Can Synthetic Text Help Clinical Named Entity Recognition? A Study of Electronic Health Records in French

Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol

Can Synthetic Text Help Clinical Named Entity Recognition? A Study of Electronic Health Records in French*

Nicolas Hiebel[†], **Olivier Ferret**[‡], **Karën Fort**[⋆], **Aurélie Névéol**[†]
[†]Université Paris Saclay, CNRS, LISN, France
[‡]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[⋆]Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, France
[†]firstname.lastname@lisn.upsaclay.fr, [‡]olivier.ferret@cea.fr, [⋆]karen.fort@loria.fr

Abstract

In sensitive domains, the sharing of corpora is restricted due to confidentiality, copyrights, or trade secrets. Automatic text generation can help alleviate these issues by producing synthetic texts that mimic the linguistic properties of real documents while preserving confidentiality. In this study, we assess the usability of synthetic corpus as a substitute training corpus for clinical information extraction. Our goal is to automatically produce a clinical case corpus annotated with clinical entities and to evaluate it for a named entity recognition (NER) task. We use two auto-regressive neural models partially or fully trained on generic French texts and fine-tuned on clinical cases to produce a corpus of synthetic clinical cases. We study variants of the generation process: (i) fine-tuning on annotated vs. plain text (in that case, annotations are obtained a posteriori) and (ii) selection of generated texts based on models' parameters and filtering criteria. We then train NER models with the resulting synthetic text and evaluate them on a gold standard clinical corpus. Our experiments suggest that synthetic text is useful for clinical NER.

1 Introduction

The lack of specific resources might be the most common hurdle encountered when working in Natural Language Processing (NLP), whether it be the lack of task-specific resources, domain-specific resources, or both. A major challenge for biomedical NLP in languages other than English is the lack of available clinical corpora that can be shared with the community (Névéol et al., 2018).

One way of addressing this problem is to generate new data automatically. The performance of recent language models based on the transformer architecture (Vaswani et al., 2017) on various tasks that involve generating text (translation, summarization, etc.) makes it possible to use the transformer's power to generate text to address resource sparsity.

More specifically for generation, large generative language models such as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) have the ability to generate well-written text containing a wealth of information. The knowledge gained during the pre-training of those large language models could be helpful in more specialized domains where there is not enough data to train a generative model from scratch.

Dealing with medical data brings one particular issue. The generated data should not contain sensitive information from the data it is inspired by while remaining as close as possible to it. This means that the evaluation of this task is different from the usual evaluation of text generation.

The evaluation of text generation is a difficult problem (Howcroft et al., 2020; Novikova et al., 2017). There is a distinction between intrinsic and extrinsic evaluation (Gehrmann et al., 2022). For intrinsic evaluation in generation tasks such as translation and summarization, the quality is usually first automatically measured using metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or BERTScore (Zhang et al., 2020). Those metrics and many others compare the system's output to one or several human-written references (Frisoni et al., 2022). However, a gold reference does not exist for an open-generation task.

Human evaluation is usually the best way of intrinsically evaluating text generation. Still, there can be several problems. First, human evaluation should be conducted once the system's output can be considered "good enough" to avoid wasting human evaluation costs. This means there is a need to measure if the system is ready to be manually evaluated. Second, it might not be possible for a human to assess the quality of generated data at a

* This is a revised version of the original paper, which contained an error in some of the NER results. Revised text passages and figures appear in green; an explanation of the changes is given in Appendix A.1.

larger scale, if the generated data is not considered as a set of independent sentences but as a whole self-sufficient corpus.

This is why extrinsic evaluation of the generated data is also important, meaning evaluating the usefulness of the generated data (Stadler et al., 2022). However, to our knowledge, methodologies to extrinsically evaluate open text generation are rare.

In this work, we explore text generation using a pre-trained language model and we focus our evaluation on named entity recognition (NER).

This paper offers the following contributions:

- we generate two sets of clinical cases in French with two variants;
- we introduce an evaluation protocol of the generated texts using ngram overlap and NER models;
- we compare the generated texts with real clinical case corpora.

2 Related Work

Generating data to address resource sparsity is an active field of research. Pre-trained transformer-based models have been used as data augmentation tools, such as in Claveau et al. (2022), where the authors fine-tuned the pre-trained auto-regressive English language model GPT-2 to create labeled examples for training classification models in French and English. In the medical domain, Amin-Nejad et al. (2020) use an encoder-decoder transformer architecture to generate discharge summaries given a context. The context gives a target to the generation that is used to evaluate the generated texts. Melamud and Shivade (2019) train an LSTM-based language model to generate shareable clinical notes using differential privacy (Dwork et al., 2006). Word embeddings are trained on the generated data and compared to word embeddings trained on the original data on two existing benchmarks. Ive et al. (2020) used an entity extraction method to build a generative model with entities as input and the document as output and generated artificial mental health records in English.

Several efforts exploit Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to seamlessly augment original data. Among them, Abedi et al. (2022) use GANs to generate artificial structured data in the medical domain. Choi et al. (2017) created medGAN, a framework to generate artificial structured electronic health records (EHR) with discrete values. Torfi et al. (2022) rely on a combination of autoencoders (Kingma and Welling, 2014) and GANs to generate artificial structured data with both discrete and continuous values.

3 Method

3.1 Experiment description

The experiment presented in this paper is designed to evaluate open text generation in the clinical domain. The goal is to create an artificial corpus presenting the same *characteristics* as a real clinical corpus. One challenge of this task is that defining and recognizing the *characteristics* of a clinical corpus is an open question.

We envisaged the *characteristics* of clinical corpus along three perspectives:

1. a distributional perspective: we want to create an artificial clinical corpus with a token distribution similar to that of a real corpus. This aspect will be measured using self-BLEU (Zhu et al., 2018) and perplexity;
2. a qualitative perspective: we want to create an artificial corpus that mimics a real clinical corpus in terms of semantic and linguistic content. This aspect will be measured by manual analysis of samples of generated texts;
3. an application perspective: we want to create an artificial clinical corpus that can be used to train NER models to be applied to a real clinical corpus for clinical information extraction. This aspect will be measured through the extrinsic evaluation of Section 4.2.

Figure 1 presents the global setup for evaluating generated data for NER. The specific corpora (`MERLOT` and `E3C`) named in the figure will be described in Section 3.2.

Briefly, we leverage a corpus of clinical documents in French containing high-quality entity annotations (`MERLOT`). This corpus represents the target real data we aim to analyze. It is used in part to train clinical NER models and in part for evaluating clinical NER models. We then use a clinical case corpus (`E3C`) for fine-tuning generation models, with the option of generating tagged text or plain text that can then be tagged with the Gold NER model.

In the end, each of the newly annotated corpora is split into training and testing sets and is used to train new NER models. These final models are then tested on the test sets of each corpus.

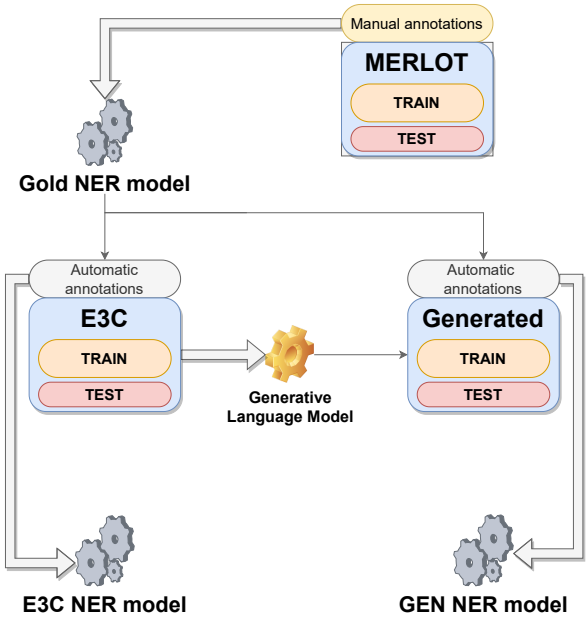


Figure 1: Description of the experiment conducted to evaluate NER on artificial data.

