Generating English Synthetic Documents with Clinical Keywords: A Privacy-Sensitive Methodology

Simon Meoni, Éric de la Clergerie, Théo Ryffel
Inria/Arkhn (Paris, France), Inria (Paris, France), Arkhn (Paris, France)
simon.meoni@arkhn.com,theo@arkhn.com, Eric.De_La_Clergerie@inria.fr

Abstract

Electronic Health Records (EHR) store valuable patient-staff interaction data. These notes, often unstructured to save healthcare personnel time, can be challenging to analyze manually. Proprietary online LLMs have demonstrated impressive results in analyzing EHR notes. However, Clinical NLP faces unique challenges due to the sensitive and specialized nature of the data. Sending patient information via external APIs poses privacy risks, and hospitals require customized NLP systems to align with their practices. Developing customized LLMs using specific training datasets is crucial to address these challenges. We propose generating synthetic training data using keywords extracted without confidential information. Furthermore, we introduce a reward mechanism that iteratively refines the quality of synthetic documents. This involves scoring synthetic candidates against real clinical reports using a semantic textual similarity score and performing an alignment step to align the model with its best-scored utterances.

Keywords: Style Transfer, Data Generation, LLM, Reinforcement Learning, Data Privacy

1. Introduction

Electronic Health Records (EHR) contain patient and healthcare staff interactions. Professionals record their impressions, observations, and various medical procedures performed. Despite the computerization of clinical documents, notes remain fairly expressive and in a free format to save time for healthcare personnel and allow for the description of unusual situations (Rosenbloom et al., 2011; Wu et al., 2022). These notes can be handy for medical professionals, but analyzing them manually is daunting. Natural Language Processing (NLP) techniques come here, as they 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 information through an external API risks patient privacy. Hospitals must maintain control over their NLP systems due to their unique practices and environments. Creating customized LLMs is an important issue.

A specific training dataset is required to develop such a model with clinical skills. Accessing real clinical data to constitute this dataset remains very complex and requires anonymization, which is time-consuming and expensive. Another option is to generate synthetic clinical notes that resemble real data and do not contain any patient identifiers (Melamud and Shivade, 2019; Ive et al., 2020). This approach reduces human intervention and is more compliant with regulation laws.

2. Contributions

This work introduces a novel method for generating synthetic documents, enforcing privacy preservation by design, only using sparsely pseudo-anonymised data. Our key contributions include:

Privacy-conscious Synthetic Document Generation: We propose a methodology that utilizes a small amount of manually anonymized data to generate synthetic documents. These documents are then used to supervise fine-tuned generators, as illustrated in Figure 1.

Incorporating Clinical Keywords: We enhance synthetic document generation by enriching prompts with privacy-safe keywords as illustrated in Figure 2. Using QuickUMLS (Soldaini and Goharian), we generate candidate documents based on keywords extracted from real Clinical reports (*CR*). The keywords guide the model to produce text that closely aligns with specified content and style criteria.

Reward Mechanism: We introduce an iterative refinement process for enhancing the quality of the synthetic documents generated by the seeded model. This method involves two main key steps:

- Scoring the synthetic candidates through comparison with private or real *CR* using an SemScore evaluator model in the private side returning only scores to the public side;
- Aligning the model with its best utterances using Direct Preference Optimization (DPO) (Rafailov et al., 2023).

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 115–123
20 May, 2024.

