Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Génération contrôlée de cas cliniques en français à partir de données médicales structurées

Hugo Boulanger*[1], Nicolas Hiebel*[2], Olivier Ferret[1], Karën Fort[3], Aurélie Névéol[2]
(1) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(2) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(3) Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, France
[1]prenom.nom@cea.fr, [2]prenom.nom@lisn.upsaclay.fr, [3]karen.fort@loria.fr

RÉSUMÉ

La génération de texte ouvre des perspectives pour pallier l'absence de corpus librement partageables dans des domaines contraints par la confidentialité, comme le domaine médical. Dans cette étude, nous comparons les performances de modèles encodeurs-décodeurs et décodeurs seuls pour la génération conditionnée de cas cliniques en français. Nous affinons plusieurs modèles pré-entraînés pour chaque architecture sur des cas cliniques en français conditionnés par les informations démographiques des patient·es (sexe et âge) et des éléments cliniques. Nous observons que les modèles encodeurs-décodeurs sont plus facilement contrôlables que les modèles décodeurs seuls, mais plus coûteux à entraîner.

ABSTRACT

Using structured health information for controlled generation of clinical cases in French.

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 conditioned generation of clinical cases in French. We fine-tune several pre-trained models for each architecture on French clinical cases conditioned by patient demographic information (gender and age) and clinical features. We observe that encoder-decoder models are easier to control than decoder-only models, but more costly to train.

MOTS-CLÉS : Génération contrôlée, textes cliniques, textes synthétiques, français.

KEYWORDS: Controlled Generation, Clinical Texts, Synthetic Texts, French.

1 Introduction

L'explosion actuelle de l'intelligence artificielle générative (Cusumano, 2023) ne doit pas faire oublier que les modèles de langue génératifs textuels ont pour compétence première de générer du texte. Les modèles actuels ont permis de pousser cette compétence à un niveau tel qu'il devient difficile de distinguer un texte produit par un humain d'un texte produit par une machine (Casal & Kessler, 2023), ouvrant ainsi la voie à de multiples applications. Dans cet article, nous considérons le cas de documents de référence ne pouvant pas être diffusés, en particulier du fait des informations à caractère personnel qu'ils contiennent, mais suffisamment génériques pour mutualiser des moyens de traitement

*. Les deux auteurs ont contribué de manière égale à ce travail. L'ordre est alphabétique.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 435–448.

à l'échelle d'une communauté. Une façon de développer de tels traitements est de travailler à partir de documents comparables dans leur nature aux documents de référence mais générés automatiquement à partir de ces derniers. Le cas des documents constitutifs des dossiers électroniques patient est à cet égard emblématique, même s'il est loin d'être unique. C'est celui que nous considérons ici.

Dans cette optique, la capacité à contrôler finement le processus de génération est central et multidimensionnel. Pour ne retenir que les principales de ces dimensions, les documents générés doivent être comparables aux documents de référence en termes de style, de structuration, de contenu tout en préservant les informations personnelles qu'ils recèlent. Si les informations directement identifiantes peuvent faire l'objet d'une désidentification robuste en amont, celle-ci ne rend pas les documents anonymes au sens du règlement général sur la protection des données (RGPD). En effet, la désidentification, qu'elle soit automatique ou manuelle, ne protège pas des possibilités de recoupements d'informations médicales, en particulier pour les pathologies rares. Si la possibilité de contrôler la génération en termes de contenu est importante du point de vue de la cohérence médicale des textes générés, elle l'est donc également sur le plan de la préservation des informations personnelles.

Dans cet article, nous proposons ainsi une méthodologie permettant d'exercer un contrôle sur la génération de texte en termes de contenu. Plus précisément, l'idée est de pouvoir conditionner la génération de comptes rendus médicaux par des profils patients. À l'instar de travaux réalisés sur la génération de profils patients synthétiques en termes de données structurées (Walonoski *et al.*, 2017), ces profils prennent la forme d'un ensemble de concepts médicaux. Cette approche, qui relève d'une problématique de génération données-vers-texte, présente l'avantage, par rapport à une approche par amorce textuelle (*prompt*), de pouvoir contrôler finement l'information servant au conditionnement. Ce dernier est mis en œuvre par l'entraînement d'un modèle de langue neuronal à l'aide d'un ensemble de couples composés chacun d'un profil patient sous forme de concepts et d'un compte rendu de référence correspondant à ce profil. Dans ce cadre, les contributions de notre article sont les suivantes :

— une méthodologie de contrôle du contenu de la génération de comptes rendus médicaux ;
— une méthode de constitution d'un ensemble d'entraînement pour la réalisation de ce contrôle ;
— deux formes de mise en œuvre de la stratégie de contrôle proposée ;
— une évaluation multidimensionnelle automatique des résultats de cette stratégie.