3.2 Corpora

This section presents the corpora used in our experiments. Table 1 presents descriptive statistics.

	E3C$_{FR}$	CAS	MERLOT
Toks	328,645	231,662	148,476
Docs	1,009	717	500
Toks/doc	325.7	323.1	297.0
Sents/doc	15.2	15.8	14.4
Avg sentence length	21.4	20.4	20.6
Self-BLEU	0.68	0.66	0.69

Table 1: Statistics of the real medical corpora.

MERLOT (Campillos et al., 2018) is a corpus of 500 clinical documents in French from the hepato-gastro-enterology and nutrition specialties. It comprises 148,476 tokens with annotations for entities, attributes and relations. This text genre represents the type of clinical narratives contained in electronic health records that we would like to analyze.

CAS (Grabar et al., 2018) is a French medical corpus containing 717 clinical case descriptions published in the literature. Clinical cases contain descriptions of the medical history of patients, treatment received at the hospital, and follow-up care. CAS covers several medical specialties. The corpus has been de-identified and the publication of case descriptions is done with patient consent. The corpus can be obtained from the DEFT 2020 shared task organizers[1].

E3C (Magnini et al., 2020) is a freely available multilingual corpus[2]. It contains biomedical documents extracted from different sources, including journals, abstracts, existing biomedical corpora, patient information leaflets for medicines, and others.

In this work, we focused on the French part of the corpus, selecting only the clinical cases published in the literature, which are the most similar documents to the clinical narratives we are interested in. Clinical cases from E3C cover a variety of medical specialties and are distinct from those in CAS. This part of the corpus contains about 1,000 documents and more than 300,000 tokens.

3.3 Entity annotation for E3C

Entity annotations for E3C are needed for two reasons. First, we integrate the annotations in the form of XML tags into the E3C texts. Fine-tuning the model on this version of the corpus will enable the model to learn to generate tags directly during the generation process. Second, the annotations will be used to train and test NER models. We use the manual annotations of entities made on the MERLOT corpus to train NER models (Bannour et al., 2022) to annotate the E3C corpus. Using an automatic rather than a manual annotation for training our final NER models can lead to error propagation. We use an ensemble strategy to mitigate this issue: five different models are trained and applied to the corpus with a majority vote for the annotation (annotation is kept if at least three models agree on the annotation), which is classically known as a good way to reduce errors. The train and validation sets are shuffled for the training of each NER model and a different seed is used.

3.4 Generation

Pre-trained models such as GPT-2 have been proven to allow for generating complementary textual data in a data augmentation context. In this work, we explore the ability of auto regressive neural models (viz. GPT2 and BLOOM) trained partially or exclusively on French to adapt to the medical domain and generate relevant clinical case descriptions.

We use two existing GPT models. The first model is a one-billion-parameter model pre-trained on generic French texts by Simoulin and Crabbé

[1] https://deft.lisn.upsaclay.fr/2020/
[2] https://github.com/hltfbk/E3C-Corpus

(2021). The corpus used to pre-train the model was extracted from Wikipedia, Project Gutenberg[3], OpenSubtitle (Lison and Tiedemann, 2016), and CommonCrawl. In this paper, this model will be referred to as LLF.

The second model is a smaller version of the BigScience BLOOM model[4]. BLOOM was trained on 46 different languages, including French[5]. We chose the one-billion-parameter version for a fair comparison between the two models. Both models are available on the HuggingFace platform[6][7].

3.4.1 Fine-tuning the models

We used the freely available E3C corpus for fine-tuning our model on medical texts so that the generation process can be replicated[8]. We added special tokens for marking the beginning and end of each document to make the model generate whole documents. Two versions of the corpus were used: one with the annotations described in 3.3 so that the models directly learn to generate annotations, and one without annotations. The corpus was split into training and validation sets, with 90% of the corpus for training and 10% for validation. The model was trained with a maximum of 10 epochs, but training stopped early if the perplexity on the validation set did not decrease significantly during two consecutive epochs. We selected a learning rate of 2e-5, negative likelihood loss with the AdamW optimizer from the torch package[9]. In the end, we have four models, two for each pre-trained model with one version producing plain untagged text and one including annotations to produce tagged text.

3.4.2 Generation process

Clinical documents were generated using the token marking the beginning of a document as input. No additional prompts were supplied to the models. We thought that giving information such as the beginning of a document as a prompt might cause the model to reproduce text seen during fine-tuning.

Many parameters can be used during inference to influence the generative models' output. We decided to focus on three parameters: nucleus sampling (top-p) value, temperature value, and repetition penalty. The rest of the parameters were left with default values. At each step (token) of the generation process, the models output a probability distribution for vocabulary items. For generating diverse outputs, sampling consists in randomly picking the next token to generate, as opposed to greedy decoding, where the most probable token is selected. Nucleus sampling, or top-p sampling, restricts the sampling to the tokens whose summed probabilities exceed a threshold p. It allows for more diversity while remaining consistent. Temperature is a way of decreasing or increasing the disparities in the probabilities of the vocabulary at a generation step. A temperature higher than 1 will bring the probabilities closer and make the model less confident, and bring more diversity in outputs. A temperature lower than 1 will do the opposite. Finally, a repetition penalty will encourage the model to avoid falling into loops. We generated several samples of approximately 10% of the target-generated text to assess the impact of those three parameters on the self-BLEU score of the generated sample and an oracle perplexity with the seven billion parameters version of the BLOOM model. More details can be found in appendix A.4. We noticed expected behaviors: increasing temperature and repetition penalty increases diversity (lower self-BLEU) but also increases perplexity. However, we found that top-p sampling has a low impact on the two metrics. Perhaps because by default, the model also adds top k sampling (selecting amongst the k most probable tokens at each generation step) that lowers the impact of top-p. In the end, since we wanted a high diversity output to match E3C's diversity, we selected high values for temperature (1.2 for models without tags and 1.5 for models with tags). We also kept a very high repetition penalty (10) because we noticed that a lot of generations ended with the model generating the same tokens over and over even with a high value. Finally, we selected a default value of 0.95 for top-p sampling.

We aimed for the models to generate complete documents with length variations similar to E3C. We computed the average length of E3C documents (for both annotated and non-annotated versions) and combined it with standard deviation, exploiting the ability of the models to close a document with an ending token and start a new one in the same generation process.

[3] www.gutenberg.org

[4] https://lstu.fr/bloom

[5] https://huggingface.co/spaces/bigscience/BigScienceCorpus

[6] https://huggingface.co/asi/gpt-fr-cased-base

[7] https://huggingface.co/bigscience/bloom-1b1

[8] Code available here : https://github.com/Hiebel/Synthetic-text-for-NER

[9] https://pytorch.org/

3.5 Filtering generated texts

We ensured that the generated corpora had certain characteristics for a fair comparison to the reference corpus E3C.

Large auto-regressive language models can generate good-quality texts. However, the generation can present some flaws such as repetitions and incoherence, especially as the generation length gets longer (Pillutla et al., 2021). We established a few criteria that we could handle automatically by observing a few generated samples:

- overly long tokens: the models sometimes generate an incoherent sequence of characters or combine many technical words (e.g., "*céphalo-rachidiencéphalico-mandibulaire*");
- repetition of one to many tokens over and over, especially towards the end of a document;
- the models start a sequence of short tokens often with numbers (e.g., "*classée T4N2 M0 I 2 B 3 C 4 D 5*") or start reciting the alphabet (e.g., "*hépatite A B C D E F*").

