Harnessing Moderate-Sized Language Models for Reliable Patient Data De-identification in Emergency Department Records: An Evaluation of Strategies and Performance

Océane Dorémus, Dylan Russon, Benjamin Contrand, Ariel Guerra-Adames, Marta Avalos-Fernandez, Cédric Gil-Jardiné, Emmanuel Lagarde

Submitted to: Journal of Medical Internet Research
on: February 28, 2024

Table of Contents

Harnessing Moderate-Sized Language Models for Reliable Patient Data De-identification in Emergency Department Records: An Evaluation of Strategies and Performance

Océane Dorémus[1]; Dylan Russon[1]; Benjamin Contrand[2]; Ariel Guerra-Adames[1]; Marta Avalos-Fernandez[2, 1]; Cédric Gil-Jardiné[3, 4]; Emmanuel Lagarde[2] PhD

[1]SISTM team INRIA Talence FR
[2]SISTM team, Bordeaux Population Health U1219 University of Bordeaux INSERM Bordeaux FR
[3]Pole of Emergency Medicine University Hospital of Bordeaux Bordeaux FR
[4]AHeaD team, Bordeaux Population Health U1219 University of Bordeaux INSERM Bordeaux FR

Corresponding Author:
Océane Dorémus
SISTM team
INRIA
Talence
FR

Abstract

Background: The digitization of healthcare, facilitated by the adoption of electronic health record (EHR) systems, has revolutionized data-driven medical research and patient care. While this digital transformation offers substantial benefits in healthcare efficiency and accessibility, it concurrently raises significant concerns over privacy and data security. Initially, the journey towards protecting patient data de-identification saw the transition from rule-based systems to more mixed approaches including machine learning for de-identifying patient data. Subsequently, the emergence of Large Language Models (LLMs) has represented a further opportunity in this domain, offering unparalleled potential for enhancing the accuracy of context-sensitive de-identification. However, despite LLMs offering significant potential, the deployment of the most advanced models in hospital environments is frequently hindered by data security issues and the extensive hardware resources required.

Objective: The objective of our study is to design, implement, and evaluate de-identification algorithms by employing fine-tuning of moderate-sized open-source language models, ensuring their suitability for production inference tasks on personal computers.

Methods: We aimed at replacing personal identifying information (PII) with generic placeholders or labeling non-PII texts as 'ANONYMOUS', ensuring privacy while preserving textual integrity. Our dataset, derived from over 425,000 clinical notes from the adult emergency department of the Bordeaux University Hospital in France, underwent independent double annotation by two experts to create a reference for model validation with 3,000 clinical notes randomly selected. Three open-source language models of manageable size were selected for their feasibility in hospital settings: Llama 2 7B, Mistral 7B, and Mixtral 8x7B. Fine-tuning utilized the quantized Low-Rank Adaptation (qLoRA) technique. Evaluation focused on PII-level (Recall, Precision and F1-Score) and clinical note-level metrics (Recall and BLEU metric), assessing de-identification effectiveness and content preservation.

Results: The generative model Mistral 7B demonstrated the highest performance with an overall F1-score of 0.9673 (vs. 0.8750 for Llama 2 and 0.8686 for Mistral 8x7B). At the clinical notes level, the same model achieved an overall recall of 0.9326 (vs. 0.6888 for Llama 2 and 0.6417 for Mistral 8x7B).This rate increased to 0.9915 for the anonymization of names with Mistral 7B. Four notes out of the total 3000 failed to be fully anonymized for names: in one case, the non-anonymized name belonged to a patient, while in the other cases, it belonged to medical staff. Beyond the fifth epoch, the BLEU score consistently exceeded 0.9864, indicating no significant text alteration due to the process.

Conclusions: Our research underscores the significant capabilities of generative NLP models, with Mistral 7B standing out for its superior ability to de-identify clinical texts efficiently. Achieving notable performance metrics, Mistral 7B operates effectively without requiring high-end computational resources. These methods pave the way for a broader availability of anonymized clinical texts, enabling their use for research purposes and the optimization of the healthcare system.

(JMIR Preprints 28/02/2024:57828)
DOI: https://doi.org/10.2196/preprints.57828

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Harnessing Moderate-Sized Language Models for Reliable Patient Data De-identification in Emergency Department Records: An Evaluation of Strategies and Performance

Océane Dorémus[1], Dylan Russon[1], Benjamin Contrand [1], Ariel Guerra-Adames [1], Avalos-Fernandez Marta [2,3], Cédric Gil-Jardiné [1,4] and Emmanuel Lagarde [1]

1 AHeaD team, Bordeaux Population Health U1219, University of Bordeaux, INSERM, Bordeaux, France
2 SISTM team, Bordeaux Population Health U1219, University of Bordeaux, INSERM, Bordeaux, France
3 SISTM team, INRIA, Talence, France
4 University Hospital of Bordeaux, Pole of Emergency Medicine, F- 33000, Bordeaux, France

Corresponding Author:

Océane Dorémus,

Bordeaux Population Health,

146 rue Léo Saignat,

Bordeaux, 33000,

France

Phone: +33 5 57 57 15 04

Email: oceane.doremus@u-bordeaux.fr

Abstract

Background: The digitization of healthcare, facilitated by the adoption of electronic health record (EHR) systems, has revolutionized data-driven medical research and patient care. While this digital transformation offers substantial benefits in healthcare efficiency and accessibility, it concurrently raises significant concerns over privacy and data security. Initially, the journey towards protecting patient data de-identification saw the transition from rule-based systems to more mixed approaches including machine learning for de-identifying patient data. Subsequently, the emergence of Large Language Models (LLMs) has represented a further opportunity in this domain, offering unparalleled potential for enhancing the accuracy of context-sensitive de-identification. However, despite LLMs offering significant potential, the deployment of the most advanced models in hospital environments is frequently hindered by data security issues and the extensive hardware resources required.

Objective: The objective of our study is to design, implement, and evaluate de-identification algorithms by employing fine-tuning of moderate-sized open-source language models, ensuring their suitability for production inference tasks on personal computers.

Methods: We aimed at replacing personal identifying information (PII) with generic placeholders or labeling non-PII texts as 'ANONYMOUS', ensuring privacy while preserving textual integrity. Our dataset, derived from over 425,000 clinical notes from the adult emergency department of the Bordeaux University Hospital in France, underwent independent double annotation by two experts to create a reference for model validation with 3,000 clinical notes randomly selected. Three open-source language models of manageable size were selected for their feasibility in hospital settings: Llama 2 7B, Mistral 7B, and Mixtral 8x7B. Fine-tuning utilized the quantized Low-Rank Adaptation (qLoRA) technique. Evaluation focused on PII-level (Recall, Precision and F1-Score) and clinical note-level metrics (Recall and BLEU metric), assessing de-identification effectiveness and content preservation.

Results: The generative model Mistral 7B demonstrated the highest performance with an overall F1-score of 0.9673 (vs. 0.8750 for Llama 2 and 0.8686 for Mixtral 8x7B). At the clinical notes level, the same model achieved an overall recall of 0.9326 (vs. 0.6888 for Llama 2 and 0.6417 for Mixtral 8x7B).This rate increased to 0.9915 for the anonymization of names with Mistral 7B. Four notes out of the total 3000 failed to be fully anonymized for names: in one case, the non-anonymized name belonged to a patient, while in the other cases, it belonged to medical staff. Beyond the fifth epoch, the BLEU score consistently exceeded 0.9864, indicating no significant text alteration due to the process.

Conclusions: Our research underscores the significant capabilities of generative NLP models, with Mistral 7B standing out for its superior ability to de-identify clinical texts efficiently. Achieving notable performance metrics, Mistral 7B operates effectively without requiring high-end computational resources. These methods pave the way for a broader availability of anonymized clinical texts, enabling their use for research purposes and the optimization of the healthcare system.

Keywords : de-identification ; Large Language Model ; electronic health records ; Transformers ; Natural Language Processing ; Nursing notes ; General Data Protection Regulation ; Fine-tuning ; French nursing notes

Introduction

The digitization of medical data has profoundly transformed healthcare, facilitating the easy and efficient sharing of patient information [1]. This digital transition, embodied by electronic health record (EHR) systems, offers promising opportunities for data-driven solutions, research, and surveillance on a pan-European scale [2]. Yet, alongside the many advantages of digitization come significant concerns about the privacy and security of sensitive patient data [3]. The European General Data Protection Regulation (GDPR) emphasizes the necessity of stringent data protection measures, particularly for health-related information [2]. Clinical notes, which often encompass identifiable patient details, must adhere to these standards to safeguard patient confidentiality [Loi informatique et liberté]. Before of any data sharing researchers face the critical task of developing and integrating methods that mask sensitive data, guaranteeing protection against any unauthorized access [4]. Our team was recently faced with this challenge in a project aimed at classifying clinical notes from emergency services to extract the necessary information for the establishment of a trauma observatory [5].

Manual de-identification of medical records is not feasible, as it is expensive in terms of personnel resources and the time required to accomplish the task. Alternatively, multiple strategies have been implemented for the automated de-identification of medical records [6,7]. These methods evolved from systems based on explicit rules, regular expressions or dictionaries [8–16], to techniques using machine learning [17–19].

In recent years, the evolution of language models, particularly those based on transformer architectures, has reshaped the landscape of natural language processing (NLP). Transformers, introduced by Vaswani et al. In 2017 [20], provided a novel approach to handling sequential data using self-attention mechanisms, thereby obviating the need for recurrent layers and significantly augmenting training efficiency. This pivotal innovation paved the way for the advent of progressively sophisticated and expansive models. Transformer-based language models of a moderate scale, particularly through customized and fine-tuned versions of the architecture BERT [21], have demonstrated high capabilities in Named Entity Recognition (NER) . Those models offer notable benefits for de-identification, thanks to their capacity to discern patterns among words and phrases. They have the ability to learn from diverse text types means they can effectively tackle various anonymization challenges, as they can be trained to erase a wide range of identifiable details across different document types.

The burgeoning of computational resources and datasets has since kindled a shift towards the construction of massive models, embedded with trillions of parameters [22–24]. As they grew in size, their generalization aptitude and versatility witnessed substantial enhancement, optimizing tasks such as de-identification. In 2023, Liu et al. underscored the potential of leveraging the GPT-4's inherent capacity for zero-shot in-context learning. A salient highlight of their methodology was its ability to maintain the original structure and meaning of the text after the removal of confidential details. While the capabilities of GPT-4 are undeniable, its application in the realm of healthcare presents serious ethical and legal dilemmas, primarily concerning

data privacy and patient confidentiality. On the one hand, due to the vastness of the model, local hosting of GPT-4 is not feasible, therefore, data should be transmitted to external servers, in this case OpenAI's infrastructure. On the other hand, considering the confidentiality of the weights, only locally hosted servers are regulatory compliant. Furthermore, considering GPT-4 is a proprietary model, organizations cannot fully control or audit the underlying mechanics or data handling processes

From a regulatory perspective, sending personal health information (PHI) externally contravenes many data protection regulations, most notably the GDPR in Europe and the Health Insurance Portability and Accountability Act (HIPAA) [25,26] in the United States. This raises not just data sovereignty issues but also infringes on patient rights, as they might not have explicitly consented for their data to be processed in external environments. Hence, while the technological feats of models like GPT-4 are commendable, their real-world applications, especially in sensitive sectors like healthcare, require careful consideration and possibly, significant adjustments to ensure full regulatory compliance and ethical integrity.