2 Travaux connexes

2.1 Génération contrainte de texte

Depuis l'avènement des premiers grands modèles de langues tels que ceux de la famille des GPT (Radford *et al.*, 2018), générer du texte ressemblant à une production humaine semble facile et le problème de la génération a évolué pour changer de cible : le but n'est plus de simplement générer du texte vraisemblable, mais de pouvoir contrôler plus finement ce qui est généré. Les textes produits par les modèles génératifs peuvent ne pas être pertinents ou présenter un contenu offensant voire dangereux (Bender *et al.*, 2021). C'est pourquoi de nombreux travaux portent sur le contrôle de la génération. Le contrôle peut concerner plusieurs aspects de la génération comme le lexique ou le style du texte (Zhang *et al.*, 2023). Plusieurs méthodes de contrôle ont été explorées, dont l'entraînement d'un modèle avec des exemples conditionnés selon des critères choisis (Keskar *et al.*, 2019) ou la modification des probabilités des tokens de sortie lors de l'inférence (Kruszewski *et al.*, 2023).

Les approches « données vers texte » (*data-to-text*) (Lin *et al.*, 2023) contraignent la génération à partir de données structurées (graphes, tableaux et, dans notre cas, les *slots*). Les architectures privilégiées sont des modèles encodeurs-décodeurs pouvant avoir des architectures internes variées, combinant des modèles pré-entraînés en encodeurs et/ou décodeurs. Il est aussi possible d'affiner directement des modèles encodeurs-décodeurs, tels que le modèle T5 (Raffel *et al.*, 2020). Les modèles de langue causaux, comme par exemple les modèles utilisant une architecture de décodeur de transformeur (Vaswani *et al.*, 2017), utilisent le contexte en début de séquence pour générer la suite de la séquence.

2.2 Génération dans le domaine biomédical

Dans le domaine biomédical, la génération de texte est notamment explorée pour produire des comptes-rendus de discussions entre médecins et patients (Eremeev *et al.*, 2023; Ben Abacha *et al.*, 2023; Asada & Miwa, 2023). L'automatisation de cette tâche pourrait en effet grandement soulager une partie de la charge de travail des médecins.

Pour répondre à la difficulté d'accès aux textes médicaux, Ive *et al.* (2020) proposent une méthodologie de génération de cas cliniques synthétiques en anglais à partir de cas cliniques réels. La génération est conditionnée par des entités extraites automatiquement des documents réels. Cependant, peu de corpus, et donc de travaux, portent sur d'autres langues que l'anglais (Névéol *et al.*, 2018). Dans cette étude, nous nous intéressons au cas du français.

3 Méthodologie générale

Comme nous l'avons esquissé dans l'introduction, l'idée directrice de ce travail est de conditionner le processus de génération par les données structurées dont le texte généré devra faire état. Bien entendu, retrouver les données de conditionnement au sein des textes générés ne peut être le seul critère d'évaluation des modèles : il suffirait en effet à ces derniers de reproduire leur entrée pour être jugés comme parfaits. Ce conditionnement doit donc intégrer une proximité de nature par rapport aux documents de référence que l'on souhaite émuler.

Comme évoqué à la section 2.1, ce double conditionnement peut se faire par un affinage a priori du modèle de langue servant à la génération ou bien par son contrôle lors de la génération. Nous avons opté pour la première solution dans la mesure où la seconde suppose d'appliquer des processus d'analyse textuelle élaborés lors de la génération pour vérifier le respect du conditionnement, ce qui est coûteux. La première solution suppose néanmoins de disposer de données d'entraînement associant données de conditionnement et textes exemples conformes à ce conditionnement. Pour ce faire, nous avons adopté une stratégie comparable à Peng *et al.* (2018) pour la génération d'histoire, reprise par Ive *et al.* (2020) pour les comptes rendus médicaux, et consistant à extraire automatiquement les données de conditionnement des textes exemples. Cette stratégie suppose bien évidemment de disposer de processus d'analyse textuelle capables d'extraire ces données de conditionnement des textes exemples avec un niveau de performance suffisamment élevé. Elle induit par conséquent un couplage étroit entre les capacités de génération et celles d'analyse, mais permet de se passer d'une annotation manuelle coûteuse. Dans le cas présent, nous nous focalisons sur les concepts médicaux et sommes donc dépendants de modèles permettant d'extraire ces concepts de comptes rendus médicaux,

mais la généricité de cette stratégie permet de prendre en compte facilement de nouveaux éléments de conditionnement, dès lors qu'ils peuvent être extraits automatiquement de textes exemples.