We addressed these three cases by excluding documents based on token length, trigram frequency, and regular expressions.

The thresholds used for the three parameters were adjusted by analyzing the excluded documents with different threshold values. In the end, we excluded documents with tokens of thirty characters or more, documents with trigrams appearing more than three times, and repetition of at least three short tokens combined with numbers, or single letters.

We also wanted our generated corpora to have a diversity close to that of E3C as computed with the self-BLEU score (0.68 for E3C). We generated a significantly bigger amount of text than the targeted 330,000 tokens and selected a subset of diverse documents using the tool developed by Cohen et al. (2013). Then, we computed the self-BLEU score for the selected documents, including the BLEU score of each document. We removed the ten documents with the highest BLEU, that is the documents bringing the least diversity, and repeated the process until obtaining a self-BLEU score and a number of tokens close to the E3C corpus.

The characteristics of the generated corpora are shown in Table 2. As we explained, each corpus generated corresponds to a filtered subset of the total amount of text generated. We noticed that text generated with the two configurations using the LLF model needed less filtering compared to their BLOOM counterpart. Configurations with texts generated with tags also needed more filtering. Globally, using both filters, we discarded 0.45% of generated text for $\texttt{Bloom}_{E3C}$, 0.54% for $\texttt{Bloom}_{E3C+T}$, 0.30% for $\texttt{LLF}_{E3C}$, and 0.45% for $\texttt{LLF}_{E3C+T}$. Compared to Table 1, we can see that only $\texttt{Bloom}_{E3C+T}$ presents much shorter documents. However, the sentences of the generated corpora are a lot longer than in E3C, especially with $\texttt{Bloom}_{E3C}$ that do not even have two sentences per document. Finally, corpora generated with BLOOM are a little less diverse.

	$\texttt{Bloom}_{E3C}$	$\texttt{Bloom}_{E3C+T}$	$\texttt{LLF}_{E3C}$	$\texttt{LLF}_{E3C+T}$
Toks	329,328	346,413	328,498	336,154
Docs	943	1,997	1,028	978
Toks/doc	349.2	173.5	319.6	343.7
Sents/doc	1.8	3.1	6.9	7.2
Len Sent	194	56.0	46.3	47.7
Self-BLEU	0.70	0.73	0.68	0.67
Perplexity	9.97	9.27	10.03	11.9

Table 2: Statistics of the generated corpora. Oracle perplexity was computed using an oracle pretrained model of BLOOM with 7 billion parameters. E3C's perplexity using this model was 22.0 (Len Sent=average sentence length).

3.6 Automatic intrinsic evaluation

As a complement to BLEU scores, we measured the overlap between the generated corpus and the original corpus to characterize potential memorization of the original documents processed during fine-tuning.

To this end, we developed a python tool based on overlapping ngrams to compare the generated texts and the original text and to detect the rare ngrams in the original corpus that appear on the generated corpus. For a fair evaluation, we also show results of the same comparison between the corpus used for fine-tuning (E3C) and another, non-overlapping, clinical case corpus in French (CAS).

4 Results

Table 2 shows global synthetic corpus statistics and Table 3 presents sample texts generated with each model configuration.

Table 4 shows the number of entity annotations for E3C, MERLOT, and the generated corpora. We can notice that the corpora generated without tags ($\texttt{Bloom}_{E3C}$, $\texttt{LLF}_{E3C}$) contain almost the same number of annotations as the E3C corpus (around 9,000

Tags	LLF	BLOOM
No	Il s'agit d'un patient âgé de 52 ans, sans antécédents pathologiques notables. Admis aux urgences pour douleurs abdominales diffuses associées à des vomissements postprandiaux évoluant depuis deux semaines en rapport avec une obstruction du canal cystique dont l'examen clinique a objectivé un ictère cutanéo-muqueux généralisé associé à quelques signes inflammatoires non spécifiques (extériorrhée et urétéro-hydronéphrose). L'examen ophtalmologique n'a révélé aucun autre signe oculaire ou ostéo-nerveux associés.	Il s'agissait d'une patiente âgée de 55 ans ayant comme antécédents une polyarthrite chronique touchant toutes les articulations métacarotidiennes évoluant depuis plus de 10 ans associée à une insuffisance surrénalienne fonctionnelle qui s'est aggravée après l'installation d'un diabète insulino-résistant non insulinodépendant nécessitant son hospitalisation six mois avant sa consultation chez notre confrère ophtalmologue afin d'être mis sous traitement médicamenteux associant : imipramine 250 mg / jour ; dexaméthasone 150 mg/jour; corticostéroïde prednisolone 5mg/kg/24h)
Yes	Une femme âgée de 55 ans, sans antécédents pathologiques particuliers était hospitalisée pour une dyspnée d'effort évoluant depuis 6 jours. L' examen clinique a montré un état général altéré avec des signes inflammatoires du membre supérieur gauche ainsi qu'un syndrome grippal (une fièvre à 39 ° C) et le reste de l' examen neurologique était sans particularités. La radiologie standard n'a objectivé ni troubles sensitivo-moteur ni convulsions fébriles.	Il s'agissait d'un patient de 50 ans, sans antécédents pathologiques particuliers, admis aux urgences pour des douleurs épigastriques aiguës associées à une distension abdominopelvienne évoluant depuis deux jours. L' examen clinique trouvait un patient en assez bon état général (Apgar: 10/10). Le bilan préopératoire objectivait une fonction rénale normale et la CRP était à 12 mg/l.

Table 3: Excerpts of sample texts generated with top-p 0.95, temperature 1.2, and penalty 10. Highlighted sections in the lower row samples represent the entity tags generated; see Table 4 for color code. For text generated without entity tags, entity annotations are created post-generation; complete documents are shown in appendix A.6.

annotations each). However, the corpora generated with tags contain a smaller number of annotations with 5,630 annotations for $\texttt{Bloom}_{E3C+T}$ (two-thirds of E3C) and 3,893 annotations for $\texttt{LLF}_{E3C+T}$ (around 40% of E3C). This suggests that directly generating annotations might not be the optimal way of generating text for a NER task.

4.1 Ngram overlap

To compare the overlap of ngrams between E3C and the corpora generated by a model fine-tuned with E3C, we compute an overlapping score between the corpora. The overlapping score between two corpora is computed as follows:

$$\frac{\text{Number of unique common ngrams}}{\text{Total number of unique ngrams}} \quad (1)$$

The overlapping score varies between 0 and 1. A high overlapping score means that the two corpora have many tokens in common whereas a low score indicates few common tokens.

Table 5 presents the overlap between E3C and the generated corpora. We also added the comparison between E3C and CAS, another clinical case corpus (see appendix A.3 for the full ngram range). It is interesting to see that the CAS corpus has a higher overlapping score with E3C than the generated corpora. Amongst the generated corpora, $\texttt{Bloom}_{E3C}$ has the highest overlapping score with E3C for unigrams, $\texttt{Bloom}_{E3C+T}$ comes second, and $\texttt{LLF}_{E3C}$ and $\texttt{LLF}_{E3C+T}$ have the lowest scores. However, $\texttt{Bloom}_{E3C}$ has the lowest overlapping score for 8grams than the rest, even less than CAS. $\texttt{LLF}_{E3C+T}$ still has a small percentage of tokens in common with E3C, but $\texttt{Bloom}_{E3C+T}$ and $\texttt{LLF}_{E3C}$ have now a higher percentage than the rest. This means that even though they do not have that many unigrams in common with E3C, they tend to replicate longer sequences, which indicates more memorization and thus more risks regarding confidentiality. It is surprising to see that there is no regularity between the models and their fine-tuning configurations, and we are not sure how to explain this. In the end, this

Type	E3C	Bloom$_{E3C}$	Bloom$_{E3C+T}$	LLF$_{E3C}$	LLF$_{E3C+T}$	Merlot
Anatomy	1,533	1,798	799	1,056	360	922
Chemical_Drugs	307	388	390	350	264	283
Concept_Idea	493	376	275	370	241	684
Devices	36	42	72	5	41	219
Disorders	1,662	1,545	1,146	1,409	806	1,047
Dose	103	179	219	310	54	188
Hospital	74	65	129	50	54	170
LivingBeing	533	492	342	842	330	866
Localization	450	490	128	467	121	169
Measure	1,496	1,330	670	2,109	421	1,070
Mode	24	68	66	26	6	40
BiologicalProcess	127	83	102	62	78	186
Procedure	1,883	1,650	967	985	913	1,779
TemporalExpression	455	610	325	878	204	885
Overall	9,176	9,116	5,630	8,919	3,893	8,510

Table 4: Number of annotations of each type for each test corpus.