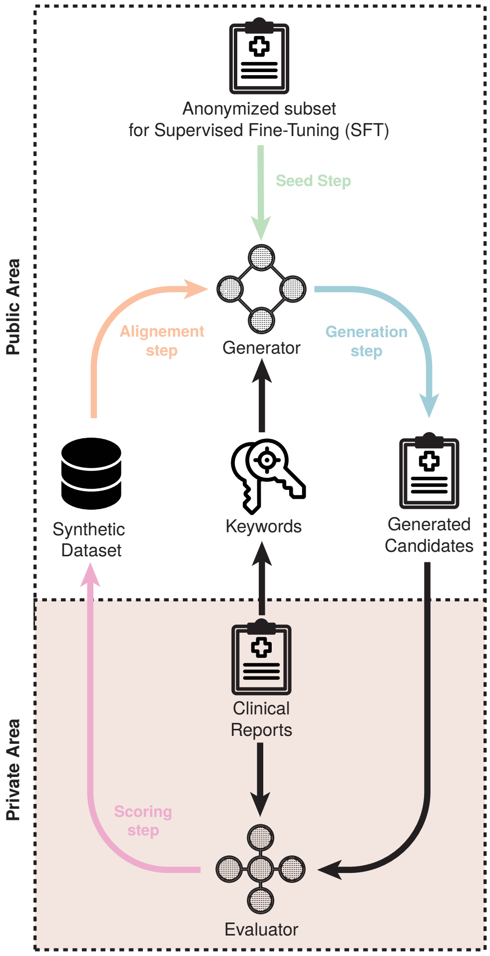

Figure 1: Our workflow is delimited by the private and public areas. Only the score is returned to the public area.

3. SemScore as Evaluation Metrics

To evaluate the quality of the responses generated by an instruction-tuned LLM, we adopt SemScore, a metric based on semantic textual similarity (STS). However, we add two differences with SEMSCORE (Aynetdinov and Akbik, 2024): (1) we embed generated document and ground truth using `all-distilroberta-v1`, fine-tuned on synthetic documents as described in section 4.3. (2) We compute the similarity between model and target responses through the cosine similarity of their embeddings. The interval $[-1, 1]$ represents the value where 1 indicates similarity and -1, semantic opposition between a pair of two sequences. Based on textual similarity, *SemScore exhibits the strongest correlation to human evaluation results, even outperforming LLM-based metrics, while not requiring special access or incurring additional cost*. We can use this metric in a real-world scenario where the model (M_{score}) must be hosted in the clinical building to measure STS between private clinical and synthetically generated reports.

4. Reward Training

The task aims to generate synthetic CRs with prompts enriched with keywords (Figure 2) extracted on real CR as illustrated in Figure 1. As preprocessing, we collect in K_{train}, keywords for each document in D_{train}, keeping in mind that these keywords do not carry confidential information.

4.1. Seed Step

Keeping associated keywords in K_{sft}, we sample, with ratio $ratio_{\text{sft}}$, a very small subset of the training dataset D_{train} to fine-tune, after deidentification, our initial generator model M_{gen}.

4.2. Generation Step

For each data point in K_{gen}, M_{gen} generates N candidate documents collected in dataset D_{gen}.

4.3. Scoring Step

The scoring step consists of two steps:

1. at the first generation step ($step = 1$), we initialize the evaluator model M_{score}, fine-tuning it with a contrastive objective on D^*_{contr} and D_{contr} correspond to a split of D^*_{gen} with the respective real documents from D_{gen}.

2. With M_{score} we score candidates from D^*_{gen} over D_{gen}, we select a pair of *chosen* and *rejected* items among N candidates where the chosen (resp. rejected) one is the candidate with the maximum (resp. minimal) score. Finally, we keep the pair where the chosen candidate is above the percentile p to obtain D_{DPO}.

4.4. Alignment Step

We align and update M_{gen} using DPO on D_{DPO} and pursue a new iteration as illustrated by Algorithm 1.

5. Experiments

Model: we use `Mistral-7B-Instruct-v0.1` (Jiang et al., 2023), a trade-off between performance and computational cost. We use QLoRA (Dettmers et al., 2023) reducing the memory footprint.

Dataset: Our dataset is derived from the Mimic-III database, applying a protocol to ensure its applicability to generate clinical narratives. The creation process includes pre-processing, keyword extraction and post-processing steps.

