Synthetic Documents for Medical Tasks: Bridging Privacy with Knowledge Injection and Reward Mechanism

Simon Meoni
Inria/Arkhn
Paris, France
simon.meoni@arkhn.com

Théo Ryffel
Arkhn
Paris, France
theo@arkhn.com

Éric de la Clergerie
Inria
Paris, France
Eric.De_La_Clergerie@inria.fr

Abstract

Electronic Health Records (EHR) store valuable patient-staff interaction data. Recent advancements in proprietary online large language models (LLMs) have shown promising capabilities in analyzing EHR notes. However, transmitting patient information through external APIs to LLMs like ChatGPT introduces privacy risks, necessitating alternative approaches that conform to hospital practices.

To address privacy concerns, we propose generating synthetic documents based on a reward-mechanism-trained model from real documents without leaking sensitive information but keeping relevant clinical knowledge. These synthetic documents may be annotated by large proprietary models or existing public ones, and used to train small specialized models that can run on constrained medical infrastructure. We validate our approach through a proof-of-concept scenario using Mimic-III, assessing the effectiveness of the generated documents through several downstream tasks: a series of ICD-9 multi-label classifications of varying complexity and a synthetic Named Entity Recognition (NER) task. The results demonstrate that synthetic documents preserve privacy and improve performance when real annotated data are sparse.

1 Introduction

Electronic Health Records (EHR) contain patient and healthcare staff interactions. Professionals record their impressions, observations, and various medical procedures performed. These notes remain fairly expressive and free to save healthcare personnel time and allow for the description of unusual situations (Rosenbloom et al., 2011; Wu et al., 2022). Natural Language Processing (NLP) techniques speed up the decision processes (Zhou et al., 2022; Wu et al., 2022). In recent years, Proprietary Online Large Language Models (LLMs) such as ChatGPT have shown impressive results using zero or few-shot techniques in analyzing these notes (Agrawal et al., 2022; Meoni et al., 2023; Hu et al., 2024). However, clinical NLP faces challenges that arise from the sensitive, confidential, and specialized nature of its data—sending such patient information through an external API raises numerous legal issues and is often impossible. Hospitals or third parties providing NLP-based medical devices (i.e., directly impacting patient care) must maintain control over their NLP systems to ensure patient safety. Therefore, the customization of open LLMs and their execution in a secure but computationally constrained environment is an important issue.

Still, specific training datasets are necessary to develop a model with clinical skills to address these challenges. To create such a dataset, obtaining real clinical data remains complicated and requires anonymization, which is time-consuming, expensive, and legally constrained. This also hinders the use of online models to annotate real data. Alternatively, we propose to create synthetic clinical notes that look like real data but do not include personally identifiable Information (PII) (Melamud and Shivade, 2019; Ive et al., 2020). This approach has several benefits: it reduces the need for human input, complies with regulations, and is suitable for annotation with external models to train local models. The local models and datasets can be shared with the community without leaking confidential information. These local models are also small enough to be hosted inside the hospital's infrastructure.

Considering these issues, we implement a novel method for generating synthetic documents, enforcing privacy preservation by design, using only a tiny seed set of pseudo-anonymised data. As a proof of concept, our key contributions include:

- **Privacy-safe Document Generation guided**

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 12–25
May 4, 2025

by Clinical Knowledge and Reward Mechanism: We present an methodology that leverages a minimal set of manually pseudo-anonymized data to train fine-tuned generative models. This process is enhanced by enriching prompts with keywords containing clinical knowledge, in our case extracted using QuickUMLS (Soldaini and Goharian, 2016), as illustrated in Section 5 and Figure 5. This extraction does not contain any PII in the sense that it contains only clinical entities (or keywords). Furthermore, we improve the quality of the synthetic documents thanks to an iterative refinement process that employs a private scorer to compare real and synthetic documents. This scorer returns only floats to the public side, ensuring privacy while enabling continuous improvement of the synthetic document quality.

- **Proof of Concept using Mimic-III:** Because it's almost impossible to evaluate our methods on real private documents, we utilize the Mimic-III clinical notes (Johnson et al., 2016) as a proxy to simulate a private healthcare environment, demonstrating our method's potential in a controlled setting. This proof of concept illustrates how our methodology could be applied in real-world hospital scenarios without compromising patient data.

- **Evaluation on downstream tasks using Mimic-III:** To assess the quality of the synthetic documents as training dataset for smaller models, we evaluate the generated data using two tasks: Multilabel Classification based on ICD-9 Codes (ICD-MC) and Synthetic Named Entity Recognition (NER). For ICD-MC, based on the codes proposed by Mullenbach et al. (2018) and Mimic-III manual annotations, we have modified this task, as described in Section 6.1, to compare the performance of the model trained with real data against the model trained with synthetic data. The NER task is conducted on annotations returned by GPT-4 on both our synthetic and real data. This allows us to compare the performance of models trained on these datasets.

2 Related Works

Synthetic Data Generation: Many recent studies focus on creating synthetic data, particularly for generating clinical data. For instance, Kweon et al. (2023) proposes to train LLMs for different purposes using synthetic clinical data generated by online LLMs. Xie et al. (2024) has developed AUG-PE, a high-quality differential privacy synthetic text generation method leveraging API access.

Furthermore, the work by Li et al. (2024) introduces Generalized Instruction Tuning (GLAN). Unlike previous approaches that rely on seed or existing datasets, GLAN uses a pre-curated taxonomy of human knowledge and capabilities as input to generate instructions across all disciplines. Inspired by their method, our work uses ontological information to extract sequences of ontology-based keywords from texts.

To assess the performance of LLM in Multiple Questions Choices in the medical field, Griot et al. (2024) developed a fictional medical benchmark to isolate the knowledge of the LLM from its test-taking abilities. Li et al. (2023a) generated a synthetic dataset of Alzheimer's Disease relative signs. As this task is relatively complex, LLM created the dataset by incorporating expert knowledge taxonomy. Finally, the Hiebel et al. (2023); Xie et al. (2024) works focus on generating a synthetic dataset of clinical cases for the NER task to study the effectiveness of real clinical data versus synthetic data.