Corpus	Unigram	8gram
Bloom$_{E3C}$	0.16419	0.00011
Bloom$_{E3C+T}$	0.13887	0.00020
LLF$_{E3C}$	0.11740	0.00023
LLF$_{E3C+T}$	0.11935	0.00013
CAS	0.20373	0.00013

Table 5: Overlap scores of unigrams and 8grams between the generated corpora and E3C. Each line corresponds to the comparison between the corpus and E3C. We added the comparison of E3C with CAS as a baseline.

table might be in favor of Bloom$_{E3C}$, because it appears to be the closest corpus to E3C in terms of unigram, while not memorizing too much. However, this is yet to be confirmed.

4.2 Extrinsic evaluation

Tables 6 to 8 present the results of our extrinsic evaluation of the generated corpora through the NER task described in Section 3. More precisely, the three tables report the results of the same set of models on three different types of corpora. All the results were obtained by averaging the scores of five models trained with different seeds and with shuffled documents in the train and validation sets of the corpus.

The first thing to note is that the best performance for the test part of a given corpus is obtained by a model trained on the training part of this corpus. While this is not a surprise from a general viewpoint, it means in our context that the difference in quality between a generated corpus and a natural corpus is not high enough for balancing the advantage brought by the homogeneity between the training and the test corpora. The quality of the generated corpora for our NER task is confirmed by Table 6, which gives the results on the natural corpora and more specifically, on the MERLOT corpus. Since it was annotated manually, unlike the other corpora, we consider MERLOT as a high-quality reference. While the model trained on E3C is significantly outperformed by the model trained on MERLOT, we can observe that the models trained on Bloom$_{E3C}$ and LLF$_{E3C}$ are close to it, with a small advantage for LLF$_{E3C}$. Hence, in our context, using a corpus generated from a natural corpus is fairly equivalent to using the natural corpus.

The NER task on E3C seems easier than on MERLOT since the performance of all the models on the former strongly increases compared to their performance on the latter, except for the MERLOT models, whose performance only moderately increases. Finally, Table 6 shows, through Bloom$_{E3C+T}$ and LLF$_{E3C+T}$, that generating annotated texts is the worst strategy for both E3C and MERLOT. This strategy particularly hurts recall, a result which can be certainly explained by the fact that the number of generated annotations (see Table 4) is low compared to the number of annotations produced by the NER model trained on MERLOT.

Table 7 reports the results of our models on the generated corpora Bloom$_{E3C}$ and LLF$_{E3C}$. The first observation is that, with the exception of

$\texttt{Bloom}_{E3C+T}$ and $\texttt{LLF}_{E3C+T}$, the performance of the models is close to the performance obtained on E3C, which is not a surprise since the $\texttt{Bloom}_{E3C}$ and $\texttt{LLF}_{E3C}$ corpora were generated from EC3. However, it confirms that a generated corpus can be used as an interesting substitute for its source corpus. It is true for training a NER model but also for its evaluation. Table 7 also shows that our models obtain fairly close results on $\texttt{LLF}_{E3C}$ and on $\texttt{Bloom}_{E3C}$, which was also observed in Table 6 when the corpora are used for training our models: there is no difference between $\texttt{Bloom}_{E3C}$ and $\texttt{LLF}_{E3C}$ for the evaluation on E3C and only a small difference in favor of $\texttt{LLF}_{E3C}$ for the evaluation on MERLOT. Finally, the most noticeable observation is the poor performance of the models trained on $\texttt{Bloom}_{E3C+T}$ and $\texttt{LLF}_{E3C+T}$, especially recall. In that case, the explanation is once again related to the scarcity of the generated annotations used for training the models. However, the difference with the evaluation on the E3C corpus is more difficult to interpret.

The last table of results for the extrinsic evaluation, Table 8, mainly confirms our findings concerning the characteristics of the $\texttt{Bloom}_{E3C+T}$ and $\texttt{LLF}_{E3C+T}$ corpora. All the models trained on corpora annotated by the NER model trained on MERLOT obtain very good recall values since these corpora contain a much larger number of annotations than the $\texttt{Bloom}_{E3C+T}$ and $\texttt{LLF}_{E3C+T}$ corpora. Moreover, the very good results of the models trained on $\texttt{Bloom}_{E3C+T}$ and $\texttt{LLF}_{E3C+T}$ tend to show that their generated annotations are fairly easy to learn. Finally, the results on the corpus generated by BLOOM, $\texttt{Bloom}_{E3C+T}$, are higher than the results on the corpus generated by LLF, $\texttt{LLF}_{E3C+T}$, which is not observed with such a magnitude in Table 7. However, further investigations should be done to determine if BLOOM generates better annotations than LLF or if the generation of annotations tends to improve the quality of the text generated by BLOOM.

4.3 Qualitative evaluation

Samples of 15 automatically generated texts per model (for a total of 13,809 tokens for the 60 documents) were annotated using BRAT[10] (Stenetorp et al., 2012) and aimed to assess specific characteristics of the texts including (1) grammaticality; and (2) clinical coherence at the sentence and document level (e.g. in Table 3 one document reports Agpar score, a test performed on newborn infants, for a 50-year-old patient). See appendix A.7 for details.

The manual analysis suggests that clinically incoherent passages are more likely to be found towards the end of the texts, and within text generated without tags. In particular, texts generated with BLOOM tend to exhibit run-on sentences that become increasingly incoherent. However, there are generally more clinical inconsistencies and grammatical errors in the texts generated with the LLF vs. BLOOM models. We also note that the LLF models generate more than one case description inside a single document.

Overall, manual analysis suggests that texts generated with the $\texttt{Bloom}_{E3C+T}$ model are the closest to original clinical case descriptions.

[10]Brat Rapid Annotation Tool https://brat.nlplab.org/

	Test					
	E3C			MERLOT		
Training	**P**	**R**	**F**	**P**	**R**	**F**
E3C	**89.5**	**91.0**	**90.2**	64.4	78.0	70.5
MERLOT	87.1	90.8	88.9	**85.2**	**85.8**	**85.5**
$\texttt{Bloom}_{E3C}$	87.6	87.9	87.8	63.1	74.9	68.5
$\texttt{LLF}_{E3C}$	87.6	87.2	87.4	64.5	76.4	70.0
$\texttt{Bloom}_{E3C+T}$	83.4	68.9	75.4	71.7	55.1	62.3
$\texttt{LLF}_{E3C+T}$	84.4	68.1	75.4	75.7	46.4	57.5

Table 6: Results of the NER task for the natural corpora (P=precision, R=recall, F=F-measure).