4 Corpus et modèles génératifs

4.1 Corpus de cas cliniques en français

Les données utilisées pour nos expériences proviennent de deux corpus de cas cliniques librement disponibles. Le premier est le corpus CAS (Grabar *et al.*, 2018), un corpus de cas cliniques désidentifiés en français [1]. Le second est le corpus E3C (Magnini *et al.*, 2020), un corpus multilingue de cas cliniques désidentifiés. Nous nous intéressons uniquement aux cas cliniques en français de ce dernier.

4.2 Construction des contraintes selon un profil patient

Nous souhaitons pouvoir contraindre la génération par des éléments cliniques afin de créer des cas cliniques cohérents. Nous avons échangé avec des cliniciens afin de définir les éléments saillants dans des cas cliniques. Ces éléments sont ensuite sélectionnés comme conditionnement de la génération. Le tableau 1 présente un exemple des différentes catégories d'éléments importants qui ont été retenus pour un cas clinique du corpus E3C. On y retrouve les informations démographiques du patient (âge et sexe), la localisation de la pathologie, des informations histologiques, différents signes ou symptômes, des traitements et procédures effectués, des résultats biologiques et des scores (mesures ou codes). En accord avec les recommandations des cliniciens, nous identifions une vingtaine de contraintes par cas, en sélectionnant si possible des éléments de chaque catégorie avec une majorité de symptômes, traitements et procédures. Cette façon de faire permet de sélectionner les informations importantes des cas cliniques selon les médecins.

Types d'éléments cliniques	**Exemple de valeurs**
Âge	54
Sexe	Masculin
Localisation	Vessie
Histologie	adénocarcinome de l'ouraque peu différencié
Signe	hématurie
Procédure	scanner CT
Traitement	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 – Exemples d'éléments de contrôle manuellement définis pour un cas clinique. Le cas clinique correspondant est présenté dans le tableau 2.

1. Contacter les auteurs pour accéder au corpus https://deft.lisn.upsaclay.fr/2020

4.3 Extraction des contraintes des documents

Concernant les données démographiques, nous nous sommes appuyés pour le corpus CAS sur les annotations présentes relatives à l'âge et au sexe des patients. Le corpus E3C ne disposant pas de ces informations, nous avons annoté les 1 009 cas cliniques en français du corpus pour obtenir l'âge et le sexe des patients. Pour les autres entités cliniques, nous annotons automatiquement les deux corpus pour que les annotations des deux corpus soient homogènes et facilitent ainsi l'apprentissage des contraintes par les modèles de génération. Pour réaliser cette annotation automatique, nous utilisons des modèles de reconnaissance d'entités cliniques entraînés sur le corpus privé MERLOT (Campillos *et al.*, 2018), qui contient des annotations manuelles pour ces entités.

Nous avons construit nos contraintes en partant des annotations démographiques manuelles et des annotations automatiques en entités cliniques. Pour chaque document, nous sélectionnons l'âge et le sexe lorsqu'ils sont renseignés. Lorsque l'âge exact n'est pas renseigné, nous utilisons les catégories d'âge issues des descripteurs obligatoires du thésaurus MeSH (Medical Subject Headings) [2].

Pour les entités cliniques, nous avons sélectionné les catégories d'annotation de MERLOT pour correspondre aux catégories discutées avec les médecins. Ainsi, nous retenons pour chaque cas clinique les dix procédures (*PROC*) et dix symptômes (*DISO*) ayant le tf.idf le plus élevé. Nous sélectionnons aussi les substances (*CHEM*) et les mesures (*MEAS*). Ces dernières sont filtrées pour ne garder que des mesures informatives (les chiffres seuls comme *6* sont par exemple annotés comme *MEAS* mais sans information supplémentaire). De cette manière, nous obtenons en moyenne 26 contraintes ($\pm$9,5) par cas clinique.

4.4 Modèles génératifs

Nous comparons les performances de deux architectures différentes pour la génération contrainte de textes cliniques : l'architecture encodeur-décodeur et l'architecture décodeur seul. Nous nous appuyons pour cela sur des modèles transformeurs pré-entraînés.

Encodeur-décodeur Cette architecture est spécialisée dans la génération de texte à partir de données structurées. Notamment, l'affinage du modèle T5 (Raffel *et al.*, 2020) s'est imposé comme une méthode standard pour ce genre de tâches. Nous avons choisi d'utiliser la version multilingue de T5, appelée mT5 (Xue *et al.*, 2021), d'un milliard de paramètres comme modèle pré-entraîné, et les modèles Flan-T5-Large (780 millions de paramètres) et Flan-T5-XL (3 milliards de paramètres) (Chung *et al.*, 2022) comme modèles affinés avec instructions.