Self-Rewarding: Reinforced Self-Training is an offline RL algorithm proposed by Gulcehre et al. (2023) for self-align LLMs generating a dataset from the initial LLM policy and using it to improve the policy via offline RL. Instruction back translation (Li et al., 2023b) is a scalable method that automatically labels human-written text with corresponding instructions by finetuning a LM on a small seed dataset and a web corpus to generate and selecting high-quality examples for further finetuning. Yuan et al. (2024) use the trained LLM to provide rewards via LLM-as-a-Judge prompting, improving both instruction following and reward provision. Lee et al. (2024) introduces Reinforcement Learning from AI Feedback (RLAIF) as an alternative, using an off-the-shelf LLM to generate preference labels. RLAIF achieves comparable or superior performance to RLHF in many tasks, such as those rated by humans.

The difference from the other approaches to generating a synthetic dataset is that our method combines LLM guided by prompts enriched with clinical knowledge, fine-tuned with a low amount of real pseudonymized data, and reinforcement learning feedback. This feedback is based on a score, which compares the real and synthetic data to ensure that they are closer to the source while maintaining privacy, as illustrated in Algorithm 1.

3 Reward-based Generation

We sketch the main steps of our reward-based generation process, illustrated with Algorithm 1.

3.1 Collecting keywords

The generation of synthetic *CRs* is guided by prompts enriched with clinical knowledge represented by non-confidential UMLS concepts (C) (Figure 6) extracted from real documents. Of course, other sources of keywords are possible. Therefore, our first processing step is to extract such keywords from each real document of dataset D_{source}, collecting them in C_{source}

3.2 Seed Step

We sample a tiny seed subset D_{sft} (i.e., supervised fine-tuning) from D_{source}, and associated keyword sequences C_{sft}, with a ratio of r%. This seed subset is assumed to be carefully pseudo-anonymized to authorize its use to finetune our initial public generator model M_{gen}. In our case, one or two hundred pseudo-anonymized documents suffice.

3.3 Generation Step

For each keyword sequence in $K_{\text{train}} = C_{\text{source}} \setminus C_{\text{sft}}$ and generation r, the generator model M_{gen} generates $N > 1$ candidate documents, collected in dataset D_{step}. This way, each synthetic document has a real counterpart based on the same sequence of keywords. In practice, we set $N = 4$.

3.4 Scoring Step

We evaluate the quality of the generated documents using SEMSCORE (Aynetdinov and Akbik, 2024), a metric based on semantic textual similarity (STS) returned by our private evaluator model M_{score}. The key point is that the M_{score} must be hosted in a private infrastructure to compare public synthetic documents with real private ones.

In Algorithm 1, we use a light orange background colour to indicate that this step takes place on the private side of the hospital building. However, being only composed of floats, the score set D_{score} can be safely declassified and returned from the private side to the public one for the Alignment step to train safely a new updated version of public M_{gen}. At the first generation step ($step = 0$), we initialize M_{score}, fine-tuning it with a contrastive objective, selecting a subset of D_0 to serve as negative examples and their real counterparts as positive examples.

Using M_{score}, we score the N candidates of each group from D^r_{step} against their counterparts in D_{train}. We keep only the best groups whose highest score is above the pth percentile. In practice, we set $p = 80$.

In each kept group, the candidate with the highest score (resp. lowest one) is selected as the *chosen* (resp. *rejected*) candidate. Finally, a dataset D_{dpo} is formed from these selected candidate pairs.

3.5 Alignment Step

Using dataset D_{dpo}, we align and update M_{gen} with *DPO* (Direct Preference Optimization) (Rafailov et al., 2023).

4 Applying Synthetic Dataset for Real Tasks

To validate the quality of the generated documents, we develop downstream tasks. In real life, the test set for such downstream tasks should be made up of real documents and manually annotated. The evaluations must be run in a private area.

5 Experiments

5.1 Base Models

We use `Mistral-7B-Instruct-v0.1` (Jiang et al., 2023) as our base generator model, a trade-off between performance and computational cost. As an evaluator model, we use `all-distilroberta-v1`.

5.2 Dataset

We use a dataset from Mimic-III as a proof of concept, involving pre-processing, keyword extraction, and post-processing.

1. *Pre-processing*: We extract from Mimic-III the clinical notes from the clinical event row. We select only the *Discharge Summaries* from these clinical notes and parse them to retrieve the *History of Patient Illness* section, using them as documents for D_{source}. On average, the documents consist of 248 words.

2. *Knowledge enrichment*: We project UMLS concepts using QuickUMLS over D_{source}. QuickUMLS is an unsupervised biomedical concept extraction based on pattern matching

Algorithm 1: Reward Training Algorithm

Input : D_{source} = initial dataset; r = sft ratio; M_{gen} = generative model; M_{score} = evaluator model; p = percentile filter value; N = number of candidates to generate;

Output : M_{gen}

```
// Running in Private Area, declassifying D_sft, C_sft
C_source ← ExtractConcepts(D_source)
D_sft, C_sft ← PseudoAnonymize(Sample(D_source, C_source, r))
D_train, K_train ← D_source \ D_sft, C_source \ C_sft
// Seed Step
M_gen ← Supervised fine-tune M_gen on pairs in (C_sft, D_sft)
for step = 0 to steps do
    // Generation Step
    D_step ← generate new N candidates with M_gen per k ∈ K_train
    // Scoring Step (Running in Private Area, declassifying D_score)
    if step = 0 then
        // Building the evaluator model
        D*_contr, D_contr ← Sample(D_0, D_train, r_contr)
        M_score ← ContrastiveTrain (M_score, neg = D*_contr, pos = D_contr)
    D_score ← score D_step over D_train with M_score
    D_dpo ← in D_score, keep a pair of candidates, then filter pairs on percentile p
    K_dpo ← filter K_train to keep keywords corresponding to candidates selected in D_dpo
    // Alignment Step
    M_gen ← DPO Alignment M_gen on (K_dpo, D_dpo)
```

that guarantees only medical concepts are extracted and no identifying information. We obtain C_{source} (cf. Section 3) used to enrich the prompts, as illustrated in Figure 6. On average, we extract 58 keywords per document.

