Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Annotation d'entités cliniques en utilisant les Larges Modèles de Langue

Simon Meoni[1, 2] Théo Ryffel[2] Éric de la Clergerie[1]
(1) Inria, Paris, France
(2) Arkhn, Paris, France
simon.meoni@arkhn.com, eric.de_la_clergerie@inria.fr

RÉSUMÉ

Dans le domaine clinique et dans d'autres domaines spécialisés, les données sont rares du fait de leur caractère confidentiel. Ce manque de données est un problème majeur lors du fine-tuning de modèles de langue. Par ailleurs, les modèles de langue de très grande taille (LLM) ont des performances prometteuses dans le domaine médical. Néanmoins, ils ne peuvent pas être utilisés directement dans les infrastructures des établissements de santé pour des raisons de confidentialité des données. Nous explorons une approche d'annotation des données d'entraînement avec des LLMs pour entraîner des modèles de moins grandes tailles mieux adaptés à notre problématique. Cette méthode donne des résultats prometteurs pour des tâches d'extraction d'information.

ABSTRACT

Annotate Clinical Data Using Large Language Model Predictions

In clinical and other specialized domains, data are scarce due to their confidential nature. This lack of data is a major problem when fine-tuning language models. Nevertheless, very large language models (LLMs) are promising for the medical domain but cannot be used directly in healthcare facilities due to data confidentiality issues. We explore an approach of annotating training data with LLMs to train smaller models more adapted to our problem. We show that this method yields promising results for information extraction tasks.

MOTS-CLÉS : Supervision Faible, Modèle de langue Large, Extraction d'information, TAL dans le domaine clinique.

KEYWORDS: Weak supervision, Large Language Model, Extraction Information Task, Clinical NLP.

1 Introduction

Les notes cliniques contiennent l'ensemble des interactions entre le patient et le personnel de santé. Les personnels de santé y indiquent notamment leurs observations et les différents actes médicaux réalisés. Malgré l'informatisation des documents cliniques, les notes doivent rester dans un format assez expressif et libre afin de faire gagner du temps au personnel soignant et de permettre la description de situations inhabituelles (Rosenbloom *et al.*, 2011). De plus, un grand nombre d'informations cruciales est contenu exclusivement dans ces notes. Selon une étude de Escudié *et al.* (2017), environ 80% des phénotypes (ensemble de caractères biologiques, physiques observables pouvant caractériser une

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 190–203.

maladie) de patients ne sont présents que dans du texte libre. Cela rend ces documents difficilement exploitables sans l'utilisation de méthodes avancées comme le deep-learning en NLP. L'utilisation de telles méthodes requiert de collecter et d'annoter un nombre important de données médicales. Cependant, Fries *et al.* (2022) met en avant le fait que les données d'apprentissage dans le domaine biomédical sont peu accessibles, peu documentées et rarement réutilisables dans un cadre commercial ou de recherche, une situation qu'il qualifie de *dataset debt*. Il montre par exemple que seulement 13% des 167 datasets analysés dans son étude sont accessibles et téléchargeables, que 22% utilisent un format standard structuré ou encore que 40% sont dans le domaine public. Ces dernières années, les larges modèles de langage (LLMs) comme GPT3 ont prouvé leur capacité à effectuer un large panel de tâches avec des performances prometteuses en contexte de zero-shot ou de few-shots. C'est une tendance très intéressante pour le développement du traitement automatique du langage naturel clinique, et ses résultats préliminaires sont prometteurs pour les tâches d'extraction d'informations (Agrawal *et al.*, 2022). Cependant, l'utilisation des LLMs soulèverait des préoccupations majeures en matière de confidentialité. En effet, il est important de garantir que le déploiement du modèle est maîtrisé et la prédiction issue du modèle doit évoluer pour s'adapter à des directives d'annotation spécifiques et changeantes. Par ailleurs, la plupart des LLMs ne sont pas disponibles librement (Scao *et al.*, 2022; Ouyang *et al.*, 2022; Thoppilan *et al.*, 2022) et à notre connaissance, seul BLOOM est open-source et déployable dans une infrastructure locale. Enfin, les ressources informatiques nécessaires pour utiliser ces modèles en inférence restent importantes, constituant de fait un obstacle de plus à leur adoption par les établissements de santé.