Décodeur seul Cette architecture est spécialisée dans la génération de texte à partir d'amorces textuelles (*prompt*). Nous avons choisi plusieurs modèles pour cette architecture : Bloom (Scao *et al.*, 2022), un modèle génératif entraîné sur plusieurs langues, et Bloomz, une variante entraînée spécialement pour réaliser différentes tâches (traduction, résumé automatique etc.). Nous prenons pour chacun de ces deux modèles deux versions en termes de taille : un et sept milliards de paramètres.

2. https://www.nlm.nih.gov/bsd/indexing/training/CHK_030.html

5 Expérimentations

5.1 Représentation des données structurées

L'utilisation de ces modèles génératifs nécessite de transformer les données structurées en format textuel. Nous avons choisi de linéariser les entrées de manière différente pour les modèles encodeurs-décodeurs et les modèles décodeurs seuls. Pour les modèles encodeurs-décodeurs, nous ajoutons devant chaque entité un token spécial lié à la classe de l'entité. Nous séparons les informations démographiques (âge, sexe) des contraintes médicales (symptôme, procédure etc.) par un token spécial *contraintes*. Pour les modèles décodeurs seuls, nous avons choisi de ne pas ajouter de tokens spéciaux. La figure 1 présente un exemple de représentation des données pour les encodeurs-décodeurs.

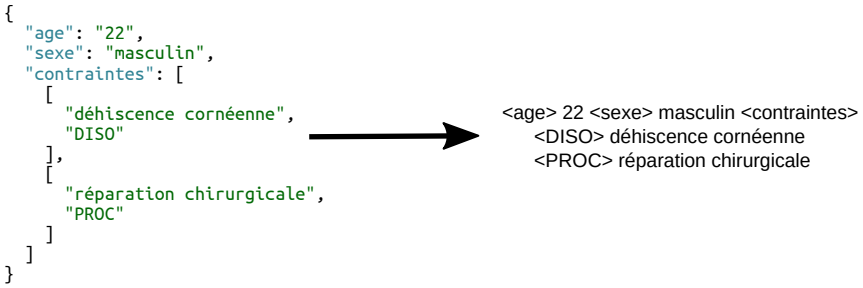

FIGURE 1 – Exemple de représentation des données pour l'architecture encodeur-décodeur.

5.2 Affinage

Le jeu d'entraînement sur lequel nous affinons nos modèles est constitué de 1 424 cas cliniques, pour un peu plus de 500 000 tokens hors contraintes. Pour l'affinage des modèles, nous choisissons de geler les poids des modèles pré-entraînés et ajoutons des matrices entraînables LoRA (Hu *et al.*, 2022). L'emplacement des matrices entraînables dépend du type de modèle. Pour les modèles encodeurs-décodeurs, nous ajoutons des matrices LoRA sur les *queries* et *values* des couches transformeurs et sur la tête de modélisation du modèle. Pour les modèles décodeurs seuls, les matrices LoRA sont ajoutées sur les couches linéaires des modèles. Nous ajoutons les tokens spéciaux aux plongements via des vecteurs initialisés aléatoirement. Le traitement des plongements lexicaux varie selon trois configurations définies comme suit :

Configuration « gelé » : les plongements sont gelés mais nous ajoutons des matrices LoRA pour leur permettre une adaptation à la tâche à un faible coût mémoire.

Configuration « dégelé » : les plongements sont dégelés, pour permettre l'adaptation à la tâche, mais à un coût plus élevé.

Configuration « partiel » : seuls les plongements des tokens spéciaux sont dégelés.

5.3 Génération des cas cliniques

Notre jeu de test est constitué de 156 cas cliniques et de leurs contraintes. Les contraintes sont données en entrée aux modèles génératifs et les cas cliniques réels servent de référence. Le décodage est

réalisé en utilisant une recherche par faisceaux (*beams*) avec 5 faisceaux, et de l'échantillonnage (*sampling*) avec un top-p de 0,90, un top-k de 100, une température de 1 et une pénalité de répétition de 3. Faire de l'échantillonnage lors de la génération signifie qu'un même modèle peut générer des textes différents avec la même entrée. Nous effectuons cinq générations pour chaque exemple de test afin de prendre en compte cette variabilité.

5.4 Métriques d'évaluation

L'évaluation automatique de la génération de texte est notoirement difficile (Novikova *et al.*, 2017). De nombreuses métriques existent, qui permettent de mesurer différents aspects de la génération de texte (Frisoni *et al.*, 2022). Nous avons sélectionné certaines d'entre elles afin de couvrir plusieurs dimensions de l'évaluation.