3. *Post-processing*: We filter out documents without keywords. We keep ordered keywords to encourage the model to follow the same narrative as the ground truth. In this way, we constitute a dataset of 4262 documents, using 70% of them (2581) as a train set (D_{train}) and 30% (1680) as a test set (D_{test}). Moreover, the D_{sft} with 4% and 6% ratios have 156 and 235 documents, respectively.

6 Evaluation on Downstream Tasks

6.1 Multilabel Classification tasks

Collecting Gold Annotations: As Mimic-III includes a set of expert-labeled ICD-9 codes (L) for each discharge summary, we use these annotations (1) to evaluate the quality of our datasets on tasks close to a real use-case (2) and test across a series of ICD-MC tasks with increasing complexity. We establish an association between these labels and the data points in D_{train} and D_{test}, respectively, 2581 and 1681 data points.

We get annotated datasets ($D_{\text{train}}, L_{\text{train}}$) and ($D_{\text{test}}, L_{\text{test}}$) by coupling documents with labels. In defining our series of ICD-MC tasks, we prioritize the most frequent k labels, denoted as class-k (see Table 1) with $k \in \{20, 50, 100, 400\}$. We subsequently refine ($D_{\text{train}}, L_{\text{train}}$) and ($D_{\text{test}}, L_{\text{test}}$) by retaining only those documents whose labels intersect with the set of **class-k** labels.

We define the refined training set as $D_{\text{gold}} = (D'_{\text{train}}, L'_{\text{train}})$ where each document in D'_{train} contains at least one label from **class-k**. Documents devoid of any intersecting labels are excluded. Table 1 presents the dataset sizes, which document the number of excerpts retained after applying these exclusion criteria.

It should be noted that the task's complexity increases with k not only because of the larger set of labels and the lower frequency of some labels but also because of the longer label set on average per document. For instance, the average length is around 6 when $k = 20$ but 11 when $k = 100$.

Constituting the Synthetic Train Datasets: As an approximation, we hypothesize that the synthetic data point from D^r_{step}, which shares the same set of UMLS keywords as its real data counterpart, can inherit the same set of ICD labels L'_{train}. This way, we easily obtain six synthetic datasets, denoted as D_{step}, corresponding to the generation steps $step \in \{0, 1, 2\}$ and seed ratios $r \in \{4\%, 6\%\}$, as shown in Table 3. Each D_{step} dataset contains four times more document data points than D_{gold}.

6.2 Named Entity Recognition (NER) Task

Annotating the Overall Dataset: Because Mimic-III does not include gold NER annotations, we use GPT-4 to automatically annotate all (synthetic and real) train and test datasets (OpenAI (2023), Appendix B.), focusing on three entity types: **problem**, **treatment** and **test**. We employ a few-shot learning approach inspired by Hu et al. (2024), using the prompt in Appendix 10. To assess whether or not the annotated entities are essentially the UMLS keywords, we evaluated the overlap between keywords and annotations and found a low 22.36% overlap.

Table 1 illustrates the distributions of labels for the ICD-MC tasks and entities for NER.

6.3 Training of Task Models

We train a series of (small) `deberta-v3-base` (He et al., 2021) models on ICD-MC tasks using either real or synthetic datasets D_{gold} or D^r_{step} over the four tasks **class-k** where k $\in \{20, 50, 100, 400\}$.

To address the quantity bias of a larger synthetic dataset, we train two baseline models, one trained with D_{gold}, and another one trained with $D_{\text{gold}\times 4}$, where each real document is oversampled $N = 4$ times, hence containing the same amount of documents as the synthetic set.

We also consider a *baseline* where only keywords (K_{train}) are used to predict labels to check that the content of the documents impacts the performance, as shown in Table 3.

We apply the same methodology for the NER task but with only D_{gold} and $D_{\text{gold}\times 4}$ as baselines.

7 Results

Table 2 presents a comparative analysis of SemScore measurements by evaluators across the different datasets generated at various steps. We observe a consistent improvement in scores with successive steps. The $M^{6\%}_{\text{gen}}$ model outperforms the $M^{4\%}_{\text{gen}}$ model, highlighting the effectiveness of alignment in refining the quality of generated documents through iterative processes. The scores indicate a trend across various models, suggesting that models trained with more real data produce higher-quality documents.

[1] O (Outside) comes from the IOB (Inside-Outside-Beginning) schema used in Named Entity Recognition task. It denotes tokens that are not part of any named entity.

	D_{gold}		D_{test}	
class-k	**# labels**	**# docs**	**# labels**	**# docs**
class-400	38602	2564	25409	1681
class-100	30015	2560	19700	1672
class-50	23323	2552	15246	1672
class-20	14619	2513	9694	1648
ner	72715	2581	47783	1681

Table 1: Multilabel classification & NER task datasets, with labels size for D_{gold}, D_{test}. The number of labels for the NER task excludes label O[1].

	steps	$M^{4\%}_{\text{score}}$	$M^{6\%}_{\text{score}}$
$M^{4\%}_{\text{gen}}$	0	67.95	65.94
	1	71.53	69.18
	2	72.25	70.12
$M^{6\%}_{\text{gen}}$	0	70.78	67.26
	1	72.54	70.78
	2	<u>76.10</u>	**74.37**

Table 2: SemScore evaluation for models M^a_{gen} with $a = r_{sft} \in \{4\%, 6\%\}$ using the different evaluators M^b_{score} with $b = r_{sft} \in \{4\%, 6\%\}$. The grey scores denote cross-evaluation where $a \neq b$.