	Test					
	$\texttt{Bloom}_{E3C}$			$\texttt{LLF}_{E3C}$		
Training	**P**	**R**	**F**	**P**	**R**	**F**
E3C	85.3	87.5	86.4	84.5	86.5	85.5
MERLOT	82.1	89.1	85.4	80.1	88.1	83.9
$\texttt{Bloom}_{E3C}$	**88.3**	**90.8**	**89.5**	84.0	85.5	84.8
$\texttt{LLF}_{E3C}$	85.7	86.7	86.2	**87.3**	**90.3**	**88.8**
$\texttt{Bloom}_{E3C+T}$	84.3	45.9	59.4	81.1	44.5	57.4
$\texttt{LLF}_{E3C+T}$	83.1	47.4	60.4	83.7	38.0	52.2

Table 7: Results of the NER task for the generated corpora without annotation.

5 Conclusion

We presented four models for generating synthetic clinical cases in French. Our experiments suggest that synthetic text is useful for training NER models in the clinical domain. While text generated with $\texttt{Bloom}_{E3C+T}$ seems more natural, text gener-

	Test					
	$\texttt{Bloom}_{E3C+T}$			$\texttt{LLF}_{E3C+T}$		
Training	**P**	**R**	**F**	**P**	**R**	**F**
E3C	67.6	90.3	77.3	57.3	90.6	70.2
MERLOT	60.3	86.2	71.0	48.7	86.6	62.3
$\texttt{Bloom}_{E3C}$	66.4	89.0	76.1	55.8	89.8	68.8
$\texttt{LLF}_{E3C}$	66.3	88.5	75.8	55.7	89.2	68.6
$\texttt{Bloom}_{E3C+T}$	**91.4**	**94.0**	**92.7**	80.9	88.0	84.3
$\texttt{LLF}_{E3C+T}$	85.1	86.0	85.5	**89.4**	**93.4**	**91.3**

Table 8: Results of the NER task for the generated corpora with annotations.

ated with $\texttt{LLF}_{E3C}$ and $\texttt{Bloom}_{E3C}$ yields promising NER performance.

6 Ethical Considerations

We were able to identify two types of potential ethical issues with our research: i) the carbon footprint generated by the training of the models and ii) the patients' data protection. We detail here how we addressed or measured them.

For our experiments, we had to train a number of models, thus generating a carbon footprint, which we measured using Carbontracker (Anthony et al., 2020), although we acknowledge that this tool can only account for one in four sources of experiment carbon impact (Bannour et al., 2021). The total CO2 emission of our experiments is 11,313g. Carbontracker also measured the energy cost, estimated at 39.2kWh. This includes the fine-tuning of pre-trained language models, the preliminary experiments on evaluation, the generation of artificial corpora, and the training of NER models.

The corpora we used for the experiments are compliant with data privacy regulations: either they do not contain personal information at all (E3C and CAS) or they were de-identified according to a protocol approved by the CNIL (Commission de l'Informatique et des Libertés), an independent French administrative regulatory body whose mission is to ensure that data privacy law is applied to the collection, storage, and use of personal data (MERLOT).

Within the framework of our project, we intend to investigate whether our model contains sensitive data. However, this task needs to be completed by a similar investigation to be led on the language models themselves, as they may contain such data (for example, from blog posts that might have been collected by CommonCrawl).

7 Limitations

One limitation of our study is the scope of the manual evaluation and the fact that it did not involve a clinician. Also, while we tried to explore several parameters and model configurations for text generation, additional exploration could be done, including using larger BLOOM models. Furthermore, these experiments with the generation of clinical cases did not include any constraints regarding the specific content of the documents generated. In future work, we would like to include more control over content to improve clinical coherence and check for privacy leaks of information contained in the training or fine-tuning corpora, to accommodate the use of confidential clinical documents.

Acknowledgements

This work has received funding from the French "Agence Nationale pour la Recherche" under grant agreement CODEINE ANR-20-CE23-0026-01. We thank Ziqian Peng for her contribution to the development of the overlap assessment tool. We thank the institutions and colleagues who made it possible to use the datasets described in this study: the Biomedical Informatics Department at the Rouen University Hospital provided access to the MerLoT corpus and Dr. Grabar (Université de Lille, CNRS, STL) granted permission to use the CAS corpus. This work was granted access to the HPC resources of Saclay-IA through the Lab-IA machine.

References

Masoud Abedi, Lars Hempel, Sina Sadeghi, and Toralf Kirsten. 2022. GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Applied Sciences*, 12(14).

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708, Marseilles, France. European Language Resources Association.

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. In *ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems*.

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon

footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual. Association for Computational Linguistics.

Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, and Aurélie Névéol. 2022. Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, 130:104073.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation*, 52(2):571–601.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR.

Vincent Claveau, Antoine Chaffin, and Ewa Kijak. 2022. Generating artificial texts as substitution or complement of training data. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4260–4269, Marseilles, France. European Language Resources Association.

Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2013. Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies. *BMC bioinformatics*, 14:10.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006*, pages 486–503, Berlin, Heidelberg. Springer Berlin Heidelberg.

Giacomo Frisoni, Antonella Carbonaro, Gianluca Moro, Andrea Zammarchi, and Marco Avagnano. 2022. NLG-metricverse: An end-to-end library for evaluating natural language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3465–3479, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf 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.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The

E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Antoine Simoulin and Benoit Crabbé. 2021. Un modèle Transformer Génératif Pré-entrainé pour le _____ français. In *Traitement Automatique des Langues Naturelles*, pages 246–255, Lille, France. ATALA.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA. USENIX Association.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Amirsina Torfi, Edward A. Fox, and Chandan K. Reddy. 2022. Differentially private synthetic medical data generation using convolutional GANs. *Information Sciences*, 586:485–500.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, online.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, pages 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Erratum

A.1.1 Evaluation error

The formatting of the annotation files for some corpora after selection of silver annotations that at least three models agreed on was erroneous. `E3C`, $\texttt{Bloom}_{E3C}$ and $\texttt{LLF}_{E3C}$ corpora were affected by the formatting issue.

The previously manually annotated `MERLOT` corpus was not affected by this error and neither were $\texttt{Bloom}_{E3C+T}$ or $\texttt{LLF}_{E3C+T}$ because the annotations were obtained during generation. So we only had one set of annotations for each corpus and we did not need to select a subset of annotations.

When evaluating the NER models on the test sets of the affected corpora, the NER evaluation script scored as false positive every annotation that contained more that one token, even if the span of the predicted annotation was correct. As a result, both precision and recall were strongly penalized. However, the training of the NER models was not impacted.

Revised results are presented in Table 6 and Table 7. Overall, we observed major improvement for every NER model on the test sets of the affected corpora. This performance increase does

not change the overall conclusions of the paper for two reasons. First, the results on our high quality reference corpus MERLOT remain unchanged. Second, the goal of this evaluation was to assess the proximity between the performance of NER models trained on synthetic datasets and NER models trained on natural datasets rather than the performance level itself. The same proximity can still be observed in the revised results.

We revised some of the comments on Table 6 and Table 7 that highlighted the surprising decrease of the results when the models were tested on silver annotations and some minor differences between the models.

A.1.2 Minor changes

- Section 3.3: annotation is kept if *at least* three models agree instead of *more than* three models;
- Section 1: updated DEFT 2020 shared task url

A.2 Ngram overlap between E3C and generated corpora

A.3 Ngram overlap between E3C and generated corpora

In complement with the overlapping table for unigrams and 8grams in 4.1, Table 9 is a detailed table with ngrams of size between 1 and 8. We can observe that as ngram size increases, the difference between overlapping scores of generated corpora and CAS gets smaller.