Generative language models significantly smaller in size (several billion parameters compared to over a trillion for GPT-4) have been recently developed and made available to the public under licenses that allow for almost unrestricted use (Llama 2 by Meta [27]) or even under open-source terms (Mistral [28]).

The objective of our study is to design, implement and evaluate de-identification methods involving proper prompt engineering and fine-tuning of three, open-source language models (Llama 2 7B, Mistral 7B and Mixtral 8x7B [29]). These models were selected for their moderate size, making them suitable for deployment on personal computers for production inference tasks.

Methods

We first attempted to perform the task using only prompt engineering and zero-shot inference. Because we failed to achieve any significant results we improved the selected models' capability to de-identify clinical texts using quantized Low-Rank Adaptation (qLoRA) [30] fine-tuning with a dataset of instruction/response pairs. In practice, the task consists in replacing personal identifying information (PII) (name, location, dates, telephone number, email, identification numbers) with generic placeholders, represented as '[XXXXX]', or, when no PII is detected, by generating the text as 'ANONYMOUS'. The ultimate goal of this procedure is to preserve text content, ensuring adherence to privacy and confidentiality requirements.

Data source, datasets allocations and annotation

Within the emergency department, triage is conducted by triage nurses. This process involves the collection of information on each patient, including medical history, current symptoms, vital signs, and personal details. It is these data that we have at our disposal in our study. For this investigation, we curated our dataset from a repository containing 425,680 clinical free-text notes (see Multimedia Appendix 1), authored by a nurse during the initial reception and triage of individuals at the Bordeaux University Hospital's adult emergency department over the period spanning from January 2013 to December 2022. A subset of 6,097 clinical notes was randomly selected and independently annotated by two experts. Any arising discrepancies were adjudicated by a third expert, thus establishing a reference database. From this curated sample of 6,097

clinical notes, 3,000 were delineated to constitute a test dataset, upon which accuracy metrics were evaluated (Figure 1). The residual 3,097 clinical notes, alongside an additional sample of 3,000 clinical notes designed using filters and keywords search to encompass a broad spectrum of identifying scenarios, comprised the validation dataset.

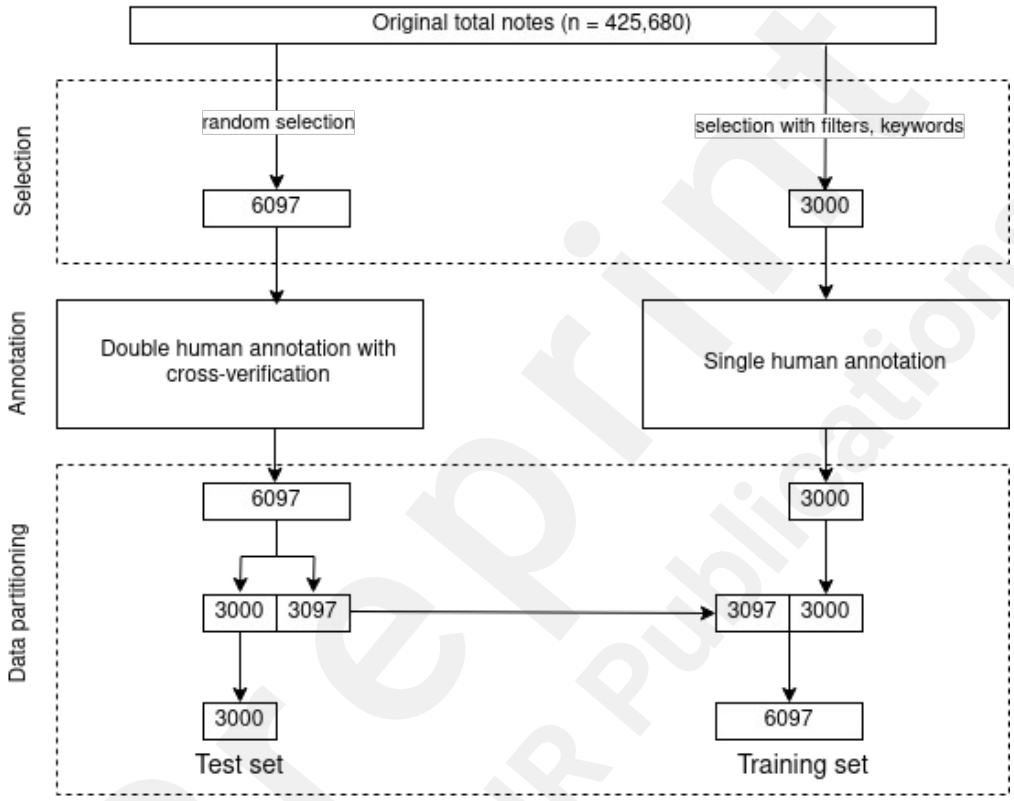


Figure 1: Data Preparation: Annotation and Splitting into Training and Test Sets

In order to further assess whether the de-identification performances of the models varies with the type of PII, we classified identifying information within clinical notes into six distinct categories (Table 1). These categories were utilized by annotators to label such information in the test dataset.

Table 1: PII categories description in medical records

Type	Code	Description
Individual Names	**NAME**	Includes both first and last names of individuals (including patients and medical staff) or of relatives, employers, or household members of the individuals, ensuring personal identification.
Dates	**DATE**	Pertains to specific dates related to medical events, appointments, or personal milestones, formatted as day/month/year.
Geographic Identifiers	**LOC**	Covers names of geographic locations like cities, medical facilities, or

		addresses, facilitating location-based identification.
Phone Numbers	**TEL**	Comprises all forms of telephone numbers for direct contact, including mobile and landline numbers.
Email Addresses	**MAIL**	Encompasses electronic mail addresses, allowing for digital communication.
Miscellaneous Identifiers	**OTHER**	A catch-all category for unique identifiers not covered by other categories, including social security numbers, medical analysis codes, and URLs for patient images.

Selected models

We have selected three language models that share the following two characteristics: being open-source and of sufficiently small size for the production phase to be implemented on affordable PC-type systems. These are Llama 2 7B, Mistral 7B and Mixtral. Llama 2 7B is developed by Meta. Launched in 2023, this is a 7-billion-parameter model, which is claimed to exhibit a good balance between performance and efficiency. We also selected the Mistral 7B model, introduced to the public in October 2023. It has demonstrated superior performance, either matching or surpassing that of Llama 2 13B in extensive benchmarks and showing comparable results to Llama 1 34B in specific domains like reasoning, mathematics, and code generation. In December 2023 the Mixtral 8x7B model was released. It is described as a Sparse Mixture of Experts (SMoE) language model. Its key innovation lies in the routing of inference tasks through one selected expert out of eight, enabled by an additional routing layer. Consequently, despite its 8x7B size with respect to fine-tuning, Mixtral achieves a significant efficiency by requiring an eightfold reduction in parameters for inference task.

Fine-tuning and inference

Each model was subjected to the same prompt/response pairs of clinical notes. The fine-tuning process was uniformly standardized across all three models, albeit with variations in batch sizes and quantization rates to accommodate our hardware constraints. The fine-tuning configuration for Mistral 7B and LlaMa 2 7B involved a batch size of 24 records per GPU, while Mixtral utilized a batch size of 20. The models were fine-tuned over 15 epochs, utilizing the AdamW optimizer with a learning rate of 5e-5 and a weight decay of 0,01. We employed the qLoRA technique, allowing for specific adjustments in selected parts of the model, such as query, key, value, output, and gates projection modules while preserving the overall architecture integrity. The LoRA configuration included rank setting of 32, a learning rate multiplier (alpha) set to 64, with a dropout of 0.1, and without any bias setting. Additionally, to optimize computational efficiency and minimize memory consumption, the models were quantized to 8-bit precision for both 7B models, and 4-bit precision for Mixtral. At every fine-tuning epoch, the inference was induced for each model.

The computational undertakings of this research were performed on a server running Ubuntu 22.04, outfitted with four A100 GPUs, collectively boasting 320 GB of VRAM.

Evaluation

In evaluating the de-identification performance of personal data within clinical notes, our analysis is structured around two primary methodologies. The first methodology operates at the PII-level, enabling us to provide estimates of recall, precision, and F1 scores that are comparable with previous work in the literature. The second methodology focuses on clinical notes as the statistical unit, enabling us to assess the variation in recall performance according to the category of PII. This latter approach needs to be complemented by the measurement of a BLEU score to assess potential modifications in the text. The assessment of the number of successful de -identifications was conducted through a comparison with the manually annotated test dataset.

PII-Based Metrics

This approach centers on treating each PII as an independent statistical unit. This perspective allows us to

gauge the precision and recall of our de-identification efforts at the most granular level. Recall in this context is conceptualized as the proportion of PIIs accurately identified and removed from the clinical notes.

$$RecallPII = \frac{numberofcorrectlyde-identifiedPIIperclinicalnotes}{totalnumberofPIIperclinicalnotes}$$

Precision, meanwhile, reflects the accuracy of our model to identifying and eliminating actual PIIs, distinguishing between correct identifications and false positives.

$$PrecisionPII = \frac{numberofcorrectlyde-identifiedPIIperclinicalnotes}{totalnumberofPIItagged}$$

The summary F1 measure is:

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

Clinical Note-Based Metrics

The second approach adopts the entire clinical note as the statistical unit of analysis. Here we evaluate the success of de-identification on a document-wide scale, marking a 'success' when every PII within a note has been successfully de-identified. Such a measure offers insight into the overall effectiveness of our de-identification protocols. Recall, in this instance, measures the ratio of fully de-identified notes to those containing any PII.

$$Recall = \frac{numberofcorrectlyde-identifiedclinicalnotesamongidentifyingclinicalnotes}{totalnumberofidentifyingclinicalnotes}$$

Because the clinical notes in the validation set are annotated by indicating the nature of the PII (according to the categories in Table 1), it is possible to detail the variations in recall by category. The relevance of precision is altered in this context, as it necessitates a different consideration of what constitutes an anonymization attempt, denoted by the presence of an anonymization tag. Instead, the potential alteration of content possibly induced by the de-identification process was measured using the BiLingual Evaluation Understudy (BLEU) score [31].

$$BLEU = BP \cdot \exp\left(\sum w_n \log p_n\right)$$

where BP is the brevity penalty, w_n the weight for each n-gram, and p_n the precision of n-grams. We set n=4 for the BLEU score calculation, aligning with common practice in NLP to capture up to four-gram coherence, thereby ensuring a comprehensive evaluation of content preservation.

Results

Data overview

Very few notes contained PIIs categorized as email addresses and 'other.' These categories are included in the training sample due to an ad-hoc selection process, which utilized filters to ensure representation, as half of the set was selected this way. Our examination of the test sample, which consists entirely of randomly selected clinical notes, reveals that names, places, and dates are the most prevalent types of PII. The categories of identifying data in the training and test sets are summarized in Table 2.