Table 3 compares F1 scores on the downstream tasks across different models and configurations, providing insights about their performance when varying task complexities and training data conditions. Notably, $M_{\text{gold}\times 4}$,trained with $D_{\text{gold}\times 4}$, outperforms the models trained with synthetic data ($M^{\{4,6\}\%}_{0,1,2}$) across all tasks. Second generation models($D^{4\%}_2$ and $D^{6\%}_2$) demonstrate performance comparable to the model trained on $D_{\text{gold}\times 4}$. In particular, for the **class-400** task, the F1 scores for $D^{4\%}_2$ and $D^{6\%}_2$ match closely those for $D_{\text{gold}\times 4}$, with only minor variations. Notably, the standard deviations for the synthetic data models are lower than those of the gold data model, indicating more consistent performance. Further-

	class-20	class-50	class-100	class-400	ner
baseline	45.7 ± 1.2	33.8 ± 2.2	26.6 ± 0.8	10.6 ± 2.0	-
D_{gold}	49.3 ± 1.8	33.3 ± 3.1	23.0 ± 3.6	04.9 ± 4.1	57.0 ± 0.2
$D_{gold \times 4}$	$\mathbf{53.7 \pm 2.3}$	42.5 ± 0.2	35.0 ± 1.3	26.4 ± 5.9	61.6 ± 0.1
$D_0^{4\%}$	49.8 ± 1.1	38.7 ± 1.1	32.2 ± 1.8	24.2 ± 2.5	-
$D_0^{6\%}$	49.9 ± 1.2	38.5 ± 1.9	31.0 ± 1.7	23.9 ± 2.4	59.6 ± 0.2
$D_1^{4\%}$	50.9 ± 0.9	41.1 ± 1.6	33.9 ± 1.8	26.9 ± 1.4	-
$D_1^{6\%}$	51.2 ± 0.9	40.7 ± 1.4	33.7 ± 2.1	24.5 ± 2.7	59.4 ± 0.2
$D_2^{4\%}$	50.6 ± 0.8	41.0 ± 1.3	34.3 ± 2.0	27.0 ± 2.0	-
$D_2^{6\%}$	51.7 ± 1.1	40.7 ± 1.0	31.9 ± 7.5	26.5 ± 2.5	59.4 ± 0.2
$D_{\{0,1,2\}}^{6\%}$	52.4 ± 0.4	$\mathbf{43.1 \pm 0.5}$	$\mathbf{37.2 \pm 0.3}$	$\mathbf{31.0 \pm 0.7}$	$\mathbf{61.7 \pm 0.1}$

Table 3: Comparative F1 Scores and standard deviation across models trained over different dataset generations. The table illustrates F1 (Micro-F1) score performance for the **class-k** and NER tasks across D_{step}^{r}, D_{gold} and the *baseline*.

more, models trained on a combination of several generations($D_{0,1,2}^{6\%}$) outperform most cases, except on the **class-20** task. This suggests increasing data diversity and quantity through dataset mixing enhances model performance in certain scenarios. Consistently across **class-k** tasks, $M_0^{\{4,6\}\%}$ models yield the lowest F1 scores. This indicates that initial generation models lack sufficient sophistication or diversity in training data to effectively capture necessary predictive features, particularly for $M_0^{4\%}$. As task complexity increases, F1 scores generally decrease for both real-based and synthetic-base models, highlighting the models' challenges in adapting to more complex interactions.

In the **class-400** task, F1 scores improve from $step = 1$ to $step = 2$, following a general trend of performance increase. The exception is in the **class-100** task, where performance decreases between $M_1^{6\%}$ and $M_2^{6\%}$.

Figure 1 presents the correlation between F1 scores and SEMSCORE computed by $M_{score}^{6\%}$ across **class-k** tasks. We observe that SEMSCORE is an effective evaluator, although with nuances. Specifically, $D_2^{6\%}$ outperforms $D_2^{4\%}$ only in **class-20**. In **class-400**, the lowest correlation is observed, suggesting that SEMSCORE 's reliability decreases as task complexity increases, likely due to label scarcity affecting training stability. In contrast, **class-20, 50, 100** show stronger correlations, emphasizing SEMSCORE effectiveness in these tasks. Though, $M_{0,1,2}^{4\%}$ consistently outperforms $M_{0,1,2}^{6\%}$, indicating that the seed may constrain the generator, leading to reduced document diversity. Further investigation is required to evaluate the impact of r on overall performance.

We also conducted ablation studies to analyze how dataset sizes and selection strategies affect the performance of encoder models for the **class-100** and NER tasks. We trained several task models using different amounts of (filtered) synthetic data generated from the $D_2^{\{4,6\}\%}$ subsets. We employed two filtering methodologies: (1) **percentile sampling**, which prioritizes the highest-scored candidates according to the SEMSCORE metric, and (2) **random sampling**, which filters documents in varying proportions.

In Figure 2, the graphs demonstrate a consistent increase in F1 scores when expanding the synthetic document set from 2,000 to 10,000 documents for both sampling methods. For **class-100**, percentile sampling shows a more pronounced improvement than random sampling, particularly at lower document counts. As the document set grows, the performance gap between the two sampling methods narrows, but percentile sampling maintains a slight edge throughout. This trend suggests that the quality of synthetic documents, measured by SEMSCORE, significantly impacts performance for this task, especially when working with smaller datasets. The observation underscores the importance of quantity and quality in synthetic data generation, with quality playing a crucial role in scenarios where data quantity is limited.

