Using Structured Health Information for Controlled Generation of Clinical Cases in French

Hugo Boulanger, Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol

Using Structured Health Information for Controlled Generation of Clinical Cases in French

Hugo Boulanger[*‡], **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, [‡]firstname.lastname@cea.fr, [⋆]karen.fort@loria.fr

Abstract

Text generation opens up new prospects for overcoming the lack of open corpora in fields such as healthcare, where data sharing is bound by confidentiality. In this study, we compare the performance of encoder-decoder and decoder-only language models for the controlled generation of clinical cases in French. To do so, we fine-tuned several pre-trained models on French clinical cases for each architecture and generate clinical cases conditioned by patient demographic information (gender and age) and clinical features. Our results suggest that encoder-decoder models are easier to control than decoder-only models, but more costly to train.

1 Introduction

The performance of current text generation models makes it difficult for humans to distinguish between natural and synthetic text (Casal and Kessler, 2023), paving the way for a wide range of applications including data augmentation and addressing resource sparsity (Claveau et al., 2021). In this article, we consider the case of reference documents that cannot be shared because of the personal information they contain but are sufficiently generic to mutualize processing resources on a community scale. One way of developing shared processes is to work with synthetic documents that are comparable in content and style to reference documents. We focus on electronic health records, though our methods can be applied to other fields with document-sharing constraints due to privacy.

Creating relevant synthetic documents is not trivial and must take several dimensions into account. As mentioned before, synthetic documents should be comparable to reference documents in terms of style, structure, and content, without leaking personal information that may be contained in the training corpora. While directly identifying information can be subject to robust upstream de-identification, this does not make documents *anonymous* according to the definition of the General Data Protection Regulation (GDPR). Indeed, de-identification, whether automatic or manual, does not prevent cross-referencing medical information, which can particularly impact privacy for rare diseases.

It is possible to leverage the abilities of current text generation models to generate synthetic documents. However, such models are not as efficient when it comes to specialized domains such as the medical domain, even more so in languages other than English. Thus, the ability to precisely control the generation process is important both for medical consistency and for preserving the privacy of the information contained in real texts.

In this article, we propose a methodology for controlling text generation in terms of content. More specifically, the goal is to condition the generation of medical reports on patient profiles. Following the example of work carried out on the generation of synthetic patient profiles in terms of structured data (Walonoski et al., 2017), these profiles take the form of a set of medical concepts. This approach, which is part of a data-to-text generation problem, has the advantage over a textual priming approach of being able to finely control the information used for conditioning. The latter is implemented by training a neural language model with a set of pairs, each composed of a patient profile in the form of concepts and a reference report corresponding to this profile. Within this framework, the contributions of our paper are as follows:

- a method for controlling the content of medical report generation;
- a method for creating a training set for carrying out this control;

[*]These authors contributed equally to this work. The order is alphabetical.

- an implementation of the strategy using language models with two different architectures[1];
- an automatic multidimensional evaluation of synthetic text.

2 Related Work

2.1 Controlled text generation

Since the advent of the first large language models (LLMs) such as those of the GPT family (Radford et al., 2018), generating text resembling human production seems easy and the problem of generation has evolved to change focus: the aim is no longer simply to generate plausible text but to be able to control more finely what we generate. The texts produced by generative models may be irrelevant, offensive, or even dangerous (Bender et al., 2021). This is why a significant amount of work is being done on generation control. Control can concern several aspects of generation, such as the lexicon or text style (Zhang et al., 2023). Several control methods have been explored, including training a model with examples conditioned according to chosen criteria (Keskar et al., 2019) or modifying the probabilities of output tokens during inference (Kruszewski et al., 2023).

The *data-to-text* (Lin et al., 2023) approaches constrain generation from structured data (graphs, tables, and, in our case, *slots*). The preferred architectures are encoder-decoder models, which can have a variety of internal architectures, combining pre-trained models as encoders and/or decoders. It is also possible to directly fine-tune encoder-decoder models, such as the T5 model (Raffel et al., 2020). Causal language models, such as those using a Transformer (Vaswani et al., 2017) decoder architecture, use the context at the start of a sequence to generate the rest of the sequence.

2.2 Biomedical text generation

In the biomedical field, text generation is being explored either to facilitate the work of doctors or to address resource sparsity due to confidentiality issues. This work falls into the second category.

Earlier methods focus on training neural models from scratch. Melamud and Shivade (2019) train an LSTM to generate shareable clinical notes using differential privacy (Dwork et al., 2006), and Ive et al. (2020) train a Transformer encoder-decoder model to generate synthetic mental health records conditioned by entities automatically extracted from real documents. However, training a model from scratch requires a substantial amount of data that is not available in languages other than English (Névéol et al., 2018).