Table 2: Enhanced distribution of PII in train and tests sets

	Train set	Test set
Clinical notes		
Non-anonymous medical notes	3442 (56.5%)	935 (31.2%)
Randomly selected medical notes	3097	3000
Ad-hoc selected medical notes	3000	-
Total count	6097	3000
PII Categories		
NAME	3016	555
LOCATION	1801	715
TELEPHONE	650	41
EMAIL	13	0
DATE	2404	607
OTHER	33	1
Total number of PII	7917	1919

Regarding the length of clinical notes, they range from 8 to 3916 characters (with an average of 443 characters) in the training set and from 3 to 2138 characters (averaging 439 characters) in the test set. A total of 935 clinical notes in the test set (31.2%) contain at least one PII.

Performance using PII-based metrics

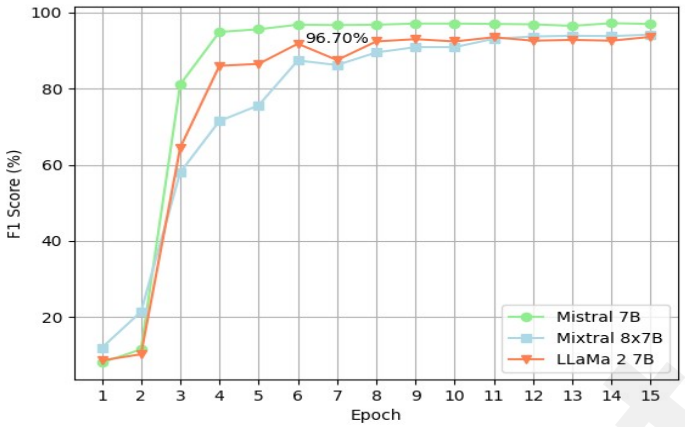


Figure 2: Plot of F1-score by epoch: PII as statistical unit.

Figure 2 plots the change in the F1 score over the 15 epochs of fine-tuning for the three respective models. The Mistral 7B model quickly reaches a performance plateau, where its F1 score stabilizes, whereas the Mixtral 8x7B and Llama2 7B models exhibit a slower rate of improvement, with both reaching a plateau in their F1 scores around the 12th epoch.

Recall analysis

The recall estimates of the three models are shown in Figure 3 and 4.

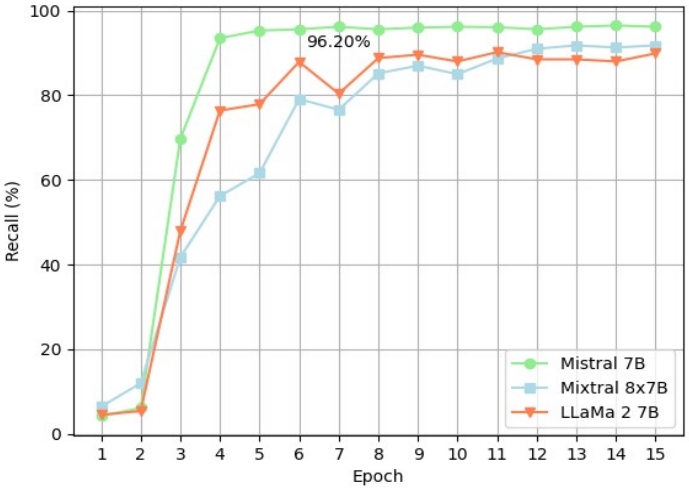


Figure 3: Plot of Recall by epoch: PII as statistical unit.

Mistral 7B and Mixtral 8x7B achieved better overall recall. The Mistral 7B and Mixtral 8x7B models demonstrated marked enhancements in their de-identification efficacy across epochs, starting from the third epoch onward. Notably, the Mistral 7B model has shown a rapid improvement in performance, achieving a performance plateau by the sixth epoch. Conversely, the Mixtral 8x7B model's improvement trajectory was more gradual, reaching a performance stable performance level by the thirteenth epoch. The overall success rate appears not to improve beyond epoch 7 for the Mistral 7B model. Consequently, in the subsequent analysis, this epoch was selected for comparing success rates across categories.

As shown in Figure 5-8, Mistral 7B consistently outperformed Mixtral 8x7B and Llama 2 across all data identification categories. Despite Mixtral's performance improving over time, it still did not surpass Mistral 7B. Using Mistral 7B, a 100% recall was observed for phone numbers (Figure 5) and recall was lower for locations (Figure 6) than for names (Figure 7).

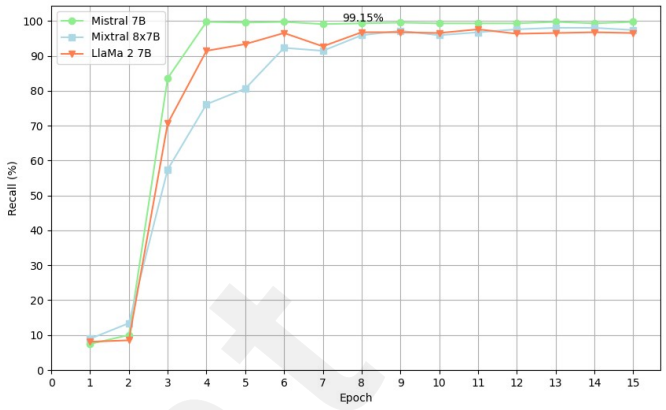
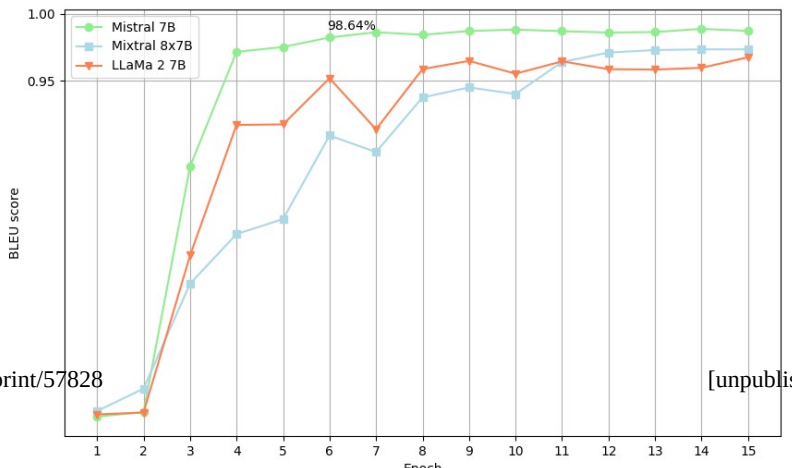


Figure 8: Plot of Recall by epoch for PII: NAME.

BLEU score

BLEU-4 scores were calculated to assess whether the models modified the texts at the note level. During the de-identification process, medical texts remained almost unchanged as demonstrated by a consistent high BLEU-4 score (Figure 9) beyond epoch 5.



Results summary at Epoch 7

The Table 3 below presents a summary of performance metrics achieved by our models at epoch 7.

Table 3: Fine-tuned models performance at Epoch 7

	Clinical notes	PII		
Model	Recall	Precision	Recall	F1
Mistral 7B	0.9326	0.9721	0.9625	0.9673
Llama 2 7B	0.6888	0.9596	0.8041	0.8750
Mixtral 8x7B	0.6417	0.9852	0.7655	0.8616

The results demonstrate that the Mistral 7B model outperforms both the Mixtral 8x7B and Llama 2 7B with a F1 score of 0.9673. When using clinical note as the statistical unit, the recall is also much higher (0.9326) for Mistral 7B than Llama 2 and Mixtral 8x7B models.

Error analysis

In epoch 7 of the Mistral 7B model, a total of 63 clinical notes were not properly anonymized, as detailed in Table 4. Among these, location (LOC) errors were the most frequent, with 44 instances. Anonymizing geographical and institutional identifiers then remains a significant challenge (with a recall of 86.1%). Specifically, 31 notes failed to anonymize health or social facilities while 12 notes were cities. Conversely, errors involving names (NAME) were significantly fewer, with only 4 instances, including one patient name and three doctors' names, resulting in a high recall of 99.8% for this category. Date-related errors (DATE) were observed in 14 notes (with a recall of 97.8%).

Table 4: Summary of de-identification errors at Epoch 7

Errors	**Count**
Total	63
Returned ANONYMOUS	29
Annotation Error	34
Errors in PII categories	
NAME	4
LOC	44
DATE	14
OTHER	1

The test dataset, comprising 3,000 clinical notes, underwent a post-hoc examination to identify any inaccuracies resulting from manual annotations that would have been detected by all 15 versions of our three

finely-tuned models, spanning epochs 1 to 15. Through this process, we were able to pinpoint 65 notes in which the model detected personally identifiable information through the medical histories that were categorized as anonymous (i.e., without identifying data, 2,066 clinical notes), in which the model detected personally identifying information that had been overlooked by human annotators.

We observed that the models outperformed human annotation in 9 clinical records from the test set. Specifically, in these 9 records, 5 locations (LOC), 3 names (NAMES), and 1 date (DATE) were omitted during manual annotation. The remaining 53 records present annotation errors from the models. Therefore, the total number of actual personally identifiable information (PII) amounts to 1928, contrary to the 1919 initially identified by our experts.

Subsequently, corrections were made to the test dataset based on these findings, and main outcomes were recomputed in an additional sensitive analysis. The metric measurements after accounting for these modifications are only slightly altered from the original results (see Multimedia Appendix 2 for the details).

Discussion

Principal Findings

In this study, we assessed the performance of three generative NLP models in the de-identification of clinical text documents. The generative model Mistral 7B demonstrated the highest performance with an overall F1-score of 0.9673. At the clinical notes level, the same model achieved an overall recall of 0.9326, with this rate increasing to 0.9915 for the anonymization of names. The evaluation was based on a test dataset of 3000 clinical notes, among which only four notes failed to be fully anonymized for names; in one case, the non-anonymized name was that of a patient. Because the method relies on the use of generative models, we also measured potential text alterations generated by the process. Beyond the fifth epoch, the BLEU score consistently exceeded 0.9864.

Strengths

Our work distinguishes itself from the existing scientific literature by employing a method that does not rely on NER and uses moderate-sized models. Instead, the use of generative LLMs allows for the production of text that is anonymized by removing PII components. This is the reason why we added metrics that use clinical notes as the statistical unit. This led us to use the BLEU metric to assess potential text alterations. Another consequence of this method is that no hyperparameters are set which made it possible to avoid the use of separate test and validation dataset partitions. The size of our training and test samples, independently annotated by two experts, constitutes a significant strength in our study. To our knowledge, no other study has used a test sample of such size (3,000 notes). Yet, it is crucial to have the means to detect rare errors if the ultimate goal is to develop a system that guarantees the anonymization of clinical texts. We deliberately limited our model selection to those whose implementation does not require powerful servers and can be executed on personal Bcomputers equipped with a consumer-grade graphics card. The largest model is Mixtral 8x7B, which has approximately eight times more parameters than the other two models. Mixtral 8x7B shares the same architecture as Mistral 7B, with the distinction that each layer consists of 8 feed-forward blocks. Although training it requires significant memory capacity, this is not the case during the inference phase, during which only two of the feed-forward blocks are utilized, selected by a network acting

as a router.

Limitations

We observed some inaccuracies in the annotation process. We try to assess the impact on metrics conducting a post-hoc analysis taking into account the corrections made by the model and found few variations. However, we cannot exclude than some mistakes remain in the test set that were not detected by the model.

