	Italian	French
AGE	0.985	0.983	0.981
CITY	0.963	0.922	0.939
COUNTRY	0.954	0.906	0.926
DATE	0.997	0.998	0.998
DOCTOR	0.944	0.955	0.952
HOSPITAL	0.981	0.975	0.915
IDNUM	0.987	0.998	0.997
ORGANIZATION	0.865	0.916	0.699
PATIENT	0.903	0.920	0.918
PHONE	0.995	0.995	0.995
PROFESSION	0.980	0.917	0.941
STREET	0.945	0.952	0.949
ZIP	0.975	0.982	0.975
Macro score average	0.960	0.955	0.937

Table 7. Turkish, Spanish, Romanian, and Arabic de-identification model outputs

Language/labels	Turkish	Spanish	Romanian	Arabic
AGE	0.988	0.980	0.984	0.980
CITY	0.979	0.958	0.889	0.867
COUNTRY	0.917	0.969	0.899	0.881
DATE	0.997	0.997	0.973	0.987
DOCTOR	0.953	0.969	0.966	0.908
EMAIL	-	0.994	0.857	-
HOSPITAL	0.942	0.976	0.935	0.988
ID	-	0.995	-	-
IDNUM	0.979	-	0.997	0.962
LOCATION	1	-	0.846	-
MEDICAL RECORD	1	0.991	0.999	-
ORGANIZATION	0.975	0.734	0.768	0.978
PATIENT	0.946	0.967	0.944	0.856
PHONE	0.982	0.981	1	0.984
PROFESSION	0.924	0.912	0.888	0.877
SEX	-	0.971	-	-
SSN	-	0.937	-	-
STREET	0.913	0.959	0.953	0.768
ZIP	0.913	0.980	0.992	0.950
FAX	-	-	0.923	-
FAMILY	1	-	-	-
Macro score average	0.963	0.957	0.930	0.922

whereas the model was quite successful in predicting the more common entities.

The low scores can be attributed to several factors. The model struggled to recognize patient and doctor names embedded in the middle of the text, despite successfully identifying those at the beginning and end. Some hospital names were partially labeled, affecting overall precision and recall. Occasionally, the model included extra tokens within labels, leading to incorrect annotations. Furthermore, although the prompt explicitly specified which labels to use, the model occasionally introduced unintended labels (e.g., time). Confusion between labels or failure to identify them also contributed to the lower performance.

6. Discussion

This work showed that expert small NER models, built with an LLM-in-the-loop development process, can deliver strong multilingual PHI de-identification while keeping inference on-premises. Across eight languages, macro-F1 scores ranged from 0.922 (Arabic) to 0.963 (Turkish), with consistently high scores for common identifiers, such as "DATE" and "PHONE." Deploying lightweight, on-premises models avoids routine transmission of clinical text to external services, which better aligns with privacy regimes, such as HIPAA and General Data Protection Regulation when paired with organizational controls (access management, audit logging, and defined data-retention policies).

Performance was not uniform across entity types or languages. "ORGANIZATION" was the most fragile category in French (0.699), and "CITY" was comparatively weaker in Romanian (0.889) and Arabic (0.867). We attribute these gaps to linguistic factors (orthography, compounding, and rich morphology), domain naming conventions (hospital and clinic aliases), tokenizer mismatch, and limited language-specific coverage in gazetteers and training corpora. Targeted remedies include (i) language-tailored augmentation that preserves morphology and diacritics, (ii) curated medical-organization gazetteers and alias tables, (iii) character-aware and subword-robust encoders; and (iv) post-processing with constrained decoding and span-consistency checks to reduce boundary errors.

The LLM-in-the-loop paradigm was most valuable during data synthesis, labeling quality assurance, and error triage, while excluding LLMs from deployment helped mitigate API-related privacy risk. Reliance on synthetic and translated corpora for non-English training limits real-world generalization. Future work must prioritize evaluation on native clinical notes where feasible, out-of-distribution stress tests (novel facilities, regional toponyms), and ablations that quantify the incremental value of each loop component versus purely supervised baselines. To strengthen the privacy posture, future work must also explore federated fine-tuning across institutions and differentially private optimization to bound memorization risks, and report operational characteristics (latency, memory footprint, and minimal hardware) to support adoption.