Several efforts exploit Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to generate structured data in the medical domain in English (Choi et al., 2017; Abedi et al., 2022; Torfi et al., 2022).

More recently, text generation is being explored to produce reports of discussions between doctors and patients, with the encoder-decoder architecture often being preferred (Eremeev et al., 2023; Ben Abacha et al., 2023; Asada and Miwa, 2023).

For French, Hiebel et al. (2023) fine-tune pre-trained auto-regressive language models to generate clinical cases with no particular constraints and propose a methodology for automatically evaluating the utility of the synthetic texts for a clinical entity recognition task.

3 Overall Method

As outlined in the introduction, we cast the task as a data-to-text generation problem where structured health data is used to shape the contents of synthetic text. Of course, finding the conditioning data within the generated texts cannot be the only criterion for evaluating the models: they would only need to reproduce their input to be judged as perfect. This conditioning must therefore be close in nature to the reference documents we wish to emulate.

As mentioned in section 2.1, this double conditioning can be achieved either by fine-tuning the language model used for generation with control elements or by steering the model during inference. We have opted for the former solution, as the latter implies applying elaborate text analysis processes during generation to check compliance with the conditioning, which is costly. The first solution, however, presupposes the availability of training data combining conditioning data and example texts conforming to this conditioning.

To this end, we have adopted a strategy comparable to Peng et al. (2018) for story generation, taken over by Ive et al. (2020) for medical reports, and consisting in automatically extracting the conditioning data from the example texts. This strategy

[1] https://github.com/HugoBoulanger/ClinicalGenerator

obviously presupposes the availability of text analysis tools capable of extracting this conditioning data from example texts with a sufficiently high level of performance. It therefore requires a close coupling between generation and analysis capabilities but eliminates the need for costly manual annotation. In the present case, we are focusing on medical concepts and are therefore dependent on models for extracting these concepts from medical reports but the genericity of this strategy means that new conditioning elements can easily be taken into account, as long as they can be automatically extracted from example texts.

4 Material and Methods

4.1 Clinical case corpora in French

The data used for our experiments come from two freely available clinical case corpora. The first corpus is the CAS corpus (Grabar et al., 2018), a corpus of de-identified clinical cases in French[2]. The second corpus is the E3C corpus (Magnini et al., 2020), a multilingual corpus of de-identified clinical cases. Our study only uses the clinical cases in French.

4.2 Defining constraints based on a patient profile

Our goal is to generate consistent clinical cases by controlling the generation using clinical elements. We have worked with clinicians to define the salient features of real clinical cases. These features are then used as constraints to generate text. Table 1 shows an example of features that were selected for a clinical case of the E3C corpus. These include patient demographics (age and gender), pathology location, histological information, various signs or symptoms, treatments and procedures performed, lab results, and scores (measures or codes). In line with clinicians' recommendations, we identify around twenty constraints per case, selecting if possible elements from each category with a majority of symptoms, treatments, and procedures. This approach ensures the selection of the salient information from the clinical cases, according to the doctors.

4.3 Extracting constraints from documents

Demographic data for the CAS corpus was directly taken from the existing corpus annotations for patient age and sex. We manually annotated the 1,009 cases from the E3C corpus to obtain equivalent demographic information for this corpus. Other clinical entities (e.g., signs and symptoms, procedures) were obtained by automatically annotating the two corpora consistently using clinical entity recognition models trained on the MERLOT private corpus (Campillos et al., 2018), which contains manual annotations for the entities of interest.

Constraint sets thus include manually annotated demographic information and automatically extracted clinical entities. For each document, we select age and gender when available. When the exact age is not provided, we use the age categories derived from the MeSH (Medical Subject Headings) thesaurus[3] check tags.

Clinical entities are selected from the MERLOT annotation categories that match the categories discussed with the doctors. For each clinical case, we select the ten procedures (*PROC*) and ten symptoms (*DISO*) with the highest tf.idf score. We also select substances (*CHEM*) and measures (*MEAS*). The latters are filtered to retain only informative measures (single digits such as *6* are annotated as *MEAS* but without additional information). Overall, we obtain an average of 26 constraints ($\pm$9.5) per clinical case.

4.4 Text generation models

We compare the performance of two different architectures for the constrained generation of clinical texts using encoder-decoder vs. decoder-only pretrained Transformer models.

Encoder-decoder This architecture aims to generate text from structured data. In particular, fine-tuning the T5 model has become a standard method for data-to-text tasks. We chose to use the multilingual version of T5, called mT5 (Xue et al., 2021), with one billion parameters as a pre-trained model, and the Small (77 million parameters), Large (780 million parameters), and XL (3 billion parameters) versions of Flan-T5 (Chung et al., 2022) as models fine-tuned with instructions.