On the other hand, there is a sharp decrease in the

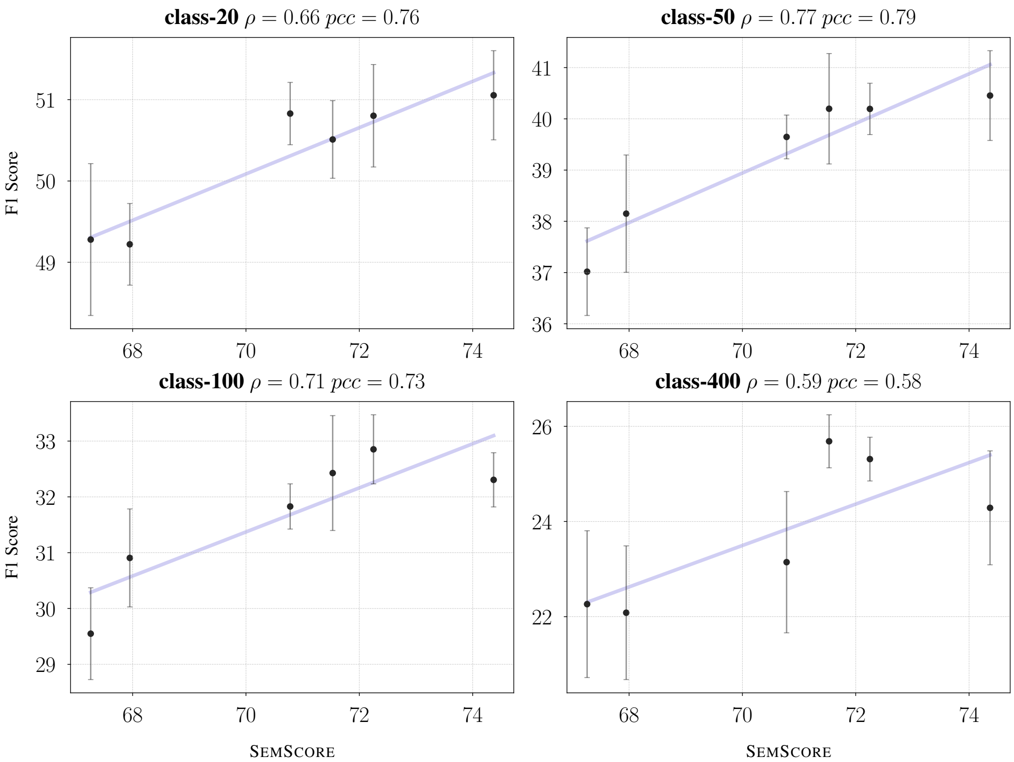

Figure 1: Correlation between SemScore and F1-score across **class-{100,400}** prediction tasks. The dots represent the model trained with D^{r}_{step}. The Spearman correlation (ρ) and Pearson correlation coefficient (pcc) indicate varying degrees of linear and rank-order association with task complexity.

performance of the NER task when $M_2^{6\%}$ is trained with the same number of documents as M_{gold} using percentile sampling. We conjecture it is partly due to the synthetic subset containing fewer annotated tokens than the gold dataset (for the same number of documents), with 510199 tokens versus 643802 tokens. To neutralize the impact of this difference, we trained a model with the same amount of annotated tokens as D_{gold}, as illustrated by a black star in Figure 2. We observe less difference between M_{gold} and $M_2^{6\%}$ (with values of 57.0 and 56.6). We hypothesize that this difference is because the distribution of D_{gold} is closer to that of the synthetic subset compared to D_{test} as illustrated in Figure 4. Furthermore, adding or removing words can affect the proportion of annotated tokens. We have not yet conducted the NER task experiment with the document generated by $M_{0,1,2}^{4\%}$ as we do not anticipate significant results for these tasks.

8 Discussion

Besides validating our privacy-safe generation process, our results have also provided crucial insights into the impact of both the quality and quantity of synthetic training data on the performance of encoder models. **It is evident that refining the generator through DPO, using clinical concepts as inputs, enhances the synthetic dataset's quality, especially when the first alignment step has been performed**. Results indicate that training models on synthetic data not only preserves but outperforms models trained on gold datasets, as illustrated in Table 3. This highlights the potential of using privacy-preserving synthetic documents to maintain high data utility while protecting sensitive information.

The accuracy of the SemScore scoring mechanism as a predictor of data quality for downstream tasks is also particularly pronounced. The nature of tasks significantly influences the predictive quality, as shown in Figure 1. The need for text closely aligned with the source material to ensure accurate identification of rarer labels was clear, highlighting SemScore's role as a critical metric in evaluating and refining the quality of synthetic documents.

While increasing the dataset size improves performance, applying selective filtering strategies, such as percentile sampling, on a larger volume further enhances results, surpassing the model trained

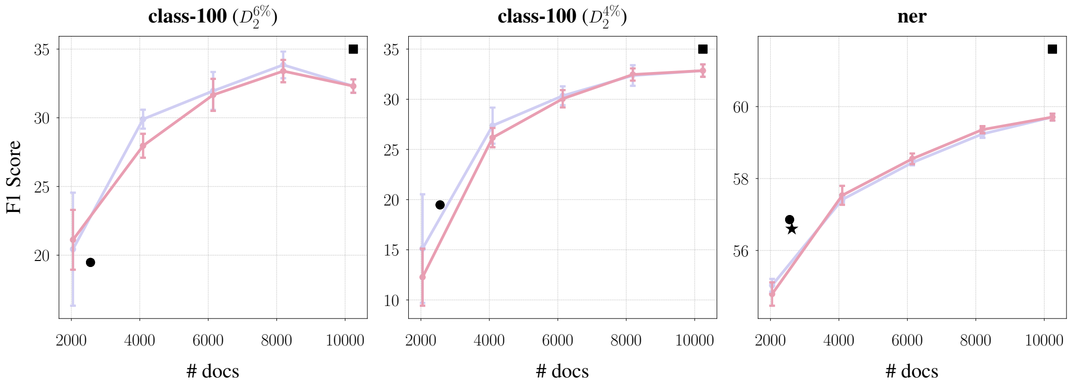

Figure 2: The figure showcases the experimental settings for training encoder models with varying quantities of synthetic data. The **pink line** (resp. **blue line**) denotes models trained on randomly sampled datasets (resp. nth-best based on SEMSCORE datasets). The black dot represents the model trained with D_{gold}, while the black square represents the model trained with $D_{\text{gold}\times 4}$.

with D_{gold}. These findings suggest that both data quantity and quality can be adjusted to optimize outcomes, as highlighted in Figure 2.

Another interesting finding is that we can concatenate the datasets generated on the different *steps* to increase performance. This is illustrated in overall tasks, where diversity is improved by using more data and simulating a more diverse dataset through the heterogeneous data quality, outperforming the model trained with $D_{\text{gold}\times 4}$.