Finally, we identified several immediate paths to broader utility, such as extending models to related biomedical entities and relations, uncertainty-aware inference to flag low-confidence spans for human review, and releasing prompts, evaluation scripts, and error taxonomies to enable reproducibility. Addressing the noted weaknesses—especially expanding native, annotated non-English resources—will make these expert small models more inclusive, robust, and clinically practical while preserving patient privacy.

7. Conclusion

This study underscores the importance of de-identification as a key method for safeguarding patient/personal health information and ensuring its ethical use in scientific research. By removing identifiable details through techniques, such as anonymization, generalization, and differential privacy, de-identification allows data to be used for diverse scientific applications, including epidemiological studies, disease modeling, and AI development, while maintaining patient privacy.

Recent advancements have demonstrated the potential of LLMs in de-identification tasks. However, challenges remain, particularly around issues of patient data security, API dependencies, and the need for domain-specific expertise in handling EHRs. Our "LLMs-in-the-loop" approach addresses these concerns by integrating small, specialized models tailored to the medical field. This method enhances both privacy and reliability, enabling the secure use of data without relying on external APIs or compromising sensitive patient information.

The multilingual nature of this research, spanning several languages, shows the adaptability and robustness of our models across diverse healthcare environments. While there are inherent risks associated with data anonymization, this study demonstrates that when properly applied, de-identification models can strike a delicate balance between protecting individual privacy and maximizing the utility of health data.

Furthermore, as the field progresses, it is crucial to establish globally recognized standards, raise awareness

of best practices, and ensure that ethical principles guide the deployment of de-identification technologies. Transparency, accountability, and a rigorous risk-benefit analysis must remain at the forefront of these efforts.

Ultimately, the findings of this study highlight the potential of expert small models developed through the LLMs-in-the-loop methodology to meet the evolving demands of healthcare research. The models presented here offer a reliable and scalable solution for future de-identification applications, advancing the capabilities of AI in healthcare while safeguarding patient privacy.

Future research should focus on refining and expanding de-identification models to cover a wider range of languages and healthcare contexts. One of the primary challenges is the scarcity of high-quality, annotated datasets in languages other than English, which limits the development of robust models for non-English speaking regions. Addressing this gap will require collaborative efforts to create and share multilingual datasets, ensuring more comprehensive language coverage. In addition, future studies could explore more advanced augmentation techniques and develop models capable of handling increasingly complex medical data types, such as clinical narratives and imaging reports. Continuous innovation in privacy-preserving methods, such as federated learning, may also prove valuable in safeguarding sensitive patient information while advancing the performance and applicability of de-identification technologies across diverse healthcare systems.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Murat Gunay
Methodology: Murat Gunay
Software: Bunyamin Keles, Raife Hizlan
Validation: Murat Gunay, Bunyamin Keles
Writing – original draft: Murat Gunay, Raife Hizlan
Writing – review & editing: Bunyamin Keles, Raife Hizlan

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data will be made available upon request to the corresponding author after the evaluation process.

References

1. Ahmed T, Al Aziz MM, Mohammed N. De-identification of electronic health record using neural network. *Sci Rep.* 2020;10(1):18600.

   doi: 10.1038/s41598-020-75544-1

2. Wood A, Denholm R, Hollings S, *et al*. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: Data resource. *BMJ*. 2021;373:n826.

   doi: 10.1136/bmj.n826

3. Güngören M, Orhan F, Kurutkan N. Mikro Rekabetcilikte Yeni Yaklasımlar: Hastanelerde Olusan Etik Iklimin Kalite ve Akreditasyon Acısından Degerlendirilmesi [New Approaches in Micro-Competitiveness: Evaluating the Ethical Climate in Hospitals in Terms of Quality and Accreditation]. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*. 2013;18(1):221-241. [Article in Turkish]

4. Varol S, Orhan F, Tuncer S, Akyuz S. Saglık kurumlarında bilgi guvenligi baglamında biyometrik sistemler [Biometric systems in the context of information security in healthcare institutions]. *Saglık Akadem Derg*. 2016;3(4):155-162.

   doi: 10.5455/sad.13-1483706096

5. Yilmaz D, Ozkoc EE, Ogutcu Ulas G. Elektronik saglik kayitlarinda farkındalık [Awareness of electronic health records]. *Hacettepe Sağlık İdaresi Derg*. 2021;24(4):777-792.