We opted for a fine-tuned LLM-based approach over a dedicated NER model due to pragmatic considerations. Our hypothesis was that a targeted human annotation process, with expert annotators pinpointing PII within texts, would be more effective than a broad NER annotation effort, given the same time investment. Focusing on essential PII elements helps us minimize the ambiguities that broader NER annotations often entail. This focus leads to improved precision and recall rates during the training phase. Furthermore, this approach is in line with the TARPON project's objectives, which prioritize the accurate removal of PII from unstructured medical texts.
The default choice for identification tasks is usually a bidirectional transformer, starting from the hypothesis that the relationship of a word with its context before and after that word allows for better comprehension of the role of those words and therefore should be more suited for NER tasks. However, this hypothesis no longer holds when dealing with generative models. Since the goal here is to generate redacted text, the provided prompt has access to the entire corrected phrase. Consequently, relative to a given word, implications cannot be considered unidirectional.

A final weakness of our method is that it uses a model fine-tuned with non-anonymous clinical texts. For this reason, we are unable to share the model's weights with the community for other uses.

Comparison with Prior Work

Comparing the performance of our models with those documented in the literature presents challenges because our models are specifically fine-tuned to anonymize French-language clinical notes. Consequently, it is not feasible to apply them to the English-language databases traditionally used for benchmarking, such as i2b2 [32], MIMIC II [33], and MIMIC III [34]. Therefore, to compare performance metrics accurately, it is necessary to assess the complexity of clinical texts from these databases against those used in our study. In the appendix, we include examples of clinical notes from our dataset to demonstrate that PIIs can appear randomly within the text, in an unstructured manner, and that these PIIs, along with the rest of the text, often include numerous abbreviations. This tendency towards abbreviation is explained by the unique demands of emergency department settings, where nurses are required to perform efficient, real-time data entry into the hospital's information system. As a result, our dataset more closely aligns with MIMIC II, which features unstructured clinical notes made by nurses, as opposed to i2b2, where each type of information is distinctly separated, preventing the amalgamation of multiple PIIs within single sentences.

As shown in Multimedia Appendix 3, our results (overall F1-score of 0.9673) are on par with previous studies on english clinical text corpus that used an algorithm including models using self-attention [17,24,35,36]. The Multimedia Appendix 4summarizes study results that examined recall variations according to PII categories. These figures consistently show that the relative weakness of these algorithms, ours included, lies in a small number of errors concerning locations. Our dataset presents additional challenges for PII identification due to the presence of multiple variations of PII, including acronyms, abbreviations, and typing errors. Specifically, of the 44 notes with failed identification, 15 involved abbreviations or acronyms, and 2 contained typing errors.

Future work

We aim to further improve the detection capabilities of PII in our medical notes by fine-tuning our model with new annotated data. This could partially be done by implementing an artificial clinical notes production

process through commercially available APIs such as GPT-4. Such a much larger LLM than ours is capable of generating realistic notes with PII and annotations. This strategy will also be employed to build and evaluate a new Mistral 7B-based model fine-tuned on only artificial clinical notes. This will enable us to address two objectives. First, if successful, this will lead to the availability of an open-source model that would be useful for the community. Second, this strategy will be implemented to train a model designed to process notes in English, then allowing for performance comparison against literature benchmark datasets such as i2b2 and MIMIC. We intend to use larger LLMs to produce artificial datasets to generate a substantial volume of new training data, ensuring equitable representation and diversity among different PII categories. This will allow us to evaluate two things: first, identifying the optimal amount of clinical notes required to achieve the highest possible accuracy and recall. Second, it will enable us to conduct a comparative study between models fine-tuned with real data to determine which are the most effective. The efficiency of the newly refined models will be evaluated using our corrected test set alongside new annotated data from various emergency services. Through this holistic approach, we aim to enhance the valorization of our models for broader usage, thus contributing to the development of privacy-preserving technologies in the healthcare domain and enhancing the security of patients' sensitive information.

Conclusions

Our research underscores the significant capabilities of generative NLP models, with Mistral 7B standing out for its superior ability to de-identify clinical texts efficiently. Achieving notable performance metrics, Mistral 7B operates effectively without requiring high-end computational resources. These methods pave the way for a broader availability of anonymized clinical texts, enabling their use for research purposes and the optimization of the healthcare system.

Acknowledgments

This study was conducted under the TARPON project by the BPH AHEAD team and the Bordeaux University Hospital's emergency department. We thank the labeling team and the University Hospital of Bordeaux for their logistical support and data access.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BLEU: BiLingual Evaluation Understudy
EHRs: Electronic Health Records
GDPR: General Data Protection Regulation
GPT: Generative Pre-trained Transformer
HIPAA: Health Insurance Portability and Accountability Act
LLM: Large Language Model
NER: Named Entity Recognition
NLP: Natural Language Processing
PHI: Personal Health Information
PII: Personal Identifying Information
qLoRA: quantized Low-Rank Adaptation
SmoE: Sparse Mixture of Experts
TARPON: Traitement Automatique des Résumés de Passages aux urgences pour un Observatoire National (in english : Automatic Processing of Emergency Department Visits Summaries for a National Observatory)

Data Availability

The data set and model’s weights ar not available because of patient privacy restrictions.

Authors' Contributions

Conceptualization and design: EL, CGJ, AFM. Annotation: BC, OD, EL, DR, CGJ. Analysis and interpretation: OD, CGJ, EL. Manuscript drafting: OD, EL, AGA. Critical revision: All authors. Provision of study material: CGJ. Supervision: EL.

Conflicts of Interest

None Declared.

Multimedia Appendix 1

In the provided document, the upper section presents an original clinical text as it was written by a nurse or hospital staff. The section below displays the same note after correction, rendered in proper French.

For this first example (Textbox 1), identifying data have been pseudonymized. The location (LOC) of the original note has been replaced with 'Clinique du Louvre'.

Textbox 1. Example of french nursing notes

Original note : tft de la cl du louvre pour pec dune fracture du col femoral droit patient sous plavix transféré du louvre pour orthopédiste pour prise en charge d'une fracture du col du fémur

Manual transcription without abbreviations: Transfert de la clinique du Louvre pour prise en charge d'une fracture du col femoral droit patient sous plavix transféré du Louvre pour orthopédiste pour prise en charge d'une fracture du col du fémur

For this second example (Textbox 2) we change the date, the name of the Doctor and the Location.

Textbox 2. Example of french nursing notes

Original note : chutes à repet trauma lombaires sans deficit plaie 2 pieds car a marchè sur debris verre (a rv pour irm cerebrale à cl du louvre)- a resulttas bacterio aurait inf urinaire pas encore ttè (sa fille a resultats en salle d'attente) 01/01/2000 22:22 - docteur Dupond, interne retrouvé au sol dans les toilettes à 17h, vu pour la derniere fois à 10h par la femme de ménage. contexte d'infection urinaire, avec ecbu positif à e.coli, sensible à la rocéphine (ecbu dans le dossier). arrive avec 39 de température.

Manual transcription without abbreviations: Chutes à répétition, traumatisme lombaires sans déficit, plaie des 2 pieds, car a marché sur des débris verre (à rendez-vous pour irm cérébrale à clinique du Louvre)- A des résultats bacterio, elle aurait infection urinaire pas encore traitée (sa fille à les résultats en salle d'attente) 01/01/2000 22:22 - docteur Dupond, interne retrouvé au sol dans les toilettes à 17h, vu pour la dernière fois à 10h par la femme de ménage. Contexte d'infection urinaire, avec ecbu positif à e.coli, sensible à la rocéphine (ecbu dans le dossier). Arrive avec 39 de température.

Multimedia Appendix 2

In the corrected data, there are 2056 non-anonymous anamneses (compared to 2066 in the original test set) and 944 anonymous anamneses (compared to 935 in the original test set), indicating that 9 anamneses were misannotated by our experts.

Recall, precision and F1 statistics

The Table 1 presents a comparative analysis of model performance, measured by recall, precision, and F1 score, at epoch 7 on both the original and post-hoc test sets. Bold values indicate improvements in the corrected versions of the Mistral 7B, Mixtral 8x7B, and LLaMa2 7B models, highlighting the effectiveness of fine-tuning.

Table 1: Comparative analysis of fine-tuned model performance at epoch 7 for original test set and post-hoc test set

Model	PII	Clinical Notes		
	Recall	Precision	Recall	F1
Mistral 7B	0.9326	0.9721	0.9625	0.9673
Mistral 7B^{p}	0.9465	0.9732	0.9630	0.9681
LLaMa2 7B	0.6888	0.9596	0.8041	0.8750
LLaMa2 7B^{p}	0.7932	0.9583	0.8026	0.8736
Mixtral 8x7B	0.6417	0.9852	0.7655	0.8616
Mixtral 8x7B^{p}	0.7620	0.9819	0.7625	0.8584

^{p}: Post-hos test set

The corrected version of Mistral 7B shows improvement in all metrics, especially precision, which increased to 0.9588. Mixtral 8x7B and LLaMa2 7B corrected models also exhibit slight enhancements, with notable increases in precision and F1 score, demonstrating the positive impact of model corrections.

Recall for PPI categories

The Table 2 provides a comparative analysis of the recall performance for different PPI categories (NAME, TEL, DATE, and LOC) evaluated at epoch 7 across both the original and post-hoc test sets for the Mistral 7B, Mixtral 8x7B, and LLaMa2 7B models, along with their corrected versions.

Table 2: Comparative analysis of fine-tuned model performance at epoch 7 for original test set and post-hoc test set. Recall PII

Model	NAME	TEL	DATE	LOC
Mistral 7B	0.9914	1.0	0.9725	0.9026
Mistral 7B^{p}	0.9914	1.0	0.9744	0.9037
LLaMa2 7B	0.9276	0.8787	0.8627	0.5553
LLaMa2 7B^{p}	0.9255	0.8787	0.8624	0.5557
Mixtral 8x7B	0.9148	0.909	0.7941	0.5464
Mixtral 8x7B^{p}	0.9106	0.909	0.7937	0.5426

^{p}: Post-hos test set

The results do not show significant differences in recall for the various PPI categories between the original and corrected versions of the test set. The recall rates across NAME, TEL, DATE, and LOC

categories remain relatively stable, indicating that the corrections made to the models have not drastically altered their ability to correctly identify these PPI categories.