L'une des approches permettant de résoudre ces enjeux consiste à distiller les LLMs en modèles de taille moins conséquente via des méthodes de supervision faible. La supervision faible a récemment attiré l'attention de la communauté scientifique car elle soulage la tâche d'annotation (Lison *et al.*, 2020, 2021). En effet, cette technique consiste à annoter automatiquement des ensembles de données à l'aide de méthodes basées sur des règles, des dictionnaires ou des techniques plus avancées, puis à entraîner le modèle sur cet ensemble de données.

2 Contributions

Notre travail étudie l'utilisation de LLMs dans la technique de distillation de connaissances par le biais de techniques de supervision faible dans le domaine clinique français, en particulier dans l'extraction d'entités cliniques. Plus précisément :

— Nous montrons que la distillation est une technique compétitive par rapport aux techniques classiques de supervision faible dans le domaine clinique français ;
— Nous proposons une approche de supervision faible qui associe les annotations provenant d'extraction de dictionnaire et les prédictions de LLM pour créer un ensemble de données d'entraînement, surpassant les annotations provenant des seules prédictions d'InstructGPT-3.

3 Travaux Connexes

Supervision Faible Le *deep learning* a connu un succès remarquable dans plusieurs domaines au-delà du NLP (Zhang *et al.*, 2022). Cependant, le principal obstacle est la collecte massive de données annotées. Pour remédier à ce problème, la supervision faible remplace l'annotation de vérité

terrain par une annotation automatique basée sur des règles heuristiques ou des règles linguistiques de contrainte. Certaines techniques appelées *supervision distante* exploitent des liens sémantiques à partir de bases de connaissances ou d'ontologies (Lison *et al.*, 2021). Karamanolakis *et al.* (2021) propose une méthode d'autoapprentissage itérative pour combiner la supervision faible classique et l'inférence du modèle d'apprentissage afin d'extraire des entités non couvertes par les règles heuristiques initiales. Dans le domaine clinique, la supervision faible a déjà été utilisée pour certains cas d'usages cliniques spécifiques (Cusick *et al.*, 2021; Fries *et al.*, 2021; Wang *et al.*, 2019).

Modèles de Langage Clinique Dans un contexte spécifique au domaine clinique, certains termes spécifiques sont sous-représentés ou absents dans le domaine général. En conséquence, la communauté en NLP clinique a entraîné des modèles de langage préentraînés sur des corpus spécifiques au domaine (Alsentzer *et al.*, 2019; Lee *et al.*, 2020; Alsentzer *et al.*, 2019), tels que MIMIC-III (Johnson *et al.*, 2016) ou des résumés Pubmed. Ces modèles peuvent être entraînés à partir de poids initialisés ou à partir d'un modèle agnostique pour le spécialiser dans le domaine clinique (Gururangan *et al.*, 2020). Cependant, les gains de performance sont marginaux par rapport au modèle de langage général. La structure du texte et les abréviations présentes dans les notes cliniques ont un impact négatif sur les performances des modèles. Au lieu de préentraîner un modèle clinique spécialisé, certains travaux ont cherché à affiner des modèles LLMs génératifs de langues générales tels que la famille de modèles GPT ou T5 sur une tâche clinique. Ces modèles généraux affinés ont prouvé leur efficacité dans la réponse aux questions cliniques, la dé-identification de documents médicaux ou encore l'extraction de relations (Lehman *et al.*, 2023). Cette approche nécessite toutefois une infrastructure importante et un réaffinage régulier si la distribution des données des rapports cliniques au sein de l'établissement de santé change. Néanmoins, certains LLMs ont été préentraînés sur des notes spécifiques au domaine clinique tel que GatorTron (Yang *et al.*, 2022), BioGPT (Luo *et al.*, 2022) ou ClinicalT5 (Lu *et al.*, 2022) et ont obtenu des performances prometteuses sur plusieurs tâches.

