# Automated anonymization of radiology reports: comparison of publicly available natural language processing and large language models

Langenbach, M.C.; Foldyna, B.; Hadzic, I.; Langenbach, I.L.; Raghu, V.K.; Lu, M.T.; ... ; Heemelaar, J.C.

EUROPEAN SOCIETY OF RADIOLOGY
European Radiology

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

# Automated anonymization of radiology reports: comparison of publicly available natural language processing and large language models

Marcel C. Langenbach[1,2*], Borek Foldyna[1], Ibrahim Hadzic[1,3], Isabel L. Langenbach[1,2], Vineet K. Raghu[1], Michael T. Lu[1], Tomas G. Neilan[1] and Julius C. Heemelaar[1,4]

## Abstract

**Purpose** Medical reports, governed by HIPAA regulations, contain personal health information (PHI), restricting secondary data use. Utilizing natural language processing (NLP) and large language models (LLM), we sought to employ publicly available methods to automatically anonymize PHI in free-text radiology reports.

**Materials and methods** We compared two publicly available rule-based NLP models (spaCy; $NLP_{ac}$, accuracy-optimized; $NLP_{sp}$, speed-optimized; iteratively improved on 400 free-text CT-reports (test set)) and one offline LLM approach (LLM-model, LLaMa-2, Meta-AI) for PHI-anonymization. The three models were tested on 100 randomly selected chest CT reports. Two investigators assessed the anonymization of occurring PHI entities and whether clinical information was removed. Subsequently, precision, recall, and F1 scores were calculated.

**Results** $NLP_{ac}$ and $NLP_{sp}$ successfully removed all instances of dates ($n = 333$), medical record numbers (MRN) ($n = 6$), and accession numbers (ACC) ($n = 92$). The LLM model removed all MRNs, 96% of ACCs, and 32% of dates. $NLP_{ac}$ was most consistent with a perfect F1-score of 1.00, followed by $NLP_{sp}$ with lower precision (0.86) and F1-score (0.92) for dates. The LLM model had perfect precision for MRNs, ACCs, and dates but the lowest recall for ACC (0.96) and dates (0.52), corresponding F1 scores of 0.98 and 0.68, respectively. Names were removed completely or majorly (i.e., one first or family name non-anonymized) in 100% ($NLP_{ac}$), 72% ($NLP_{sp}$), and 90% (LLM-model). Importantly, $NLP_{ac}$ and $NLP_{sp}$ did not remove medical information, while the LLM model did in 10% ($n = 10$).

**Conclusion** Pre-trained NLP models can effectively anonymize free-text radiology reports, while anonymization with the LLM model is more prone to deleting medical information.

## Key Points

***Question*** *This study compares NLP and locally hosted LLM techniques to ensure PHI anonymization without losing clinical information.*

***Findings*** *Pre-trained NLP models effectively anonymized radiology reports without removing clinical data, while a locally hosted LLM was less reliable, risking the loss of important information.*

***Clinical relevance*** *Fast, reliable, automated anonymization of PHI from radiology reports enables HIPAA-compliant secondary use, facilitating advanced applications like LLM-driven radiology analysis while ensuring ethical handling of sensitive patient data.*

**Keywords** Natural language processing, Electronic health records, Machine learning, Data anonymization, Diagnostic imaging

*Correspondence:
Marcel C. Langenbach
mlangenbach@mgh.harvard.edu
Full list of author information is available at the end of the article



Springer

## Introduction

Standard radiology reports, whether in free-text or structured format, frequently contain sensitive personal health information (PHI) such as dates, medical record numbers (MRNs), contact information, and accession numbers (ACC) even if the report header, including the patient's name and date of birth, has been anonymized. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) and established hospital data-sharing regulations strictly prohibit the sharing of patients' PHI for secondary data use [1]. Secondary data use includes data sharing in the context of research, collaborations, or teaching. Beyond this conventional usage, large language models (LLMs) such as LLaMa 2 (Meta AI) or the generative pre-trained transformer 4 (GPT-4, OpenAI) [2, 3] have opened up new avenues for automated processing and analysis of radiology reports, offering capabilities such as summarization, structuring, and feature extraction [4, 5]. However, utilizing LLMs for medical reports often requires uploading these reports to online servers as the available offline LLMs have disadvantages regarding accessibility or performance. HIPAA and hospital regulations present a substantial challenge when using LLMs in clinical or research contexts, particularly for analyzing large volumes of radiology reports, where manual anonymization is not feasible. The manual anonymization of extensive datasets is a time-consuming process prone to errors and incomplete anonymization [6].

