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

Reconnaissance d'entités cliniques en *few-shot* en trois langues

Marco Naguib[1] Aurélie Névéol[1] Xavier Tannier[2]
(1) Université Paris-Saclay, CNRS, LISN, 91405 Orsay cedex, France
(2) Sorbonne Université, Inserm, Université Sorbonne Paris Nord, LIMICS, 75006 Paris, France
marco.naguib@lisn.upsaclay.fr, aurelie.neveol@lisn.upsaclay.fr,
xavier.tannier@sorbonne-universite.fr

RÉSUMÉ

Les grands modèles de langues deviennent la solution de choix pour de nombreuses tâches de traitement du langage naturel, y compris dans des domaines spécialisés où leurs capacités *few-shot* devraient permettre d'obtenir des performances élevées dans des environnements à faibles ressources. Cependant, notre évaluation de 10 modèles causaux (auto-régressifs) et 16 modèles masqués montre que, bien que les modèles causaux utilisant des prompts puissent rivaliser en termes de reconnaissance d'entités nommées (REN) en dehors du domaine clinique, ils sont dépassés dans le domaine clinique par des *taggers* biLSTM-CRF plus légers reposant sur des modèles masqués. De plus, les modèles masqués ont un bien moindre impact environnemental que les modèles causaux. Ces résultats, cohérents dans les trois langues étudiées, suggèrent que les modèles à apprentissage *few-shot* ne sont pas encore adaptés à la production de REN dans le domaine clinique, mais pourraient être utilisés pour accélérer la création de données annotées de qualité.

ABSTRACT

Few-shot learning for clinical entity recognition in three languages.

Large language models have become the preferred solution for many natural language processing tasks, including specialized domains where their few-shot capabilities should deliver high performance in low-resource environments. However, our evaluation of 10 auto-regressive models and 16 masked models shows that while prompt-based auto-regressive models can compete in named entity recognition (NER) outside the clinical domain, they are outperformed within the clinical domain by lighter biLSTM-CRF taggers based on masked models. Additionally, masked models have a much lower environmental impact than auto-regressive models. These consistent results across the three languages studied suggest that few-shot learning models are not yet suited for NER production in the clinical domain but could be used to expedite the creation of quality annotated data.

MOTS-CLÉS : Apprentissage en *few-shot* ; modèles de langues ; reconnaissance d'entités nommées.

KEYWORDS: *few-shot* learning ; large language models ; named entity recognition.

1 Introduction

Les documents cliniques représentent d'importantes sources d'informations (Demner-Fushman *et al.*, 2009), souvent présentées sous forme de texte non structuré (Escudié *et al.*, 2017). L'extraction efficace des informations de ces documents vers une forme plus structurée peut améliorer la recherche

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 169–197.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

clinique, la surveillance de la santé publique et l'aide à la décision clinique automatique (Wang *et al.*, 2018).

La reconnaissance d'entités nommées (REN) constitue une étape cruciale de cette extraction d'informations. Elle consiste à identifier et à typer des mentions d'intérêt dans un texte. Dans le cadre de l'extraction d'informations cliniques, il s'agit notamment des entités cliniques telles que les maladies ou les médicaments. L'extraction de ces entités peut grandement faciliter la normalisation des concepts (Cho *et al.*, 2017; Wajsbürt *et al.*, 2021; Sung *et al.*, 2022) et l'interprétation du profilage et du phénotypage des patients (Gérardin *et al.*, 2022). Alors que la REN dans le domaine général (identification des entités telles que les personnes et les lieux) a été largement étudiée dans la communauté du traitement automatique des langues (TAL), la REN clinique est souvent considérée comme plus complexe : les entités cliniques sont souvent exprimées en jargon ou en termes ambigus, et les textes cliniques présentent une structure grammaticale non standard (Luo *et al.*, 2020; Leaman *et al.*, 2015).

Les modèles de langues sont progressivement devenus l'approche principale pour aborder la REN (Li *et al.*, 2022; Wang *et al.*, 2022). Des travaux antérieurs se sont concentrés sur la REN générale (Devlin *et al.*, 2019) ainsi que sur la REN clinique (Gérardin *et al.*, 2022; Sun *et al.*, 2021). Ces travaux peuvent être principalement divisés en deux approches, selon le type de modèles de langues utilisés.

La première approche consiste à utiliser des **modèles de langues masqués (MLM)** pré-entraînés. Ce type de modèles est d'abord pré-entraîné pour prédire des mots masqués sélectionnés au hasard dans de grands corpus de textes à l'aide d'une représentation vectorielle dense de chaque token (mot, par exemple) dans le texte (Devlin *et al.*, 2019; Peters *et al.*, 2018). Pour utiliser ces modèles pour la REN, on apprend généralement une projection linéaire à transformer les représentations vectorielles des mots en étiquettes désignant les entités nommées dans la phrase. En parallèle, on ajuste (*fine-tune)* également les paramètres du modèle de langues pour la tâche de REN. Cette approche a fait l'objet d'une grande attention de la part de la communauté, et est devenue la solution de référence pour la construction de systèmes de REN robustes.

Toutefois, cette approche rencontre deux principaux obstacles dans le contexte de la REN clinique. Tout d'abord, en raison de la nature sensible des documents cliniques, les corpus publics sont rares, soumis à des licences restrictives et peu disponibles dans des langues autres que l'anglais. Cela contraint la communauté à utiliser des solutions construites sur des MLM pré-entraînés principalement sur des corpus de domaine général, ce qui peut entraîner des problèmes de changement de domaine (*domain shift*). Deuxièmement, pour que l'entraînement soit efficace, de grands corpus de textes annotés dans le domaine d'intérêt sont nécessaires (Jia *et al.*, 2019; Liu *et al.*, 2021). Or, les campagnes d'annotation de REN clinique sont très coûteuses en temps et en ressources, nécessitant un haut niveau d'expertise du domaine pour être menées à bien (Luo *et al.*, 2020; Névéol *et al.*, 2014; Doğan *et al.*, 2014; Báez *et al.*, 2020). De plus, en raison de la diversité des cas cliniques, les données annotées pour une application biomédicale ne sont pas nécessairement transférables à une autre. D'où la nécessité de développer des approches de REN clinique efficaces en termes de données, également connues sous le nom de REN en *few-shot* (en peu d'exemples).

La deuxième approche, plus récente, consiste à utiliser des **modèles de langues causaux (CLM)** pré-entraînés. Ces modèles, considérablement plus grands, sont pré-entraînés sur des corpus (souvent plus importants) en tant que modèles génératifs et auto-régressifs. En d'autres termes, le modèle reçoit en entrée une série de tokens ou *prompt* et estime la série de tokens suivante la plus probable. Pour exploiter ces modèles de langues dans des tâches telles que la REN, il est possible de formuler la tâche

en langage naturel dans un *prompt*. Le *prompt* est conçu de manière à ce que la continuité du texte implique la résolution de la tâche. Le modèle de langue est ensuite utilisé pour prédire cette continuité. Ce processus est souvent appelé « *in-context learning* » (ICL) (Brown *et al.*, 2020). Éventuellement, il est possible de créer un *prompt* comprenant quelques exemples résolus de la tâche pour d'autres instances (dans ce cas, des instances annotées en entités nommées, spécifiques à la tâche), avant la nouvelle instance de test (Lee *et al.*, 2022). Le modèle produit ainsi une estimation de l'étiquetage en entités nommées le plus probable pour l'instance de test. Alors que les MLM ont été étudiés pour la REN en *few-shot* (Du *et al.*, 2021), les CLM semblent plus naturellement adaptés à ce contexte. L'apprentissage ICL a en fait démontré un succès particulier avec les CLM dans l'apprentissage en *few-shot*, montrant des résultats prometteurs dans un large éventail de tâches de TAL (Shin *et al.*, 2022; Wei *et al.*, 2022; Srivastava *et al.*, 2023).

Cependant, la supériorité des CLM sur les MLM pour la REN en *few-shot* est discutable. De nombreux efforts étudiant l'apprentissage en « *few-shot* » avec des CLM choisissent les *prompts* en fonction de leurs performances sur de grands jeux de données de validation (Brown *et al.*, 2020; Tam *et al.*, 2021; Radford *et al.*, 2021; Qin & Eisner, 2021). Cela pose un problème car il a été démontré que l'ICL dépendait fortement de la structure du *prompt* : un petit changement dans la formulation de la tâche, les exemples présentés, l'ordre des exemples ou le format d'étiquetage peut affecter la performance. Par conséquent, faire ces choix en supposant l'existence d'un grand jeu de données de validation annotées conduit à des résultats qui s'avèrent trop optimistes (Perez *et al.*, 2021) et impossibles à trouver dans un cadre réel de quelques exemples annotées. Deuxièmement, la plupart de ces études se sont principalement concentrées sur la langue anglaise et sur des modèles basés sur GPT (Wang *et al.*, 2023b; Ashok & Lipton, 2023; Hu *et al.*, 2023b; Jimenez Gutierrez *et al.*, 2022). Cela peut conduire à des *prompts* trop adaptés à cette langue et à ce modèle de langue. Il est donc nécessaire de mener une étude systématique, indépendante du modèle, sur l'élaboration des *prompts* dans le contexte clinique et pour des langues autres que l'anglais. Les contributions de ce travail sont les suivantes :

1. Nous présentons et comparons les techniques de *prompting* appliquées à la REN les plus récentes lorsqu'elles sont appliquées à la REN clinique dans trois langues : l'anglais, le français et l'espagnol. À notre connaissance, il s'agit du premier travail axé sur les *prompts* de REN pour les langues autres que l'anglais, et du premier travail comparant les *prompts* pour la REN clinique.
2. Nous accordons une attention particulière aux *prompts* de balisage, une technique proposée récemment (Wang *et al.*, 2023b), et nous mesurons les améliorations qu'elle apporte.
3. Nous offrons une comparaison juste avec les MLM les plus performants dans un cadre *few-shot*, à travers les langues, les modèles et les structures de *prompts*, lorsqu'ils sont appliqués à la REN clinique.
4. Nous menons des expériences facilement reproductibles, en utilisant des méthodes faciles à mettre en œuvre, exclusivement sur des ensembles de données et des modèles de langues publiquement accessibles.

2 Etat de l'art

Reconnaissance d'entité nommées en *few-shot* avec les MLM pré-entraînés La méthode classique pour utiliser les MLM dans la REN est de les utiliser comme encodeurs. Habituellement, une couche d'étiquetage REN est entraînée à partir de zéro pour projeter l'encodage du texte dans l'éti-

quetage REN de ses tokens (Devlin *et al.*, 2019). D'autres approches adaptent les MLM au contexte *few-shot*. L'apprentissage de métrique (Fritzler *et al.*, 2019; Yang & Katiyar, 2020; Huang *et al.*, 2021a) propose d'entraîner les systèmes à apprendre une métrique sur l'espace de sortie, permettant de classer de nouvelles instances en fonction de leur distance par rapport aux instances étiquetées. L'encodage des types d'entités (Aly *et al.*, 2021; Ma *et al.*, 2022a; Hou *et al.*, 2020) exploite les noms ou descriptions des types d'entités pour mieux les étiqueter.