9 Conclusion

We deliver a method for generating synthetic privacy-safe documents. Our method consists of (1) initializing the model with a small number of pseudo-anonymized documents, which reduces the need for human input, and (2) employing a private evaluator to score the generated document against real documents, preserving the confidentiality of the data while ensuring proximity between real and synthetic documents. Our study shows that models trained on small gold datasets face the practical limitations of current NLP systems when handling complex tasks. Scaling the amount of high-quality and diverse synthetic documents is a way to address these limitations. It can outperform models trained on real data under certain configurations, thereby validating the approach of generating on-demand data to overcome data scarcity and privacy issues. These findings facilitate the sharing of high-fidelity synthetic datasets. Furthermore, such datasets may be then annotated using (proprietary) LLMs or via large-scale manual annotation. Finally, the proposed solution is more ethical for patients. It focuses on privacy concerns and is motivated by the opening of clinical data for research advancements.

10 Limitations

Currently, evaluation is limited to multi-label classification and NER tasks. Expanding testing to more complex tasks that require reasoning and domain-specific knowledge, such as medical question answering, could give more insights into the applicability and robustness of our method.

By design, Personal Identifiable Information are absent from our synthetic documents but there exist some slight risks of re-identification from some specific sequences of UMLS keywords. Adding some noise to such sequences should solve the issue.

The economical cost for generating large synthetic datasets may also be an issue (see Appendix A.) for some healthcare providers, even if it occurs in public environments. Investigating the efficacy of smaller generation models could make this technology more accessible, especially for hospitals or clinics with limited budgets.

We are exploring alternative reinforcement learning techniques, such as KTO (Ethayarajh, 2024), to exploit all the generated data rather than only selecting example pairs filtered by percentile with DPO. We are also considering simpler RL methods like ORPO (Hong et al., 2024) and SimPO (Meng et al., 2024).

Finally, we wish to investigate more accurate evaluation metrics than just SemScore, in particular, by combining them with other sophisticated metrics, such as style transfer or document quality (Jin et al., 2022).

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Few-Shot Clinical Information Extractors.

Ansar Aynetdinov and Alan Akbik. 2024. SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity. *arXiv preprint*. ArXiv:2401.17072 [cs].

Maxime Griot, Jean Vanderdonckt, Demet Yuksel, and Coralie Hemptinne. 2024. Multiple Choice Questions and Large Languages Models: A Case Study with Fictional Medical Data. *arXiv preprint*. ArXiv:2406.02394 [cs].

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced Self-Training (ReST) for Language Modeling. *arXiv preprint*. ArXiv:2308.08998 [cs].

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

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

Jiwoo Hong, Noah Lee, and James Thorne. 2024. {ORPO}: Monolithic Preference Optimization without Reference Model. volume Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11170–11189.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for Natural Language Processing. *npj Digital Medicine*, 3(1):1–9. Publisher: Nature Publishing Group.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*. ArXiv:2310.06825 [cs].

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205. Place: Cambridge, MA Publisher: MIT Press.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035. Publisher: Nature Publishing Group.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2023. Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes. *arXiv preprint*. ArXiv:2309.00237 [cs].

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green Algorithms: Quantifying the Carbon Footprint of Computation. *Advanced Science*, 8(12):2100707.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26874–26901. PMLR.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. 2024. Synthetic Data (Almost) from Scratch: Generalized Instruction Tuning for Language Models. *arXiv preprint*. ArXiv:2402.13064 [cs].

Rumeng Li, Xun Wang, and Hong Yu. 2023a. Two Directions for Clinical Data Generation with Large Language Models: Data-to-Label and Label-to-Data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7129–7143, Singapore. Association for Computational Linguistics.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023b. Self-Alignment with Instruction Backtranslation. *arXiv preprint*. ArXiv:2308.06259 [cs].

Oren Melamud and Chaitanya Shivade. 2019. Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large Language Models as Instructors: A Study on Multilingual Clinical Entity Extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, Toronto, Canada. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. *arXiv preprint*. ArXiv:1802.05695 [cs, stat].

OpenAI. 2023. GPT-4: Generative Pre-trained Transformer 4.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

S. Trent Rosenbloom, Joshua C. Denny, Hua Xu, Nancy Lorenzi, William W. Stead, and Kevin B. Johnson. 2011. Data from clinical notes: A perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186.

Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: a fast, unsupervised approach for medical concept extraction. *MedIR workshop, sigir*, pages 1–4.

Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael T. C. Poon, Natalie Fitzpatrick, Adam P. Levine, Luke T. Slater, Alex Handy, Andreas Karwath, Georgios V. Gkoutos, Claude Chelala, Anoop Dinesh Shah, Robert Stewart, Nigel Collier, Beatrice Alex, William Whiteley, Cathie Sudlow, Angus Roberts, and Richard J. B. Dobson. 2022. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *npj Digital Medicine*, 5(1):1–15. Publisher: Nature Publishing Group.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. 2024. Differentially Private Synthetic Data via Foundation Model APIs 2: Text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54531–54560. PMLR.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-Rewarding Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57905–57923. PMLR.

Nina Zhou, Qiucheng Wu, Zewen Wu, Simeone Marino, and Ivo D. Dinov. 2022. DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes. *Journal of Medical Systems*, 46(12):96.

Appendix A. Carbon Footprint

The algorithm 1 with the experimentation protocol detailed in 5 runs in 40h on 4 GPUs NVIDIA A100 PCIe, and draws 46.34 kWh. Based in France, this has a carbon footprint of 2.38 kg CO2e, which is equivalent to 2.59 tree-months (calculated using green-algorithms.org v2.2 (Lannelongue et al., 2021)).

Appendix B. Mimic-III Usage

PhysioNet provides MIMIC-III under the PhysioNet Credentialed Health Data License 1.5.0. We employed Azure GPT, a HIPAA-certified platform for our large language model annotations. Importantly, Azure GPT does not store processed data. We received approval to waive abuse monitoring for Azure GPT, ensuring that no additional parties could access the data or prompts used in our experiments.