Natural language processing (NLP) represents a suite of computational linguistics methods designed to analyze and transform raw text, holding the potential to streamline the anonymization of reports [7]. However, the applicability of publicly available NLP methods in the context of PHI anonymization in radiology reports has yet to be investigated. Furthermore, there is little data on the accuracy of offline LLMs in anonymized radiology reports. The development of secure and reliable methods to exclude PHI from radiology reports could facilitate HIPAA-compliant secondary data use.

Hence, the primary objective of this study is to compare the performance of publicly available NLP techniques and a locally hosted LLM to remove PHI from large volumes of free-text radiology reports without the removal of clinical information. This innovative approach seeks to ensure HIPAA compliance, thereby enhancing the utility and ethical handling of sensitive patient data for secondary data use like LLM-driven radiology applications.

## Materials and methods

The study was conducted within a single academic hospital network (Massachusetts General Hospital, Boston, MA, United States) and received ethical approval from the institutional review board (IRB no. 2023P002169) with a waiver of written informed consent. The radiology report dataset was randomly sampled from an existing institutional radiology database of women with breast cancer (diagnosis based on ICD-10 codes), encompassing 2608 diagnostic chest CT scans between 2018 and 2023. Every patient only had a single scan within this database. Reports were signed by at least two different radiologists at different experience levels (i.e., residents, fellows, and attendings) with in total more than 40 different readers. Reports were free-text reports without using templates for report generation.

### Workflow

We employed a structured workflow to identify and eliminate PHI from free-text radiology reports, utilizing three distinct models. Comprehensive details of the models and elaboration on several NLP technical terms are displayed in a visual glossary (Fig. 1).

The NLP models leverage the publicly available spaCy toolkit (version 3.6, Explosion). spaCy contains a range of pre-trained models with functions such as Named Entity Recognition (NER). NER can be combined with basic NLP functions such as regular expressions to formulate rules to detect PHI [8]. Detailed background information on spaCy model development, architecture, and benchmarking is published on the spaCy website (www.spacy.io).

The first NLP model, $NLP_{ac}$, uses NER functions based on state-of-the-art transformer-based architectures, enabling superior performance in capturing context and complex language. However, the transformer-based model exhibits a higher computational complexity and resource requirement than the more 'lightweight' second model, $NLP_{sp}$. The NER function of $NLP_{sp}$ is trained on word embeddings, a simpler method of understanding word meanings. This method is faster in processing and analysis. Therefore, $NLP_{ac}$ requires a graphics processing unit (GPU)-enabled workstation, while $NLP_{sp}$ can be executed on any standard workstation.

As an alternative to a rule-based NLP method, we used the LLaMa-2 LLM, developed by Meta AI (LLM-model). LLaMa-2 is a family of pre-trained and fine-tuned LLMs, LLaMa-2 and LLaMa-2-Chat, at scales of 7, 13, and 70 billion parameters [2]. This LLM is open-source, available for offline use, and, therefore, fully compliant with HIPAA regulations for the analysis of patient reports. The models were executed on the same Linux workstation equipped with one Nvidia Titan RTX A6000 GPU using Llama.cpp [9] and LangChain [10], ensuring that any patient information, anonymized or not, is not transferred outside of the hospital system. We used a quantized version of the 13 billion model (available at https://huggingface.co/TheBloke/Llama-2-13B-chat-GGUF/blob/main/llama-2-13b-chat.Q5_K_M.gguf), and we set the