Reconnaissance d'entité nommées en *few-shot* avec les CLM pré-entraînés Récemment, la construction de *prompts* a suscité l'intérêt de la communauté (Brown *et al.*, 2020; Liu *et al.*, 2023). Les travaux connexes se sont concentrés sur l'étude de la formulation du *prompt* (Wei *et al.*, 2022; Ashok & Lipton, 2023; Vilar *et al.*, 2023; Wang *et al.*, 2023b), également connue sous le nom de « *prompt engineering* », d'autres travaux ont proposé une optimisation continue du *prompt* (Ma *et al.*, 2022b; Layegh *et al.*, 2023; Hu *et al.*, 2023a).

Il n'existe pas de méthode standard, largement adoptée, pour construire les *prompts* de REN (Liu *et al.*, 2023). Trois familles de *prompts* principales émergent : Le ***prompting* contraint** tente de mieux formuler la tâche REN en contraignant la génération à remplir des patrons spécifiques créés à la main, généralement adaptés aux MLMs (Cui *et al.*, 2021; Shen *et al.*, 2023; Ye *et al.*, 2023; Schick & Schütze, 2021). Les ***prompts* de listage** consistent simplement à faire prédire au modèle de langues les entités sous forme de liste (Ashok & Lipton, 2023). Les ***prompts* de balisage** ont été étudiés plus récemment par (Wang *et al.*, 2023b). Ils font en sorte que le modèle de langues entoure les mentions d'entités avec des balises spéciales.

Reconnaissance d'entité nommées clinique en *few-shot* Peu d'études se sont concentrées sur la REN en *few-shot* clinique basée sur les CLM. Dans (Hu *et al.*, 2023b), GPT-3 et ChatGPT sont évalués sur la tâche i2b2/VA 2010 (Uzuner *et al.*, 2011) dans un contexte *few-shot*. Dans (Jimenez Gutierrez *et al.*, 2022), GPT-3 est évalué sur un ensemble de tâches d'extraction d'informations biomédicales, y compris le NCBI-Disease (Doğan *et al.*, 2014). Une autre approche intéressante consiste à affiner en partie (Liao *et al.*, 2023) un CLM de domaine général sur des textes cliniques (Han *et al.*, 2023; Toma *et al.*, 2023), et à *prompter* le CLM qui en résulte. Les MLMs ont également été explorés pour la REN en *few-shot* dans le domaine biomédical (Ge *et al.*, 2023). L'apprentissage de métrique (Yang & Katiyar, 2020) et l'encodage des types (Aly *et al.*, 2021; Ma *et al.*, 2022a) ont été étudiés, ainsi que d'autres approches telles que l'apprentissage actif (Kormilitzin *et al.*, 2021), le pré-apprentissage supervisé (Huang *et al.*, 2021b) et le *in-context learning* basé sur les MLMs (Lee *et al.*, 2022).

3 Expérimentation

Corpus utilisés Afin d'évaluer les modèles, nous utilisons 14 corpus annotés en entités nommées, accessibles publiquement. Pour chaque langue, nous avons choisi deux corpus hors domaine et deux ou trois corpus dans le domaine clinique, visant des ressources comparables (même genre, mêmes types, mêmes schémas d'annotation) entre les langues, dans la mesure du possible. Nous utilisons les sous-ensembles officiels d'entraînement, de validation et de test de chaque corpus, quand ceux-ci sont disponibles.

WikiNER (Nothman *et al.*, 2013) est un corpus multilingue annoté en entités nommées, extrait de

Wikipédia fin 2010 dans neuf langues, annotant automatiquement les hyperliens vers personnes, lieux ou organisations. Nous utilisons les versions anglaise, française et espagnole.

CoNLL-2002 (Tjong Kim Sang, 2002) et **CoNLL-2003** (Tjong Kim Sang & De Meulder, 2003) sont des corpus multilingues annotés manuellement en entités nommées de types personnes, lieux et organisations, publiés pour les tâches partagées CoNLL. Nous utilisons les données espagnoles de 2002 (une collection d'articles tirés de la presse espagnole) et les données anglaises de 2003 (articles de presse de Reuters).

Quaero French Press (Grouin *et al.*, 2011) est un corpus annoté manuellement d'émissions radiophoniques francophones, avec des annotations pour 5 types d'entités : les personnes, lieux, organisations, fonctions et installations.

E3C (Magnini *et al.*, 2021) est un corpus multilingue européen de textes cliniques collectés à partir de multiples sources telles que PubMed [1] et SciELO [2]. Nous utilisons les versions anglaise, française et espagnole de ce corpus, annotées sémantiquement en types d'entités, acteurs, parties du corps, événements, RMLs (mesures et résultats de tests) et entités cliniques.

La tâche partagée **n2c2-2019** (Luo *et al.*, 2020) se concentre sur la normalisation des concepts médicaux à partir du corpus MCN (Luo *et al.*, 2019), composé de compte-rendus d'hospitalisation tirés de deux établissements hospitaliers dans le Massachusetts. Nous utilisons les identifiants uniques de concept (CUI) pour associer les mentions aux groupes sémantiques UMLS. (Lindberg *et al.*, 1993; McCray *et al.*, 2001).

Le corpus **NCBI-Disease** (Doğan *et al.*, 2014) corpus rassemble des résumés PubMed où les mentions de maladies sont annotées en quatre types selon leur syntaxe : maladies spécifiques, classes de maladies, mentions composites et modificateurs.

QuaeroFrenchMed (Névéol *et al.*, 2014) se compose de deux parties : **EMEA**, une collection de notices patient concernant des médicaments commercialisés en Europe, et **MEDLINE**, de titres d'articles scientifiques indexés dans MEDLINE. Ces deux parties sont annotées en 10 types d'entités nommées, correspondant aux groupes sémantiques UMLS.

Le corpus **The Chilean Waiting List** (Báez *et al.*, 2020) contient ordonnances anonymisées pour des consultations à partir de la liste d'attente dans les hôpitaux publics chiliens, annotées manuellement avec 10 types d'entités : abréviations, parties du corps, résultats cliniques, procédure de diagnostic, maladies, membres de la famille, résultats de laboratoire ou de test, procédures de laboratoire, médicaments, procédures, signes ou symptômes et procédures thérapeutiques. Notons que ces types peuvent être redondants (par exemple, toutes les procédures de diagnostic sont également annotées en tant que procédures).

Configuration de l'apprentissage en *few-shot* Pour simuler le contexte de *few-shot*, nous fournissons aux modèles seulement quelques exemples annotés, représentant l'ensemble des exemples autorisés pour l'apprentissage, le *prompting* et la validation. Dans cette étude, nous choisissons de nous concentrer principalement sur $k = 100$ phrases, ce qui correspond à une à deux heures d'annotation dans le domaine clinique (Névéol *et al.*, 2014; Campillos *et al.*, 2018). Nous utilisons une graine aléatoire fixe p pour choisir k exemples parmi tous ceux disponibles dans le corpus réel. Dans la section 5.2, nous discutons de l'effet du choix de k et du choix de p.

1. http://pubmed.ncbi.nlm.nih.gov/
2. https://scielo.org/

En outre, nous testons les modèles les plus performants avec l'entièreté des annotations à disposition pour une comparaison à la *skyline*.

Modèles de langues Nous évaluons 10 modèles causaux et 16 modèles masqués sur les tâches de REN précisées ci-dessus. Ces modèles sont listés dans le tableau de résultats 1 et décrits en plus de détail dans l'annexe 1. Alors que le français et l'espagnol sont couverts par de certains des modèles causaux, nous pouvons observer que l'anglais est omniprésent. La plupart des modèles de type BERT sont monolingues, à l'exception des modèles multilingues mBERT et XLM-RoBERTa.

REN avec des modèles de langues masqués Comme mentionné dans la section 2, les modèles de langues masqués ont été adaptés à l'apprentissage en *few-shot* dans des architectures adaptées aux contextes *few-shot*. Cependant, dans ce travail, nous souhaitons comparer la nouvelle approche CLM à l'utilisation standard et plus répandue des MLM sans adaptation pour le contexte *few-shot*. Nous utilisons NLStruct (Wajsbürt, 2021), une bibliothèque Python open-source [3] qui met en œuvre l'approche standard du *fine-tuning*. NLStruct utilise les représentations fournies par le modèle de langues pour encoder l'entrée, puis utilise un décodeur LSTM bidirectionnel et un CRF pour prédire itérativement les entités présentes dans l'entrée encodée, comme décrit par Gérardin *et al.* (2022). Cette démarche permet à NLStruct de traiter efficacement les entités imbriquées, très présentes dans certains des corpus d'étude. Nous entraînons le modèle pendant 20 epochs sur 80 % des données et utilisons les 20 % restants pour valider l'*early stopping*.

REN avec des modèles de langues causaux Dans nos expériences, nous invitons les modèles à baliser les mentions, et non de les lister. Nous discutons de ce choix plus en détail dans la section 5.2. La partie supérieure de la figure 1 montre un exemple de *prompt* de balisage, en mettant en évidence les différentes sections de celui-ci. Ci-dessous, nous décrivons 9 caractéristiques de formulation du *prompt* et de sélection des exemples qui y figurent.

1. **Langue du *prompt*** : Par défaut, nous construisons les *prompts* en anglais, car il s'agit de la langue la plus répandue dans tous les corpus d'apprentissage. Cette caractéristique consiste à faire plutôt aligner la langue du *prompt* sur celle de la phrase de test.
2. **Phases supplémentaires** : Par défaut, nous présentons 5 phrases annotées dans les annotées. Cette caractéristique permet de présenter 5 phrases supplémentaires (soit 10 phrases au total). La partie 5.2 discute de ce choix, ainsi que de la possibilité de présenter plus de démonstrations dans le *prompt*.
3. **Auto-vérification** : Par défaut, nous sélectionnons les 5 (ou 10) phrases les plus proches de la phrase test en termes de distance TF-IDF. Les mentions étiquetées par le modèle sont alors considérées comme les prédictions finales du modèle. Cette fonctionnalité sélectionne plutôt les 5 phrases contenant le plus d'entités du type ciblé et les présente dans un *prompt* initial. Intuitivement, ce *prompt* se traduit par un rappel plus élevé et une précision plus faible. Un deuxième *prompt* d'« auto-vérification » est ensuite construit sur les prédictions initiales du modèle afin d'éliminer les faux positifs. Un exemple de *prompt* d'auto-vérification est présenté dans la partie inférieure de la figure 1. Le nombre de démonstrations suit celui du *prompt* principal.