Adéquation aux contraintes - *Exactitude* Cette mesure sert à évaluer la capacité du modèle à respecter les contraintes qui lui sont imposées. Nous calculons la proportion de contraintes respectées dans les textes générés par rapport au nombre total de contraintes imposées.

Qualité de la langue - *Perplexité* La perplexité évalue l'adéquation des données textuelles avec la distribution de probabilité d'un modèle de langue. Nous utilisons un modèle spécifique au français, GPTFR (Simoulin & Crabbé, 2021). Pour cette métrique, nous souhaitons que la perplexité obtenue sur les données générées se rapproche de la perplexité obtenue sur les données réelles (égale à 19 ici pour le corpus d'entraînement).

Diversité des textes générés - *Self-BLEU* Le score Self-BLEU (Zhu *et al.*, 2018) est la moyenne des scores BLEU de toutes les phrases d'un corpus entre elles. Ainsi, un corpus redondant aura un score Self-BLEU élevé tandis qu'un corpus varié aura un score plus faible.

Proximité avec le corpus naturel - *Corpus-BLEU* Corpus-BLEU (Yu *et al.*, 2017) est une mesure de proximité entre deux corpus et correspond à la moyenne des scores BLEU entre chaque phrase du corpus généré et toutes les phrases du corpus naturel. Nous calculons Corpus-BLEU en comparant les cas cliniques du corpus de test avec les textes générés.

Proximité avec le cas clinique correspondant aux contraintes - *BLEU* Le score BLEU (Papineni *et al.*, 2002) est calculé entre le texte généré et le cas clinique réel d'où proviennent les contraintes. Il mesure la proximité avec les données réelles de façon plus spécifique que le score Corpus-BLEU.

6 Résultats

6.1 Génération des cas cliniques

Le tableau 2 présente des exemples de textes générés à partir d'un ensemble de contraintes par un modèle encodeur-décodeur (Flan-T5-XL gelé) et un modèle décodeur seul (Bloomz 1b1 dégelé). Le tableau 3 présente quant à lui l'évaluation automatique des cas cliniques générés avec les différentes architectures étudiées. Parmi nos baselines, la simple copie des entités de conditionnement (« Copie ») obtient comme attendu une exactitude de 100 %, mais aussi une perplexité très grande. La baseline « Corpus » correspond à une copie du corpus de test dans laquelle nous avons enlevé les retours à la ligne. Cette modification explique pourquoi les scores BLEU et corpus-BLEU ne sont pas

Contraintes extraites automatiquement (hors balises)	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
Cas clinique réel	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.
Génération Bloomz 1b1 dégelé	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].
Génération Flan-T5-XL gelé	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 – Exemples de textes générés par deux modèles à partir de contraintes automatiquement extraites d'un cas clinique réel.

parfaits et, de façon plus surprenante, fait baisser la perplexité de 30,5 à 19,5. Le score d'exactitude révèle quant à lui des limites de nos données et du calcul d'exactitude. La majorité de ces erreurs concerne le sexe du patient, lorsqu'il n'est pas marqué via l'accord du terme « patient » en genre ou l'utilisation du qualificatif « masculin » ou « féminin ». Les autres erreurs proviennent majoritairement de reformulations ou d'erreurs dans les contraintes.

Les résultats des modèles montrent plusieurs tendances. Les modèles encodeurs-décodeurs ayant bénéficié d'une période d'affinage avec instructions, les modèles Flan, obtiennent globalement de meilleurs résultats que les modèles pré-entraînés sans instructions. Les modèles Flan ont en outre