6. Hughes J. *De-Identification of PHI According to the HIPAA Privacy Rule*. Healthtech Security. Published October 15, 2021. Available from: https://www.techtarget.com/healthtechsecurity/feature/De-Identification-of-PHI-According-to-the-HIPAA-Privacy-Rule [Last accessed on 2023 Apr 13].

7. Act A. *Health Insurance Portability and Accountability Act of 1996*. Vol. 104. Public Law; 1996. p. 191. Available from: https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf [Last accessed on 2025 Sep 17].

8. Office of the Assistant Secretary for Planning and Evaluation. Standards for privacy of individually identifiable health information [45 CFR Parts 160 and 164]. Available from: https://aspe.hhs.gov/standards-privacy-individually-identifiable-health-information

   doi: 10.1016/j.jbi.2012.12.003

9. Office for Civil Rights HH. Standards for privacy of

individually identifiable health information. Final rule. *Fed Regist*. 2002;67(157):53181-53273.

10. Toscano F, O'Donnell E, Unruh MA, *et al*. Electronic health records implementation: Can the European union learn from the United States? *Eur J Public Health*. 2018;28 Suppl 4:pcky213.401.

    doi: 10.1093/eurpub/cky213.401

11. *Guidance on De-Identification of Protected Health Information*. 2012. Available from: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf [Last accessed on 2023 Jul 17].

12. Standards for Privacy of Individually Identifiable Health Information [45 CFR Parts 160 and 164]. 2013. Available from: https://aspe.hhs.gov/standards-privacy-individually-identifiable-health-information [Last accessed on 2023 Jul 17].

13. Neamatullah I, Douglass MM, Lehman LW, *et al*. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*. 2008;8:32.

    doi: 10.1186/1472-6947-8-32

14. Paul T, Rana MKZ, Tautam PA, *et al*. Investigation of the utility of features in a clinical de-identification model: A demonstration using EHR pathology reports for advanced NSCLC patients. *Front Digit Health*. 2022;4:728922.

    doi: 10.3389/fdgth.2022.728922

15. Garfinkel SL. De-Identification of Personal Information. National Institute of Standards and Technology; 2015.

    doi: 10.6028/nist.ir.8053

16. Wu H, Toti G, Morley KI, *et al*. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc*. 2018;25(5):530-537.

    doi: 10.1093/jamia/ocx160

17. Stubbs A, Uzuner Ö. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *J Biomed Inform*. 2015;58:S78-S91.

    doi: 10.1016/j.jbi.2015.05.009

18. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. A novel COVID-19 data set and an effective deep learning approach for the de-identification of italian medical records. *IEEE Access*. 2021;9:19097-19110.

    doi: 10.1109/ACCESS.2021.3054479

19. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc*. 2020;27(3):491-497.

    doi: 10.1093/jamia/ocz192

20. Ong JCL, Seng BJ, Law JZ, *et al*. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: Current state and future directions. *Cell Rep Med*. 2024;5(1):101356.

    doi: 10.1016/j.xcrm.2023.101356

21. Gunasekeran DV, Tham YC, Ting DS, Tan GS, Wong TY. Digital health during COVID-19: Lessons from operationalising new models of care in ophthalmology. *Lancet Digit Health*. 2021;3(2):e124-e134.

    doi: 10.1016/S2589-7500(20)30287-9

22. Ting DS, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. *Nat Med*. 2020;26(4):459-461.

    doi: 10.1038/s41591-020-0824-5

23. Verdicchio M, Perin A. When doctors and AI interact: On human responsibility for artificial risks. *Philos Technol*. 2022;35(1):11.

    doi: 10.1007/s13347-022-00506-6

24. Dai SC, Xiong A, Ku LW. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. *arXiv*. Preprint posted online 2023.

    doi: 10.48550/arXiv.2310.15100

25. De Paoli S. Can Large Language Models emulate an inductive Thematic Analysis of semi-structured interviews? An exploration and provocation on the limits of the approach and the model. *arXiv*. Preprint posted online 2023.

    doi: 10.48550/arXiv.2305.13014

26. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text- annotation tasks. *Proc Natl Acad Sci U S A*. 2023;120(30):e2305016120.