3. https://github.com/percevalw/nlstruct

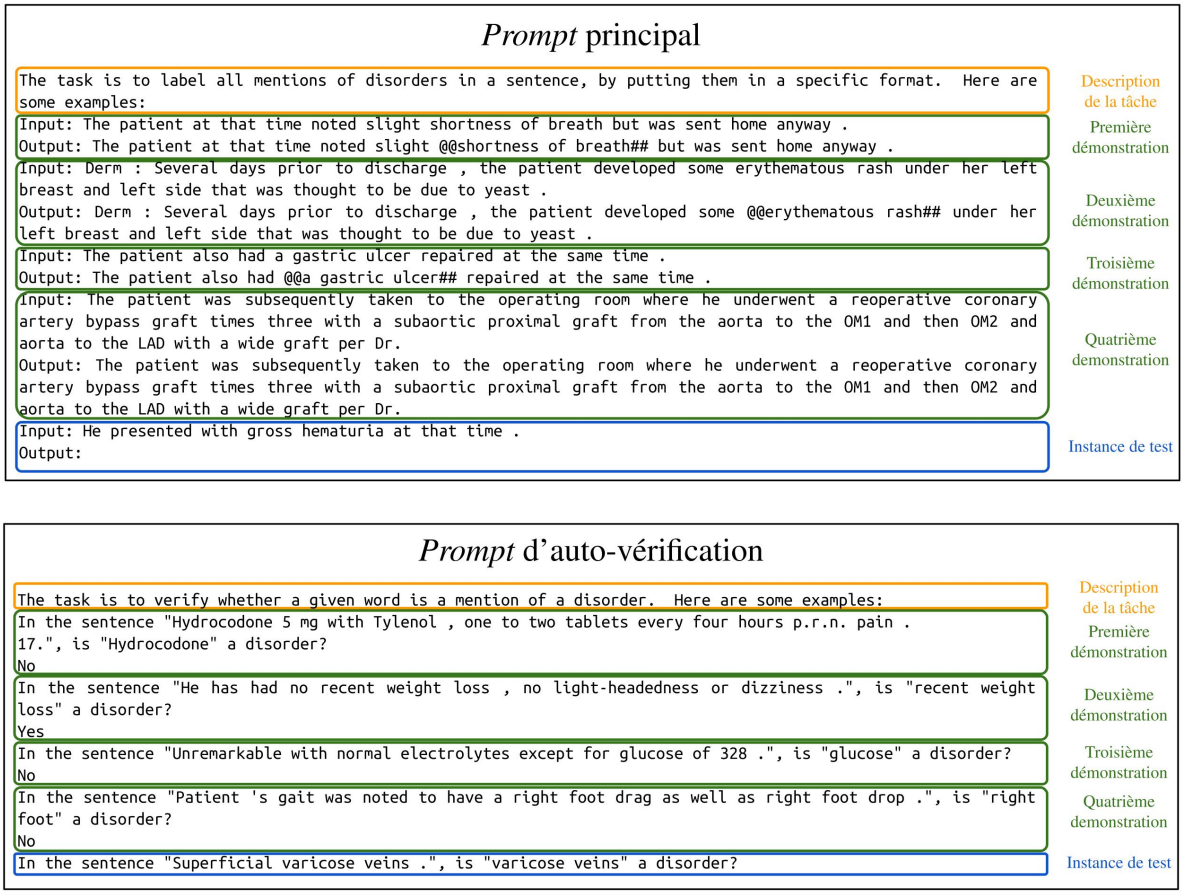


FIGURE 1 – Exemple d'un *prompt* de balisage, utilisée dans l'expérience principale (en haut) et d'un *prompt* d'auto-vérification (en bas) pour détecter les mentions DISO dans **n2c2-2019**

4. **Baliseurs** : Par défaut, nous suivons Wang *et al.* (2023b) qui invite le modèle à entourer les mentions de @@ et ##. Cette caractéristique l'invite plutôt à entourer les mentions de guillemets « et ».
5. **S'addresser à un spécialiste** : Par défaut, la première phrase est la description de la tâche présentée dans la figure 1. Cette caractéristique fait commencer le *prompt* par *You are an excellent <specialist>. You can identify all the mentions of <entity-type> in a sentence, by putting them in a specific format. Here are some examples you can handle :* à la place. Le *<specialist>* est un *linguist* ou un *clinician*, suivant le domaine de la tâche.
6. **Inclure les descriptions des types dans le *prompt*** : Cette caractéristique ajoute une description d'une phrase pour chaque type d'entité. Les descriptions complètes des entités utilisées figurent à l'annexe 3.
7. **Phrase d'introduction pour l'instance de test** : Par défaut, les démonstrations sont immédiatement suivies de l'instance de test. Cette fonctionnalité consiste à la précéder par *Identify all the mentions of <entity-type> in the following sentence, by putting <begin-tag> in front and a <end-tag> behind each of them.*.
8. **Demander une réponse longue pour l'auto-vérification** : Par défaut, le *prompt* d'auto-vérification demander *Yes* (respectivement *No*) comme réponse. Cette fonctionnalité demande *<mention> is a(n) <entity-type>, yes.* (respectivement *<mention> is not a(n) <entity-type>, no.*).

9. **Format dialogue** : Cette fonction remplace les *Input :* et *Output :* du *prompt* par des tirets pour imiter un format de dialogue.

Les performances de l'*in-context learning* varient considérablement en fonction de la formulation exacte du *prompt* (Lu *et al.*, 2022; Min *et al.*, 2022). En outre, le choix optimal de chacune de ces caractéristiques peut varier en fonction du modèle utilisé. Par exemple, intuitivement, les modèles dont le pré-entraînement est fortement concentrés sur la langue anglaise ont tendance à être plus performants avec un *prompt* en anglais qu'avec un *prompt* dans la langue du corpus.

Notre système vise à rechercher la meilleure combinaison de caractéristique pour chaque modèle, mais un *grid search* sur ces caractéristique nécessiterait $2^9 = 512$ expériences pour chaque modèle et pour chaque corpus. Afin de construire un système plus léger, nous avons choisi d'effectuer un *greedy search*. Nous itérons sur les caractéristiques dans cet ordre, en testant la valeur qui n'est pas celle par défaut et en la conservant si elle est plus performante que la valeur par défaut. Dans la section 5.2, nous comparons cette approche à un *grid search* pour un modèle sur un corpus.

De nombreux travaux sur l'apprentissage en « *few-shot* » avec les CLM optimisent les *prompts* sur de grands jeux de données de validation (Brown *et al.*, 2020; Tam *et al.*, 2021; Radford *et al.*, 2021; Qin & Eisner, 2021). Cela conduit à des résultats qui se révèlent (Perez *et al.*, 2021) trop optimistes. Une comparaison équitable entre les MLM et les CLM devrait les comparer avec l'accès au même (petit) nombre d'instances annotées, ce qui correspond à notre $k = 100$. Dans ce contexte d'absence d'entraînement, nous suivons Perez *et al.* (2021) en optimisant ces caractéristiques par une validation LOOCV (leave-one-out cross-validation).

Mesures Nous évaluons la performance des modèles à l'aide de deux mesures. Pour des raisons de simplicité, nous évaluons les modèles sur la base d'un score de performance global, la **micro-F1**. Il est calculé comme la micro-moyenne des F1-mesures de la détection de chaque type d'entité. Nous mesurons également l'**empreinte carbone** de chacune des approches. Nous utilisons GreenAlgorithms v2.2 (Lannelongue *et al.*, 2021) [4] pour estimer l'empreinte carbone de chaque expérience, sur la base de facteurs tels que la durée d'exécution, le matériel informatique et le lieu de production de l'électricité utilisée par notre installation informatique.

4 Résultats et discussion

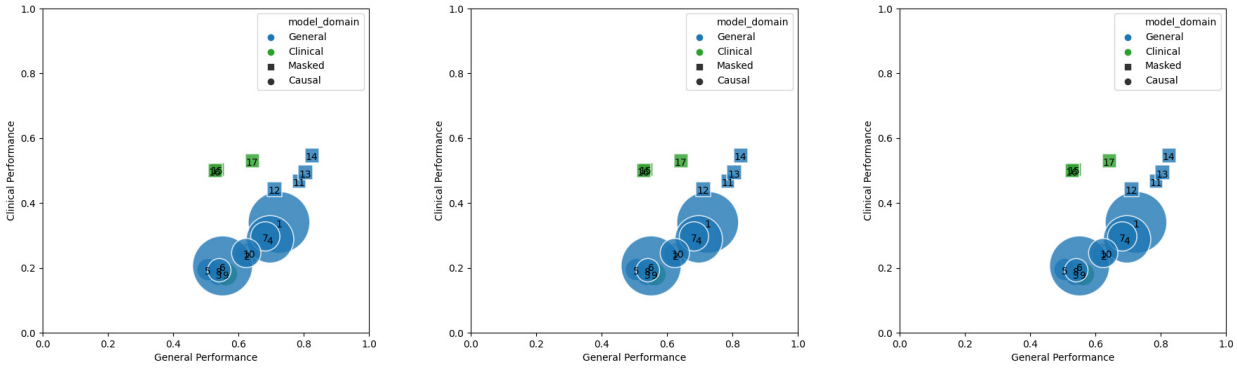
Mesures Le tableau 1 et la figure 2 décrivent les performances des modèles testés. L'annexe 4 détaille les estimations des émissions carbone pour toutes nos expériences. En particulier, nous estimons que l'expérience utilisant Mistral-7B sur CoNLL-2003 a généré 41 g d'équivalent CO2. (6 g pour l'optimisation du *prompt* et 35 g pour l'inférence sur l'ensemble de test). LLaMA-2-70B, environ 10 fois plus grand, est estimé avoir généré 191 g d'équivalent CO2. (44 g pour l'optimisation du *prompt* et 147 g pour l'inférence sur l'ensemble de test). L'expérience sur le modèle de langue masqué BERT-large est quant à elle estimée avoir généré 6 g d'équivalent CO2. (2 g pour le fine-tuning et l'entraînement et 4 g pour l'inférence sur l'ensemble de test).

Au total, on estime que les expériences décrites dans cet article ont généré environ 27 kg d'équivalent CO2 (25 kg pour les expériences principales et 2 kg pour l'ablation).