Multimedia Appendix 3

Table 5: Comparison with previous work

Authors	Corpus (N)	Language	Precision	Recall	F1	Notes
Grouin et al. (2013) [37]	Cardiology set (62)	FR	94.8[b]	89.4[b]	92.1[b]	Medina-CFR
			89.1[b]	86.5[b]	87.8[b]	Medina-RB
	Foetopathology set (10)		75.4[b]	58.5[b]	65.9[b]	Medina-CFR
			72.0[b]	72.6[b]	72.3[b]	Medina-RB
Chazar et al. (2014) [38]	French discharge letters (508)	FR	79.6	98.1	87.9	Pattern matching. Use a list of authorized words
Catelli et al. (2020) [39]	SIRM (50)	IT	-	-	85.61	Bert-base (IT) cased
		IT	-	-	94.49	mBERT Cased
		IT	-	-	83.17	Bi-LSTM-CRF : BPEemb (IT) + Fair (IT)
		EN|IT	-	-	86.19	Bi-LSTM-CRF : Multi-BPEemb + Fair multi fast
Berg et al. (2020) [40]	neurology clinical unit (-)	SW	96.07c.	92.82[c]	94.4[c]	token binary evaluation optimized for F1 model
Ahmed et al. (2020) [17]	i2b2 (1304)	EN	98.74	95.85	97.28	GRU
			99.01	95.12	97.08	GRU-GRU
			98.74	95.27	96.98	LSTM-GRU
			98.03	98.41	98.22	Self-attention
	MIMIC-II (486)	EN	81.82	66.23	73.21	GRU
			80.00	71.00	75.23	GRU-GRU
			85.14	68.18	75.72	LSTM-GRU
			89.20	82.90	85.90	Self-attention
	MIMIC-III (891)	EN	99.94	100	99.97	GRU
			99.93	99.99	99.96	GRU-GRU
			99.94	99.99	99.96	LSTM-GRU
			99.95	98.78	99.36	Self-attention
Syed et al. (2022) [41]	i2b2 (-)	EN	94.89	95.96	93.84	Input Embeddings+Bi-LSTM+CRF
	CoNLL-2003 (-)		92.99	93.50	93.25	
	internal corpus (100)		93.86	93.37	94.31	without mixed domain pre-training
			96.23	94.51	95.36	with mixed domain pre-training
Meaney et al.(2022)	2014-i2b2 (486)	EN	96.69	96.81	96.75	Roberta-Large
			96.62	96.27	96.44	Albert-XXLarge fine-tuned

			95.10	95.33	95.22	Roberta-Base fine-tuned
			95.53	95.34	95.43	Bert-Large fine-tuned
			93.87	93.85	93.86	Albert-Base fine-tuned
			93.80	94.40	94.10	Bert-Base fine-tuned
Tchouka et al. (2022) [42]	HNFC (375)	FR.	94.6a	94.9a	94.7a	NER hybrid system
Liu Z. et al. (2023)	2014-i2b2 (50)	EN	-	-	-	Accuracy 0.99 Explicit prompt GPT-4
			-	-	-	Accuracy 0.929 Explicit prompt ChatGPT
Liu L. et al. (2023) [36]	2014-i2b2 (-)	EN	98.92a,c	97.66[a,c]	98.29[a,c]	BiLSTM-CRF (RoBERTA).
	CardiacAI (40)		95.19[a,c]	93.47[a,c]	94.32[a,c]	BiLSTM-CRF (RoBERTA).
	CardiacAI (60)		94.87[a,c]	95.26[a,c]	95.07[a,c]	BiLSTM-CRF (RoBERTA).
Liu J. et al. (2023) [24]	OpenDeid (700)	EN	95.58[a,b]	92.42[a,b]	93.97[a,b]	fine-tuned BioBERT
			95.82[a,b]	91.98[a,b]	93.86[a,b]	fine-tuned Clinical BioBERT
			95.87[a,b]	92.22[a,b]	94.01[a,b]	fine-tuned Discharge Summary BioBERT
			97.84[a,b]	95.92[a,b]	96.87[a,b]	fine-tuned Discharge Summary BioBERT + cascading rules
			95.59[a,b]	89.35[a,b]	92.37[a]	LSTM GloVe+PMC+word2vec-OpenDeID corpus word embeddings.
Our Model	French nursing notes (3000)	FR	97.32	96.30	96.73	fine-tuned Mistral 7B model with LoRA

Multimedia Appendix 4

Table 1 : Comparison of Recall for PII with previous work

Authors	Method	L.	NAME		DATE		LOC		TEL		Info
			R (%)	N	R (%)	N	R (%)	N	R (%)	N	
Grouin et al [37]	MEDINA-RB	FR	90.7 – 92.7	314[s]	87.1	238[s]	12.5-100	81[s]	100	8	NAME : Last Name (205) and First Name (109) LOC : HOSPITAL (43), TOWN(22), ZIP(8), ADDRESS(8)
	MEDINA-CRF	FR	88.3-89,0		94.6		12.5-75.0		75.0		
Tchouka et al [42]	FlauBERT-MEDINA	EN	99.8	-	86.7	-	57.3-95.1	-	97.9	-	LOC : ORGANIZATION + LOCATION
L Liu et al [36]	BiLSTM-CRF (RoBERTA)	EN	95.83	528	96.92	65	50.0	6	85.71	35	Dataset Cardaic AI (N = 600)
Our work	Mistral-7B + qLoRA	FR	99.14	555	97.25	607	90.26	715	100	100	

S : sum of entities

References

1. Menachemi N, Collum T. Benefits and drawbacks of electronic health record systems. Risk Manag Heal Policy. 2011;4:47–55.

2. European Parliament C. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Parliament, Council; 2016. p. 1–88.

3. Organization WH. mHealth: New horizons for health through mobile technologies: second global survey on eHealth. World Health Organization; 2012.

4. El Emam K. Methods for the de-identification of electronic health records for genomic research. Genome Med. 2011;3:25.

5. Chenais G, Gil-Jardiné C, Touchais H, Avalos Fernandez M, Contrand B, Tellier E, et al. Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study. JMIR AI. 2023;2:e40843.

6. 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 [Internet]. 2010;10. Available from: http://dx.doi.org/10.1186/1471-2288-10-70

7. Negash B, Katz A, Neilson CJ, Moni M, Nesca M, Singer A, et al. De-identification of Free Text Data containing Personal Health Information: A Scoping Review of Reviews. Int J Popul Data Sci [Internet]. 2023;8. Available from: http://dx.doi.org/10.23889/ijpds.v8i1.2153

8. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak [Internet]. 2006;6. Available from: http://dx.doi.org/10.1186/1472-6947-6-12

9. Berman JJ. Concept-Match Medical Data Scrubbing. Arch Pathol Amp Lab Med. 2003;127:680–

10. Friedlin FJ, McDonald CJ. A Software Tool for Removing Patient Identifying Information from Clinical Documents. J Am Med Inform Assoc. 2008;15:601–10.

11. 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;121:176–86.

12. Morrison FP, Li L, Lai AM, Hripcsak G. Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes? J Am Med Inform Assoc. 2009;16:37–9.

13. Neamatullah I, Douglass MM, Lehman LH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak [Internet]. 2008;8. Available from: http://dx.doi.org/10.1186/1472-6947-8-32

14. Ruch P, Baud RH, Rassinoux A-M, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. Proc AMIA Symp. 2000. p. 729–33.

15. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp. 1996;333–7.

16. Thomas SM, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp. 2002. p. 777–81.

17. Ahmed T, Aziz MMA, Mohammed N. De-identification of electronic health record using neural network. Sci Rep [Internet]. 2020;10. Available from: http://dx.doi.org/10.1038/s41598-020-75544-1

18. Guo Y, Gaizauskas RJ, Roberts I, Demetriou G, Hepple M. Identifying Personal Health Information Using Support Vector Machines. 2006. Available from: https://api.semanticscholar.org/CorpusID:16833759

19. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc. 2016;24:596–606.

20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. 2023.

21. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019;

22. OpenAI, :, Achiam J, Adler S, Agarwal S, Ahmad L, et al. GPT-4 Technical Report. 2023;

23. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: Harmlessness from AI Feedback. 2022;

24. Liu J, Gupta S, Chen A, Wang C-K, Mishra P, Dai H-J, et al. OpenDeID Pipeline for Unstructured Electronic Health Record Text Notes Based on Rules and Transformers: Deidentification Algorithm Development and Validation Study. J Med Internet Res. 2023;25:e48145.

25. Law P. Health insurance portability and accountability act of6. Available from: http://www.eolusinc.com/pdf/hipaa.pdf

26. Liu Z, Huang Y, Yu X, Zhang L, Wu Z, Cao C, et al. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4 [Internet]. arXiv; 2023 [cited 2024 Feb 5]. Available from: http://arxiv.org/abs/2303.11032

27. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023.

28. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D de las, et al. Mistral 7B. 2023.

29. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of Experts. 2024.

30. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs. 2023;

31. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine

translation. Proc 40th Annu Meet Assoc Comput Linguist [Internet]. USA: Association for Computational Linguistics; 2002. p. 311–8. Available from: https://doi.org/10.3115/1073083.1073135

32. Informatics for Integrating Biology & the Bedside (i2b2) [Internet]. Available from: https://www.i2b2.org/

33. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Crit Care Med. 2011;39:952–60.

34. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.

35. Meaney C, Hakimpour W, Kalia S, Moineddin R. A Comparative Evaluation Of Transformer Models For De-Identification Of Clinical Text Data. 2022.

36. Liu L, Perez-Concha O, Nguyen A, Bennett V, Blake V, Luxan B, et al. De-identifying free text data in electronic medical records: a web-based application based on human in the loop deep learning (Preprint). Interact J Med Res. 2023;12.

37. Grouin C, Zweigenbaum P. Automatic de-identification of French clinical records: comparison of rule-based and machine-learning approaches. Stud Health Technol Inf. 2013;192:476–80.

38. Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J-B, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inf. 2014;83:303–12.

39. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. Appl Soft Comput. 2020;97:106779.

40. Berg H, Henriksson A, Dalianis H. The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. 2020. p. 1–11.

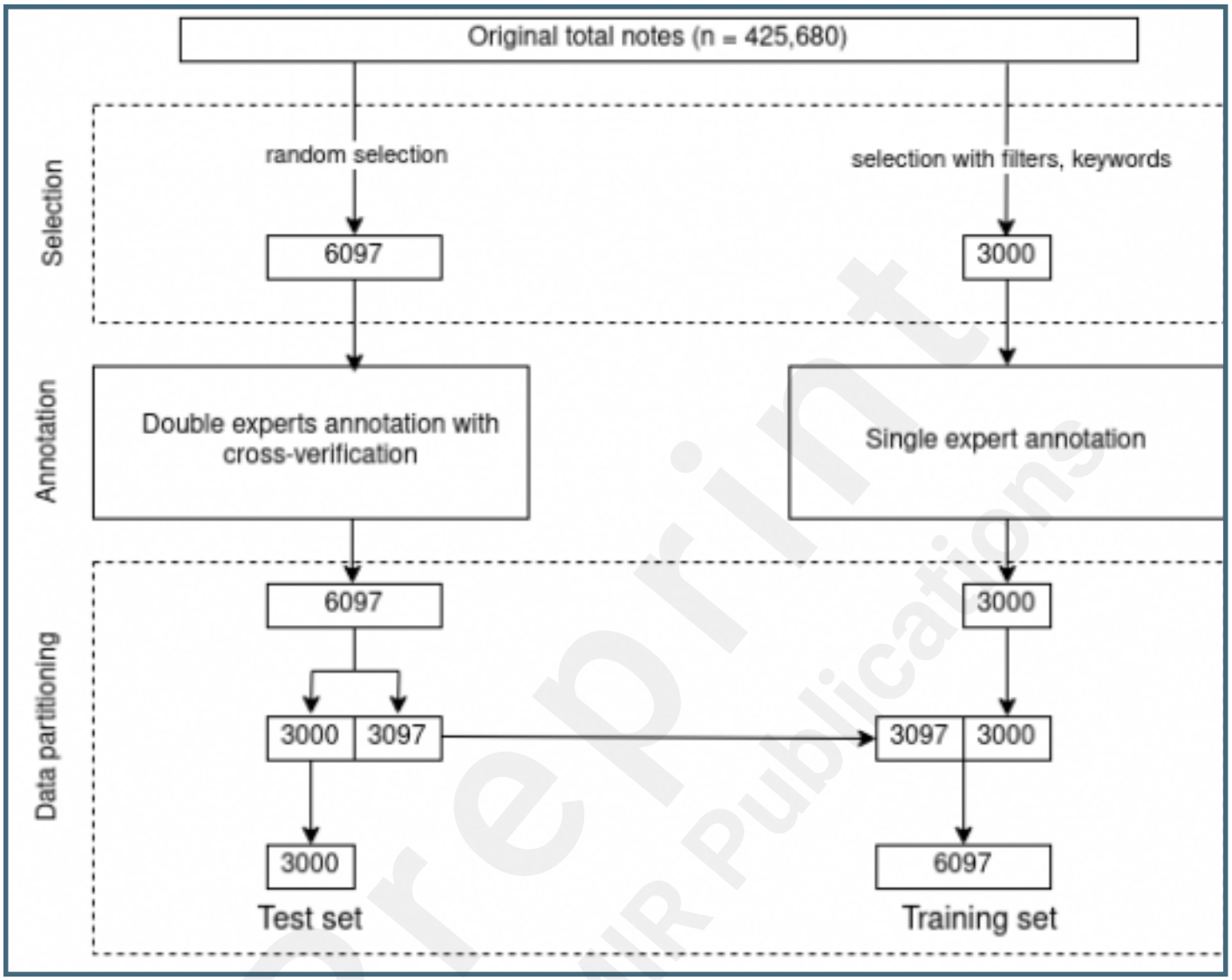
41. Syed M, Sexton K, Greer M, Syed S, VanScoy J, Kawsar F, et al. DeIDNER model: A neural network named entity recognition model for use in the DE-identification of clinical notes. Proc 15th Int Jt Conf Biomed Eng Syst Technol. SCITEPRESS - Science and Technology Publications; 2022.