Algorithm 1: Reward Training Algorithm

Input : D_{train} = train dataset; $ratio_{\text{sft}}$ = sft ratio; $ratio_{gen}$ = gen ratio; M_{gen} = generative model; M_{score} = evaluator model; p = percentile filter value; N = number of candidates to generate;

Output : M_{gen}

```
K_train ← ExtractKeywords(D_train)
D_sft, K_sft ← Anonymize(Sample(D_train, K_train, ratio_sft))
D_gen, K_gen ← Sample (D_train, K_train, ratio_gen)
// Seed Step
M_gen ← Supervised fine-tune M_gen on pairs in (K_sft, D_sft)
for step = 1 to steps do
    // Generation Step
    D*_gen ← generate new N candidates with M_gen per k ∈ K_gen
    // Scoring Step
    if step = 1 then
        // Building the evaluator model
        D*_contr, D_contr ← Sample(D*_gen, D_gen, ratio_contr)
        M_score ← ContrastiveTrain (M_score, neg=D*_constr, pos=D_constr)
    endif
    D_score ← score D*_gen over D_gen with M_score over the candidates generated
    D_dpo ← for each data point in D_score, keep a pair of candidates, then filter pairs on percentile p
    // Alignement Step
    M_gen ← DPO Alignment M_gen on (K_gen, D_dpo)
endfor
```

<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, preserving the order. You must adopt a medical telegraphic style, abbreviated, characterized by concise and direct language.
Keywords: *cirrhosis c, portal, esophageal varices, SBP, angioectasias, gout, liver, note, fractured, left wrist, hip, note, admissions, asymptomatic, range, PRBCs, angioectasias, estrogen, bleeding, hospital course, SBP, guaiac, stool, L wrist, L hip, consulted, L wrist, leg, said, surgical, pantoprazole, gtt, morphine, hip pain, PRBCs, transfer, sat, hip pain, esp, feeling, note, iron, stools, stools* [/INST]

Figure 2: Example of prompt with injected keywords

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. we use them as data points for our D_{train}. On average, the data points consist of a 248-word excerpt.

2. *keywords extraction*: We project UMLS concepts using QuickUMLS over D_{train}. QuickUMLS is an unsupervised biomedical concept extraction based on pattern matching that guarantees only medical concepts are extracted and no identifying information. We obtain K_{train} (cf. Section 4) used to enrich the prompts, as illustrated in Figure 2. On average, we extract 58 keywords per data point.

3. *post-processing*: We filter out data points without keywords. We keep the keywords ordered to force the model to follow the same narrative as the ground truth. In this way, we constitute a dataset of 5602 excerpts as data points. We use 70% (either 3921 data points) of these data points as a train set (D_{train}) and 30% (either 1680 data points) as a test set (D_{test}).

Evaluation: To monitor and evaluate M_{gen} progression, we also train a model (M_{ref}) supervised fine-tune overall D_{train}. M_{ref} is used as a witness and reference, trained without privacy concerns. We compare the performance of M_{gen} and M_{ref} along the different *step* as described in Algorithm 1. Additionally, we calculate a baseline where we compute SemScore between the real D_{test} and K_{test} as illustrated in Table 1.

6. Results and Discussion

Our experimental setup aimed to evaluate the performance of our model trained with the method described in section 4 with different $ratio_{\text{sft}} \in \{4\%, 6\%\}$ (i.e 4% and 6% is equal to 156 and 235 data points, respectively) against M_{ref}, a reference model fine-tuned with the full D_{train}. To gauge the different fine-tuned scenarios, we use two M_{score} fine-tuned as described in Section 4.3 on D_{test}.

We observe monotonous score improvements over steps. $M_{\text{gen}}^{6\%}$ model even outperforms at step 2 the score of M_{ref}, highlighting the relative efficiency of alignment in refining the generated documents' quality over successive iterations. Moreover, $M_{\text{score}}^{4\%}$ trained on lower-quality synthetic data tends to overestimate the higher-quality generated documents. This overestimation is observed in both $M_{\text{ref}}^{100\%}$ and $M_{\text{gen}}^{6\%}$. However, the same trends have been observed with any evaluator.

These improvements can be attributed to various factors. The scoring mechanism allows for a focused learning approach, where a model iteratively learns from the chosen examples and adjusts away from the rejected ones. Such a dynamic refinement process effectively distils the desired style and content characteristics along the steps.