4. http ://calculator.green-algorithms.org/

	#	Modèle	Anglais					Français					Espagnol			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Approches few-shot																
Causal	1	LLAMA-2-70B	0.728	0.721	0.312	0.309	0.400	0.740	0.400	0.483	0.201	0.312	0.805	0.616	0.021	0.339
	2	Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374
	3	BLOOM-7B1	0.524	0.557	0.279	0.113	0.151	0.148	0.206	0.320	0.197	0.120	0.470	0.419	0.051	0.117
	4	Falcon-40B	0.686	0.708	0.280	0.279	0.305	0.662	0.456	0.378	0.279	0.283	0.720	0.543	0.072	0.267
	5	GPT-J-6B	0.521	0.493	0.167	0.179	0.238	0.423	0.244	0.334	0.080	0.177	0.005	0.142	0.021	0.162
	6	OPT-66B	0.608	0.495	0.227	0.157	0.234	0.624	0.406	0.019	0.206	0.283	0.166	0.273	0.043	0.204
	7	Vicuna-13B	0.657	0.708	0.355	0.236	0.300	0.677	0.350	0.399	0.207	0.326	0.744	0.250	0.040	0.213
	8	Vicuna-7B	0.594	0.489	0.259	0.147	0.172	0.591	0.277	0.439	0.152	0.296	0.659	0.569	0.042	0.151
	9	Medalpaca-7B	0.537	0.586	0.272	0.138	0.132	0.529	0.142	0.259	0.162	0.252	0.581	0.490	0.088	0.220
	10	Vigogne-13B	0.593	0.655	0.252	0.176	0.309	0.515	0.250	0.464	0.099	0.142	0.580	0.561	0.010	0.198
Masked	11	mBERT	0.768	0.804	0.624	0.378	0.401	0.801	0.728	0.741	0.588	0.428	0.812	0.760	0.324	0.432
	12	XLM-R-large	0.786	0.826	0.637	0.462	0.471	0.811	0.781	0.762	0.629	0.531	0.797	0.781	0.325	0.528
	13	BERT-large	0.776	0.835	0.626	0.435	0.422	-	-	-	-	-	-	-	-	-
	14	RoBERTa-large	0.790	0.862	0.626	0.462	0.552	-	-	-	-	-	-	-	-	-
	15	Bio_ClinicalBERT	0.528	0.542	0.621	0.469	0.420	-	-	-	-	-	-	-	-	-
	16	ClinicalBERT	0.462	0.597	0.622	0.480	0.397	-	-	-	-	-	-	-	-	-
	17	MedBERT	0.613	0.673	0.607	0.478	0.504	-	-	-	-	-	-	-	-	-
	18	CamemBERT-large	-	-	-	-	-	0.829	0.793	0.768	0.661	0.577	-	-	-	-
	19	FlauBERT-large	-	-	-	-	-	0.826	0.778	0.760	0.635	0.542	-	-	-	-
	20	DrBERT-4GB	-	-	-	-	-	0.587	0.599	0.730	0.602	0.486	-	-	-	-
	21	CamemBERT-bio	-	-	-	-	-	0.782	0.761	0.779	0.636	0.549	-	-	-	-
	22	BETO	-	-	-	-	-	-	-	-	-	-	0.794	0.732	0.352	0.522
	23	PatanaBERT	-	-	-	-	-	-	-	-	-	-	0.802	0.769	0.343	0.487
	24	TulioBERT	-	-	-	-	-	-	-	-	-	-	0.804	0.798	0.340	0.482
	25	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	0.804	0.758	0.354	0.578
	26	BSC-Bio	-	-	-	-	-	-	-	-	-	-	0.804	0.775	0.358	0.552
Skyline en utilisant toutes les données à disposition																
		RoBERTa-large	0.919	0.939	0.718	0.712	0.815	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	0.928	0.834	0.828	0.748	0.713	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	0.918	0.881	0.411	0.736

TABLE 1 – Mesures micro-F1 obtenues. Nous évaluons les modèles masqués monolingues uniquement dans les langues sur lesquelles ils ont été entraînés.



(a) Anglais (b) Français (c) Espagnol

FIGURE 2 – Performance dans les domaines général vs. clinique des modèles étudiés

Comparaison des performances des modèles Nous avons comparé différents modèles de langues masqués (MLM) et causaux (CLM) pour la reconnaissance d'entités nommées en nous concentrant sur les contextes à faibles ressources, typiques des applications biomédicales. Les résultats montrent que les modèles masqués de type BERT, bien que plus petits et nécessitant théoriquement une plus grande quantité de données d'apprentissage, surpassent systématiquement les CLM. Cette performance s'accompagne d'un impact environnemental beaucoup plus faible (les émissions de CO2 sont 10 à 50 fois inférieures pour les MLM par rapport aux CLM), et d'une plus grande consistance (par exemple, sur la tâche généraliste WikiNER en anglais, les 4 modèles du domaine général testés ont

obtenu des scores F1 compris entre 0,768 et 0,79). Par ailleurs, les MLMs spécialisés dans le domaine biomédical apportent peu d'amélioration significative dans les tâches spécialisées. Ce commentaire doit cependant être pondéré par la différence de taille entre les modèles : tous les modèles spécialisés ont seulement 110 millions de paramètres.

La reconnaissance d'entités nommées basée sur des représentations de type BERT a reçu beaucoup d'attention ces dernières années, et est sans aucun doute plus mature que l'utilisation des CLMs pour cette tâche. Nous avons exploré les techniques de reconnaissance d'entités nommées basées sur les CLMs existantes dans la littérature, avec nos connaissances actuelles. De nouvelles approches pourraient améliorer les performances à l'avenir, mais cette tâche reste difficile pour un modèle génératif en raison de ses contraintes syntaxiques et d'évaluation spécifiques. Ces résultats ne reflètent pas nécessairement les performances sur d'autres tâches, comme la classification.

Usage pratique des modèles de langues pour la REN à peu de ressources Nos expériences indiquent que les modèles de langues pour la reconnaissance des entités nommées cliniques ont actuellement des performances sous-optimales. Même les modèles MLM, *fine-tunés* simplement avec le peu de données à disposition, ne rivalisent pas avec les modèles entièrement supervisés. Les grands modèles entraînés avec l'ensemble de chaque corpus d'entraînement surpassent systématiquement les meilleurs résultats en *few-shot*, de 5 à 16 % pour le domaine général et de 8 à 48 % pour le domaine biomédical (*skylines* Table 1). Cependant, les performances peuvent suffire pour une utilisation en pré-annotation, accélérant ainsi l'annotation manuelle, par exemple dans un contexte d'*online learning* ou d'*active learning*.

Bruit aléatoire Dans les expériences MLM, les paramètres de la couche d'étiquetage REN ajoutée au modèle pré-entraîné sont initialisés de manière aléatoire. De même, dans les expériences CLM, les démonstrations dans les *prompts* sont ordonnées de façon aléatoire et les exemples négatifs dans les *prompts* d'auto-vérification sont sélectionnés de manière aléatoire, introduisant potentiellement du bruit dans nos mesures de performance. Répliquer toutes les expériences renforcerait nos conclusions (Reimers & Gurevych, 2017), mais il serait coûteux (25kg de CO2eq et 56 heures de calcul par réplication). Le grand nombre de modèles testés et de tâches traitées peut toutefois conforter les principales observations de cet article. Par exemple, nous utilisons l'ordre presque stochastique (ASO)[5] (Dror *et al.*, 2019) avec $\alpha = 0,05$ pour mesurer la significativité de la supériorité des MLM sur les CLM pour chaque ensemble de données séparément. Les MLM ne montrent pas toujours une supériorité significative sur les CLM pour la REN dans le domaine général (0,54 et 0,121 respectivement pour WikiNER anglais et CoNLL2003). Pour la REN clinique, les MLM sont nettement supérieurs aux CLM : les MLM dominent stochastiquement les CLM (ϵ_{min}=0) pour tous les corpus cliniques.

Conclusion Cette étude a évalué les performances de deux types de modèles de langues pour la reconnaissance d'entités en *few-shot* dans trois langues. Nos expériences révèlent que la performance du *few-shot learning* est significativement plus faible dans le domaine clinique que dans le domaine

5. Étant donné les scores de performance de deux algorithmes A et B, chacun étant exécuté plusieurs fois avec des paramètres différents, l'ASO calcule une valeur spécifique au test (ϵ_{min}) qui indique à quel point l'algorithme A est loin d'être significativement meilleur que l'algorithme B. Si la distance ϵ_{min} = 0,0, on peut affirmer que A domine stochastiquement B avec le niveau de signification prédéfini. La littérature interprète généralement ϵ_{min} < 0,5 comme un indicateur de supériorité significative de A sur B.

général. Alors que les modèles de langues masqués surpassent les modèles de langues causaux (avec une F1-mesure plus élevée et des émissions de CO2 plus faibles), leur utilisation devrait être restreinte à la pré-annotation plutôt qu'à l'extraction d'informations efficace.

Références

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JINDI D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical BERT embeddings. In A. RUMSHISKY, K. ROBERTS, S. BETHARD & T. NAUMANN, Éds., *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

ALY R., VLACHOS A. & MCDONALD R. (2021). Leveraging type descriptions for zero-shot named entity recognition and classification. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1516–1528, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.120.

ASHOK D. & LIPTON Z. (2023). Promptner : Prompting for named entity recognition.

BÁEZ P., VILLENA F., ROJAS M., DURÁN M. & DUNSTAN J. (2020). The Chilean waiting list corpus : a new resource for clinical named entity recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, p. 291–300, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.clinicalnlp-1.32.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

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**, 571–601.

CARRINO C. P., LLOP J., PÀMIES M., GUTIÉRREZ-FANDIÑO A., ARMENGOL-ESTAPÉ J., SILVEIRA-OCAMPO J., VALENCIA A., GONZALEZ-AGIRRE A. & VILLEGAS M. (2022). Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, p. 193–199, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bionlp-1.19.

CAÑETE J., CHAPERON G., FUENTES R., HO J.-H., KANG H. & PÉREZ J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

CHO H., CHOI W. & LEE H. (2017). A method for named entity normalization in biomedical articles : Application to diseases and plants. *BMC Bioinformatics*, **18**. DOI : 10.1186/s12859-017-1857-8.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds.,

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

CUI L., WU Y., LIU J., YANG S. & ZHANG Y. (2021). Template-based named entity recognition using BART. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1835–1845, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.161.

DEMNER-FUSHMAN D., CHAPMAN W. W. & MCDONALD C. J. (2009). What can natural language processing do for clinical decision support ? *Journal of Biomedical Informatics*, **42**(5), 760–772. Biomedical Natural Language Processing, DOI : https ://doi.org/10.1016/j.jbi.2009.08.007.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DOĞAN R. I., LEAMAN R. & LU Z. (2014). Ncbi disease corpus : A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, **47**, 1–10. DOI : https ://doi.org/10.1016/j.jbi.2013.12.006.

DROR R., SHLOMOV S. & REICHART R. (2019). Deep dominance - how to properly compare deep neural models. In A. KORHONEN, D. R. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1 : Long Papers*, p. 2773–2785 : Association for Computational Linguistics. DOI : 10.18653/v1/p19-1266.

DU S. S., HU W., KAKADE S. M., LEE J. D. & LEI Q. (2021). Few-shot learning via learning the representation, provably.

ESCUDIÉ J.-B., RANCE B., MALAMUT G., KHATER S., BURGUN A., CELLIER C. & JANNOT A.-S. (2017). A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease : a case study on autoimmune comorbidities in patients with celiac disease. *BMC medical informatics and decision making*, **17**(1), 1–10.

FRITZLER A., LOGACHEVA V. & KRETOV M. (2019). Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, p. 993–1000, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3297280.3297378.

GAO L., BIDERMAN S., BLACK S., GOLDING L., HOPPE T., FOSTER C., PHANG J., HE H., THITE A., NABESHIMA N. *et al.* (2020). The pile : An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv :2101.00027*.

GE Y., GUO Y., DAS S., AL-GARADI M. A. & SARKER A. (2023). Few-shot learning for medical text : A review of advances, trends, and opportunities. *Journal of Biomedical Informatics*, **144**, 104458. DOI : https ://doi.org/10.1016/j.jbi.2023.104458.

GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In N. IDE, A. MEYERS, S. PRADHAN & K. TOMANEK, Éds., *Proceedings of the 5th Linguistic Annotation Workshop*, p. 92–100, Portland, Oregon, USA : Association for Computational Linguistics.