	Modèle	Exactitude↑	Perplexité	Self-BLEU-4↓	Corpus-BLEU-4↑	BLEU-4↑
Baselines	Copie	100	194,3	14,4	25,5	1,1
	Corpus	98,8	19,5	33,4	97,4	97,5
	Bloom 1b1 gelé∗	s/o	11,5±1,5	86,1±0,4	64,8±0,4	s/o
	Bloom 1b1 dégelé∗	s/o	10,2±0,9	82,9±0,4	60,6±0,5	s/o
	Bloom 7b1 gelé∗	s/o	8,4±2,8	79,3±1,2	57,1±0,5	s/o
Encodeurs-décodeurs	mT5-large gelé	78,0±0,6	13,6±0,2	53,5±0,5	55,8±0,5	12,0±0,1
	mT5-large dégelé	73,6±0,8	13,4±0,2	53,8±0,4	56,4±0,3	10,9±0,2
	mT5-large partiel	75.7±0.3	13,2±0,4	55.1±0.5	56.5±0.2	11.6±0.3
	Flan-T5-large gelé	81,5±1,1	14,8±0,4	52,8±0,4	55,3±0,4	12,0±0,1
	Flan-T5-large dégelé	80,3±1,0	15,6±0,5	51,9±0,2	55,0±0,4	11,7±0,2
	Flan-T5-large partiel	80,9±0,9	16,1±0,2	50,9±0,3	54,3±0,4	11,6±0,1
	Flan-T5-XL gelé	84,2±0,8	14,9±0,2	50,2±0,2	54,5±0,2	12,8±0,1
	Flan-T5-XL dégelé	85,3±0,8	14,9±0,2	49,0±0,1	53,8±0,4	12,9±0,2
	Flan-T5-XL partiel	82,2±1,3	15,4±0,2	50,3±0,2	54,6±0,3	12,0±0,2
Décodeurs	Bloom 1b1 gelé	40,5±3,9	8,8±0,2	62,5±5,8	42,3±11,1	4,7±1,0
	Bloom 1b1 dégelé	29,6±0,9	9,3±0,4	63,6±4,7	50,4±9,7	4,0±0,5
	Bloom 7b1 gelé	43,5±2,5	9,9±0,6	54,0±2,1	47,5±2,0	5,8±1,0
	Bloomz 1b1 gelé	45,4±4,2	9,2±0,2	61,9±7,6	41,8±11,0	5,2±1,3
	Bloomz 1b1 dégelé	32,1±1,7	9,6±0,2	65,7±6,0	47,0±13,2	4,3±0,7
	Bloomz 7b1 gelé	39,8±3,0	9,9±0,2	55,0±1,9	49,8±1,5	5,4±0,4

TABLE 3 – Évaluation des données générées à partir des contraintes provenant du jeu de test. Modèles baselines marqués par « ∗ » : entraînement et génération sans contrainte.

l'avantage d'être plus rapidement affinés à taille égale, avec une période d'entraînement de 16 h pour Flan-T5-large contre 60 h pour mT5-large. Nous observons, comme attendu, que les modèles Flan-T5-XL sont les plus performants des modèles encodeurs-décodeurs testés. Ces derniers génèrent des textes plus variés (Self-BLEU), et présentent la meilleure exactitude. Les textes générés ressemblent le plus aux références (BLEU) et la perplexité et le Corpus-BLEU sont meilleurs que ceux de la version plus petite du modèle. Le Corpus-BLEU reste cependant assez stable quel que soit le modèle initial et le traitement des plongements lexicaux. Il est cependant à noter que les modèles mT5 obtiennent une perplexité inférieure, probablement due au fait que le modèle initial soit multilingue tandis que les modèles Flan-T5 n'ont vu de français que sur des tâches de traduction.

Nous observons que les décodeurs seuls obtiennent de moins bons résultats que les encodeurs-décodeurs. Les modèles décodeurs sont également nettement plus instables d'une génération à une autre, avec des écarts-types importants au niveau de l'exactitude, du Self-BLEU et du Corpus-BLEU. Au niveau de la perplexité, ces modèles obtiennent des scores plus faibles et s'éloignent donc du corpus d'entraînement. Le modèle permettant de calculer la perplexité étant un décodeur seul, l'architecture commune biaise potentiellement les décodeurs pour cette métrique. En revanche, le temps d'entraînement des décodeurs est beaucoup plus court : 10 à 15 minutes pour les modèles à un milliard de paramètres et 30 minutes pour les modèles à sept milliards de paramètres.

Nous pouvons également identifier quelques bonnes pratiques concernant le pré-entraînement des modèles et la configuration des plongements lexicaux. Les modèles ayant bénéficié d'un pré-entraînement avec instructions sont globalement plus performants que les modèles avec un pré-entraînement simple

sur une tâche de modélisation de la langue. Cela s'observe principalement pour l'exactitude et le score BLEU. Nous pouvons aussi observer que les modèles dégelés ont de moins bonnes performances que les modèles gelés. Cependant, nous avons remarqué que les modèles dégelés convergent plus rapidement, en temps et en époques d'entraînement. Les résultats plus faibles de Flan-T5-large gelé sont peut être le résultat d'un sous-entraînement.

6.2 Impact environnemental

Modèle	Entraînement	Génération	Perplexité	Total
mT5-large	7,26	0,75	0,01	8,02
flan-T5-large	1,95	0,75	0,01	2,71
flan-T5-XL	7,26	0,75	0,01	8,02
Bloom(z) 1b1	0,03	0,78	0,01	0,82
Bloom(z) 7b1	0,05	0,64	0,01	0,70

TABLE 4 – Impact environnemental en $kgCO_2$éq des expériences finales pour chaque modèle. Chaque ligne somme les émissions des différentes configurations associées. Les émissions totales sont de 20,27 $kgCO_2$éq.