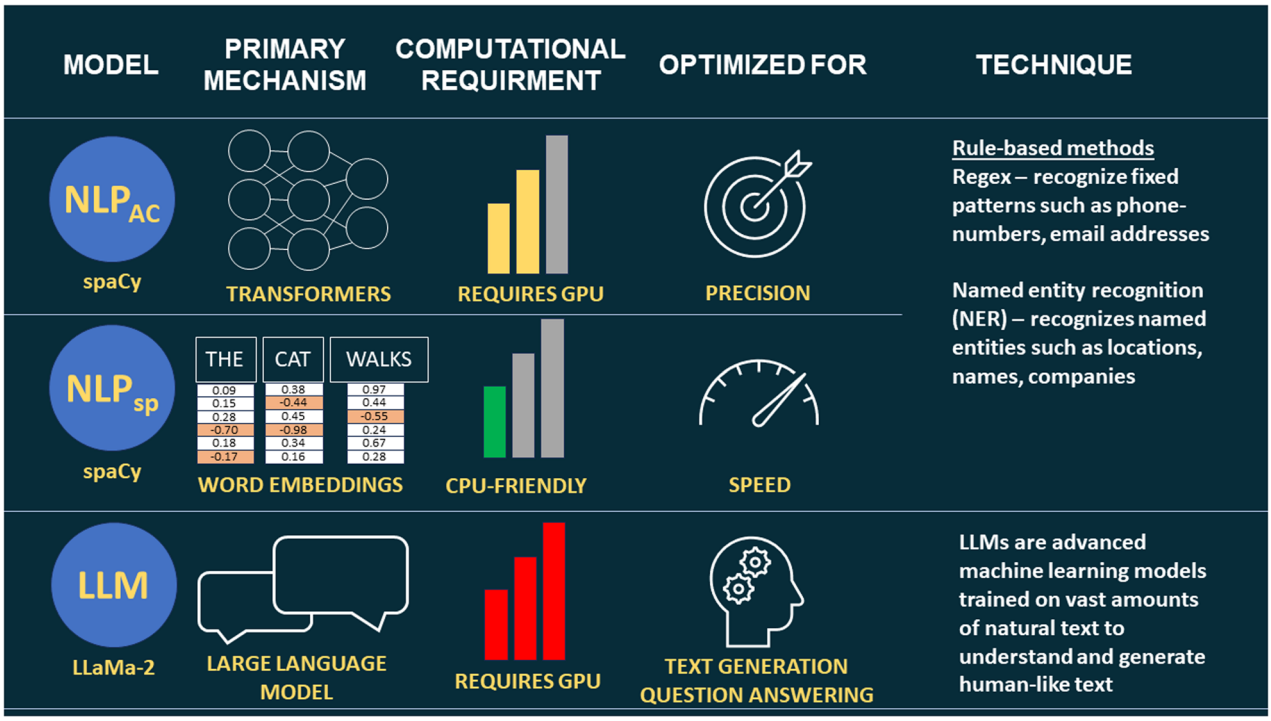


**Fig. 1** Visual glossary of models, characteristics, and techniques. A comprehensive visual glossary illustrating the models, their characteristics, and the techniques used in the study, highlighting their specific features and methodologies employed for PHI anonymization in radiology reports. GPU, graphics processing unit; LLM, large language model; NER, named entity recognition; NLP, natural language processing; PHI, personal health information

batch size to 512, context length to 4096, maximum amount of tokens to 8192, temperature to 0, and offload all layers to the GPU. An iterative approach to developing the questions used to prompt the LLM was applied. The radiology reports from the test set were used for prompt optimization.

The rules for NLP models were iteratively improved to recognize PHI using 400 randomly sampled free-text radiology chest CT reports from our institutional radiology database. Details on the prompt design and the specific prompt used for the LLM model are provided in the supplements (Supplemental Text S1). It is important to note that the LLM was neither trained nor modified on specific datasets to perform anonymization.

To evaluate the performance of all three models in anonymization, we employed an identical test set consisting of 100 newly selected chest CT radiology reports. A scheme of the workflow is displayed in Fig. 2. These reports were randomly sampled from our institutional database. Processing time was quantified by calculating the speed in reports per second using a background CPU code. We recorded the average processing time for each model. The script for the NLP models and the LLM model are available on GitHub (https://github.com/ibro45/NLP-vs-LLM-Medical-Report-Anonymization).

### Report evaluation

Two independent investigators (M.C.L., I.L.L.) with 7 years and 3 years of experience in diagnostic radiology adjudicated the radiology reports before automated anonymization. During an initial screening of the test set, all instances of PHI were marked. The PHI within the test set included dates, medical records numbers (MRN), ACCs, and provider names (Supplemental Table S1). Subsequently, the investigators assessed the successful anonymization of the NLP- or LLM-processed reports in consensus, blinded to the anonymization model. This assessment was performed by manually comparing the processed reports to the original reports based on the PHI entities mentioned above. For the provider names, we employed a 4-point Likert scale with 4—complete removal, 3—majority removed (i.e., one first or family name non-anonymized), 2—minority removed (i.e., one or more names non-anonymized), and 1—no anonymization. For all other PHI, a single fragment indicated incomplete anonymization. Furthermore, the readers screened the processed report for removal of essential clinical information or removal or the removal of report fragments that could potentially lead to falsification of the clinical results. As essential clinical information, every text fragment that was required to understand the report