GUPTA S., GARDNER M. & SINGH S. (2023). Coverage-based example selection for in-context learning. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational*

Linguistics : EMNLP 2023, p. 13924–13950, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.930.

GÉRARDIN C., WAJSBÜRT P., VAILLANT P., BELLAMINE A., CARRAT F. & TANNIER X. (2022). Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, **128**, 102311. DOI : https ://doi.org/10.1016/j.artmed.2022.102311.

HAN T., ADAMS L. C., PAPAIOANNOU J.-M., GRUNDMANN P., OBERHAUSER T., LÖSER A., TRUHN D. & BRESSEM K. K. (2023). Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv :2304.08247*.

HOU Y., CHE W., LAI Y., ZHOU Z., LIU Y., LIU H. & LIU T. (2020). Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1381–1393, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.128.

HU N., ZHOU X., XU B., LIU H., XIE X. & ZHENG H.-T. (2023a). Vpn : Variation on prompt tuning for named-entity recognition. *Applied Sciences*, **13**(14). DOI : 10.3390/app13148359.

HU Y., AMEER I., ZUO X., PENG X., ZHOU Y., LI Z., LI Y., LI J., JIANG X. & XU H. (2023b). Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv :2303.16416*.

HUANG J., LI C., SUBUDHI K., JOSE D., BALAKRISHNAN S., CHEN W., PENG B., GAO J. & HAN J. (2021a). Few-shot named entity recognition : An empirical baseline study. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 10408–10423, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.813.

HUANG J., LI C., SUBUDHI K., JOSE D., BALAKRISHNAN S., CHEN W., PENG B., GAO J. & HAN J. (2021b). Few-shot named entity recognition : An empirical baseline study. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 10408–10423, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.813.

JIA C., LIANG X. & ZHANG Y. (2019). Cross-domain NER using cross-domain language modeling. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2464–2474, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1236.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L. *et al.* (2023). Mistral 7b. *arXiv preprint arXiv :2310.06825*.

JIMENEZ GUTIERREZ B., MCNEAL N., WASHINGTON C., CHEN Y., LI L., SUN H. & SU Y. (2022). Thinking about GPT-3 in-context learning for biomedical IE ? think again. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2022*, p. 4497–4512, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-emnlp.329.

JOHNSON A. E., POLLARD T. J., SHEN L., LEHMAN L.-W. H., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., ANTHONY CELI L. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, **3**(1), 1–9.

KORMILITZIN A., VACI N., LIU Q. & NEVADO-HOLGADO A. (2021). Med7 : A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, **118**, 102086. DOI : https ://doi.org/10.1016/j.artmed.2021.102086.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). DrBERT : A robust pre-trained model in French for biomedical and clinical domains. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16207–16221, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LANNELONGUE L., GREALEY J. & INOUYE M. (2021). Green algorithms : quantifying the carbon footprint of computation. *Advanced science*, **8**(12), 2100707.

LAURENÇON H., SAULNIER L., WANG T., AKIKI C., VILLANOVA DEL MORAL A., LE SCAO T., VON WERRA L., MOU C., GONZÁLEZ PONFERRADA E., NGUYEN H. *et al.* (2022). The bigscience roots corpus : A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, **35**, 31809–31826.

LAYEGH A., PAYBERAH A. H., SOYLU A., ROMAN D. & MATSKIN M. (2023). Contrastner : Contrastive-based prompt tuning for few-shot ner. *arXiv preprint arXiv :2305.17951*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for French. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

LEAMAN R., KHARE R. & LU Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, **57**, 28–37. DOI : https ://doi.org/10.1016/j.jbi.2015.07.010.

LEE D.-H., KADAKIA A., TAN K., AGARWAL M., FENG X., SHIBUYA T., MITANI R., SEKIYA T., PUJARA J. & REN X. (2022). Good examples make a faster learner : Simple demonstration-based learning for low-resource NER. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2687–2700, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.192.

LI J., SUN A., HAN J. & LI C. (2022). A survey on deep learning for named entity recognition. *IEEE Trans. on Knowl. and Data Eng.*, **34**(1), 50–70. DOI : 10.1109/TKDE.2020.2981314.

LIAO B., MENG Y. & MONZ C. (2023). Parameter-efficient fine-tuning without introducing new latency. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4242–4260, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.233.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The unified medical language system. *Yearbook of medical informatics*, **2**(01), 41–51.

LIU P., YUAN W., FU J., JIANG Z., HAYASHI H. & NEUBIG G. (2023). Pre-train, prompt, and predict : A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, **55**(9), 1–35.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

LIU Z., XU Y., YU T., DAI W., JI Z., CAHYAWIJAYA S., MADOTTO A. & FUNG P. (2021). Crossner : Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(15), 13452–13460. DOI : 10.1609/aaai.v35i15.17587.

LU Y., BARTOLO M., MOORE A., RIEDEL S. & STENETORP P. (2022). Fantastically ordered prompts and where to find them : Overcoming few-shot prompt order sensitivity. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8086–8098, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.556.

LUO Y.-F., HENRY S., WANG Y., SHEN F., UZUNER O. & RUMSHISKY A. (2020). The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, **27**(10), 1529–e1. DOI : 10.1093/jamia/ocaa106.

LUO Y.-F., SUN W. & RUMSHISKY A. (2019). Mcn : A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, **92**, 103132. DOI : https ://doi.org/10.1016/j.jbi.2019.103132.

MA J., BALLESTEROS M., DOSS S., ANUBHAI R., MALLYA S., AL-ONAIZAN Y. & ROTH D. (2022a). Label semantics for few shot named entity recognition. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Findings of the Association for Computational Linguistics : ACL 2022*, p. 1956–1971, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.155.

MA R., ZHOU X., GUI T., TAN Y., LI L., ZHANG Q. & HUANG X. (2022b). Template-free prompt tuning for few-shot NER. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5721–5732, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.420.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2021). The e3c project : European clinical case corpus. *Language*, **1**(L2), L3.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MCCRAY A., BURGUN A. & BODENREIDER O. (2001). Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, **84**, 216–20. DOI : 10.3233/978-1-60750-928-8-216.

MIN S., LYU X., HOLTZMAN A., ARTETXE M., LEWIS M., HAJISHIRZI H. & ZETTLEMOYER L. (2022). Rethinking the role of demonstrations : What makes in-context learning work ? In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048–11064, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.759.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The quaero french medical corpus : A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, p. 24–30.

NOTHMAN J., RINGLAND N., RADFORD W., MURPHY T. & CURRAN J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, **194**, 151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources, DOI : https ://doi.org/10.1016/j.artint.2012.03.006.

PENEDO G., MALARTIC Q., HESSLOW D., COJOCARU R., CAPPELLI A., ALOBEIDLI H., PANNIER B., ALMAZROUEI E. & LAUNAY J. (2023). The refinedweb dataset for falcon llm : outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv :2306.01116*.

PEREZ E., KIELA D. & CHO K. (2021). True few-shot learning with language models. In A. BEYGELZIMER, Y. DAUPHIN, P. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems*.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1202.

QIN G. & EISNER J. (2021). Learning how to ask : Querying LMs with mixtures of soft prompts. 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. 5203–5212, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.410.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G. & SUTSKEVER I. (2021). Learning transferable visual models from natural language supervision. In M. MEILA & T. ZHANG, Éds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 de *Proceedings of Machine Learning Research*, p. 8748–8763 : PMLR.

REIMERS N. & GUREVYCH I. (2017). Reporting score distributions makes a difference : Performance study of LSTM-networks for sequence tagging. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 338–348, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1035.

SCHICK T. & SCHÜTZE H. (2021). It's not just size that matters : Small language models are also few-shot learners. 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. 2339–2352, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.185.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHEN Y., TAN Z., WU S., ZHANG W., ZHANG R., XI Y., LU W. & ZHUANG Y. (2023). PromptNER : Prompt locating and typing for named entity recognition. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 12492–12507, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.698.

SHIN S., LEE S.-W., AHN H., KIM S., KIM H., KIM B., CHO K., LEE G., PARK W., HA J.-W. & SUNG N. (2022). On the effect of pretraining corpora on in-context learning by a large-scale

language model. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5168–5186, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.380.

SRIVASTAVA A., RASTOGI A., RAO A. & CO AUTHORS (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

SUÁREZ P. J. O., ROMARY L. & SAGOT B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv :2006.06202*.

SUN C., YANG Z., WANG L., ZHANG Y., LIN H. & WANG J. (2021). Biomedical named entity recognition using bert in the machine reading comprehension framework. *Journal of Biomedical Informatics*, **118**, 103799. DOI : https ://doi.org/10.1016/j.jbi.2021.103799.

SUNG M., JEONG M., CHOI Y., KIM D., LEE J. & KANG J. (2022). BERN2 : an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, **38**(20), 4837–4839. DOI : 10.1093/bioinformatics/btac598.

TAM D., R. MENON R., BANSAL M., SRIVASTAVA S. & RAFFEL C. (2021). Improving and simplifying pattern exploiting training. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4980–4991, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.407.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in OPUS. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 2214–2218, Istanbul, Turkey : European Language Resources Association (ELRA).

TJONG KIM SANG E. F. (2002). Introduction to the CoNLL-2002 shared task : Language-independent named entity recognition. In *COLING-02 : The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147.

TOMA A., LAWLER P. R., BA J., KRISHNAN R. G., RUBIN B. B. & WANG B. (2023). Clinical camel : An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv :2305.12031*.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2023). CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In C. SERVAN & A. VILNAT, Éds., *18e Conférence en Recherche d'Information et Applications*
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, p. 323–334, Paris, France : ATALA. HAL : hal-04130187.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

UZUNER Ö., SOUTH B. R., SHEN S. & DUVALL S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, **18**(5), 552–556.

VASANTHARAJAN C., TUN K. Z., THI-NGA H., JAIN S., RONG T. & SIONG C. E. (2022). Medbert : A pre-trained language model for biomedical named entity recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, p. 1482–1488. DOI : 10.23919/APSIPAASC55919.2022.9980157.

VILAR D., FREITAG M., CHERRY C., LUO J., RATNAKAR V. & FOSTER G. (2023). Prompting PaLM for translation : Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15406–15427, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.859.

WAJSBÜRT P. (2021). *Extraction and normalization of simple and structured entities in medical documents*. Theses, Sorbonne Université. HAL : tel-03624928.

WAJSBÜRT P., SARFATI A. & TANNIER X. (2021). Medical concept normalization in french using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics*, **114**, 103684. DOI : https ://doi.org/10.1016/j.jbi.2021.103684.

WANG B. & KOMATSUZAKI A. (2021). GPT-J-6B : A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

WANG G., LIU X., YING Z., YANG G., CHEN Z., LIU Z., ZHANG M., YAN H., LU Y., GAO Y. *et al.* (2023a). Optimized glycemic control of type 2 diabetes with reinforcement learning : a proof-of-concept trial. *Nature Medicine*, p. 1–10.

WANG S., SUN X., LI X., OUYANG R., WU F., ZHANG T., LI J. & WANG G. (2023b). Gpt-ner : Named entity recognition via large language models.