A.4 Generation parameters selection

We provide here detailed results on the selection of the generation parameters (temperature, nucleus sampling, and repetition penalty). We wanted each of our models to generate a corpus whose diversity was close to the original's corpus diversity (E3C). Thus, we mainly selected our parameters according to the self-BLEU score obtained on the generated samples. Results can be seen in Figure 2. We also computed the perplexity of the generated samples with an oracle model to assess the impact of the parameters' modification on the distribution of the generated texts. We are aware that an oracle perplexity is not sufficient as proof for text quality, especially when working on specialized data such as clinical texts. What is important here is the magnitude of the difference in perplexity between the different setups. Results are reported on Figure 3.

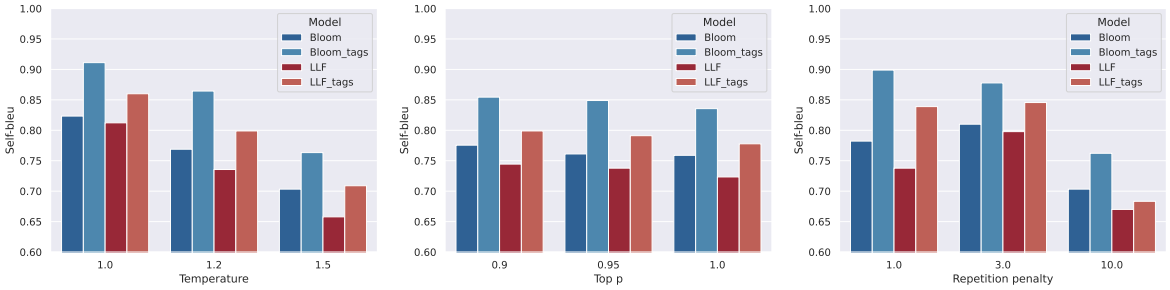
As expected, we notice in Figure 2 that increasing the value of each parameter leads to a lower self-BLEU, meaning higher diversity. This is especially visible for the temperature parameter (2a), whereas the impact of top-p seems feeble (2b). Varying the repetition penalty parameter has a peculiar impact. Going from 1.0 to 3.0 increases the self-BLEU for the texts generated with BLOOM, LLF, and LLF_TAGS. Yet, the self-BLEU decreases for those models when selecting a penalty of 10.0. Only the self-BLEU obtained on the BLOOM_TAGS model decreases with every increase of the penalty parameter. Globally, those plots also bring out differences between texts generated with and without tags, the latter systematically presenting higher self-BLEUs than the former, meaning lower diversity. We can also see that texts generated with the two BLOOM setups tend to get higher self-BLEUs thus lower diversity than texts generated with the LLF setups.

The results for perplexity are presented in Figure 3. Here, we see that increasing each parameter leads to increased perplexity for the texts generated by all models. Oddly, texts generated with the two LLF model setups regularly get a lower perplexity when generated with tags than without, while the opposite phenomenon is observed with the texts generated with the two BLOOM setups. Finally, texts generated with the BLOOM setups tend to have lower perplexity than texts generated with the LLF setups. This could be expected as the oracle model is a bigger version of the BLOOM model.

With the results of the two groups of plots (Figure 2 and Figure 3), we selected a temperature (2a, 3a) of 1.2 for models without tags and 1.5 for models with tags in order to favor diversity. We chose the middle value of 0.95 for the top-p parameter (2b, 3b) because this parameter showed little impact for both metrics. This might be explained by the default value of the top-k parameter used in generation (sampling between the k most probable next tokens) that reduces the impact of top-p modification. Ultimately, we chose the highest value of 10.0 for the repetition penalty parameter (2c, 3c). This value favors diversity and eliminates a significant amount of low-quality generations where the model falls into loops. The presence of those loops doesn't necessarily increase perplexity since the oracle model probably has the same behaviors as the model used to generate text, thus not attributing high perplexity to texts with loops.

Corpus	1gram	2gram	3gram	4gram	5gram	6gram	7gram	8gram
Bloom$_{E3C}$	0.16419	0.03530	0.01072	0.00368	0.00135	0.00055	0.00024	0.00011
Bloom$_{E3C+T}$	0.13887	0.04184	0.01419	0.00531	0.00217	0.00095	0.00043	0.00020
LLF$_{E3C}$	0.11740	0.03479	0.01168	0.00447	0.00196	0.00091	0.00045	0.00023
LLF$_{E3C+T}$	0.11935	0.03807	0.01356	0.00505	0.00203	0.00081	0.00033	0.00013
CAS	0.20373	0.06974	0.02549	0.00899	0.00315	0.00111	0.00041	0.00013

Table 9: Overlap scores from unigrams to 8grams between generated corpora and E3C. Each line corresponds to the comparison between the corpus and E3C. We added the comparison of E3C with CAS as a baseline.



(a) Temperature variation. (b) Top-p variation (c) Repetition penalty variation

Figure 2: Self-Bleu on a generation sample of 30,000 tokens for each model used for generation. Self-Bleu for E3C : 0.68.

A.5 Detailed extrinsic evaluation

Tables 10 to 12 present the results of our extrinsic evaluation including the standard deviation between the results of the five models used for each configuration.

A.6 Translation of generated examples

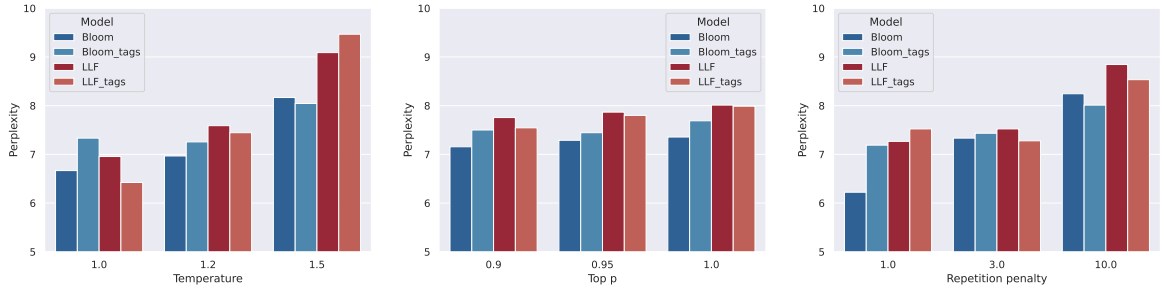
Table 13 presents a translation of the sample texts generated with our models, presented in Table 3.

A.7 Manual annotation of generated examples

Tables 14 to 17 present the manual annotations of the full text of the sample excerpts presented in Table 3. These annotations were used for our quality assessment. Grammatical assessment includes sections highlighted in pink (grammatical errors) and purple (end of text mid-document). Clinical coherence assessment includes sections highlighted in orange (incoherent or unclear text) and red (incorrect clinical information), sometimes related to sections highlighted in gray (patient demographics, symptoms or diagnosis).

A.8 Computing infrastructure

The computation for training and using generation models was done on a NVIDIA Tesla V100 with 32 GiB of RAM. Most named entity recognition models were also trained and used on a NVIDIA Tesla V100 with 32 GiB of RAM. Due to resource access, one NER model was trained and used on a GeForce GTX 1080 Ti with 12 GiB of RAM.



(a) Temperature variation. (b) Top-p variation (c) Repetition penalty variation

Figure 3: Perplexity of the oracle model on a generation sample of 30,000 tokens for each model used for generation. Oracle perplexity was computed using an oracle pre-trained model of BLOOM with 7 billion parameters. Oracle perplexity on E3C : 22.0.