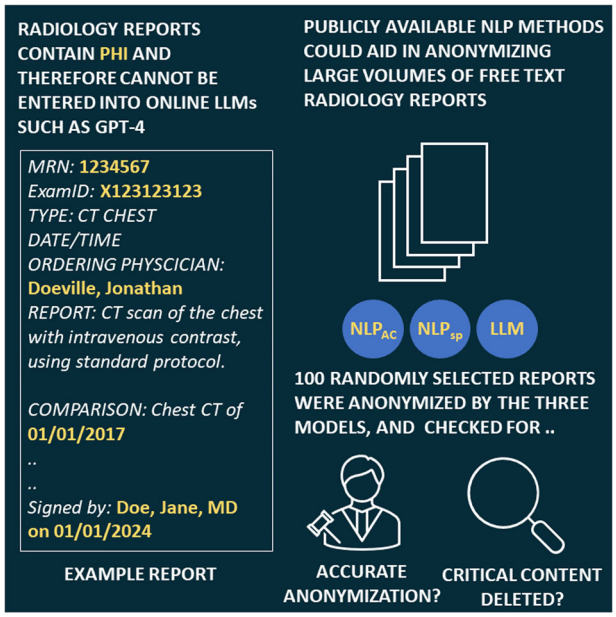


**Fig. 2** Example of PHI in radiology reports and study scheme. This figure presents an example of PHI in radiology reports and outlines the study scheme. ACC, accession number; CT, computed tomography; MRN, medical record number; GPT-4, generative pre-trained transformer 4

or any parts was defined (e.g., description, diagnosis, side, image, or series number).

### Statistical evaluation

Quantitative variables were described with mean and SD or median, minimum, and maximum qualitative variables with absolute and relative frequency (%). Categorical variables were compared using the Fisher exact test, and continuous variables were compared using the one-way ANOVA test for normal distribution or the Kruskal–Wallis *t*-test for asymmetric distribution. We further calculated precision ($P$ = true positives/(true positives + false positives), recall ($R$ = true positives/(true positives + false negatives), and the harmonic mean of precision and recall, F1-score (F1) = $2 \times ((P \times R)/(P + R)$ for all models and PHIs to evaluate the performance (Fig. 3). Statistical analysis was performed using R, version 4.3.1, on RStudio, version 2023.06.0 + 421 (https://cran.r-project.org/).

## Results

Using the NLP models, all radiology reports ($n = 100$) were successfully processed, while the LLM model failed processing in four cases (4%) due to an error in recognizing the report structure. $NLP_{sp}$ had the fastest processing speed with 1.4 ms per report, followed by $NLP_{ac}$ (2.4 s per report) and the LLM model (27.4 s per report). $NLP_{sp}$ was executed on an 8-core CPU workstation, and $NLP_{ac}$ and the LLM model on a GPU workstation.

The reports in the test set contained 333 instances of dates, 6 occurrences of MRNs, and 92 instances of ACCs. Overall, 235 provider names were observed throughout the reports.

The NLP models removed 100% of dates ($n = 333$), MRNs ($n = 6$), and ACCs ($n = 92$) from the reports. While $NLP_{ac}$ had no instances of false positive or false negative PHI identifications, $NLP_{sp}$ misclassified report content as a date in 28 reports (28%). Provider names were entirely removed in 35% ($n = 35$) and majorly in 65% ($n = 65$) of the reports by $NLP_{ac}$, in detail removal of 171 of 235 (73%) instances. With $NLP_{sp}$, provider names were eliminated completely in 16% ($n = 16$), majorly in 56% ($n = 56$), and minorly in 28% ($n = 28$). In total, 131 of 235 (56%) of the observed names. Using the LLM model successfully erased all PHI instances of MRN ($n = 6$, 100%) and ACC ($n = 89$, 96%) but was not able to successfully anonymize most of the dates ($n = 193$, 59%) with missed occurrences ($n = 134$, 41%) in 65 reports (68%). Overall, this resulted in an insufficient anonymization of 65 reports (68%). When analyzing the provider names, we observed a complete elimination in 69 cases (69%), major removal in 17 patients (17%), and minor removal in $n = 10$ (10%) using the LLM model, leading to a removal of 201/231 (87%) provider name instances. Comparing all three models, we found a significant difference regarding the anonymization of the date and the provider names (both $p \leq 0.001$) and no significant differences in successful anonymization for MRN ($p = 0.996$) and ACC ($p = 0.978$) (Table 1).

$NLP_{ac}$ achieved perfect (1.00) precision, recall, and F1-score for the removal of dates, ACCs, and MRNs. For names, precision and recall were 1.00 and 0.73, respectively, leading to an F1-score of 0.84. $NLP_{sp}$ also achieved a perfect F1-score for the removal of MRNs and ACCs and an F1-score of 0.92 for the removal of dates. However, precision and recall were lower for names, 1.00 and 0.56, respectively, with an F1-score of 0.72. Using the LLM model, precision, recall, and F1-score for anonymization of MRNs were 1.00, while ACC had a precision of 1.00 with a recall of 0.96, resulting in an F1 of 0.98. While the LLM model had excellent precision for anonymizing dates and names (1.00 and 0.95), PHI was more often missed. Therefore, the recall was lower for dates and names (0.52 and 0.87), and $F1_{LLM}$ scores were 0.68 and 0.91, respectively (Table 2).