WANG Y., TONG H., ZHU Z. & LI Y. (2022). Nested named entity recognition : A survey. *ACM Trans. Knowl. Discov. Data*, **16**(6). DOI : 10.1145/3522593.

WANG Y., WANG L., RASTEGAR-MOJARAD M., MOON S., SHEN F., AFZAL N., LIU S., ZENG Y., MEHRABI S., SOHN S. & LIU H. (2018). Clinical information extraction applications : A literature review. *Journal of Biomedical Informatics*, **77**, 34–49. DOI : https ://doi.org/10.1016/j.jbi.2017.11.011.

WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V., ZHOU D. *et al.* (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, **35**, 24824–24837.

WORKSHOP B., SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

YANG Y. & KATIYAR A. (2020). Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6365–6375, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.516.

YE F., HUANG L., LIANG S. & CHI K. (2023). Decomposed two-stage prompt learning for few-shot named entity recognition. *Information*, **14**(5). DOI : 10.3390/info14050262.

ZHANG S., ROLLER S., GOYAL N., ARTETXE M., CHEN M., CHEN S., DEWAN C., DIAB M., LI X., LIN X. V. *et al.* (2022). Opt : Open pre-trained transformer language models. *arXiv preprint arXiv :2205.01068*.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. *et al.* (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv :2306.05685*.

ZHU Y., KIROS R., ZEMEL R., SALAKHUTDINOV R., URTASUN R., TORRALBA A. & FIDLER S. (2015). Aligning books and movies : Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, p. 19–27.

5 Annexes

5.1 Modèles évalués

	#	Modèle	Nombre de paramètres	Taille des données d'entraînement	Corpus d'entraînement
Causal	1	LLAMA-2-70B[en] (Touvron *et al.*, 2023)	70B	2 billions de tokens	Un mélange de corpus disponibles publiquement, principalement en anglais
	2	Mistral-7B[?] (Jiang *et al.*, 2023)	7B	Non divulgué	Non divulgué
	3	BLOOM-7B1[en][fr][es] (Workshop *et al.*, 2022)	7B	1,6 To	ROOTS (Laurençon *et al.*, 2022), un mélange de corpus et de données pseudo-crawlées dans 59 langues
	4	Falcon-40B[en][fr][es]	40B	1 billion de tokens	RefinedWeb (Penedo *et al.*, 2023), un ensemble de données de Web filtrées et dédoublonnées
	5	GPT-J-6B[en] (Wang & Komatsuzaki, 2021)	6B	825 Gio	The Pile (Gao *et al.*, 2020), un mélange de corpus publics et de données Web en anglais
	6	OPT-66B[en] (Zhang *et al.*, 2022)	66B	180 milliards de tokens	Données crawlées sur le Web, principalement en anglais
	7	Vicuna-13B[en]* (Zheng *et al.*, 2023)	13B	125K conversations	LLAMA 2, affiné sur des conversations collectées sur ShareGPT.com, principalement en anglais
	8	Vicuna-7B[en]* (Zheng *et al.*, 2023)	7B	125K conversations	LLAMA 2, affiné sur des conversations collectées sur ShareGPT.com, principalement en anglais
	9	Medalpaca-7B[en]* (Han *et al.*, 2023)	7B	400K paires Q.R.	LLAMA 2, affiné sur des paires de questions-réponses médicales semi-générées en anglais
	10	Vigogne-13B[fr][en]*	13B	52K instructions	LLAMA 2, affiné sur des instructions en anglais automatiquement traduites en français
Masked	11	mBERT[en][fr][es] (Devlin *et al.*, 2019)	110M	Non divulgué	Un corpus comprenant 104 langues construit à partir de sources non divulguées
	12	XLM-R-large[en][fr][es] (Conneau *et al.*, 2020)	355M	2,5 To	Données CommonCrawl filtrées contenant 100 langues
	13	BERT-large[en] (Devlin *et al.*, 2019)	345M	3,3 milliards de mots	BookCorpus (Zhu *et al.*, 2015), un corpus composé de livres non publiés et de Wikipédia en anglais.
	14	RoBERTa-large[en] (Liu *et al.*, 2019)	355M	160 Gio	BooksCorpus (Zhu *et al.*, 2015), Wikipédia en anglais, et données Web crawlées
	15	Bio_ClinicalBERT[en] (Alsentzer *et al.*, 2019)	110M	2 millions de notes cliniques	MIMIC-III (Johnson *et al.*, 2016), une base de données contenant des dossiers médicaux électroniques de patients en soins intensifs hospitalisés
	16	ClinicalBERT[en] (Wang *et al.*, 2023a)	110M	1,2 milliard de mots	non divulgué
	17	MedBERT[en] (Vasantharajan *et al.*, 2022)	110M	57 millions de mots	Plusieurs corpus publics (y compris N2C2 (Luo *et al.*, 2020)) et articles médicaux crawlés depuis Wikipédia
	18	CamemBERT-large[fr] (Martin *et al.*, 2020)	335M	64 milliards de tokens	OSCAR (Suárez *et al.*, 2020), un corpus de données Web en français
	19	FlauBERT-large[fr] (Le *et al.*, 2020)	335M	13 milliards de tokens	Un mélange de Wikipédia français, de livres français et de données Web français
	20	DrBERT-4GB[fr] (Labrak *et al.*, 2023)	110M	1 milliard de mots	Un mélange de corpus biomédicaux disponibles publiquement en français (dont QuaeroFrenchMed (Névéol *et al.*, 2014)).
	21	CamemBERT-bio[fr] (Touchent *et al.*, 2023)	110M	413 millions de mots	Un mélange de corpus biomédicaux disponibles publiquement en français (dont E3C (Magnini *et al.*, 2021)).
	22	BETO[es] (Cañete *et al.*, 2020)	110M	3 milliards de mots	Wikipédia en espagnol et données espagnoles d'OPUS (Tiedemann, 2012)
	23	PatanaBERT[es]	110M	Non divulgué	Espagnol
	24	TulioBERT[es]	110M	Non divulgué	Espagnol
	25	BSC-BioEHR[es] (Carrino *et al.*, 2022)	110M	1,1 milliard de tokens	Un mélange de corpus biomédicaux, y compris des documents EHR et des données crawlées en espagnol
	26	BSC-Bio[es] (Carrino *et al.*, 2022)	110M	963 millions de tokens	Un mélange de corpus biomédicaux et de données crawlées en espagnol

TABLE 2 – Caractérisation des modèles de langage utilisés dans nos expériences en termes de paramètres et de corpus d'entraînement. Les modèles marqués avec [en] (respectivement [fr], [es]) sont fortement entraînés en anglais (respectivement en français, en espagnol). Les CLMs marqués avec * sont des versions affinées d'autres CLMs.

5.2 Ablation

Pour mieux comprendre la contribution de chaque étape de notre approche, nous avons mené une série d'expériences complémentaires.

***Prompts* de listage** Dans cette section, nous comparons les *prompts* de balisage adoptés aux *prompts* de listage. Dans les *prompts* de listage, les démonstrations listent simplement les mentions étiquetées. Le séparateur de liste est optimisé (de la même manière que les baliseurs) entre une virgule et un retour à la ligne. Le cas échéant, les phrases introductives demandent de lister les entités (et non de les baliser). Les résultats présentés dans le tableau 3 corroborent davantage notre choix de nous concentrer uniquement sur les *prompts* de balisage.

	Anglais					Français					Espagnol			
Modèle	WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Prompts de listage														
Mistral-7B	0.659	0.533	0.417	0.281	0.340	0.676	0.083	0.451	0.169	0.403	0.697	0.620	0.211	0.273
Prompts de balisage														
Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374

TABLE 3 – F1-mesures obtenues avec les *prompts* de listage vs. balisage

Échantillon et taille de l'échantillon Nous avons testé notre approche avec différents échantillons et différentes tailles d'échantillon pour un MLM : XLM-RoBERTa-large, et un CLM : Mistral-7B. Les résultats sont présentés dans le tableau 4. On peut noter que, bien que l'écart-type par rapport à p soit assez élevé, une différence significative est systématiquement observée entre les deux modèles sur des échantillons de même taille. Nous observons également que, lorsque le nombre d'instances annotées diminue, les performances du MLM chutent plus rapidement que celles du CLM.

	CoNLL2003			n2c2		
100 exemples annotés						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.646	0.626	0.714	0.291	0.178	0.215
XLM-R-large	0.826	0.814	0.786	0.462	0.478	0.526
50 exemples annotés						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.615	0.648	0.637	0.278	0.176	0.106
XLM-R-large	0.697	0.77	0.714	0.431	0.476	0.35
25 exemples annotés						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.509	0.599	0.52	0.152	0.252	0.116
XLM-R-large	0.487	0.588	0.637	0.393	0.361	0.283

TABLE 4 – 1-mesures obtenues avec différents échantillons et différentes tailles d'échantillons.

***Grid search* des caractéristiques** Afin d'évaluer la qualité du *greedy search* adopté pour trouver la meilleure combinaison de caractéristiques à incorporer dans le *prompt*, nous comparons cette méthode à un *grid search* naïf sur ces caractéristiques. Nous testons les 512 combinaisons des 9 caractéristiques identifiées, pour Mistral-7B sur CoNLL2003. Les scores trouvés par LOOCV varient entre 0,0 et 0,656 avec une valeur moyenne de 0,387 et une médiane de 0,46. La combinaison la plus

performante est : *phrases supplémentaires*, *auto-vérification*, *phrase d'introduction pour l'instance de test* et *demander une réponse longue pour l'auto-vérification*, qui est précisément la combinaison que nous avons trouvée initialement par le biais d'un *greedy search*, qui est environ 20 fois plus rapide et moins consommateur.

Nombre de démonstrations Le choix de limiter le nombre d'exemples annotés présentés aux modèles causaux, dans le *prompt*, à 10 maximum est dicté par deux contraintes.

Tout d'abord, les modèles imposent une limite sur le nombre de *tokens* passés en entrée, sur lesquels l'attention est calculée. La plupart des modèles utilisés ont 2048 *tokens* comme limite, mais Mistral-7B autorise jusqu'à 8096 *tokens*. Cette limite se traduit par une limite du nombre de phrases présentables dans le *prompt*, entre 40 et 50 pour Mistral-7B et entre 10 et 15 pour les modèles moins permissifs, selon les corpus et les *tokenizers*. Par exemple, considérons la tâche de détection de parties du corps dans la partie française d'E3C. Si l'on utilise Mistral-7B (et son *tokenizer*), la limite pratique est autour de 40 exemples, ce qui fait un *prompt* de 7779.5 *tokens* en moyenne. Si l'on utilise Bloom-7b1 (et son *tokenizer*), elle se situe autour de 11 exemples, ce qui fait 1851.5 *tokens* en moyenne.

En outre, l'amélioration apportée par l'ajout d'exemples ne semble pas conséquente, comme on peut l'observer dans le tableau 5, qui montre les résultats obtenus avec Mistral-7B en triplant le nombre d'exemples annotés. Notons que les améliorations marginales obtenues en triplant le nombre d'exemples, viennent à un coût considérable, surtout dans le contexte d'une complexité quadratique en fonction de la longueur du *prompt*.