    doi: 10.1073/pnas.2305016120

27. Islam T, Goldwasser D. Discovering Latent Themes in Social Media Messaging: A Machine-in-the-Loop Approach Integrating LLMs. *arXiv*. Preprint posted online 2024.

    doi: 10.48550/arXiv.2403.10707

28. Pham DK, Vo BQ. Towards Reliable Medical Question Answering: Techniques and Challenges in Mitigating Hallucinations in Language Models. *arXiv*. Preprint posted online 2024.

    doi: 10.48550/arXiv.2408.13808

29. Umphrey R, Roberts J, Roberts L. Investigating Expert-in-the-Loop LLM Discourse Patterns for Ancient Intertextual Analysis. *arXiv*. Preprint posted online 2024.

    doi: 10.48550/arXiv.2409.01882

30. Keles B, Gunay M, Caglar SI. LLMs-in-the-loop Part-1: Expert Small AI Models for Bio-Medical Text Translation. *arXiv*. Preprint posted online 2024.

    doi: 10.48550/arXiv.2407.12126

31. Khin K, Burckhardt P, Padman R. A Deep Learning

Architecture for De-identification of Patient Notes: Implementation and Evaluation. *arXiv*. Preprint posted online 2018.

doi: 10.48550/arXiv.1810.01570

32. Morrison FP, Sengupta S, Hripcsak G. Using a pipeline to improve de-identification performance. *AMIA Annu Symp Proc*. 2009;2009:447-451.

33. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *J Biomed Inform*. 2015;58:S11-S19.

doi: 10.1016/j.jbi.2015.06.007

34. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*. 2007;14(5):550-563.

doi: 10.1197/jamia.M2444

35. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc*. 2017;24(3):596-606.

doi: 10.1093/jamia/ocw156

36. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. *BMC Med Res Methodol*. 2012;12:109.

doi: 10.1186/1471-2288-12-109

37. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Med Res Methodol*. 2010;10:70.

doi: 10.1186/1471-2288-10-70

38. Liu Z, Chen Y, Tang B, *et al*. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *J Biomed Inform*. 2015;58 Suppl: S47-S52.

doi: 10.1016/j.jbi.2015.06.009

39. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. *J Biomed Inform*. 2015;58 Suppl: S30-S38.

doi: 10.1016/j.jbi.2015.06.015

40. Nadkarni PM, Ohno-Machado L, Chapman WW. *Natural language processing: An introduction*. *J Am Med Inform Assoc*. 2011;18(5):544-551.

doi: 10.1136/amiajnl-2011-000464

41. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Annu Fall Symp*. 1996:333-337.

42. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol*. 2004;121(2):176-186.

doi: 10.1309/E6K33GBPE5C27FYU

43. He B, Guan Y, Cheng J, Cen K, Hua W. CRFs based de-identification of medical records. *J Biomed Inform*. 2015;58:S39-S46.

doi: 10.1016/j.jbi.2015.08.012

44. Lafferty J, McCallum A, Pereira F. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. Williamstown, MA: ACM; 2001.

45. Kocaman V, Talby D, Ul Hak H. RWD143 Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets. *Value in Health*. 2023;26(12):S532.

doi: 10.1016/j.jval.2023.09.2860

46. Liu Z, Huang Y, Yu X, *et al*. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. *arXiv*. Preprint posted online 2023.

doi: 10.48550/arXiv.2303.11032

47. Stubbs A, Kotfila C, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *J Biomed Inform*. 2015;58:S67-S77.

doi: 10.1016/j.jbi.2015.07.001

Appendix

A. The prompt used to obtain benchmarks with GPT-4o

prompt = f""" You are tasked with extracting Protected Health Information (PHI) from clinical notes. Your job is to identify and mark specific entities within the text. Here are the entities you need to look for:

<entities>

AGE (Identifies the age number or age-related information. Example: In "88 years old," 88 would be marked as AGE. In "in his 50's," 50's would be marked as AGE.)

CITY (Identifies the name of a city.)

COUNTRY Identifies the name of a country.)

DATE (Identifies specific dates or years. Example: In "He was admitted on 03/29/2089," 03/29/2089 would be marked as DATE. In "His surgery was in the 1980s," 1980s would be marked as DATE. In "His record was marked on 2089-08-24," 2089-08-24 would be marked at DATE.)