Crucially, using the NLP models, in no case essential clinical information was deleted from any report based on evaluation by both readers. With the LLM model, 14 reports showed a loss of relevant information during the anonymization process. Besides the four cases (4%) that were not processed correctly during the anonymization process, relevant medical information was deleted in ten

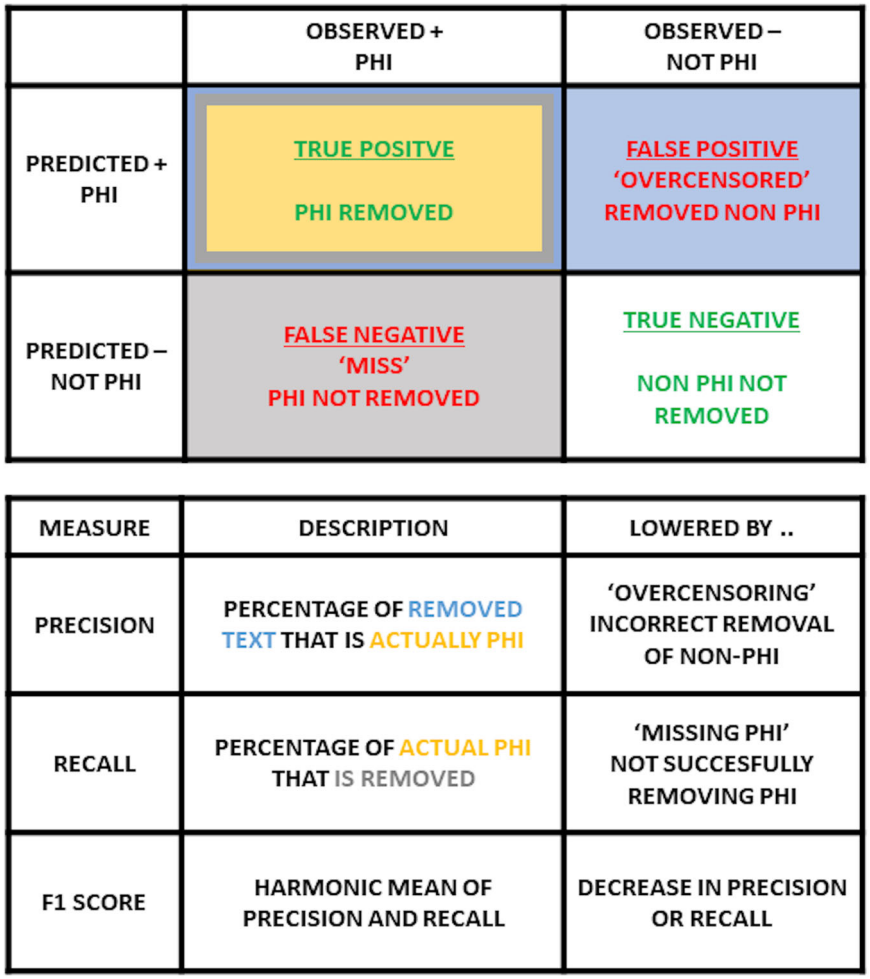


**Fig. 3** Definitions of true/false positives and negatives for PHI data and calculation of precision and recall. This figure defines true positives, false positives, true negatives, and false negatives for PHI data. It also explains the calculation of precision and recall, including factors that can lower these metrics. PHI, personal health information

**Table 1** Comparison of the three different models ($NLP_{ac}$, $NLP_{sp}$, and LLM-model) regarding the anonymization of the PHI, analyzed per report

| Successful anonymization, *n* (%) | $NLP_{ac}$ (reports, $n = 100$) | $NLP_{sp}$ (reports, $n = 100$) | LLM-model (reports, $n = 96$) | *p*-value |
|---|---|---|---|---|
| MRN | 6/6 (100%) | 6/6 (100%) | 6/6 (100%) | 0.996 |
| ACC | 92/92 (100%) | 92/92 (100%) | 89/92 (96.7%) | 0.978 |
| Date | 100/100 (100%) | 100/100 (100%) | 31/96 (32.2%) | < 0.001 |
| Provider names | | | | < 0.001 |
| Minor | 0/100 (0%) | 28/100 (28.0%) | 10/96 (10.4%) | |
| Major | 65/100 (65.0%) | 56/100 (56.0%) | 17/96 (17.7%) | |
| Complete | 35/100 (35.0%) | 16/100 (16.0%) | 69/96 (71.9%) | |
| Complete relevant medical information | 100/100 (100%) | 100/100 (100%) | 86/96 (89.6%) | < 0.001 |