Comparing different data ratios further reveals the nuanced impact of training data volume on model performance. It underscores the efficiency of DPO in leveraging available data regardless of the seed dataset size to achieve superior outcomes.

	steps	$M_{\text{score}}^{4\%}$	$M_{\text{score}}^{6\%}$
baseline	-	48.43	49.35
$M_{\text{ref}}^{100\%}$	-	**74.48**	72.48
$M_{\text{gen}}^{4\%}$	0	67.95	65.94
	1	71.53	69.18
	2	72.25	70.12
$M_{\text{gen}}^{6\%}$	0	70.78	67.26
	1	72.54	70.78
	2	76.10	**74.37**

Table 1: SemScore evaluation for models M_{gen}^{a} with $a = r_{sft} \in \{4\%, 6\% \, 100\%\}$ using the different evaluators M_{score}^{b} with $b = r_{sft} \in \{4\%, 6\%\}$. The grey scores denote cross-evaluation where $a \neq b$.

7. Related Works

Synthetic Data Generation: Recent works tend to generate synthetic data with privacy concerns (Li et al., 2023a; Hiebel et al., 2023; Xie et al., 2024; Li et al., 2024). For instance, (Kweon et al., 2023) proposes to train LLMs for different purposes using synthetic clinical data generated by online LLMs. This way, (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.

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, proposed by (Li et al., 2023b), is a scalable method that automatically labels human-written text with corresponding instructions by finetuning a language model 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, leading to improvements in both instruction following and reward provision.

Our method differs from the methods described above in these differences :

- only the score is accessible to the learner, preserving the privacy of real EHR.
- only public medical keywords extracted from EHRs are used to generate synthetic data
- the SemScore evaluator can be easily hosted in a clinical environment and the generator LLM may be shared with external actors.

8. Future Directions

This study has laid the groundwork for generating synthetic documents enforcing privacy protection. It leverages a small anonymized seed dataset for supervised fine-tuning alongside keyword-augmented prompts and refinement steps based on synthetic candidates to reduce human intervention. Despite its promise, shortcomings and openings need to be addressed.

As we can annotate privacy-free generated documents using online models for NER and EL tasks, we can train models for downstream tasks using the generated data and compare them with models trained on real data to reinforce our evaluation. Moreover, We envision advancing our methodology by exploring a mixture of evaluation metrics incorporating more sophisticated evaluators and experimenting with alternative reinforcement learning such as KTO (Ethayarajh et al., 2024), or IPO (Azar et al., 2023). These would rely on classical metrics in style transfer and embrace notions of document quality (Jin et al., 2022). Such advancements could streamline the generation process, reduce the computational cost, and enhance synthetic documents' overall quality and applicability in privacy-sensitive applications.

9. Bibliographical References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Few-Shot Clinical Information Extractors. ArXiv: 2205.12689.

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

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A General Theoretical Paradigm to Understand Learning from Human Preferences. ArXiv:2310.12036 [cs, stat].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv:2305.14314 [cs].

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. ArXiv:2402.01306 [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:2308.08998 [cs].

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.

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. ArXiv:2303.16416 [cs].

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: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.

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:2309.00237 [cs].

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

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: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: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.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. ArXiv:2305.18290 [cs].

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. QuickUMLS: a fast, unsupervised approach for medical concept extraction.

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. ArXiv:2403.01749 [cs].

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-Rewarding Language Models. ArXiv:2401.10020 [cs].

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. 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 3: An example of prompt for the Figures 4,5,6

```
    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 98% on RA. Access 3PIV's 2,
20's, 18. Pt also found to be in ARF.
.
Currently, pt reports that he is anxious.
.
```

Figure 4: An example of a ground truth associated with the prompt 3

```
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 97% on 2L. As a work up she was given
 ↪  antibiotics
for possible E.Coli pneumonia. She was found to have hypotension on
admission to the floor and received dobutamine.
```

Figure 5: An example of synthetic data generated with Figure 3 with a SemScore at 52

```
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 99% RA. On ABG: pH 7.31,
PaCO2 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 6: An example of synthetic data generated with Figure 3 with a SEMSCORE at 79