DEVICE (Identifies serial numbers, item code, or product code of a medical device mentioned. Example: In " The AA 737 pacemaker was implanted," AA 737 would be marked as DEVICE.)

DOCTOR (Identifies the name of a doctor or healthcare professional. Only the name should be marked, not the title, such as " Dr.," " M.D.")

HOSPITAL (Identifies the name of a hospital or nursing home.)

IDNUM (Identifies identification numbers, such as medical record or patient numbers.)

LOCATION (Identifies specific locations related to healthcare, excluding city or country.)

MEDICALRECORD (Identifies medical record numbers or similar identifiers.)

ORGANIZATION (Identifies names of organizations or institutions.)

PATIENT (Identifies the patient's name. Only the name should be marked, not titles like "Mr." or " Mrs.")

PHONE (Identifies phone numbers, including fax numbers.)

PROFESSION (Identifies professions or job title.)

STATE (Identifies the name of a state or region.)

STREET (Identifies street addresses.)

USERNAME (Identifies usernames or account IDs.)

ZIP (Identifies postal or zip codes.)

</entities >

I will provide you with a clinical note. Your task is to process this note and mark all instances of the PHI entities listed above.

Here is the clinical note:

{clinical_note}

Instructions for marking PHI entities:

* Carefully read through the entire clinical note.

* Identify any text that matches one of the PHI entity types listed above.

* For each identified PHI entity, mark the beginning and end of the relevant text chunk using the following format:

BEGINNER_ LABEL CHUNK ENDNER where ENTITY LABEL is one of the entity types from the list, and CHUNK is the actual text containing the PHI.

* While marking, DO NOT EDIT OR CHANGE the original clinical text, only put marks described above.

Here are a few examples of correct markup: Original text:

Mrs. Linda Martinez, a 45-year-old architect, having MR \#:2775283 for an evaluation on 2023-05-10. Her insulin pump model ZX900 was assessed by Dr. Michael Brown, M. D. The patient's condition has improved since the 1990s, but she mentioned feeling unwell for the past 6 months. MF381/1183 was referenced during her visit, which lasted approximately 5 hours and concluded at 10:05:03. She was discharged on 20/10/2023.

Marked text:

Mrs. BEGINER_PATIENT Linda Martinez ENDNER, a BEGINER_AGE 45 ENDNER year- old BEGINER_ PROFESSION architect ENDNER, having MR\#: BEGINER_MEDICALRECORD 2775283 ENDNER for an evaluation on BEGINER_DATE 2023-05-10 ENDNER. Her insulin pump model BEGINER_DEVICE ZX900 ENDNER was assessed by Dr. BEGINER_DOCTOR Michael Brown ENDNER, M. D. The patient's condition has improved since the BEGINNER_DATE 1990s ENDNER, but she mentioned feeling unwell for the past 6 months. BEGINER_IDNUMMF381/1183 ENDNER was referenced during her visit, which lasted approximately 5 hours and concluded at 10:05:03. She was discharged on BEGINER_DATE 20/10/2023 ENDNER.

Important notes:

* Be sure to process the entire clinical note and mark all instances of PHI entities.
* If a chunk of text could belong to multiple entity types, choose the most specific or appropriate one.
* Do not mark information that is not part of the specified PHI entity types.
* Preserve the original text exactly as it appears, including any spelling errors or formatting.
* Label the data, ensuring that professional titles or suffixes, such as "M. D.," "Ph. D.," or similar, are not removed. These titles must be preserved exactly as they appear in the text, without alteration or omission, and should NEVER be inside the label.
* Apostrophe "s" ('s) should not be included within the label when associated with Names. Only the person's name should be inside the label, and the apostrophes should remain outside the marked text. However, apostrophe s' is allowed within the DATE label when referring to a decade (e.g., 80's).
* Mark only specific calendar dates as DATE. Do not mark relative time expressions like "6 months," "1 year ago," "5 weeks," "5 wks," "yesterday," "today," "days," or similar units of time (months, years, weeks), as they do not represent actual dates.
* Mark only actual dates as DATE. Do not mark time-related expressions, such as "10:05:03," "10 am," or durations like "5 hours" as DATE, since they refer to times or durations rather than specific calendar dates.
* Fax numbers should be treated as PHONE entities and marked the same way as phone numbers.

Please process the provided clinical note and return it with all PHI entities appropriately marked.

"""