Decoder only This architecture aims to generate text from textual prompts. We have chosen several models for this architecture. The Bloom (Scao et al., 2022) model, a generative model trained on several languages, and the Bloomz model, a variant

[2]Corpus can be accessed with permission from the authors https://deft.lisn.upsaclay.fr/2020

[3]https://www.nlm.nih.gov/bsd/indexing/training/CHK_030.html

Type of clinical feature	Sample value
Age	54
Sex	Masculin
Localisation	Vessie
Histology	adénocarcinome de l'ouraque peu différencié
Sign	hématurie
Procedure	scanner CT
Treatment	chimiothérapie par Méthotrexate-Vinblastine-Endoxan-Cisplatine
Score	T III A (selon la classification de Sheldon)
Bio	une négativité pour les cytokératines (ck) 7 et 20

Table 1: Sample control data based on manual analysis of a clinical case. The source case is shown in Table 2. We show in Appendix A.1 an English version based on the automatic translation of the document (Tables 5 and 6).

specially trained to perform different tasks (translation, automatic summarization, etc.). For each of these two models, we consider two versions in terms of size: one billion and seven billion parameters.

5 Experiments

5.1 Structured data representation

The use of these generative models requires the conversion of structured data into text format. We have chosen to linearize the inputs differently for the encoder-decoder models and the decoder-only models. For the encoder-decoder models, a special token representing the entity type is added before each entity. We separate demographic information (age, sex) from medical constraints (symptom, procedure, etc.) with a special token *contraintes* (*constraints*). For decoder-only models, no special tokens are used. Figure 1 shows an example of data representation for encoder-decoders.

5.2 Fine-tuning

The training set used to fine-tune our models comprises 1,424 clinical cases, containing over 500,000 tokens excluding constraints. For fine-tuning, we freeze the weights of the pre-trained model and add LoRA trainable matrices (Hu et al., 2022). The location of the trainable matrices depends on the type of model. For encoder-decoder models, we add LoRA matrices on the *queries* and *values* of the Transformer layers and the model head. For decoder-only models, LoRA matrices are added to the linear layers of the models. Special tokens are added to the embeddings via randomly initialized vectors. The processing of word embeddings varies according to two configurations defined as follows:

"Frozen" configuration: embeddings are frozen but we add LoRA matrices to enable adaptation to the task at a low memory cost.

"Unfrozen" configuration: the embeddings are unfrozen, to enable adaptation to the task, but at a higher cost.

We show the total number of parameters and the number of trainable parameters for each model in Table 7 in Appendix A.2.

5.3 Automatically generating clinical cases

Our test set consists of 156 clinical cases and their constraints. The constraints are given as input to the generative models and the real clinical cases are used as a reference when computing evaluation metrics. Decoding is performed using a beam search with five beams. We use sampling with a top-p of 0.9, a temperature of 1, and a repetition penalty of 3. Using sampling means that the same model might generate different texts from the same input. We run five generations for each test example to account for this variability.

5.4 Evaluation metrics

Automatic evaluation of text generation is notoriously difficult (Novikova et al., 2017). Numerous metrics exist to measure different aspects of text generation (Frisoni et al., 2022). Our metric selection aims to cover several dimensions of evaluation.

Fit to constraints - *Accuracy* This measure is used to assess the model's ability to implement the constraints. We calculate the proportion of constraints respected in generated texts in relation to the total number of constraints imposed.

Language quality - *Perplexity* Perplexity evaluates how well the textual data matches the probability distribution of a language model. We use a

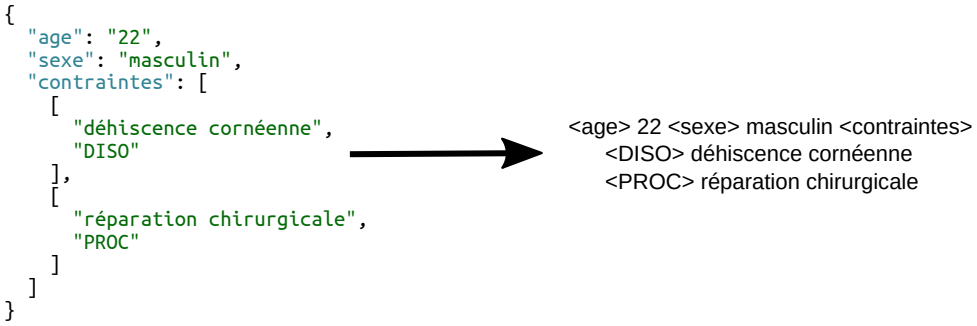

Figure 1: Example of data representation for encoder-decoder architecture (see Figure 2 in Appendix A.1 for its translation).

model specific to French, GPTFR (Simoulin and Crabbé, 2021). For this metric, we want the perplexity obtained on the generated data to be close to the perplexity obtained on the real data (equal to 19.5 for the training corpus).