Anonymization of the provider names was defined as either complete removal (i.e., all name fragments removed), majority removed (i.e., one first or family name non-anonymized), and minority removed (i.e., one or more entire name non-anonymized). Results were compared using the Fisher exact test
*NLP* natural language processing, *LLM* large language model, *MRN* medical record number, *ACC* accession number

**Table 2** Precision, recall, and F1-score for PHI data (MRN, ACC, date, and provider names) presented for $NLP_{ac}$, $NLP_{sp}$, and the LLM-model

| | | Precision | Recall | F1 score |
|---|---|---|---|---|
| $NLP_{ac}$ | MRN | 1.00 | 1.00 | 1.00 |
| | ACC | 1.00 | 1.00 | 1.00 |
| | Date | 1.00 | 1.00 | 1.00 |
| | Provider names | 1.00 | 0.73 | 0.84 |
| $NLP_{sp}$ | MRN | 1.00 | 1.00 | 1.00 |
| | ACC | 1.00 | 1.00 | 1.00 |
| | Date | 0.86 | 1.00 | 0.92 |
| | Provider names | 1.00 | 0.56 | 0.72 |
| LLM-model | MRN | 1.00 | 1.00 | 1.00 |
| | ACC | 1.00 | 0.96 | 0.98 |
| | Date | 1.00 | 0.52 | 0.68 |
| | Provider names | 0.95 | 0.87 | 0.91 |

*NLP* natural language processing, *LLM* large language model, *MRN* medical record number, *ACC* accession number

reports (10%). Of these reports, in nine cases, findings or measurements (e.g., nodule size) were wrongly anonymized, and one report was shortened due to the wrong identification of medical information (Supplemental Table S2). Comparing the three models resulted in a significant difference in whether relevant medical information was deleted ($p < 0.001$). A comparison table (Supplemental Table S3) is included in the supplements to illustrate an example input radiology report and the anonymized outputs from the three models.

## Discussion

Our study demonstrates that reliable automated anonymization of free-text radiology reports is achievable using publicly available rule-based NLP models or LLMs. Notably, pre-trained NLP models effectively anonymized free-text radiology reports while maintaining the report content's integrity. Anonymization with the LLM model was more prone to delete relevant non-PHI medical data and therefore and therefore should not be preferred. The successful anonymization of PHI offers various possibilities for HIPAA-compliant advanced processing and secondary data use of large volumes of radiology reports.

Rapid improvement in the field of language models suggests that LLMs will become even more important in radiology, for research purposes, and in clinical care [11, 12]. To date, the use of LLMs on clinical data, whether radiology reports or other patient-related datasets, e.g., discharge letters, pathology reports, or visit notes, is limited by potential PHI within these datasets. PHIs are very well defined in the HIPAA and institutional regulations and can occur in various forms in all medical documents. The correct and complete identification of PHIs is complex, especially in free-text reports. To utilize the full potential of LLMs, especially in the context of radiology reports, fast and reliable anonymization of PHI is necessary to prevent a breach of data policies to patients and protect their data [13]. Based on the existing literature, the anonymization of radiology data leveraged for research purposes in LLMs was secured either by using report facsimiles, or by manually deleting PHI, or the anonymization was not explicitly stated in the manuscripts [14–16]. Facsimiles or manual deletion impedes the use of LLMs for large data volumes and limits the optimal applicability. Further, a manual de-identification is time-consuming and, especially with large volumes, involves the risk of possible errors, leading to insufficient anonymization and, therefore, violating HIPAA regulations.

From a research perspective, automated anonymization might allow for processing and analyzing large report volumes in the context of LLMs, for example, to screen for a specific diagnosis, impression, or any other relevant feature stated in the report. This enables multiple options in a straightforward approach to retrospectively create databases of specific cohorts for subsequent analysis.

$NLP_{ac}$ showed the most robust results, with a de-identification rate of 100% for the most sensitive PHI categories (i.e., patient names, dates, MRNs, ACCs). The lighter $NLP_{sp}$ showed lower recall in anonymization of the provider names and precision regarding the date whilst maintaining a 100% anonymization rate for this most sensitive PHI. A possible explanation would be that a transformer-based model can better understand context than $NLP_{sp}$, which uses word embedding. The trade-off is that the latter can be run on any workstation. Both models performed better compared to previous studies, where simpler algorithms were employed for anonymizing radiology reports. These previous models could only anonymize PHI with an F1 between 0.892 and 0.964 [7, 17]. To achieve a 100% anonymization rate of the provider names, a possible approach would be more fine-tuning on regularly occurring names. However, this could lead to a potential deletion of report content, as proper names (e.g., Fleischner or Agatston) might be removed, too. To reduce this overlap, institution-specific training on local provider names or to recognize the names in the context of the report might be an option.