42. Tchouka Y, Couchot J-F, Coulmeau M, Laiymani D, Rahmani A. De-Identification of French Unstructured Clinical Notes for Machine Learning Tasks [Internet]. 2022. Available from: https://hal.science/hal-03720808

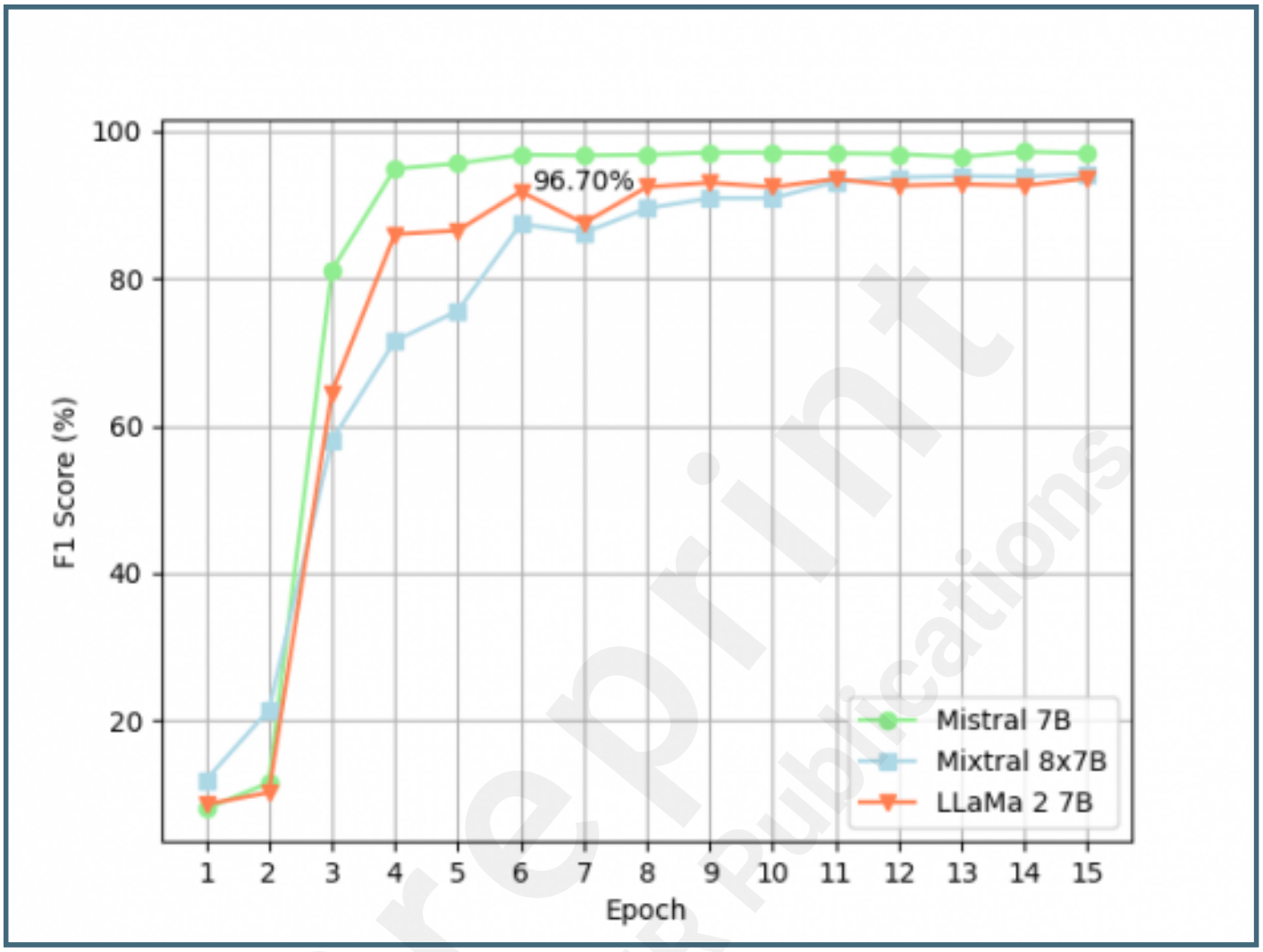
Supplementary Files

Figures

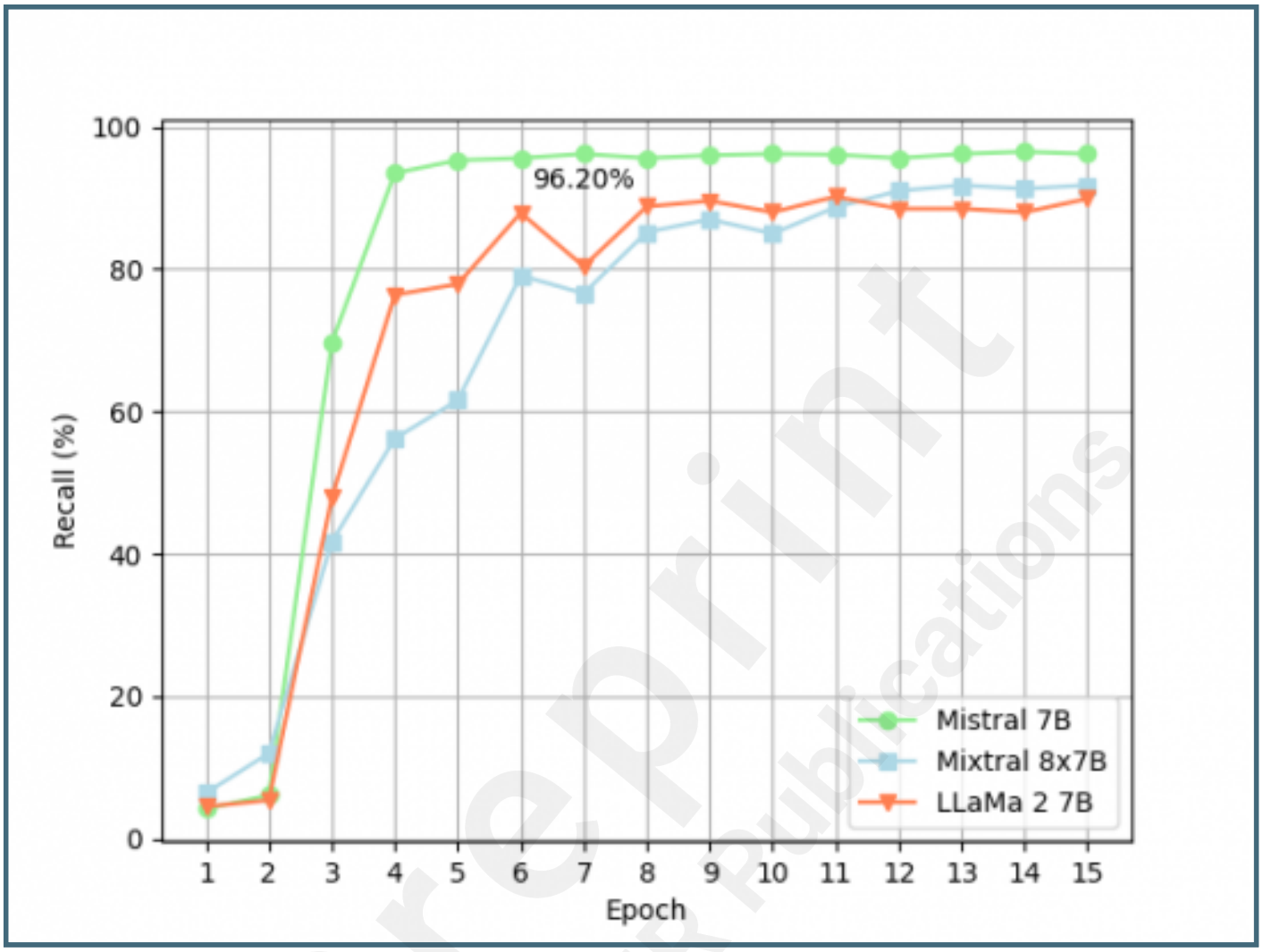
Data Preparation: Annotation and Splitting into Training and Test Sets.