De plus, l'apprentissage en contexte avec des LLMs de langue générale telles que InstructGPT-3 (Ouyang *et al.*, 2022), où aucun entraînement n'a été effectué, donne de bons résultats (Agrawal *et al.*, 2022; Brown *et al.*, 2020) et surpasse les modèles spécialisés de plus petites tailles sur plusieurs tâches cliniques.

Méthodes basées sur les instructions L'apprentissage basé sur les instructions pour les modèles de langage génératif traite une tâche comme un problème de modélisation de langage où un modèle de langage prédit les tokens suivant une instruction textuelle (ou *prompt*) donnée en entrée (Sainz *et al.*, 2021). Dans ce paradigme, au lieu d'affiner un modèle pour une tâche donnée (*"pré-entraînement, affinage et prédiction"*), nous voulons manipuler le comportement d'un modèle de langage pré-entraîné en utilisant un prompt approprié pour obtenir la sortie souhaitée (*"pré-entraînement, instructions et prédiction"*). L'avantage majeur de cette méthode est qu'un modèle de langage pré-entraîné de manière non supervisée peut être utilisé pour de nombreuses tâches (Liu *et al.*, 2023). De ce fait, le *prompt engineering* a été développé pour explorer la méthode d'instruction la plus adaptée appliquée à un modèle de langue pour résoudre une tâche donnée. Parmi ces méthodes, l'apprentissage en contexte (*in-context learning*) est l'une des méthodes les plus populaires pour l'extraction d'informations, la réponse aux questions ou l'analyse des sentiments. Dans le domaine clinique, certains travaux existent sur la récupération d'informations et la réponse aux questions. Le prompt contient trois composants : le modèle ou le format des exemples, l'ensemble des exemples et l'ordre des prompts tel que présenté dans la Figure 4. Le but est de fournir dans le prompt quelques exemples d'entraînement concaténés

avec l'exemple de test. Cependant, les exemples choisis, leur agencement ainsi que le format ont un impact en termes de performance (Zhao *et al.*, 2021), si bien que ces trois composants doivent être calibrés pour optimiser les performances.

4 Méthode

4.1 Annoter et distiller des connaissances via une supervision faible

Extraire des annotations à partir de la sortie LLM Notre étude s'inspire largement de la méthode développée par Agrawal *et al.* (2022). Dans ces travaux, ils évaluent la performance d'InstructGPT-3 (Ouyang *et al.*, 2022) pour des tâches de NLP clinique en anglais. InstructGPT-3 obtient des résultats prometteurs, dépassant de loin certains modèles de langues spécialisés pour plusieurs tâches d'extraction d'information. De plus, les auteurs introduisent trois nouveaux ensembles de données pour évaluer les informations cliniques dans un contexte few-shot. Enfin, un nouveau concept, nommé *requête guidée* (Figure 4) permet de pré-structurer la sortie et faciliter son exploitation via des résolveurs (ou des fonctions de *string matching*).

Notre travail se distingue par les contributions suivantes :

1. notre domaine d'étude principal est la distillation de connaissances via une supervision faible et l'amélioration de cette technique en combinant des annotations de LLMs et des méthodes basées sur de l'extraction par dictionnaire ;
2. nous appliquons nos méthodes au domaine clinique français sur lequel les données annotées sont plus rares, alors que le travail initial a été réalisé en anglais ;
3. notre travail se base sur les directives de l'ensemble de données E3C (Magnini *et al.*, 2020) pour réaliser l'extraction des entités cliniques.