Diversity of generated texts - *Self-BLEU* The Self-BLEU (Zhu et al., 2018) score is the average of the BLEU scores of all the sentences in a corpus. Thus, a redundant corpus will have a high Self-BLEU score while a varied corpus will have a lower score.

Proximity to natural corpus - *Corpus-BLEU* Corpus-BLEU (Yu et al., 2017) is a measure of proximity between two corpora and corresponds to the average BLEU score between each sentence in the generated corpus and all sentences in the natural corpus. We calculate Corpus-BLEU by comparing the clinical cases in the test corpus with the generated texts.

Proximity with the clinical case corresponding to the constraints - *BLEU* The BLEU (Papineni et al., 2002) score is calculated between the generated text and the actual clinical case from which the constraints originate. It measures proximity to real data in a more specific way than the Corpus-BLEU score.

6 Results

6.1 Evaluation of synthetic clinical cases

Table 2 shows examples of texts generated from a set of constraints by an encoder-decoder model (Flan-T5-XL frozen) and a decoder-only model (Bloomz 1b1 unfrozen). Table 3 shows the automatic evaluation of clinical cases generated with the different architectures studied. Among our baselines, the simple copy of the conditioning entities (Copy) obtains, as expected, an accuracy of 100 %, but also a very high perplexity. The Corpus baseline corresponds to a copy of the test corpus in which we have removed the line breaks. This change explains why the BLEU and corpus-BLEU scores are not perfect and, more surprisingly, reduces perplexity from 30.5 to 19.5. The accuracy score, meanwhile, reveals the limitations of our data and accuracy calculation. The majority of these errors concern the sex of the patient, when this is not indicated by the gender agreement of the term "patient" or the use of the qualifier "male" or "female". Other errors are mainly due to rephrasing or errors in constraints.

The results show several trends. The first trend, which was expected but is confirmed by Table 3, is the positive correlation between the size of the models, both for encoder-only and encoder-decoder models, and their results: larger models obtain better results. When comparing encoder-decoder models of equal size (large), a model that has benefited from a training period with instructions, a Flan model, tends to obtain better overall results than a model pre-trained without instructions, especially for the unfrozen configuration. The Flan models also have the advantage of being fine-tuned more quickly for the same size, with a training period of 16 h for Flan-T5-large versus 60 h for mT5-large. As expected, the Flan-T5-XL models were the best-performing of the encoder-decoders tested. They generate more varied texts (Self-BLEU) and have the best accuracy. The texts generated most closely resemble the references (BLEU) and the perplexity values are better than those of the smaller versions of the model. It should be noted that mT5 models achieve lower perplexity —probably because the initial model is multilingual, whereas Flan-T5 models only saw French on translation tasks— and better Corpus-BLEU. Finally, Flan-T5 models are closer to the Corpus baseline than mT5 models in terms of perplexity, which was not *a priori* obvi-