Nous choisissons donc de limiter le prompt à 5/10 exemples, choisis selon le critère choisi (proximité TF-IDF à la phrase de référence ou nombre d'entité présentes). Au lieu de sélectionner les exemples de façon indépendante, Gupta *et al.* (2023) proposent de les sélectionner de façon interdépendante, afin d'améliorer la représentativité du *prompt*. Cette piste serait intéressante à implémenter dans notre système dans de futurs travaux.

	Anglais					Français					Espagnol			
Modèle	WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
5/10 exemples														
Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374
15/30 exemples														
Mistral-7B	0.763	0.692	0.453	0.263	0.377	0.782	0.355	0.587	0.237	0.396	0.785	0.751	0.163	0.413

TABLE 5 – F1-mesures obtenues avec 5/10 exemples vs. avec 15/30 exemples

5.3 Description des types d'entités nommées

Etiquette	Nom du type (en singulier)	Description
PER	person names (a person's name)	These are names of persons such as real people or fictional characters.
FAC	facilities (a facility)	These are names of man-made structures such as infrastructure, buildings and monuments.
LOC	locations (a location)	These are names of geographical locations such as landmarks, cities, countries and regions.
ORG	organizations (an organization)	These are names of organizations such as companies, agencies and political parties.
FUNC	functions and jobs (a function or a job)	These are words that refer to a profession or a job.
ACTI	activities and behaviors (an activity or behavior)	These are words that refer to human activities, behaviors or events as well as governmental or regulatory activities.
ANAT	anatomy (an anatomy)	These are words that refer to the structure of the human body, its organs and their position, such as body parts or organs, systems, tissues, cells, body substances and embryonic structures.
CHEM	chemicals and drugs (a chemical or a drug)	These are words that refer to a substance or composition that has a chemical characteristic, especially a curative or preventive property with regard to human or animal diseases, such as drugs, antibiotics, proteins, hormones, enzymes and hazardous or poisonous substances.
CONC	concepts and ideas (a concept or an idea)	These are words that refer to a concept or an idea, such as a classification, an intellectual product, a language, a law or a regulation.
DEVI	medical devices (a device)	These are words that refer to a medical device used to administer care or perform medical research.
DISO	disorders (a disorder)	These are words that refer to an alteration of morphology, function or health of a living organism, animal or plant, such as congenital abnormalities, dysfunction, injuries, signs or symptoms or observations.
GENE	genes and molecular sequences (a gene or a molecular sequence)	These are words that refer to a gene, a genome or a molecular sequence.
GEOG	geographical areas (a geographical area)	These are words that refer to a country, a region or a city.
LIVB	living beings (a living being)	These are words that refer to a living being or a group of living beings, such as a person or a group of persons, a plant or a category of plants, an animal or a category of animals.
OBJC	objects (an object)	These are words that refer to anything animate or inanimate that affects the senses, such as physical manufactured objects.
OCCU	occupations (an occupation)	These are words that refer to a professional occupation or discipline.
ORGA	organizations (an organization)	These are words that refer to an organization such as healthcare related organizations.

TABLE 6 – Description des types d'entités nommées utilisées dans nos expériences sur l'anglais.

Etiquette	Nom du type (en singulier)	Description
PHEN	phenomema (a phenomemon)	These are words that refer to a phenomenon that occurs naturally or as a result of an activity, such as a biologic function.
PHYS	physiology (a physiology)	These are words that refer to any element that contributes to the mechanical, physical and biochemical functioning or organization of living organisms and their components.
PROC	procedures (a procedure)	These are words that refer to an activity or a procedure that contributes to the diagnosis or treatment of patients, the information of patients, the training of medical personnel or biomedical research.
EVENT	events (an event)	These are words that refer to actions, states, and circumstances that are relevant to the clinical history of a patient such as pathologies and symptoms, or more generally words like "enters", "reports" or "continue".
TIMEX3	time expressions (a time expression)	These are time expressions such as dates, times, durations, frequencies, or intervals.
RML	results and measurements (a result or a measurement)	These are test results, results of laboratory analyses, formulaic measurements, and measure values.
ACTOR	actors (an actor)	These are words that refer patients, healthcare professionals, or other actors relevant to the clinical history of a patient.
Abbreviation	abbreviations (an abbreviation)	These are words that refer to abbreviations.
Body_Part	body parts (a body part)	These are words that refer to organs and anatomical parts of persons.
Clinical_Finding	clinical findings (a clinical finding)	These are words that refer to observations, judgments or evaluations made about patients.
Diagnostic_Procedure	diagnostic procedures (a diagnostic procedure)	These are words that refer to tests that allow determining the condition of the individual.
Disease	diseases (a disease)	These are words that describe an alteration of the physiological state in one or several parts of the body, due to generally known causes, manifested by characteristic symptoms and signs, and whose evolution is more or less predictable.
Family_Member	family members (a family member)	These are words that refer to family members.
Laboratory_or_Test_Result	laboratory or test results (a laboratory or test result)	These are words that refer to any measurement or evaluation obtained from a diagnostic support examination.
Laboratory_Procedure	laboratory procedures (a laboratory procedure)	These are words that refer to tests that are performed on various patient samples that allow diagnosing diseases by detecting biomarkers and other parameters. Blood, urine, and other fluids and tissues that use biochemical, microbiological and/or cytological methods are considered.
Medication	medications (a medication)	These are words that refer to medications or drugs used in the treatment and/or prevention of diseases, including brand names and generics, as well as names for groups of medications.
Procedure	procedures (a procedure)	These are words that refer to activities derived from the care and care of patients.
Sign_or_Symptom	signs or symptoms (a sign or symptom)	These are words that refer to manifestations of a disease, determined by medical examination or perceived and expressed by the patient.
Therapeutic_Procedure	therapeutic procedures (a therapeutic procedure)	These are words that refer to activities or treatments that are used to prevent, repair, eliminate or cure the individual's disease.
CompositeMention	composite mentions of diseases (a composite mention of diseases)	These are words that refer to mentions of multiple diseases, such as "colorectal, endometrial, and ovarian cancers".
DiseaseClass	disease classes (a disease class)	These are words that refer to classes of diseases, such as "an autosomal recessive disease".
Modifier	modifiers (a modifier of diseases)	These are words that refer to modifiers of diseases, such as "primary" or "C7-deficient".
SpecificDisease	diseases (a disease)	These are words that refer to specific diseases, such as "diastrophic dysplasia".

TABLE 7 – Description des types d'entités nommées utilisées dans nos expériences sur l'anglais, suite.

Etiquette	Nom du type (en singulier)	Description
PER	de noms de personnes (un nom de personne)	Il s'agit des noms de personnes, qu'elles soient réelles ou fictives.
FAC	de productions humaines (une production humaine)	Il s'agit des noms de structures faites par les humains comme des infrastructures, des bâtiments ou des monuments.
LOC	de lieux (un lieu)	Il s'agit des noms de lieux comme des endroits, villes, pays ou régions.
ORG	d'organisations (une organisation)	Il s'agit des noms d'organisations comme des entreprises, des agences ou des partis politiques.
FUNC	de fonctions et métiers (une fonction ou un métier)	Il s'agit de mots qui se rapportent à une activité professionnelle.
ANAT	d'anatomie (une partie du corps)	Il s'agit d'une entité se rapportant à la structure du corps humain, ses organes et leur position. Il s'agit principalement des parties du corpus ou organes, des appareils, des tissus, des cellules, des substances corporelles et des organismes embryonaires.
CHEM	de médicaments et substances chimiques (un médicament ou une substance chimique)	Il s'agit d'une substance ou composition présentant des propriétés chimiques caractéristiques, en particulier des propriétés curatives ou préventives à l'égard des maladies humaines ou animales. Il s'agit principalement des médicaments disponibles en pharmacie, des antibiotiques, des proteines, des hormones, des substances dangereuses, des enzymes.
DEVI	de matériel (un matériel)	Il s'agit d'un matériel utilisé pour administrer des soins ou effectuer des recherches médicales.
DISO	de problèmes médicaux (un problème médical)	Il s'agit d'une altération de la morphologie, des fonctions, ou de la santé d'un organisme vivant, animal ou végétal. Il peut s'agir de malformations, de maladies, de blessure, de signe ou symptome ou d'une observation.
GEOG	de zones géographiques (une zone géographique)	Il s'agit d'un pays, une région, ou une ville.
LIVB	d'êtres vivants (un être vivant)	Il s'agit d'un être vivant ou groupe d'êtres vivants. Il peut s'agir d'une personne ou d'un groupe de personnes, d'une plante ou d'une catégorie de végétaux, d'un animal ou d'une catégorie d'animaux.
OBJC	d'objets (un objet)	Il s'agit de tout ce qui, animé ou inanimé, affecte les sens. Ici, il s'agit principalement d'objets physiques manufacturés.
PHEN	de phénomènes (un phénomène)	Il s'agit d'un phénomène qui se produit naturellement ou à la suite d'une activité. Il s'agit principalement de fonctions biologiques.
PHYS	de physiologie (une physiologie)	Il s'agit de tout élément contribuant au fonctionnement ou à l'organisation mécanique, physique et biochimique des organismes vivants et de leurs composants.
PROC	de procédures (une procédure)	Il s'agit d'une activité ou procédure contribuant au diagnostic ou au traitement des patients, à l'information des patients, la formation du personnel médical ou à la recherche biomédicale.
EVENT	d'événements (un événement)	Il s'agit d'une action, d'un état ou d'une circonstance qui est pertinent pour l'histoire clinique d'un patient. Il peut s'agir de pathologies et symptômes, ou plus généralement de mots comme "entre", "rapporte" ou "continue".
TIMEX3	d'expressions temporelles (une expression temporelle)	Il s'agit d'expressions temporelles comme des dates, heures, durées, fréquences, ou intervalles.
RML	de résultats et mesures (un résultat ou une mesure)	Il s'agit de résultats d'analyses de laboratoire, de mesures formelles, et de valeurs de mesure.
ACTOR	d'acteurs (un acteur)	Il s'agit de patients, de professionnels de santé, ou d'autres acteurs pertinents pour l'histoire clinique d'un patient.

TABLE 8 – Description des types d'entités nommées utilisées dans nos expériences sur le français.