The applied LLM model, LLaMa-2, showed incomplete anonymization of ACCs, MRNs, and dates in 68% of the cases, insufficient processing in 4%, and deletion of medical information in 10%. These results indicate that using LLaMa-2 in the context of anonymization is not feasible at this moment. More extended prompt engineering might help to improve the anonymization rate

and could be applied for further evaluations, but our intention in this first proof-of-concept study was to be as generalizable as possible. Further, the processing time of the LLM model, even on our used multi-core GPU system, was very long at 27.4 s compared to the NLP models in the millisecond range. This could limit the applicability of larger datasets with a much longer processing time.

In our study, prompt engineering for the LLM model involved iterative refinement to balance specificity and generality, ensuring comprehensive PHI identification while minimizing the removal of relevant medical information. We designed the prompt based on the local specifications of the radiology reports at our institution to explicitly instruct the model on recognizing and anonymizing key PHI entities such as dates, names, MRNs, and ACCs as appearing in our report structure. Future research is required to explore the impact of different LLMs and different prompt variations to further optimize LLM performance in medical text anonymization.

A machine learning-based method that has previously been published in multilayer bi-directional transformer encoder (BERT) algorithms to anonymize clinical text [18, 19]. These NLP models have superior contextual awareness by being able to read "bi-directional". However, these models need to be fine-tuned to a specific task using large amounts of annotated data, require elaborate computational resources and expertise, and have a limited number of words they can process. To the best of our knowledge, thus far, no BERT model has been specifically trained and tested for anonymization of radiology reports.

At the time of this study, LLaMA-2 was the most advanced open-source LLM among models with a relatively small model size (i.e., blow 15B parameters), which is why it was selected as the comparison model for our anonymization analysis. Since the completion of our research, LLaMA-3 has been released, offering enhanced capabilities and performance improvements over its predecessor [20]. Future studies could explore the potential of LLaMA-3 in medical text anonymization, as its advanced architecture might yield better results in preserving clinical information while ensuring comprehensive anonymization. This rapid evolution in LLMs highlights the importance of continually reassessing the tools used for sensitive tasks such as PHI anonymization.

One limitation of this study is that our models have only been investigated for reports of our hospital network. Hospital-specific styles of PHI, like ACCs or MRNs, might require slight adjustments that are easily implacable with basic programming. Therefore, an external validation of our anonymization algorithm would be favorable. Also, application in more diverse cohorts could increase the rate of name misclassifications. Due to our institutional restrictions, we were not allowed to use GPT-4 for any patient-related data, even if this data is anonymized. Therefore, we used the locally installed and processable model LLaMa-2 for external validation of the datasets. We restricted our analysis to radiology reports and chest CTs only. Extending these models to other kinds of reports, scans, and modalities or adapting to specific requirements is feasible and paths the way for all medical specialties. All used CT scans were from women with breast cancer, with the need to validate the algorithm in a more variety of reports and diagnoses. However, these reports do not differ from other radiology reports in our hospital network, and therefore, we assume that they are a representative sample set in this proof-of-concept study setting. The models were executed on specific workstations, allowing for no generalizability of the observed speed measurements. These can vary based on the computation resources. Finally, our methods were not able to completely remove all the provider names from free-text reports. Provider names, however, are not considered PHI by HIPAA, and the majority of instances were completely or majorly deleted.

## Conclusion

Our generalizable NLP-based anonymization pipeline enables fast and reliable HIPAA-compliant removal of PHI from free-text radiology reports, facilitating safe secondary data use. The LLM LLaMa-2 was more prone to anonymize non-PHI information, did not achieve 100% anonymization, and is therefore not recommended.

### Abbreviations

| | |
|---|---|
| ACC | Accession number |
| GPT-4 | Generative pre-trained transformer 4 |
| GPU | Graphics processing unit |
| HIPAA | Health insurance portability and accountability act of 1996 |
| LLM | Large language model |
| MRN | Medical record number |
| NER | Named entity recognition |
| NLP | Natural language processing |
| PHI | Personal health identifier |

## Supplementary information

The online version contains supplementary material available at https://doi.org/10.1007/s00330-024-11148-x.

### Funding

The authors state that this work has not received any funding.

## Compliance with ethical standards

### Guarantor

The scientific guarantor of this publication is Marcel C. Langenbach.