Automatically extracted constraints	âge: 54 ; sexe: masculin ; contraintes: hématurie isolée, examen tomodensitométrique, masse, 4 cm, adénocarcinome peu différencié, de type III, bilan d' extension, cystoprostatectomie radicale totale, lymphadénectomie iliaque, obturatrice, omphalectomie, entérocystoplastie de substitution, adénocarcinome de l'ouraque peu différencié, très localement mucosécrétant, ulcéré, carcinome transitionnel, grade III, Antigène Carcino-Embryonnaire, Leu-M1, CD 15, cytokératines, épithélium vésical, classification de Sheldon, Méthotrexate, Vinblastine, Endoxan, Cisplatine
Real clinical case	Un homme de 54 ans a consulté pour hématurie isolée. Une échographie, puis un examen tomodensitométrique, démontraient une masse de 4 cm de diamètre, au centre nécrotique, antérieure au dôme vésical, envahissant uniquement la graisse adjacente (Figure 1A.). Les biopsies, faites à la cystoscopie, révèlaient un adénocarcinome peu différencié, avec des zones d'épithélioma transitionnel de type III (Figure 1B). Le bilan d' extension était négatif. Le patient bénéficiait d'une cystoprostatectomie radicale totale avec lymphadénectomie iliaque et obturatrice, associée à une omphalectomie et une entérocystoplastie de substitution selon Studer. L'examen anatomo-pathologique complet de la pièce opératoire, révèlait un adénocarcinome de l'ouraque peu différencié, invasif, très localement mucosécrétant (Figure 1C). Il était ulcéré dans la lumière vésicale, et contenait des zones de carcinome transitionnel de grade III, ainsi que des cellules argentaffines. Deux poches de lymphonoeuds et la graisse environnante, étaient envahis. L'immunohistochimie montrait un positivité des cellules néoplasiques pour l'Antigène Carcino-Embryonnaire (ACE), Leu-M1 (CD 15), et une négativité pour les cytokératines (ck) 7 et 20. L'épithélium vésical était positif pour les ck 7. Il s'agissait d'un stade T III A (selon la classification de Sheldon). Le patient subissait une chimiothérapie par Méthotrexate - Vinblastine - Endoxan -Cisplatine par voie systémique. Avec 35 mois de recul, il va bien.
Synthetic Bloomz 1b1 unfrozen	Il s'agit d'un patient âgé de 54 ans sans antécédents pathologiques particuliers qui a consulté pour une hématurie isolée. L'examen clinique trouvait un patient en bon état général, apyrétique, avec des signes d'hypertension artérielle et une sensibilité abdominale à la palpation. La biologie était normale. L'échographie abdominale objectivait une masse tissulaire hétérogène du flanc gauche mesurant 4 cm de grand axe. L'IRM montrait une masse tissulaire homogène au niveau du flanc gauche (Figure 1). L'étude histologique de la pièce opératoire confirmait le diagnostic d'adénocarcinome peu différencié de type III selon la classification de Sheldon [1]. Le bilan d'extension par résonance magnétique (RMN) ne retrouvait pas de métastase cérébrale ou pulmonaire. Une cystoprostatectomie radicale totale associée à une lymphadénectomie iliaque et obturatrice ainsi qu'une omphalectomie et entérocystoplastie de substitution ont été réalisées. L'étude anatomopathologique de la pièce opératoire avait confirmé le diagnostic d'adénocarcinome de l'ouraque peu différencié évoluant vers un carcinome transitionnel stade III selon la classification de Sheldon [1].
Synthetic Flan-T5-XL frozen	Il s'agit d'un patient âgé de 54 ans, sans antécédents pathologiques particuliers, qui a présenté une hématurie isolée évoluant depuis 2 mois. L'examen tomodensitométrique a objectivé une masse de 4 cm au niveau de la paroi postérieure de l'ouraque (Figure 1). Le diagnostic d'adénocarcinome peu différencié de type III a été retenu. Le bilan d' extension était négatif. Une cystoprostatectomie radicale totale associée à une lymphadénectomie iliaque non obturatrice et une omphalectomie avec entérocystoplastie de substitution a été réalisée. L'étude anatomopathologique de la pièce opératoire a confirmé le diagnostic d'adénocarcinome de l'ouraque peu différencié comportant un aspect très localement mucosécrétant et ulcéré. Cet aspect est compatible avec un carcinome transitionnel de grade III (Antigène Carcino-Embryonnaire: CD34, Leu-M1 et CD 15). Les cellules tumorales exprimaient les cytokératines spécifiques de l'épithélium vésical (classification de Sheldon). Un traitement associant Méthotrexate, Vinblastine, Endoxan et Cisplatine a été débuté.

Table 2: Sample clinical cases generated by two models based on constraints that were automatically extracted from a reference clinical case. An automatic English translation is shown in Table 6 in Appendix A.

ous since an instructed-based language model is not necessarily the best starting point for training a base text generator. This is particularly true for the Flan-T5-small models, without an evident explanation.

We observe that encoder-decoder models perform better than decoder-only models. Decoder-only models are also more unstable from one generation to another, with large standard deviations in Accuracy, Self-BLEU, and Corpus-BLEU, especially for the smallest models. In terms of perplexity, these models achieve lower scores and thus,

	Generation method	Accuracy↑	Perplexity	Self-BLEU-4↓	Corpus-BLEU-4↑	BLEU-4↑
Baselines	Copying constraints	100	194.3	14.4	25.5	1.1
	Copying natural corpus	98.8	19.5	33.4	97.4	97.5
	Bloom 1b1 frozen∗	s/o	11.5±1.5	86.1±0.4	64.8±0.4	s/o
	Bloom 1b1 unfrozen∗	s/o	10.2±0.9	82.9±0.4	60.6±0.5	s/o
	Bloom 7b1 frozen∗	s/o	8.4±2.8	79.3±1.2	57.1±0.5	s/o
Encoder-decoder Encoder-decoder	mT5-large frozen	78.0±0.6	13.6±0.2	53.5±0.5	55.8±0.5	12.0±0.1
	mT5-large unfrozen	73.6±0.8	13.4±0.2	53.8±0.4	56.4±0.3	10.9±0.2
	Flan-T5-small frozen	61.6±0.4	18.9±0.4	49.1±0.4	47.1±0.4	6.6±0.1
	Flan-T5-small unfrozen	61.5±0.7	17.6±0.5	51.2±0.4	50.0±1.4	7.0±0.2
	Flan-T5-large frozen	81.5±1.1	14.8±0.4	52.8±0.4	55.3±0.4	12.0±0.1
	Flan-T5-large unfrozen	80.3±1.0	15.6±0.5	51.9±0.2	55.0±0.4	11.7±0.2
	Flan-T5-XL frozen	84.2±0.8	14.9±0.2	50.2±0.2	54.5±0.2	12.8±0.1
	Flan-T5-XL unfrozen	85.3±0.8	14.9±0.2	49.0±0.1	53.8±0.4	12.9±0.2
Decoder	Bloom 1b1 frozen	40.5±3.9	8.8±0.2	62.5±5.8	42.3±11.1	4.7±1.0
	Bloom 1b1 unfrozen	29.6±0.9	9.3±0.4	63.6±4.7	50.4±9.7	4.0±0.5
	Bloom 7b1 frozen	43.5±2.5	9.9±0.6	54.0±2.1	47.5±2.0	5.8±1.0
	Bloomz 1b1 frozen	45.4±4.2	9.2±0.2	61.9±7.6	41.8±11.0	5.2±1.3
	Bloomz 1b1 unfrozen	32.1±1.7	9.6±0.2	65.7±6.0	47.0±13.2	4.3±0.7
	Bloomz 7b1 frozen	39.8±3.0	9.9±0.2	55.0±1.9	49.8±1.5	5.4±0.4