Etiquette	Nom du type (en singulier)	Description
PER	nombres de personas (un nombre de persona)	Estos son nombres de personas, ya sean reales o personajes ficticios.
FAC	instalaciones (una instalación)	Estos son nombres de estructuras hechas por el hombre como infraestructura, edificios y monumentos.
LOC	lugares (un lugar)	Estos son nombres de ubicaciones geográficas como hitos, ciudades, países y regiones.
ORG	organizaciones (una organización)	Estos son nombres de organizaciones como empresas, agencias y partidos políticos.
ACTI	actividades y comportamientos (una actividad o comportamiento)	Estas son palabras que se refieren a actividades humanas, comportamientos o eventos, así como actividades gubernamentales o regulatorias.
ANAT	anatomía (una anatomía)	Estas son palabras que se refieren a la estructura del cuerpo humano, sus órganos y su posición, como partes del cuerpo u órganos, sistemas, tejidos, células, sustancias corporales y estructuras embrionarias.
CHEM	productos químicos y medicamentos (un producto químico o un medicamento)	Estas son palabras que se refieren a una sustancia o composición que tiene una característica química, especialmente una propiedad curativa o preventiva con respecto a las enfermedades humanas o animales, como medicamentos, antibióticos, proteínas, hormonas, enzimas y sustancias peligrosas o venenosas.
CONC	conceptos e ideas (un concepto o una idea)	Estas son palabras que se refieren a un concepto o una idea, como una clasificación, un producto intelectual, un idioma, una ley o un reglamento.
DEVI	dispositivos médicos (un dispositivo)	Estas son palabras que se refieren a un dispositivo médico utilizado para administrar atención o realizar investigaciones médicas.
DISO	trastornos (un trastorno)	Estas son palabras que se refieren a una alteración de la morfología, la función o la salud de un organismo vivo, animal o vegetal, como anomalías congénitas, disfunción, lesiones, signos o síntomas u observaciones.
GENE	genes y secuencias moleculares (un gen o una secuencia molecular)	Estas son palabras que se refieren a un gen, un genoma o una secuencia molecular.
GEOG	áreas geográficas (un área geográfica)	Estas son palabras que se refieren a un país, una región o una ciudad.
LIVB	seres vivos (un ser vivo)	Estas son palabras que se refieren a un ser vivo o un grupo de seres vivos, como una persona o un grupo de personas, una planta o una categoría de plantas, un animal o una categoría de animales.
OBJC	objetos (un objeto)	Estas son palabras que se refieren a cualquier cosa animada o inanimada que afecte los sentidos, como objetos físicos fabricados.
OCCU	ocupaciones (una ocupación)	Estas son palabras que se refieren a una ocupación o disciplina profesional.
ORGA	organizaciones (una organización)	Estas son palabras que se refieren a una organización, por ejemplo organizaciones relacionadas con la salud.
PHEN	fenómenos (un fenómeno)	Estas son palabras que se refieren a un fenómeno que ocurre naturalmente o como resultado de una actividad, por ejemplo una función biológica.

TABLE 9 – Description des types d'entités nommées utilisées dans nos expériences sur l'espagnol.

Etiquette	Nom du type (en singulier)	Description
PHYS	fisiología (una fisiología)	Estas son palabras que se refieren a cualquier elemento que contribuya al funcionamiento mecánico, físico y bioquímico o la organización de los organismos vivos y sus componentes.
PROC	procedimientos (un procedimiento)	Estas son palabras que se refieren a una actividad o un procedimiento que contribuye al diagnóstico o tratamiento de pacientes, la información de pacientes, la capacitación del personal médico o la investigación biomédica.
EVENT	eventos (un evento)	Estas son palabras que se refieren a acciones, estados y circunstancias que son relevantes para la historia clínica de un paciente, como patologías y síntomas, o más generalmente palabras como "entra", "reporta" o "continúa".
TIMEX3	expresiones de tiempo (una expresión de tiempo)	Estas son expresiones de tiempo como fechas, horas, duraciones, frecuencias o intervalos.
RML	resultados y mediciones (un resultado o una medida)	Estos son resultados de análisis de laboratorio, mediciones formales y valores de medición.
ACTOR	actores (un actor)	Estas son palabras que se refieren a pacientes, profesionales de la salud u otros actores relevantes para la historia clínica de un paciente.
Abbreviation	abreviaciones (una abreviación)	Estas son los casos de siglas y acrónimos.
Body_Part	partes del cuerpo (una parte del cuerpo)	Estas son palabras que se refieren a òrganos y partes anatómicas de personas.
Clinical_Finding	hallazgos clínicos (un hallazgo clínico)	Estas son palabras que se refieren a observaciones, juicios o evaluaciones que se hacen sobre los pacientes.
Diagnostic_Procedure	procedimientos diagnósticos (un procedimiento diagnóstico)	Estas son palabras que se refieren a exámenes que permiten determinar la condición del individuo.
Disease	enfermedades (una enfermedad)	Estas son palabras que describen una alteración del estado fisiológico en una o varias partes del cuerpo, por causas en general conocidas, manifestada por síntomas y signos característicos, y cuya evolución es más o menos previsible.
Family_Member	miembros de la familia (un miembro de la familia)	Estas son palabras que se refieren a miembros de la familia.
Laboratory_or_Test_Result	resultados de exámenes de laboratorio u otras pruebas (un resultado de un examen de laboratorio u otra prueba)	Estas son palabras que se refieren a cualquier medición o evaluación obtenida a partir de un exámen de apoyo diagnóstico.
Laboratory_Procedure	procedimientos de laboratorio (un procedimiento de laboratorio)	Estas son palabras que se refieren a exámenes que se realizan en diversas muestras de pacientes que permiten diagnosticar enfermedades mediante la detección de biomarcadores y otros parámetros. Se consideran los análisis de sangre, orina, y otros fluidos y tejidos que emplean métodos bioquímicos, microbiológicos y/o citológicos.
Medication	medicamentos o drogas (un medicamento o una droga)	Estas son palabras que se refieren a medicamentos o drogas empleados en el tratamiento y/o prevención de enfermedades, incluyendo marcas comerciales y genéricos, así como también nombres para grupos de medicamentos.
Procedure	procedimientos (un procedimiento)	Estas son palabras que se refieren a actividades derivadas de la atención y el cuidado de los pacientes.
Sign_or_Symptom	signos o síntomas (un signo o un síntoma)	Estas son palabras que se refieren a manifestaciones de una enfermedad, determinadas mediante la exploración médica o percibidas y expresadas por el paciente.
Therapeutic_Procedure	procedimientos terapéuticos (un procedimiento terapéutico)	Estas son palabras que se refieren a actividades o tratamientos que es empleado para prevenir, reparar, eliminar o curar la enfrmedad del individuo.

TABLE 10 – Description des types d'entités nommées utilisées dans nos expériences sur l'espagnol, suite.

5.4 Empreinte carbone

	#	Modèle	Anglais					Français					Espagnol			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Approches few-shot																
Causal	1	LLAMA-2-70B	46	44	126	233	54	85	131	129	273	284	41	76	114	344
	2	Mistral-7B	4	6	12	24	8	5	8	14	13	25	7	5	11	27
	3	BLOOM-7B1	4	6	10	26	9	8	13	9	26	20	4	8	8	18
	4	Falcon-40B	49	45	56	176	45	31	58	75	162	129	33	25	82	99
	5	GPT-J-6B	7	6	8	23	7	5	8	13	21	17	6	6	13	28
	6	OPT-66B	73	50	120	253	96	38	64	138	273	240	57	52	106	247
	7	Vicuna-13B	10	11	20	52	11	11	12	18	33	40	10	11	22	51
	8	Vicuna-7B	6	8	14	17	6	5	10	10	24	14	8	6	13	27
	9	Medalpaca-7B	8	4	17	24	10	7	14	11	19	21	5	8	15	26
	10	Vigogne-13B	14	14	29	37	11	13	20	26	36	39	11	14	32	44
Masked	11	mBERT	2	1	2	2	2	2	2	2	1	1	1	2	1	2
	12	XLM-R-large	2	2	2	1	2	2	2	2	2	2	1	1	1	2
	13	BERT-large	2	1	2	2	2	-	-	-	-	-	-	-	-	-
	14	RoBERTa-large	1	2	2	2	2	-	-	-	-	-	-	-	-	-
	15	Bio_ClinicalBERT	2	2	1	2	1	-	-	-	-	-	-	-	-	-
	16	ClinicalBERT	1	1	2	2	1	-	-	-	-	-	-	-	-	-
	17	MedBERT	2	2	1	1	1	-	-	-	-	-	-	-	-	-
	18	CamemBERT-large	-	-	-	-	-	1	1	1	2	2	-	-	-	-
	19	FlauBERT-large	-	-	-	-	-	2	2	2	2	2	-	-	-	-
	20	DrBERT-4GB	-	-	-	-	-	2	2	2	2	2	-	-	-	-
	21	CamemBERT-bio	-	-	-	-	-	1	2	2	2	2	-	-	-	-
	23	BETO	-	-	-	-	-	-	-	-	-	-	2	1	1	1
	23	PatanaBERT	-	-	-	-	-	-	-	-	-	-	2	2	2	2
	24	TulioBERT	-	-	-	-	-	-	-	-	-	-	1	2	2	1
	25	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	2	2	2	2
	26	BSC-Bio	-	-	-	-	-	-	-	-	-	-	2	2	2	2
Skyline en utilisant toutes les données à disposition																
		RoBERTa-large	647	68	5	12	24	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	595	15	4	5	8	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	579	41	3	21

TABLE 11 – Ce tableau présente les émissions carbone (en g) de l'optimisation de chaque modèle sur sur le jeu de validation de chaque corpus. Pour les CLMs , il s'agit du *greedy search* sur les potentiels caractéristiques du *prompt* via la validation croisée. Pour les MLM, cela correspond au *fine-tuning* des paramètres du modèle (et à l'apprentissage des couches de classification introduites).

	#	Modèle	Anglais					Français					Espagnol			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Approches few-shot																
Causal	1	LLAMA-2-70B	812	147	36	196	33	508	11	13	92	47	514	201	11	198
	2	Mistral-7B	234	35	8	59	21	148	3	4	27	20	261	50	2	32
	3	BLOOM-7B1	220	33	8	44	16	255	3	5	38	29	261	47	2	46
	4	Falcon-40B	600	109	26	144	46	722	9	19	155	70	752	154	9	157
	5	GPT-J-6B	146	17	4	53	20	245	2	6	14	26	154	40	3	53
	6	OPT-66B	765	139	33	185	63	971	12	27	179	93	993	196	12	217
	7	Vicuna-13B	314	47	11	63	24	363	5	8	61	46	502	67	4	74
	8	Vicuna-7B	146	17	4	53	20	246	2	6	14	26	155	65	3	53
	9	Medalpaca-7B	192	24	5	39	14	98	2	2	17	13	172	53	1	21
	10	Vigogne-13B	322	49	11	65	24	245	5	6	44	33	361	68	3	66
Masked	11	mBERT	14	4	<1	2	<1	15	1	<1	1	1	13	2	<1	2
	12	XLM-R-large	14	4	<1	2	<1	15	1	<1	1	1	13	2	<1	2
	13	BERT-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	14	RoBERTa-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	15	Bio_ClinicalBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	16	ClinicalBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	17	MedBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	18	CamemBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	19	FlauBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	20	DrBERT-4GB	-	-	-	-	-	17	1	<1	1	1	-	-	-	-
	21	CamemBERT-bio	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	22	BETO	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	23	PatanaBERT	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	24	TulioBERT	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	25	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	26	BSC-Bio	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
Skyline en utilisant toutes les données à disposition																
		RoBERTa-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	13	2	<1	2

TABLE 12 – Ce tableau présente les émissions carbone (en g) de l'inférence, en utilisant chaque modèle, sur sur le jeu de test de chaque corpus.