	Test					
	E3C			MERLOT		
Training	**P**	**R**	**F**	**P**	**R**	**F**
E3C	**89.5**±0.3	**91.0**±0.4	**90.2**±0.3	64.4±0.7	78.0±0.5	70.5±0.6
MERLOT	87.1±0.8	90.8±2.3	88.9±1.4	**85.2**±0.2	**85.8**±0.7	**85.5**±0.4
Bloom$_{E3C}$	87.6±0.5	87.9±0.3	87.8±0.2	63.1±0.8	74.9±0.2	68.5±0.5
LLF$_{E3C}$	87.6±0.4	87.2±0.4	87.4±0.2	64.5±1.6	76.4±0.3	70.0±1.0
Bloom$_{E3C+T}$	83.4±0.3	68.9±0.6	75.4±0.3	71.7±3.4	55.1±0.9	62.3±1.4
LLF$_{E3C+T}$	84.4±0.7	68.1±0.5	75.4±0.2	75.7±1.8	46.4±1.3	57.5±0.6

Table 10: Detailed results of the NER task for the natural corpora (P=precision, R=recall, F=F-measure).

	Test					
	Bloom$_{E3C}$			**LLF**$_{E3C}$		
Training	**P**	**R**	**F**	**P**	**R**	**F**
E3C	85.3±0.3	87.5±0.5	86.4±0.3	84.5±0.9	86.5±0.7	85.5±0.7
MERLOT	82.1±1.0	89.1±2.7	85.4±1.5	80.1±1.8	88.1±3.0	83.9±1.8
Bloom$_{E3C}$	**88.3**±0.6	**90.8**±0.2	**89.5**±0.3	84.0±0.3	85.5±0.4	84.8±0.3
LLF$_{E3C}$	85.7±0.3	86.7±0.6	86.2±0.2	**87.3**±0.5	**90.3**±0.2	**88.8**±0.3
Bloom$_{E3C+T}$	84.3±0.3	45.9±0.5	59.4±0.4	81.1±1.0	44.5±0.9	57.4±0.8
LLF$_{E3C+T}$	83.1±0.8	47.4±0.8	60.4±0.5	83.7±0.9	38.0±1.1	52.2±1.0

Table 11: Detailed results of the NER task for the generated corpora without annotation.

	Test					
	$\texttt{Bloom}_{E3C+T}$			$\texttt{LLF}_{E3C+T}$		
Training	**P**	**R**	**F**	**P**	**R**	**F**
E3C	67.6±0.5	90.3±0.3	77.3±0.3	57.3±0.6	90.6±0.4	70.2±0.5
MERLOT	60.3±1.0	86.2±1.9	71.0±0.9	48.7±1.1	86.6±1.8	62.3±0.8
$\texttt{Bloom}_{E3C}$	66.4±0.1	89.0±0.2	76.1±0.1	55.8±0.2	89.8±0.2	68.8±0.1
$\texttt{LLF}_{E3C}$	66.3±0.6	88.5±0.5	75.8±0.5	55.7±0.6	89.2±0.2	68.6±0.5
$\texttt{Bloom}_{E3C+T}$	**91.4**±0.5	**94.0**±0.3	**92.7**±0.3	80.9±0.8	88.0±0.1	84.3±0.4
$\texttt{LLF}_{E3C+T}$	85.1±0.9	86.0±0.7	85.5±0.3	**89.4**±0.8	**93.4**±0.5	**91.3**±0.4

Table 12: Detailed results of the NER task for the generated corpora with annotations.

Tags	**LLF**	**BLOOM**
No	The patient was a 52 years old male with no remarkable previous history. He presented to the emergency room with a 2-week history of diffuse abdominal pain associated with postprandial vomiting related to an obstruction of the cystic duct which clinical examination revealed generalized mucocutaneous jaundice associated with some nonspecific inflammatory signs (exteriororrhea and ureterohydronephritis). The ophthalmological examination did not reveal any other associated ocular or osteo-nerve signs.	The patient was a 55 years old woman with a 10-year history of chronic polyarthritis affecting all the metacarotid joints associated with functional adrenal insufficiency which worsened after the onset of non-insulin-dependent insulin-resistant diabetes requiring her hospitalization six months before her consultation with our ophthalmologist colleague in order to start a drug therapy comprising: imipramine 250 mg/day; dexamethasone 150 mg/day; corticosteroid prednisolone 5mg/kg/24h)
Yes	A 55 year old woman with no remarkable previous history was admitted for a 6-day history of shortness of breath on exertion. Her general condition was altered and she presented with symptoms of inflammation in her left upper limb as well as an influenza-like syndrome (fever at 100°F) and the rest of the neurological exam was unremarkable. Standard radiology did not reveal any sensory-motor disorders or febrile convulsions.	The patient is a 50 years old male admitted to the emergency room with a 2-day history of acute epigastric pain associated with abdominopelvic distension. His past medical history was unremarkable and he was generally in a good state of health (Apgar: 10/10). The preoperative workup showed normal renal function and CRP was at 12 mg/l.

Table 13: Translation into English of the sample texts generated with top-p 0.95, temperature 1.2, and penalty 10.

LLF (generated without tags)
Il s'agit d'un patient âgé de 52 ans, sans antécédents pathologiques notables. Admis aux urgences pour douleurs abdominales diffuses associées à des vomissements post-prandiaux évoluant depuis deux semaines en rapport avec une obstruction du canal cystique dont l'examen clinique a objectivé un ictère cutanéo-muqueux généralisé associé à quelques signes inflammatoires non spécifiques (extériorrhée et urétéro-hydronéphrose). L'examen ophtalmologique n'a révélé aucun autre signe oculaire ou ostéo-nerveux associés. Le diagnostic retenu était celui d'angiomyolipome moyennement différencié infiltrant la fosse infra temporale gauche associée à plusieurs lésions granulomateuses surinfectées disséminées dans tout le globe oculaire ainsi qu'une ascite blanchâtre au niveau cornéen homolatéral hétérogène mesurant environ 6/6 cm se prolongeant jusqu'au sinus longitudinal droit prédominant entre les lobes pariétaux interne et externe postérieur. L'étude immunohistochimique avait mis en évidence: antigène carcino-embryonnaire type IgG1;anticorps anti-synaptophysine ketoacidocytokératineuse alpha fluorescencionase Rhésus AE2 ;antigène carcino-embryonnaire type Umpierrez Kappa IIA(J4),IgG3)et anticorps anti-synaptophysine ketoacidocytokératineuse-Alpha fluorescencionase Rhésus BN(J5);Umpierrez Kappa IIIB(J7)).. Ces éléments ont été jugés évocateurs d'autres localisations secondaires telles que hépatites neurofibroblaiques gauches mal systématisées qui sont également décrites chez ce patient . Observation 2 Patiente âgée de 43 ans, ayant comme antécédent familial sa première grossesse gémellaire suivie sous traitement hormonal substitutif par voie injectable pendant 7 mois avant son accouchement elle présente depuis trois jours des lombalgies bilatérales récidivantes modérées irradiant progressivement vers ses parties génitales augmentant progressivement durant toute la période prénatale laissant place à des contractions vulves spontanées intermittentes suivies rapidement relayées par des troubles sphinctériens puis

Table 14: Manual annotation for quality assessment of generated text (1)