Table 3: Evaluation of synthetic text generated from the constraints of the test set. Baseline models marked with "∗": training and generation without constraints.

deviate from the training corpus. As the model used to calculate perplexity is also a decoder, the common architecture potentially biases the decoders for this metric. On the other hand, decoder training time is much shorter: 10 to 15 minutes for billion-parameter models and 30 minutes for seven-billion-parameter models.

We can also identify some good practices regarding model pre-training and word embedding configuration. Models that have benefited from fine-tuning with instructions perform better overall than models with pre-training on a language modeling task. This is mainly true for accuracy and the BLEU score. We can assume that the type of instructions used for this fine-tuning – more precisely, whether these instructions are directly related or not to text generation tasks – may have an influence on the performance of these models but this analysis is beyond the scope of this article. We can also observe that frozen models perform better than unfrozen models. This observation could be considered surprising since the unfrozen models are supposed to have better adaptation capabilities but their heterogeneity in terms of parameters (LoRA and word embeddings matrices) is perhaps the source of these results.

6.2 Environmental impact

Model	Fine-tuning	Generation	Perplexity	Total
mT5-large	4.84	0.5	0.01	5.35
flan-T5-small	0.76	0.08	0.01	0.85
flan-T5-large	1.3	0.5	0.01	1.81
flan-T5-XL	4.84	0.5	0.01	5.35
Bloom(z) 1b1	0.03	0.78	0.01	0.82
Bloom(z) 7b1	0.05	0.64	0.01	0.70

Table 4: Environmental impact of the final experiments for each model, in $kgCO_2e$. Each line sums the emissions for different associated configurations. The total emissions reach 14.87 $kgCO_2e$.

Table 4 presents the greenhouse gas emissions of the experiments in terms of $kgCO_2e$. The environmental impact is essentially linked to the training of encoder-decoder models, which takes longer and requires more GPUs for larger models. These estimations were computed using the Machine-Learning Impact calculator presented in (Lacoste et al., 2019) with emission values for France (0.101 $kgCO_2e/kWh$) found in (Moro and Lonza, 2018).

7 Conclusion

In this study, we generate French clinical cases conditioned on structured clinical data. We com-

pare models with different architectures, encoder-decoder and decoder-only, which we fine-tune on a corpus of clinical cases using LoRA matrices. We propose an evaluation methodology based on a set of automatic measures: accuracy, perplexity, Self-BLEU, Corpus-BLEU, and BLEU. We observe that models with encoder-decoder architecture achieve better results on the task of generation from structured data, but with more costly training. Our experiments suggest that the best training strategy is to add LoRA matrices to the word embeddings rather than unfreezing them, although this does lengthen training.

The computing power available in a hospital setting limits the possibility of using larger and/or heavier models. The smallest size encoder-decoder model, Flan-T5-Small (77 million parameters), fits on the smaller Nvidia P6000 GPUs for fine-tuning and inference and obtains better performances than the larger decoder models. Small encoder-decoder models should be used if this type of resource is available for multiple hours. Decoders are more suitable if time on the GPUs is limited. However, it would be necessary to generate several candidates and filter them to compensate for the irregularity of these models.

Quantization might also be a solution for lightening computational loads, provided that quantized models achieve comparable results to their regular counterparts.

7.1 Limitations

The set of measures we have put in place gives us a fairly good view of what our models generate. There are, however, limits to using only accuracy, especially as calculated, to describe the fidelity of information transcription. Accuracy here seeks an exact match between the constraints and the text. Any reformulation of the model is therefore discarded, even though it may be correct. Moreover, using this measure alone does not give us any information on potential additions of information or entities by the models. In this study, we have exclusively used automatic metrics for the evaluation of generated texts. It is difficult to manually assess the quality of generated texts without clinical knowledge. Manual evaluation by clinical experts would enable us to estimate the medical consistency of generated texts more reliably. Finally, we have found that generations from the same model can be unstable. Filtering texts to keep the best candidate could improve results (Hiebel et al., 2023).