Appendix C. ICD-9 label distribution

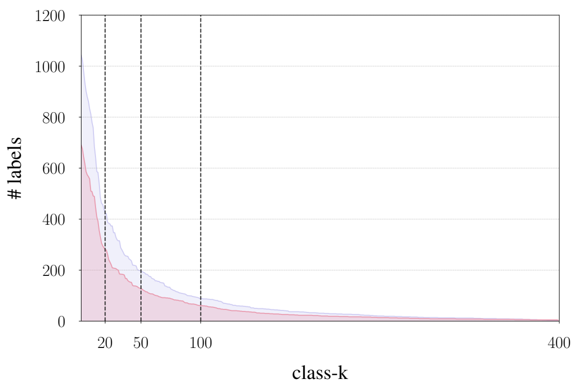

Figure 3: Rank-frequency distribution of labels within the D_{gold} (blue) and D_{test} (pink) datasets, plotted over the **class-k**.

Appendix D. Name Entities Proportion

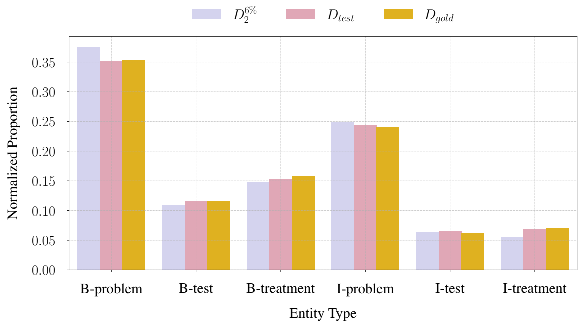

Figure 4: Normalized proportion of entity type in D_{gold}, D_{test} and $D_2^{6\%}$.

Appendix E. Workflow

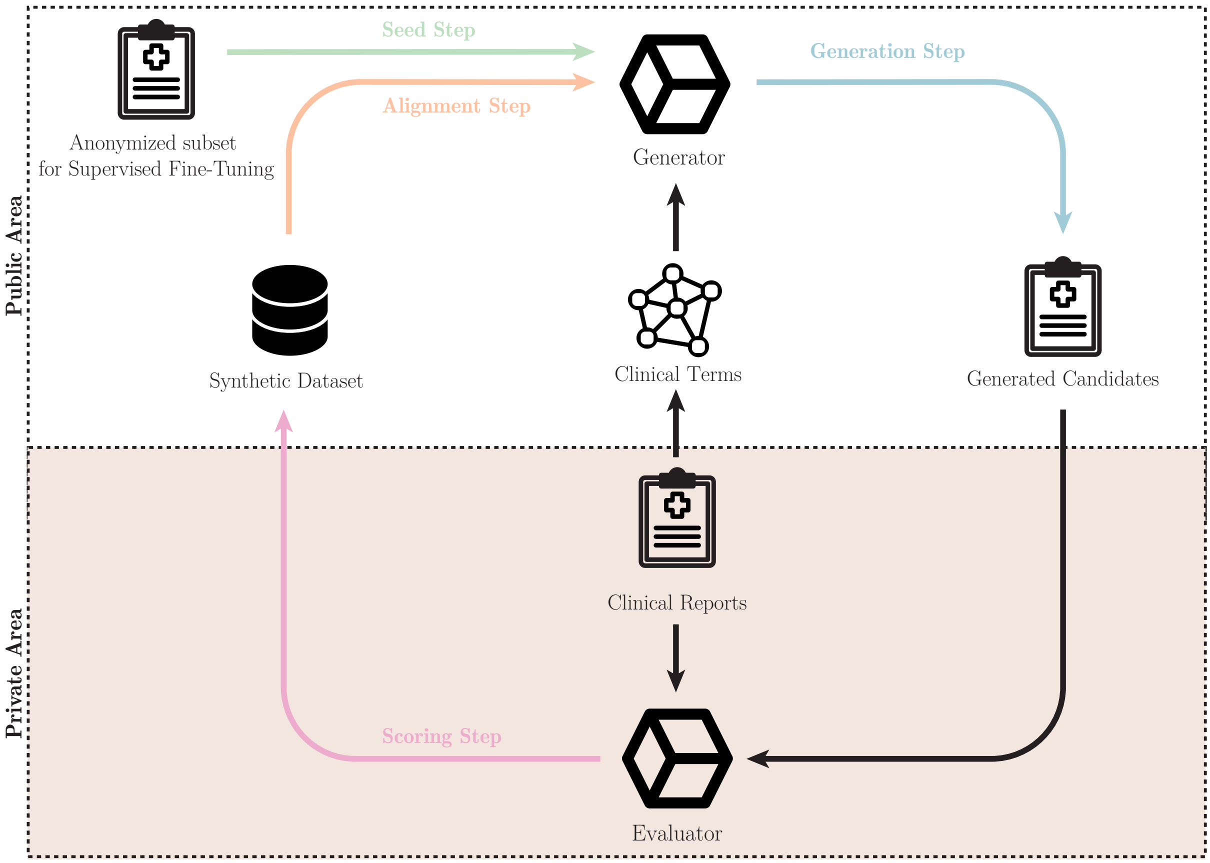

Figure 5: workflow of our approach

Appendix F. Outputs

<s>[INST]As a doctor, you must write an original 'History of Present Illness' (HPI) section for a discharge summary. Your response should capture the essence of a patient's health journey and recent medical experiences, while strictly using all the provided keywords conserving the order. You must adopt a medical telegraphic style, abbreviated, characterized by concise and direct language.
Keywords: *metastatic, RCC, pancreas, reports, chills, tylenol, reports, rib pain, lying, chills, reports, dark stools, fever, zosyn, headache, contacts, anxious, pain, dysuria, joint pain, rash, hypotensive, asymptomatic, given, lactate, baseline, guaiac, stool, saw, stent, pancreatic, blood, tomorrow, treated, cholangitis, given, sat, ARF, reports, anxious*[/INST]

Figure 6: An example of prompt for the Figures 7,9,8

This is a 67 y.o male with h.o metastatic RCC to the pancreas, recent ICU course for UGIB (12units pRBCs) who reports sudden intermittent chills since wednesday for which he took tylenol. Pt also reports R.side gnawing rib pain, while lying in bed before the onset of chills. In addition, pt reports dark stools for the last few days which started after taking "iron pills". Pt states he went to [**Hospital1 2436**] ED because of a fever of 101.3, however he felt better and did not want to wait to be seen. He returned to [**Hospital1 2436**] today and was transferred to [**Hospital1 18**] after a dose of zosyn, HCT 25. Pt denies headache/dizziness/blurred vision, URI/cough, sick contacts, CP, +palp when anxious, -abd pain/n/v/d/brbpr, dysuria/hematuria, joint pain, rash, paresthesias.