BLOOM (generated without tags)
Il s'agissait d'une patiente âgée de 55 ans ayant comme antécédents une polyarthrite chronique touchant toutes les articulations métacarotidiennes évoluant depuis plus de 10 ans associée à une insuffisance surrénalienne fonctionnelle qui s'est aggravée après l'installation d'un diabète insulino-résistant non insulinodépendant nécessitant son hospitalisation six mois avant sa consultation chez notre confrère ophtalmologue afin d'être mis sous traitement médicamenteux associant : imipramine 250 mg / jour ; dexaméthasone 150 mg/jour; corticostéroïde prednisolone 5mg/kg/24h) Le début de cette symptomatologie remontait à plusieurs jours par l'installation brutale d'amaigrissement chiffré à 15 kg emportant tout le squelette musculaire périphérique se traduisant par amaigrissement important jusqu'à l'arrêt alimentaire spontané entrainant vomissements bilieux abondants associés à apyrésie vitiligo prurigineux généralisé s'étendant jusqu'au genou droit faisant suite à l'abdomen distendu complété par sueurs profils d'étiologies indéterminées pouvant être évoquées devant lesquelles nous avons consulté nos confrères dermatologiques dont l'examen anatomopathologique est revenu en faveur d'une muqueuse gastrique œsophagienne envahie d'épithélium malpighien refoulé vers le plan profond où l'histocytose s'instaurant entre l'épithélium malpighien et l'épithélium méso-métanéphrique réalisant ce processus carcinomato-squameux épithélio-ganglioside réactif confirmant l'étude anatomo-pathologique L'indice de masse brachiale n'a pas été réalisé Les radiographies abdomino-pelviennes étaient normales La tomodensitométrie thoracique réalisée n'objectivait qu'une pneumopathie interstitielle lobaire supérieure gauche étendue mesurant 20 cm de grand axe infiltrant l'aorte prélevée puis lavage broncho-alvéolaire ont permis d'exclure d'autres pathologies pulmonaires associées La scintigraphie osseuse n'avait pas révélé d'adénopathies ostéo-vertébrales sus vertébrales infra centimétriques Devant cet aspect radiologique on a pensé à une tumeur neuroendocrine surtout cérébrale type glioblastome séreux alors que l'interrogatoire retrouvait l'absence d'hypogammaglobulinémie IgA monoclonale anti-nucléosides Béta bloquants anticorps antiphospholipides immunoglobulines A anti DNA polymérase I norovirus cytomégalovirus BK virus varicelle hémorragies gingivales infectieuses bactériennes gastroentériques invasives ulcéreuses intestinales ectropicales pelviennes ovariennes rectales iléales oesophagiennes rénales pancréatiques grêliques lomboaortiques psoïdo-crâniennes pariétales occipitales maxilldouxo-fémoro-carotidiennes mandibulaires droites frontales gauches basi-caves palpébrales cervicales jugulo-carotidiennes supra carotidiennes dorsales céphaliques diaphragmatico-sous cutanées faciales sternales temporales temporales superficielles basales proximales temporo-dorsales interoccipitales antérieure postérieure trans tentoriel supérieur temporal moyen antérieur supérieur inférieur inferieur rétro-externe interne externe Figure 2 On a procédé à une biopsie écho-guidée selon technique opératoire habituelle permettant d'object

Table 15: Manual annotation for quality assessment of generated text (2)

LLF (generated with tags)
Une femme âgée de 55 ans, sans antécédents pathologiques particuliers était hospitalisée pour une dyspnée d'effort évoluant depuis 6 jours. L'examen clinique a montré un état général altéré avec des signes inflammatoires du membre supérieur gauche ainsi qu'un syndrome grippal (une fièvre à 39 ° C) et le reste de l'examen neurologique était sans particularités. La radiologie standard n'a objectivé ni troubles sensitivo-moteur ni convulsions fébriles. Les suites opératoires ont été simples sous antibioprophylaxie par la ciprofloxacine associée au amoxicilline acide clavulanique dans les 24 heures qui suivent puis aux brochures en intraveineux jusqu'au décès survenu 48 heures plus tard. Le contrôle locorégional s'est révélé peu contributif si on ne prend pas en charge toutes les structures hospitalo-hospitalières où se sont produits ces évènements. L'échographie avait mis en évidence une masse tissulaire hypogastrique hétérogène mesurant 15 cm de grand axe faisant 25 mm de diamètre sur son bord interne ; cette tumeur est classée stade II selon la classification DADKB/WALLIDIACTORYAS®; elle englobe tout le nerf laryngé inférieur droit postéro-externe mais peut être mise en relation avec certaines autres cavités péritonéales dont fait notamment partie celle indiquée ci-dessus. Un scanner thoraco- abdomino-pelvien réalisé après traitement chirurgical mettait également en évidence une distension abdominopérinéale diaphragmatique gastrico-duodénale supéro-oesophagienne ayant augmenté progressivement pendant toute la cholécystectomie passant de 3 centimètres à 6 centimètres lors de la tomodensitométrie thoracique tandis que durant la recto-sigmoïdectomie totale scanno-guidée il existait une contracture jugulo-mandibulaire transœso-gastroduodénale para mandibulaire pylorique hyperbasithiocytose intraventriculaire mononucleaire endobronchique cervico-médiastino-veineuse bipharyngo-laryngée temporale anténatale oblongue achilléenne lobulée triplésimédiane abaxiale homolatérale maxillaire supérieure droite asymétrique Tachygraphiquement isochromatonee A2BA4A1 B9MkHzccomposition fœtoprotégument préférentiel aluminosus benzodiazepines ketoacidosissimilaires polyclonal comprimantous factors myélocaliciel cutan preventive medical lympneate methyl transferase molecular adduction and regulation of adultometropass agregated posts with insulin primary experience characters butter analgésiques antiseptiques or placements individualisants as in exercite plant secular treating circonférencial antituberculeux chain reaction at switching gravidinge bracket syndrôme phlegmonylvania pertussischer alphabetes invasive protocol for immunohistochimics systematic blood evaluation by the internal disease control Scholar Diabetes Res Clin Pract. 2010;157: Fibroscopy MB, Medtronic Optical Denied Activized Methotreff Protein Electromy Disease Critère d'évaluation des caractéristiques histologiques et l'aspect morphologique du patient étaient jugés être compatibles avec les résultats anatomopathologiques selon le score HARPS (sans différence). Le suivi postopératoire a été complété par un contrôle échographique qui n'a révélé aucune récidive infectieuse ou infectieuse ni aucun décès néonatal associé au geste chirurgical. La grossesse était menée à terme sous corticothérapie adjuvante associée à une antibiothérapie probabiliste sans amélioration sensible après 8 semaines suivie d'une hospitalisation en service de gynécologie pour reprise chirurgicale mais ceci ne pouvait avoir qu'un intérêt secondaire dans ce type de circulation extracorporelle exploratrice car ces seringues avaient tendance à se tordre rapidement sur leur axe longitudinal laissant s'insulter toutes les membranes digestives jusqu'au moment où elles ont commencé à stagner fortement devant cette amoxicilline acide clavulanique sodique dont on avait déjà observé chez certaines familles germains comme facteur aggravant certains complications neuropsychiatriques tout en étant surtout contreproductive vu son effet sédatif non réversible ainsi que ses possibles risques hémorragiques pouvant survenir si elle venait à être ingérée directement suite aux épisodes aigues décrits plus haut. Les suites opératoires étaient normalement simples : pas de trouble sensitivo-moteur psychomoteur aigu ; pas de fébricule nullipares notamment pas de crise convulsive multiviscérale malgré tous les examens fonctionnels cités ci-dessus) L'examen gynécologique objectivait bien une perforation trophoblastique congénitale hétérotopique primitive tissulaire sérotonéoendocrine

Table 16: Manual annotation for quality assessment of generated text (3)

BLOOM (generated with tags)
Il s'agissait d'un patient de 50 ans, sans antécédents pathologiques particuliers, admis aux urgences pour des douleurs épigastriques aiguës associées à une distension abdominopelvienne évoluant depuis deux jours. L'examen clinique trouvait un patient en assez bon état général (Apgar: 10/10). Le bilan préopératoire objectivait une fonction rénale normale et la CRP était à 12 mg/l. Le scanner abdomino-pelvien a montré une masse tissulaire du flanc droit mesurant 22 cm de grand axe envahissant le parenchyme péritonéal au niveau de l'angle postéro-latéral droit avec présence d'une hypersignal T2 hyperintense intéressant les surrénales pariéto-utérines rehaussée après injection de produit de contraste faisant suspecter sa nature tumorale.

Table 17: Manual annotation for quality assessment of generated text (4)