7.2 Ethical Considerations

The clinical documents used for fine-tuning the generation models (E3C and CAS) do not contain personal information. Thus, there is no additional risk of generating sensitive information with our models fine-tuned on those documents. The documents used for training clinical entity recognition models (MERLOT) 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. In this work, we only use the models' annotations on the E3C and CAS corpus.

Acknowledgments

This work has received funding from the French "Agence Nationale pour la Recherche" under grant agreement CODEINE ANR-20-CE23-0026-01.

This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council.

Calculations involving decoder-only models were performed using HPC resources from GENCI–IDRIS (Grant 2023-AD011014538).

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).

Masaki Asada and Makoto Miwa. 2023. BioNART: A biomedical non-AutoRegressive transformer for natural language generation. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 369–376, Toronto, Canada. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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.

J. Elliott Casal and Matt Kessler. 2023. Can linguists distinguish between chatgpt/ai and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, 2(3):100068.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Vincent Claveau, Antoine Chaffin, and Ewa Kijak. 2021. La génération de textes artificiels en substitution ou en complément de données d'apprentissage. In *TALN 2021 - 28e Conférence sur le Traitement Automatique des Langues Naturelles*, volume 1, pages 37–49, Lille, France. ATALA.

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.

Maksim Eremeev, Ilya Valmianski, Xavier Amatriain, and Anitha Kannan. 2023. Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2373–2390, Toronto, Canada. Association for Computational Linguistics.

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.

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.

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.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, Online.

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.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Germán Kruszewski, Jos Rozen, and Marc Dymetman. 2023. disco: a toolkit for distributional control of generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 144–160, Toronto, Canada. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2023. A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

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*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy. 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.

Alberto Moro and Laura Lonza. 2018. Electricity carbon intensity in european member states: Impacts on ghg emissions of electric vehicles. *Transportation Research Part D: Transport and Environment*, 64:5–14. The contribution of electric vehicles to environmental challenges in transport. WCTRS conference in summer.

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.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Teven Le Scao et al. 2022. Bloom: A 176b-parameter open-access multilingual language model.

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.

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2017. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 2852–2858. AAAI Press.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. volume 56, New York, NY, USA. Association for Computing Machinery.

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 *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Translation of Tables and Figures

Figure 2 presents the translation of the example of data representation shown in Figure 1.

Tables 5 and 6 present an automatic translation of the natural document with the corresponding constraints and generated samples that were presented in Tables 1 and 6. The automatic translation was done with DeepL[4].

A.2 Model Sizes

Table7 present the total number of parameters and the trainable parameters for each model.

[4] www.deepl.com

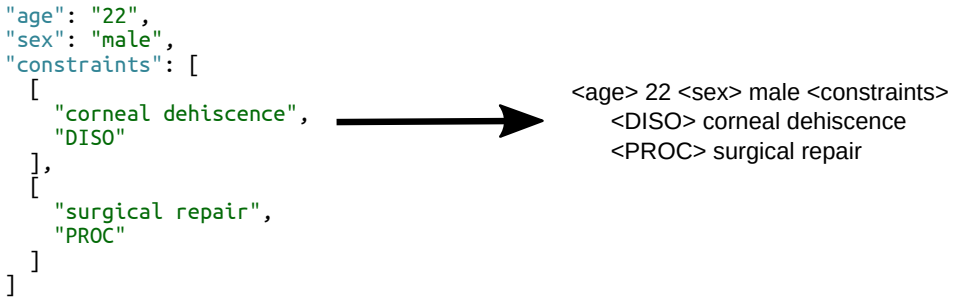

Figure 2: Example of data representation for encoder-decoder architecture (translation of Figure 1).

Type of clinical feature	**Sample value**
Age	54
Sex	Male
Localisation	Bladder
Histology	poorly differentiated adenocarcinoma of the urachus
Sign	hematuria
Procedure	CT scan
Treatment	methotrexate-vinblastine-endoxan-cisplatin chemotherapy
Score	T III A (according to Sheldon's classification)
Bio	negative for cytokeratins (ck) 7 and 20

Table 5: Sample control data based on manual analysis of a clinical case (translation of Table 1).