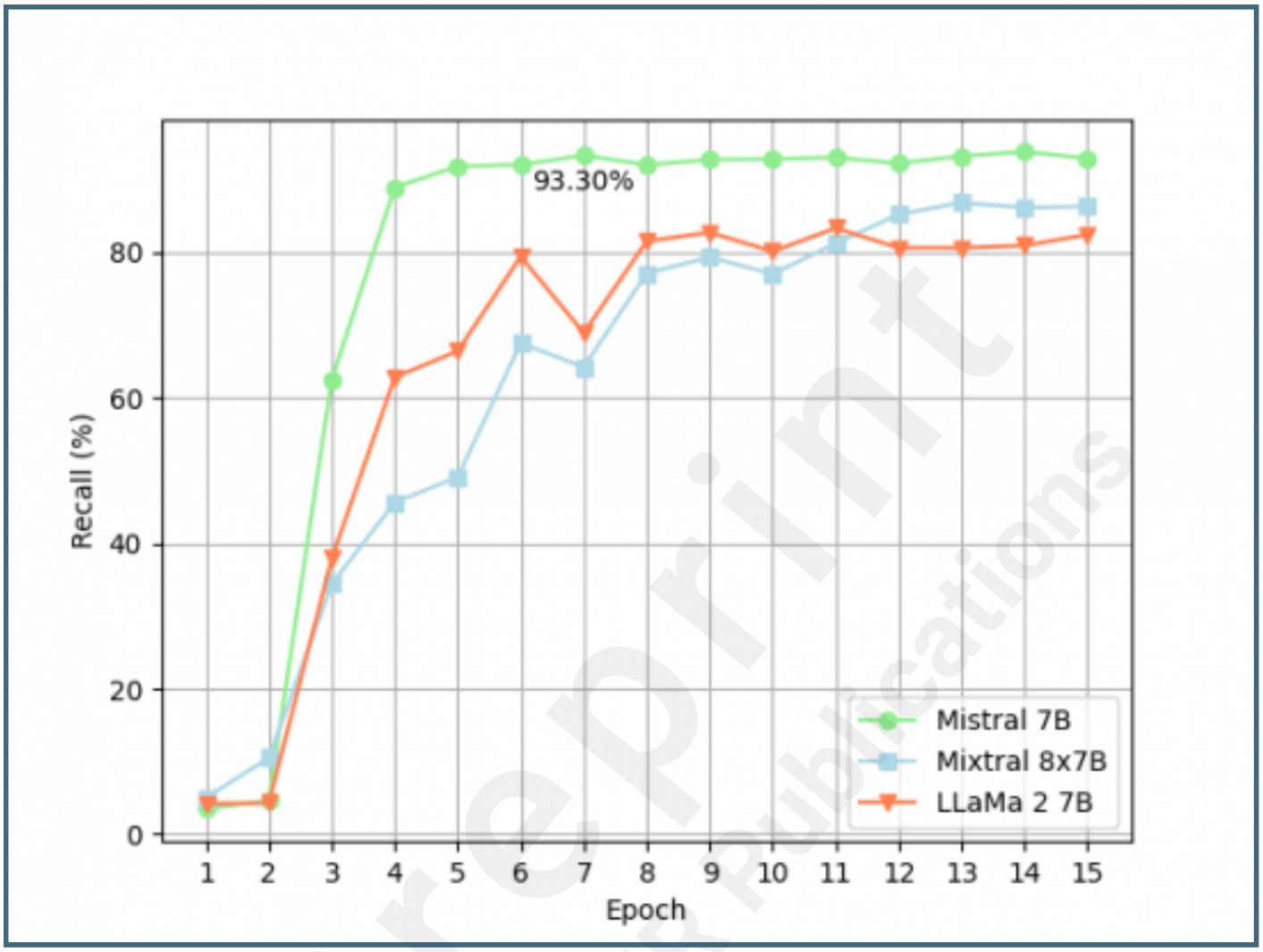
Plot of F1-score by epoch : PII as statistical unit.



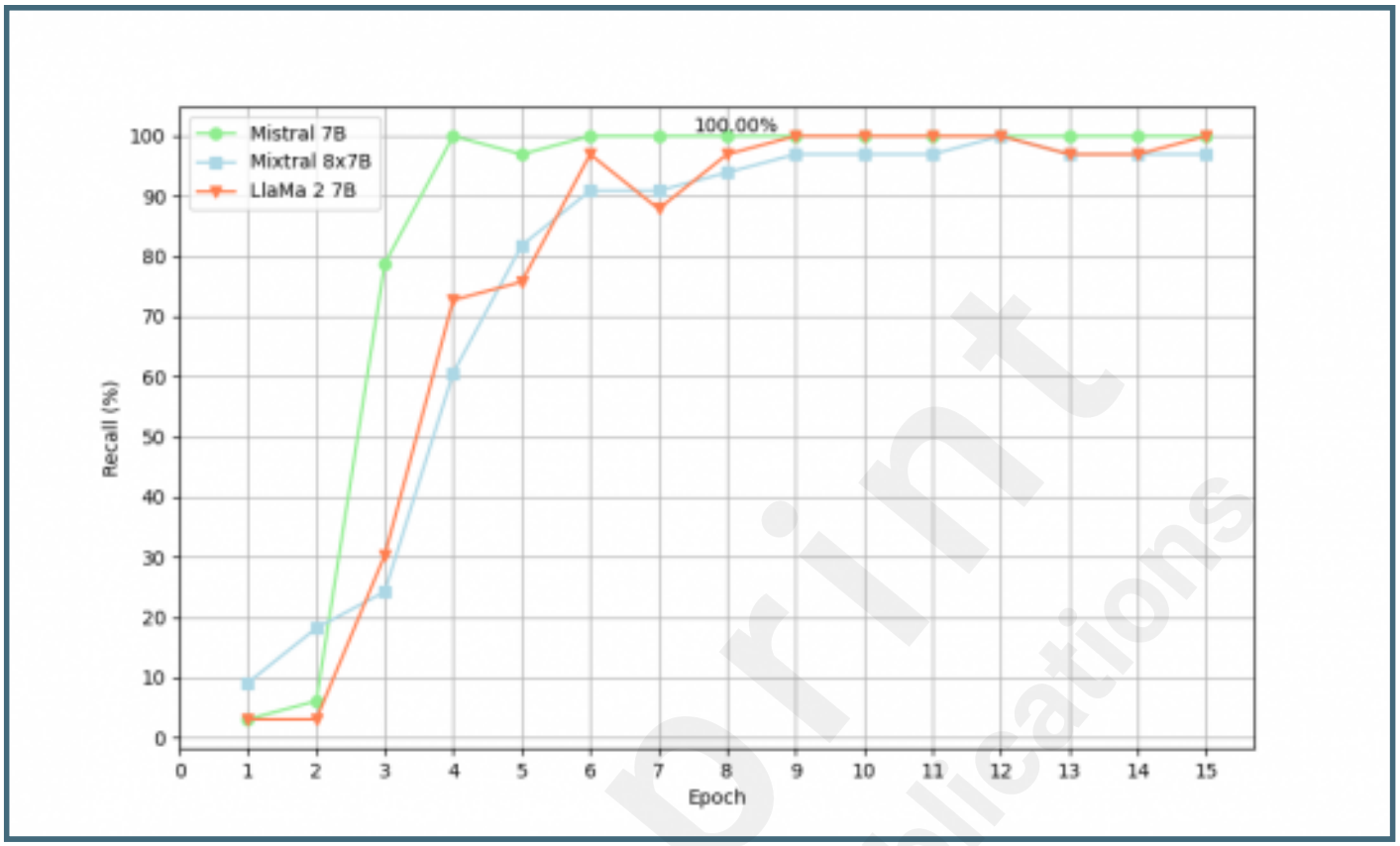
Plot of Recall by epoch : PII as statistical unit.



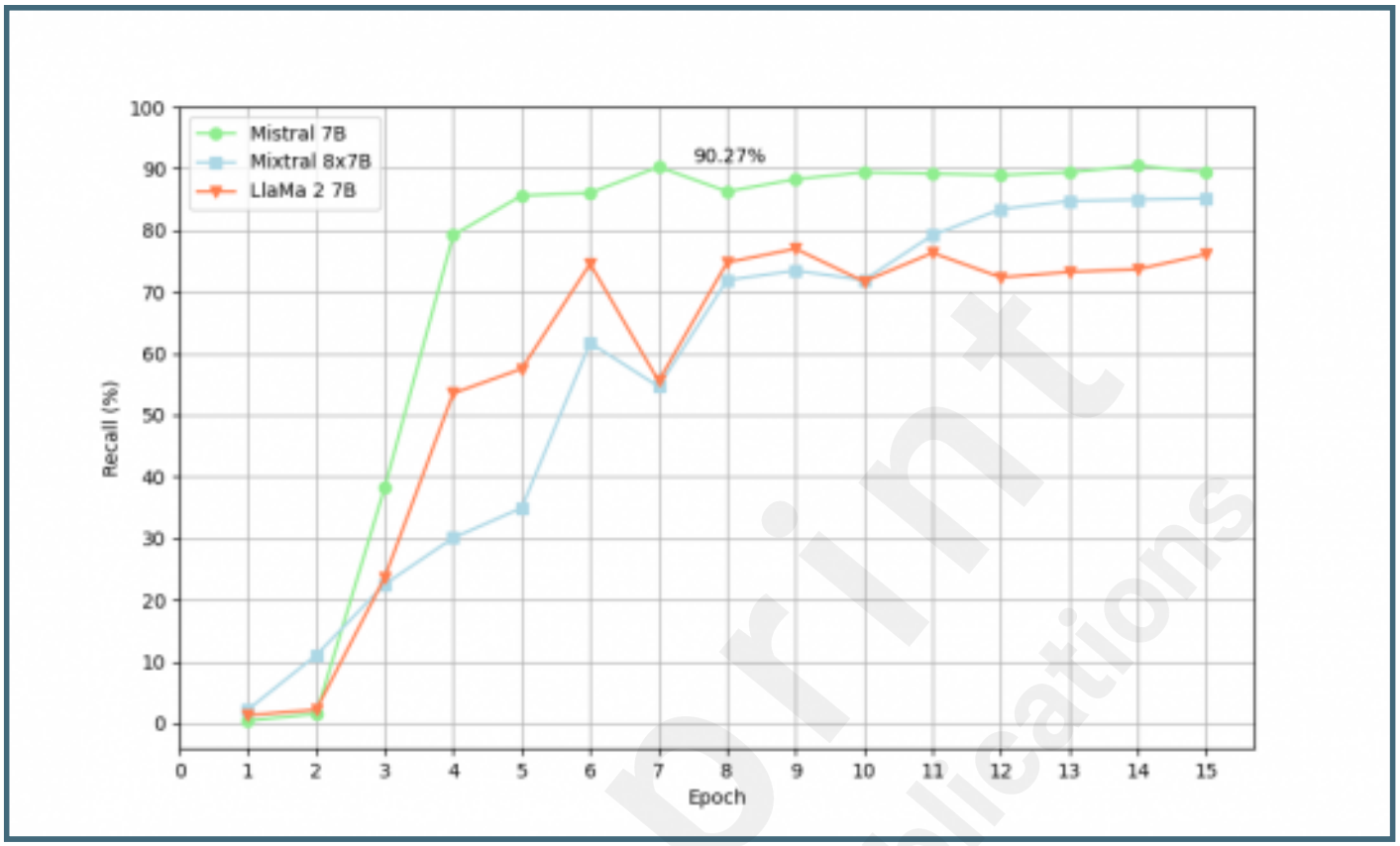
Plot of Recall by epoch : clinical note as statistical unit.