La compilation des émissions en $kgCO_2$éq peut être retrouvée dans le tableau 4. L'impact environnemental est essentiellement lié à l'entraînement des modèles encodeurs-décodeurs, qui est plus long et requiert plus de GPU pour les modèles plus grands. Ces évaluations ont été réalisées avec le MachineLearning Impact calculator présenté dans (Lacoste *et al.*, 2019) avec les valeurs d'émission de la France (0,101 $kgCO_2$éq/kWh) se trouvant dans (Moro & Lonza, 2018).

6.3 Limites

L'ensemble de mesures que nous avons mis en place nous permet d'avoir une vision assez bonne sur ce que nos modèles génèrent. Il y a néanmoins des limites à n'utiliser que l'exactitude, en particulier telle qu'elle est calculée, pour décrire la fidélité de la retranscription des informations. L'exactitude recherche ici une correspondance exacte entre les contraintes et le texte. Toute reformulation du modèle est donc écartée bien qu'elle puisse être correcte. De plus, utiliser cette mesure seule ne nous donne pas d'information sur de potentiels ajouts d'informations ou d'entités par les modèles. Dans cette étude, nous avons exclusivement utilisé des métriques automatiques pour l'évaluation des textes générés. Il est difficile d'évaluer manuellement la qualité des textes générés sans connaissances cliniques. Une évaluation manuelle par des experts cliniques permettrait d'estimer la cohérence médicale des textes générés de manière plus fiable. Nous avons enfin constaté que les générations par un même modèle peuvent être instables. Un filtrage des textes pour garder le meilleur candidat pourrait améliorer les résultats (Hiebel *et al.*, 2023).

7 Conclusion

Dans cette étude, nous générons des comptes rendus médicaux en français conditionnés par des données cliniques structurées. Nous comparons des modèles d'architectures différentes, des encodeurs-décodeurs et des décodeurs seuls, que nous affinons sur un corpus de cas cliniques à l'aide de matrices LoRA. Nous proposons une méthodologie d'évaluation fondée sur un ensemble de mesures automatiques : exactitude, perplexité, Self-BLEU, Corpus-BLEU et BLEU. Nous observons que les modèles à architecture encodeurs-décodeurs obtiennent de meilleurs résultats sur la tâche de génération à partir de données structurées, mais avec un entraînement plus coûteux. Concernant les différentes stratégies d'affinage au niveau des plongements lexicaux, la meilleure stratégie consiste à ajouter des matrices LoRA sur les plongements lexicaux et non de les dégeler, bien que cela allonge l'apprentissage. La puissance de calcul disponible dans un cadre hospitalier limite la possibilité d'utiliser des modèles plus gros et/ou plus lourds. D'après nos résultats, les architectures décodeurs sont plus légères et donc plus adaptées. Il faudrait cependant générer plusieurs candidats et les filtrer pour compenser l'irrégularité de ces modèles. Il serait aussi intéressant d'explorer les performances de modèles à architecture encodeur-décodeur plus petits que ceux testés dans cette étude (le modèle Flan-T5-small ne contient par exemple que 80 millions de paramètres). La quantification (*quantization*) des modèles pourrait aussi être une solution pour réduire la charge de calcul, à condition que les modèles quantifiés donnent des résultats comparables à leurs homologues standards.

Remerciements

Ce travail a été réalisé dans le cadre d'un projet de l'Agence Nationale de la Recherche, CODEINE (artificial text COrpus DEsIgNed Ethically), ANR-20-CE23-0026-01. Il a été réalisé grâce au super-calculateur Factory-IA financé par le Conseil Régional d'Île-de-France.

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011014538 attribuée par GENCI.

Références

ASADA M. & MIWA M. (2023). BioNART : A biomedical non-AutoRegressive transformer for natural language generation. In D. DEMNER-FUSHMAN, S. ANANIADOU & K. COHEN, Éds., *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, p. 369–376, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.bionlp-1.34.

BEN ABACHA A., YIM W.-W., FAN Y. & LIN T. (2023). An empirical study of clinical note generation from doctor-patient encounters. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2291–2302, Dubrovnik, Croatie : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.168.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (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, p. 610–623, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3442188.3445922.

CAMPILLOS L., DELÉGER L., GROUIN C., HAMON T., LIGOZAT A.-L. & NÉVÉOL A. (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. DOI : 10.1007/s10579-017-9382-y.

CASAL J. E. & KESSLER M. (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. DOI : https ://doi.org/10.1016/j.rmal.2023.100068.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI Y., WANG X., DEHGHANI M., BRAHMA S. *et al.* (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv :2210.11416*.