Automatically extracted constraints	age: 54 ; sex: male ; constraints: isolated hematuria, CT scan, mass, 4 cm, poorly differentiated adenocarcinoma, of type III, extension work-up, total radical cystoprostatectomy, iliac and obturator lymphadenectomy, omphalectomy, replacement enterocystoplasty, adenocarcinoma of the urachus, very locally mucosecretory, ulcerated, transitional cell carcinoma, grade III, Carcinoembryonic Antigen, Leu-M1, CD 15, cytokeratins, bladder epithelium, Sheldon's classification, methotrexate, vinblastine, endoxan, cisplatin
Real clinical case	A 54-year-old man presented with isolated hematuria. An ultrasound scan, followed by a CT scan, showed a mass 4 cm in diameter, with a necrotic center, anterior to the bladder dome, invading only the adjacent fat (Figure 1A.). Cystoscopic biopsies revealed a poorly differentiated adenocarcinoma, with areas of type III transitional epithelioma (Figure 1B). The extension work-up was negative. The patient underwent total radical cystoprostatectomy with iliac and obturator lymphadenectomy, combined with omphalectomy and Studer replacement enterocystoplasty. Complete anatomopathological examination of the surgical specimen revealed a poorly differentiated, invasive, very locally mucosecretory adenocarcinoma of the urachus (Figure 1C). It was ulcerated in the bladder lumen, and contained areas of grade III transitional cell carcinoma, as well as argentaffin cells. Two lymph node pockets and surrounding fat were invaded. Immunohistochemistry showed neoplastic cells to be positive for Carcinoembryonic Antigen (CEA), Leu-M1 (CD 15), and negative for cytokeratins (ck) 7 and 20. The bladder epithelium was positive for ck 7. This was stage T III A (according to Sheldon's classification). The patient was undergoing systemic methotrexate-vinblastine-endoxan-cisplatin chemotherapy. With 35 months' follow-up, he is doing well.
Synthetic Bloomz 1b1 unfrozen	This is a 54-year-old patient with no particular pathological history who consulted for isolated hematuria. Clinical examination revealed a patient in good general condition, apyretic, with signs of arterial hypertension and abdominal tenderness to palpation. Laboratory tests were normal. Abdominal ultrasound revealed a heterogeneous tissue mass on the left flank measuring 4 cm in long axis. MRI showed a homogeneous tissue mass in the left flank (Figure 1). Histological examination of the operative specimen confirmed the diagnosis of type III poorly differentiated adenocarcinoma according to Sheldon's classification [1]. Magnetic resonance imaging (MRI) did not reveal any brain or lung metastases. A total radical cystoprostatectomy associated with an iliac and obturator lymphadenectomy and a substitution omphalectomy and enterocystoplasty were performed. Anatomopathological study of the surgical specimen confirmed the diagnosis of poorly differentiated adenocarcinoma of the urachus, progressing to stage III transitional cell carcinoma according to Sheldon's classification [1].
Synthetic Flan-T5-XL frozen	This was a 54-year-old patient with no specific pathological history who presented with isolated hematuria evolving for 2 months. CT examination revealed a 4 cm mass in the posterior wall of the myringotomy (Figure 1). A diagnosis of type III poorly differentiated adenocarcinoma was made. The extension work-up was negative. Total radical cystoprostatectomy combined with non-obturator iliac lymphadenectomy and omphalectomy with replacement enterocystoplasty was performed. Anatomopathological study of the surgical specimen confirmed the diagnosis of poorly differentiated adenocarcinoma of the urachus, with a very locally mucosecretory and ulcerated appearance. This appearance is compatible with a grade III transitional cell carcinoma (Carcinoembryonic Antigen: CD34, Leu-M1 and CD 15). Tumor cells expressed cytokeratins specific to the bladder epithelium (Sheldon classification). Treatment with Methotrexate, Vinblastine, Endoxan and Cisplatine was initiated.

Table 6: Sample clinical cases generated by two models based on constraints that were automatically extracted from a reference clinical case (translation of Table 2).

Model	Total parameters	Trainable parameters	Percentage trainable
mT5-large frozen	1.2 B	9.6 M	0.8%
mT5-large unfrozen	1.5 B	518 M	34.5%
Flan-T5-small frozen	78.3 M	1.3 M	1.7%
Flan-T5-small unfrozen	94.3 M	33.8 M	35.8%
Flan-T5-large frozen	787 M	4.3 M	0.5%
Flan-T5-large unfrozen	819 M	69.7 M	8.5%
Flan-T5-XL frozen	2.9 B	7.9 M	0.3%
Flan-T5-XL unfrozen	2.9 B	139 M	4.7%
Bloom(z) 1b1 frozen	1.1 B	6.7 M	0.6%
Bloom(z) 1b1 unfrozen	1.5 B	390 M	26.8%
Bloom(z) 7b1 frozen	7.1 B	17.8 M	0.3%

Table 7: Parameter count as reported by the PEFT library used for fine-tuning. We report the same numbers for Bloom and Bloomz because the models have the same architecture and the same amount of parameters. Shift of total parameters in unfrozen models are due to tied embeddings being counted twice.