### Conflict of interest

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest. The authors of this manuscript declare relationships with the following companies: M.C.L. is funded by the German Research Foundation

(DFG), project number 502109212. B.F. has received unrelated research grants from the NHI/NHLBI. AstraZeneca, Eli Lilly, and MedTrace. M.C.L. and B.F. are members of the Scientific Editorial Board of *European Radiology* (section: Cardiac). They did not participate in the selection or review processes for this article. M.T.L. reports research support to his institution from the American Heart Association, AstraZeneca, Ionis, Johnson & Johnson Innovation, Kowa Pharmaceuticals America, MedImmune, the National Academy of Medicine, the NIH/NHLBI, and the Risk Management Foundation of the Harvard Medical Institutions Incorporated outside of the submitted work. All other authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

### Statistics and biometry
One of the authors has significant statistical expertise. No complex statistical methods were necessary for this paper.

### Informed consent
Written informed consent was waived by the Institutional Review Board.

### Ethical approval
Institutional Review Board approval was obtained.

### Study subjects or cohorts overlap
The results of this cohort have not been previously reported.

### Methodology

- Retrospective
- Observational study
- Performed at one institution

### Author details
[1]Cardiovascular Imaging Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [2]Institute for Diagnostic and Interventional Radiology, University Hospital Cologne, Cologne, Germany. [3]Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, The Netherlands. [4]Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands.

Received: 17 April 2024 Revised: 23 August 2024 Accepted: 23 September 2024
Published online: 31 October 2024

### References
1. U.S. Government Printing Office (1996) Public law 104–191—Health Insurance Portability and Accountability Act of 1996. GPO, Washington, DC
2. Touvron H, Martin L, Stone K et al (2023) Llama 2: open foundation and fine-tuned chat models. Preprint at https://doi.org/10.48550/arXiv.2307.09288
3. OpenAI (2023) GPT-4 technical report. https://doi.org/10.48550/arXiv.2303.08774
4. Adams LC, Truhn D, Busch F et al (2023) Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 307:e230725. https://doi.org/10.1148/radiol.230725
5. Patel SB, Lam K (2023) ChatGPT: The future of discharge summaries? Lancet Digital Health 5:e107–e108. https://doi.org/10.1016/S2589-7500(23)00021-3
6. Kushida CA, Nichols DA, Jadrnicek R et al (2012) Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care 50:S82–S101. https://doi.org/10.1097/MLR.0b013e3182585355
7. Chambon PJ, Wu C, Steinkamp JM et al (2023) Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. J Am Med Inform Assoc 30:318–328. https://doi.org/10.1093/jamia/ocac219
8. Honnbial M, Montani I (2017) spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Sentometrics Research. https://sentometrics-research.com/publication/72/. Accessed 20 Sep 2023
9. Gerganov G (2023) llama.cpp. *GitHub* https://github.com/ggerganov/llama.cpp. Accessed 15 Nov 2023
10. Chase H (2022) LangChain. *GitHub* https://github.com/langchain-ai/langchain. Accessed 15 Nov 2023
11. López-Úbeda P, Martín-Noguerol T, Luna A (2023) Radiology in the era of large language models: the near and the dark side of the moon. Eur Radiol. https://doi.org/10.1007/s00330-023-09901-9
12. Silverberg M (2023) Preparing radiology trainees for AI and ChatGPT. https://www.rsna.org/news/2023/july/radiology-trainees-ai-and-chatgpt. Accessed 19 Sep 2023
13. Lecler A, Duron L, Soyer P (2023) Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging 104:269–274. https://doi.org/10.1016/j.diii.2023.02.003
14. Sun Z, Ong H, Kennedy P et al (2023) Evaluating GPT4 on impressions generation in radiology reports. Radiology 307:e231259. https://doi.org/10.1148/radiol.231259
15. Lyu Q, Tan J, Zapadka ME et al (2023) Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 6:9. https://doi.org/10.1186/s42492-023-00136-5
16. Gertz RJ, Bunck AC, Lennartz S et al (2023) GPT-4 for automated determination of radiologic study and protocol based on radiology request forms: a feasibility study. Radiology 307:e230877. https://doi.org/10.1148/radiol.230877
17. Stubbs A, Uzuner Ö (2015) Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. J Biomed Inform 58:S20–S29. https://doi.org/10.1016/j.jbi.2015.07.020
18. Mao J, Liu W (2019) Hadoken: a BERT-CRF model for medical document anonymization. In: Proceedings of the Iberian languages evaluation forum co-located with 35th Conference of the Spanish Society for Natural Language Processing. IberLEF@SEPLN, Bilbao, pp 720–726
19. García-Pablos A, Perez N, Cuadros M (2020) Sensitive data detection and classification in Spanish clinical text: experiments with BERT. European Language Resources Association, Marseille
20. Meta AI LLaMa-3.1 model card. *GitHub* https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md. Accessed 21 Aug 2024

## Publisher's Note