CUSUMANO M. A. (2023). Generative ai as a new innovation platform. *Communications of the ACM*, **66**(10), 18—-21. DOI : 10.1145/3615859.

EREMEEV M., VALMIANSKI I., AMATRIAIN X. & KANNAN A. (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)*, p. 2373–2390, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.133.

FRISONI G., CARBONARO A., MORO G., ZAMMARCHI A. & AVAGNANO M. (2022). NLG-metricverse : An end-to-end library for evaluating natural language generation. In N. CALZOLARI, C.-R. HUANG, H. KIM, J. PUSTEJOVSKY, L. WANNER, K.-S. CHOI, P.-M. RYU, H.-H. CHEN, L. DONATELLI, H. JI, S. KUROHASHI, P. PAGGIO, N. XUE, S. KIM, Y. HAHM, Z. HE, T. K. LEE, E. SANTUS, F. BOND & S.-H. NA, Éds., *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3465–3479, Gyeongju, Corée du Sud : International Committee on Computational Linguistics.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Bruxelles, Belgique : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2023). Can synthetic text help clinical named entity recognition ? a study of electronic health records in French. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2320–2338, Dubrovnik, Croatie : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.170.

HU E. J., YELONG SHEN, WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). LoRA : Low-rank adaptation of large language models. In *International Conference on Learning Representations*, En ligne.

IVE J., VIANI N., KAM J., YIN L., VERMA S., PUNTIS S., CARDINAL R., ROBERTS A., STEWART R. & VELUPILLAI S. (2020). Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, **3**. DOI : 10.1038/s41746-020-0267-x.

KESKAR N. S., MCCANN B., VARSHNEY L. R., XIONG C. & SOCHER R. (2019). CTRL : A conditional transformer language model for controllable generation. *CoRR*, **abs/1909.05858**.

KRUSZEWSKI G., ROZEN J. & DYMETMAN M. (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)*, p. 144–160, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-demo.14.

LACOSTE A., LUCCIONI A., SCHMIDT V. & DANDRES T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv :1910.09700*.

LIN Y., RUAN T., LIU J. & WANG H. (2023). A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, p. 1–20. DOI : 10.1109/TKDE.2023.3304385.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2020). The E3C Project : Collection and Annotation of a Multilingual Corpus of Clinical Cases. In J. MONTI, F. DELL'ORLETTA & F. TAMBURINI, Éds., *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769 de *CEUR Workshop Proceedings*, Bologne, Italie : CEUR-WS.org.

MORO A. & LONZA L. (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, DOI : https ://doi.org/10.1016/j.trd.2017.07.012.

NÉVÉOL A., DALIANIS H., VELUPILLAI S., SAVOVA G. & ZWEIGENBAUM P. (2018). Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of Biomedical Semantics*, **9**(1), 12. DOI : 10.1186/s13326-018-0179-8.

NOVIKOVA J., DUŠEK O., CERCAS CURRY A. & RIESER V. (2017). Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2241–2252, Copenhague, Danemark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1238.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphie, Pennsylvanie, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PENG N., GHAZVININEJAD M., MAY J. & KNIGHT K. (2018). Towards controllable story generation. In M. MITCHELL, T.-H. K. HUANG, F. FERRARO & I. MISRA, Éds., *Proceedings of the First Workshop on Storytelling*, p. 43–49, Nouvelle-Orléans, Louisiane : Association for Computational Linguistics. DOI : 10.18653/v1/W18-1505.

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVER I. (2018). Improving language understanding by generative pre-training.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, **21**(1).

SCAO T. L. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. DOI : 10.48550/ARXIV.2211.05100.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle Transformer Génératif Pré-entrainé pour le _____ français. In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *Traitement Automatique des Langues Naturelles*, p. 246–255, Lille, France : ATALA. HAL : hal-03265900.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. VON LUXBURG, S. BENGIO, H. M. WALLACH, R. FERGUS, S. V. N. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, p. 5998–6008.

WALONOSKI J., KRAMER M., NICHOLS J., QUINA A., MOESEL C., HALL D., DUFFETT C., DUBE K., GALLAGHER T. & MCLACHLAN S. (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. DOI : 10.1093/jamia/ocx079.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mT5 : A massively multilingual pre-trained text-to-text transformer. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 483–498, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.41.

YU L., ZHANG W., WANG J. & YU Y. (2017). Seqgan : sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, p. 2852–2858 : AAAI Press.

ZHANG H., SONG H., LI S., ZHOU M. & SONG D. (2023). A survey of controllable text generation using transformer-based pre-trained language models. volume 56, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3617680.

ZHU Y., LU S., ZHENG L., GUO J., ZHANG W., WANG J. & YU Y. (2018). Texygen : A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, p. 1097–1100, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3209978.3210080.