.

At [**Hospital1 18**], pt found to be hypotensive to 75/40, asymptomatic. He was given 3L IVF, lactate 6.8. HCT 22.8 from a baseline of 35 a few weeks ago. He was found to have black, guaiac +stool. GI saw pt, pt s/p stent to pancreatic ampulla, ?blocked from blood. Plan is to transfuse, ERCP tomorrow. ED also treated for possible cholangitis/sepsis and pt was given dose of vanco. Vitals 99.2, BP 99/66 HR 88 sat 9820's, 18. Pt also found to be in ARF.

.

Currently, pt reports that he is anxious.

.

Figure 7: An example of a ground truth associated with the prompt of Figure 6

77 yo M w/ hx of metastatic RCC of the pancreas (pancreo-duodenal
mass reports R/o chills since starting [**6-2**] on [**2160-5-18**] (day 7-10 from Hx. He has had fevers with taking tylenol at that time
(Ibuprofen 400mg, acetaminophen 650mg PRN for R/o rib pain and
steal). On [**2160-5-22**] he was lying in bed and became extremely
ill with chills. He reports this started a few minutes ago. Also
reports dark stools. His fever is 102F (38.3C), he is on 2g IV
zosyn (given IVF 2L NS, as he is V/5) for pneumonia with HC
reported by the Wife. He reports severe headache. He is not
compliant with PCP but has noticed many recent contacts with
his S/M. He is very anxious for relief of pain. He has some
dysuria and no joint pain. He has no rash. He is in the ED
actively pacing. He is hypotensive. He is BP 90/45 with 3+ pti
on his left hand. PRI all his vitals are stable otherwise at
BP 92/44 HR 81 RR 14 96 O2 Sat 99PaCO2 28.1, PaO2 113. As a baseline (was done in ED), guaiac
in stool was positive. He was sent to CT with ortho in ED. Saw
his PCR [**Last Name (Only) **] [**Date**] [**Time (only) **] 5:15 and showed a 6.2mm x 5.1mm pancreatic
tail mass (blood in the head of the pancreas with no dilation
distally). He is scheduled for pancreatic stent placement
tomorrow. He was treated for acute cholangitis (e.g. given 3L NS
and 1g of IVF) and was given 1L NS to help with ARF. He
reports that he is more anxious for relief of pain.

Figure 8: An example of synthetic data generated with the prompt of Figure 6, with a high SEMSCORE at 79

Ms. [**Known lastname **] is a 69 year old woman with a history of metastatic renal cell carcinoma (RCC) with distant metastases to the pancreas
who presents with fever, chills, abdominal pain since [**2119-1-7**]. She reports taking 5 grams of tylenol daily at home
which she also reports taking in the past week. She also reports having
rib pain while lying in the [**location unclear **]. She reports chills
and having dark stools, but denies any associated fever or rigors.
She was recently started on zosyn for headache and nausea and
vomiting. She denies any contact with sick person, recent travel
or animals, or recent change in her work or home environment.

She is an anxious, non-smoker with a past medical history of pancreatic
insufficiency and hypertension who presents with abdominal pain and
lower extremity edema. She denies dysuria, joint pain or rash. She
is not hypotensive with a BP of 125/80 at her arrival to the
Emergency Department. She is asymptomatic for lactate at the time of
this admission to the floor. On review of her baseline admission guaiac
stool testing is not available. On her last outpatient visit to her
oncologist at [**Hospital1 **] on [**1-7**], she was seen to have a stent placed for
cholangitis and was given an ERCP. She was treated for cholangitis
at the time and was given some ARF on [**1-7**].

Her oxygen saturation was 97for possible E.Coli pneumonia. She was found to have hypotension on
admission to the floor and received dobutamine.

Figure 9: An example of synthetic data generated with the prompt of Figure 6, with a low SEMSCORE at 52

Task
Your task is to generate a JSON list detailing specific healthcare-related entities in an input text. The entities to be identified are: 'medical problems', 'treatments', and 'tests'. Each entry in the JSON should specify the entity and its type.
Entity Markup Guide
- "problem" for medical problems.
- "treatment" for treatments.
- "test" for medical tests.
Examples
Example Input 1: At the time of admission, he denied fever, diaphoresis, nausea, chest pain or other systemic symptoms.
Example Output 1: ["entity": "fever", "type": "problem", "entity": "diaphoresis", "type": "problem", "entity": "nausea", "type": "problem", "entity": "chest pain", "type": "problem"]
Example Input 2: He had been diagnosed with osteoarthritis of the knees and had undergone arthroscopy years prior to admission.
Example Output 2: ["entity": "osteoarthritis of the knees", "type": "problem", "entity": "arthroscopy", "type": "test"]
Example Input 3: After the patient was seen in the office on August 10, she persisted with high fevers and was admitted on August 11 to Cottonwood Hospital.
Example Output 3: ["entity": "high fevers", "type": "problem"]
Example Input 4: HISTORY OF PRESENT ILLNESS: The patient is an 85-year-old male who was brought in by EMS with a complaint of a decreased level of consciousness.
Example Output 4: ["entity": "a decreased level of consciousness", "type": "problem"]
Example Input 5: Her lisinopril was increased to 40 mg daily.
Example Output 5: ["entity": "lisinopril", "type": "treatment"]
Input Text: `[INPUT]`
Output Text:

Figure 10: The prompt for annotating documents for the synthetic NER task