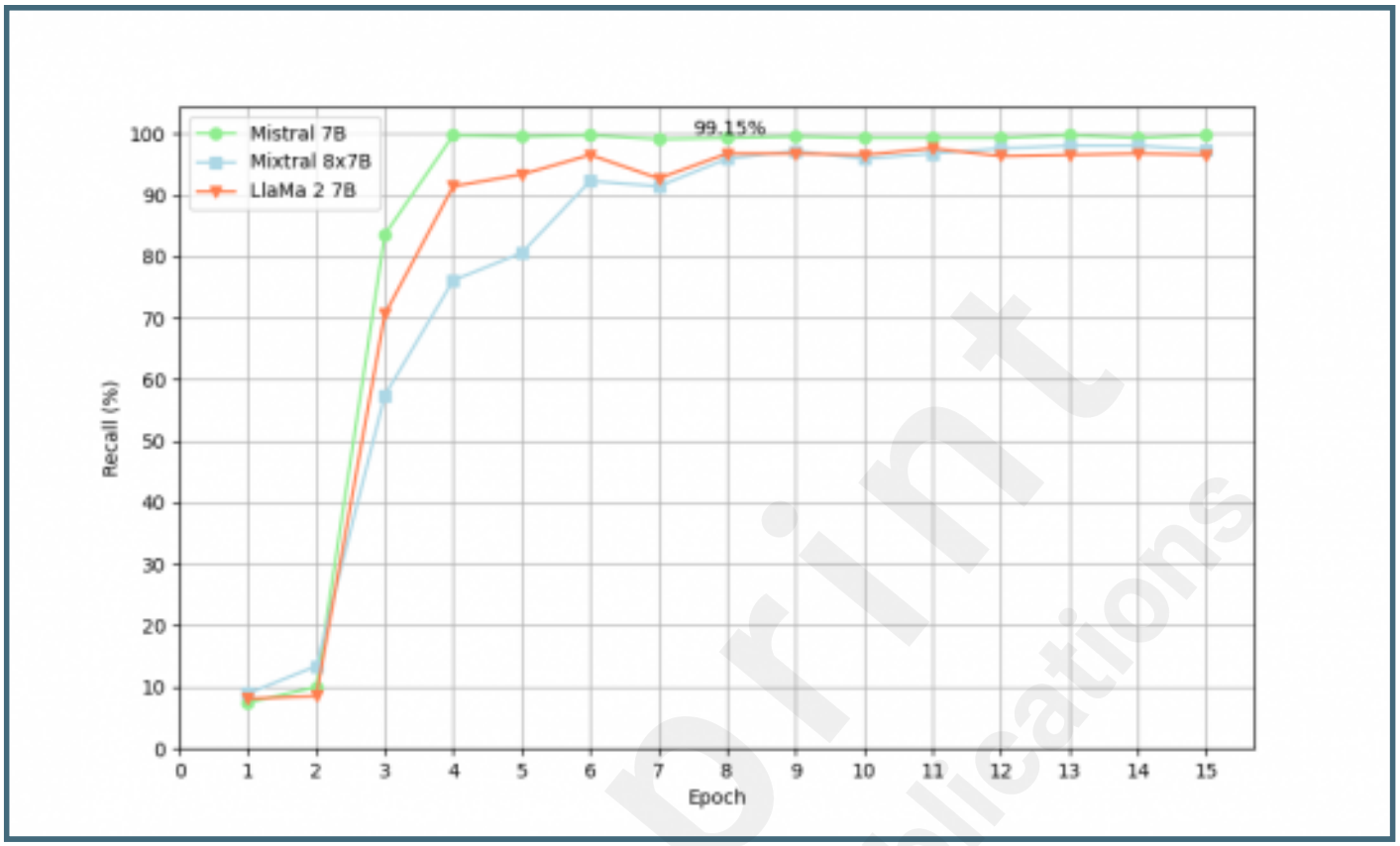
Plot of Recall by epoch for PII : TEL.



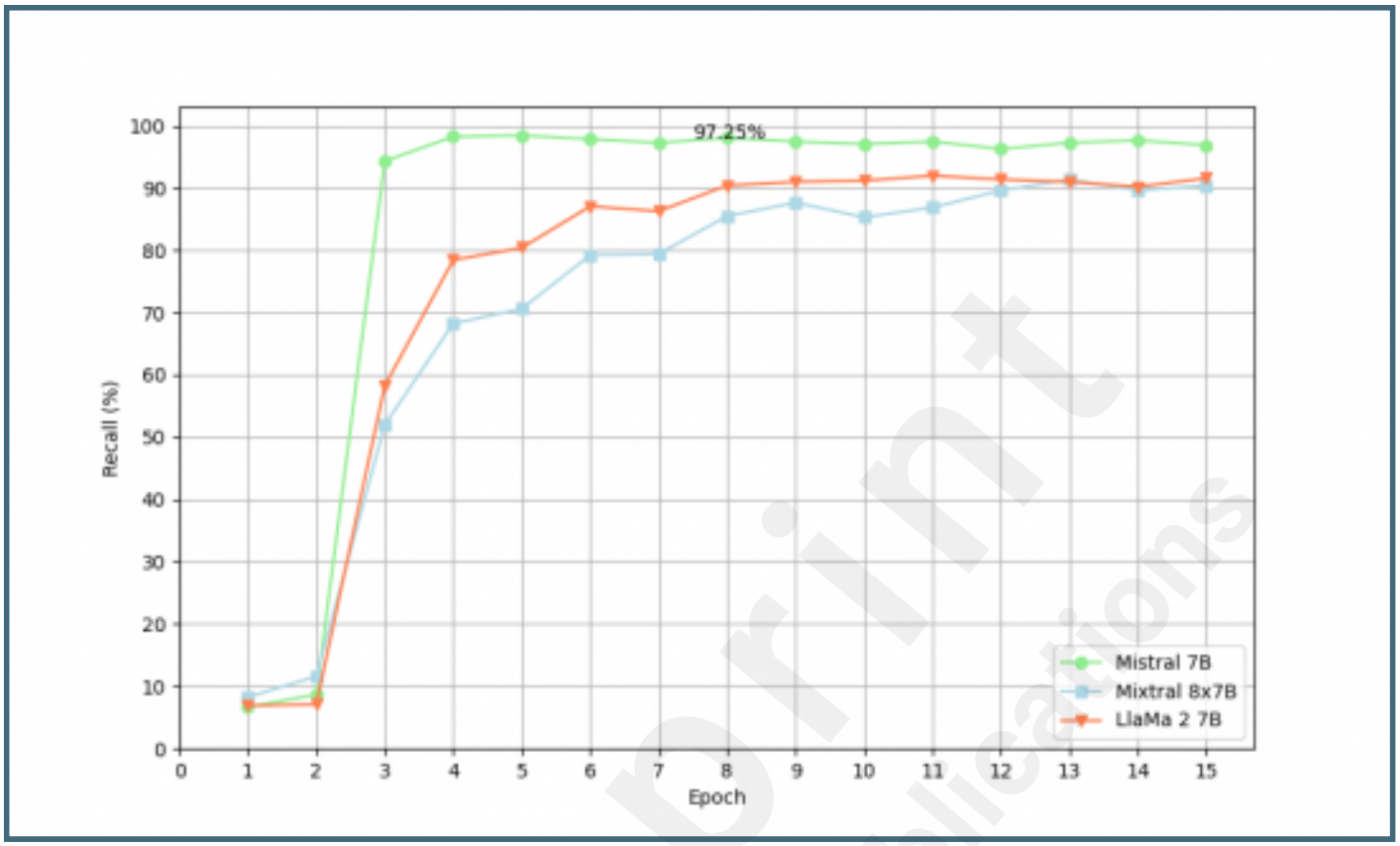
Plot of Recall by epoch for PII : LOC.



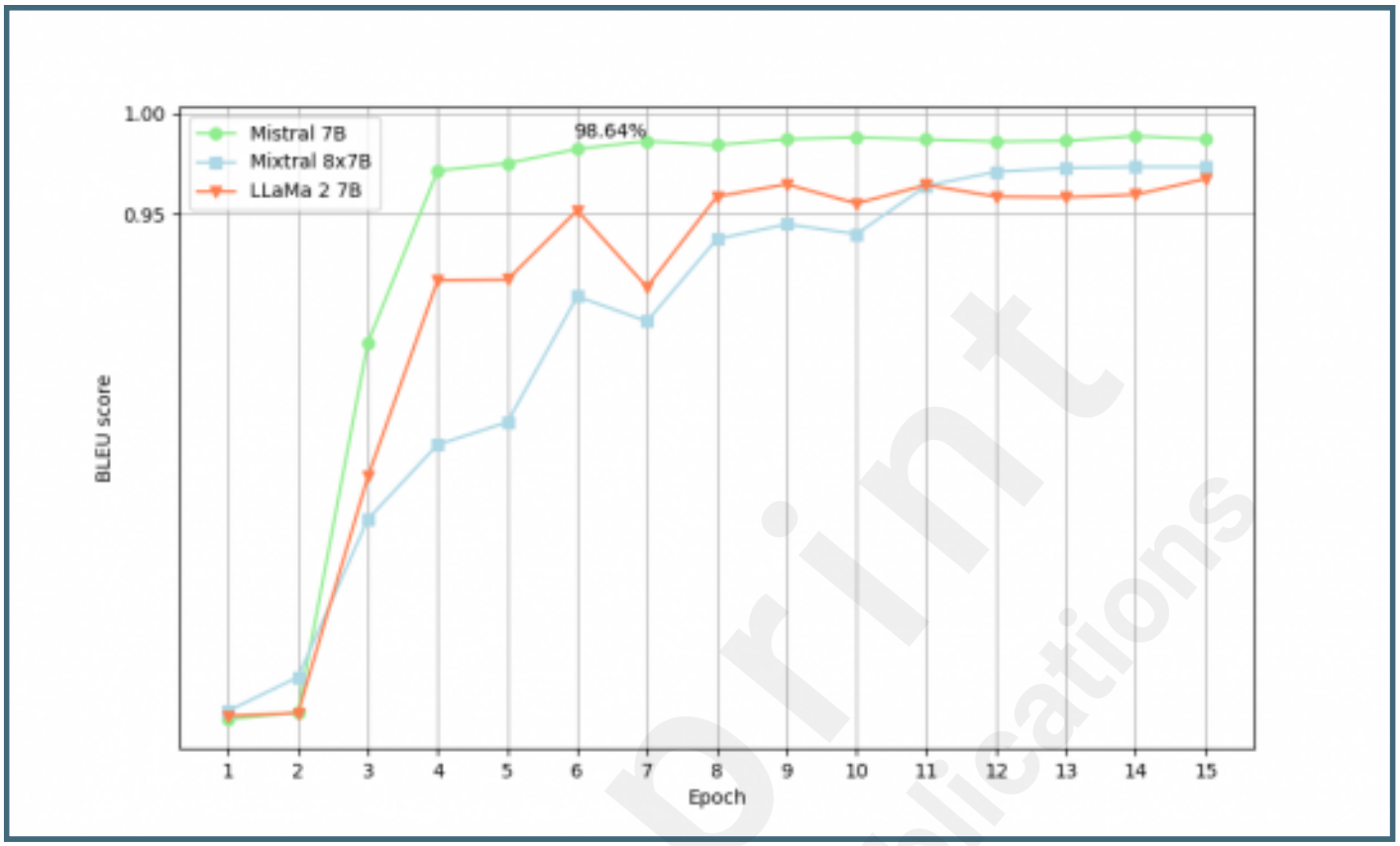
Plot of Recall by epoch for PII : NAME.



Plot of Recall by epoch for PII : DATE.



Plot of BLEU score by epoch : clinical note as statistical unit.



Multimedia Appendixes

Examples of french nursing notes.
URL: http://asset.jmir.pub/assets/1225fd668d368cc83054469f5a46c510.docx

Analysis of Performance Evaluation on corrected Test set.
URL: http://asset.jmir.pub/assets/a4ec12704de6f18e3bddbe4d96db13d0.docx

Comparative Table of Statistical Results from Previous Studies.
URL: http://asset.jmir.pub/assets/2e3d42df59d4331e6e4757123e85b2f1.docx

Comparative Table of Recall Across PII Categories from Previous Studies.
URL: http://asset.jmir.pub/assets/819f4d68be5f975eee31a56f5b6d2e4c.docx