Dans ce travail, InstructGPT-3 n'est pas entraîné et est utilisé tel quel ; nous nous contentons d'interroger le modèle, aucune étape d'affinage supplémentaire n'est réalisée et nous n'avons accès qu'aux paramètres d'inférence tels que la température, le top p, la pénalité de fréquence ou de présence. Nous réglons la *température* et le *top p* à 0 pour contrôler le hasard et avoir un comportement déterministe. Pour ne pas pénaliser les répétitions, nous réglons la *pénalité de présence* et la *pénalité de fréquence* à 0. Nous utilisons un modèle InstructGPT-3 (text-davinci-003) (Ouyang *et al.*, 2022) pour inférer toutes les annotations pour toutes nos expériences. Nous donnons en entrée du modèle une instruction que nous complétons par l'exemple à prédire (Figure 4). La sortie d'InstructGPT-3 est une chaîne de caractères que nous structurons afin d'aligner les entités prédites avec le texte initial (Figure 1).

La tâche consiste à annoter les mots (ou tokens) d'une phrase x avec un ensemble d'étiquettes $L := \{O, B_{\text{clin}}, I_{\text{clin}}\}$ où O dénote un mot du texte sans étiquette, B_{clin} le premier mot d'une entité clinique et I_{clin} les mots suivants selon le format *IOB* (Ramshaw & Marcus, 1995). Le but est d'analyser une phrase $x = [x_1, \ldots, x_n] \in \Sigma^*$ et d'identifier pour chaque token x_i la bonne étiquette O, B_{clin}, I_{clin} ainsi que l'offset de caractères. La prédiction du modèle est donc $\hat{y} = [y_1, y_2, \ldots, y_n] \in Y$ où une annotation y_i est définie comme $y_i := \langle x_i, s, e, l \rangle$ avec s l'offset de début, e l'offset de fin et $l \in L$.

Comme mentionné ci-dessus, une méthode basée sur des instructions nécessite de concaténer un template t avec notre phrase x pour donner notre instruction telle que $p = concat(t, x)$. Nous produisons notre sortie $o \in \Sigma^*$ à partir de notre modèle LLM Φ tel que $o = \Phi(p, \theta_h)$, où θ_h

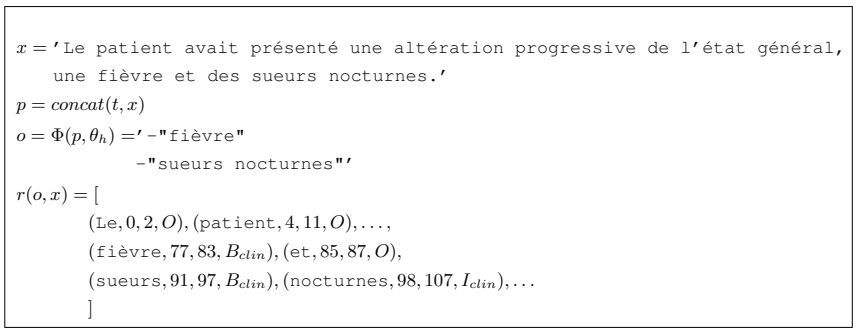


FIGURE 1 – Illustration des étapes de prédiction et de structuration de notre méthode sur un exemple. Le template t est illustré dans la Figure 4.

représente l'ensemble des hyperparamètres (*température*, *top p*, *pénalité de fréquence*, *pénalité de présence*) et o est une chaîne de caractères.

Nous structurons la sortie o en utilisant une fonction de *string matching* retournant un ensemble d'étiquettes $\hat{y} = r(o, x)$, où $r : \Sigma^* \times \Sigma^* \rightarrow Y$ est le résolveur appliquant la fonction de *string matching*.

Distillation de connaissances via la supervision faible Enfin, les annotations générées via la prédiction d'InstructGPT-3 sont utilisées comme ensemble d'entraînement pour affiner un modèle de langage plus petit pour la tâche cible.

4.2 Instructions

Nous fournissons des instructions de mise en contexte avec trois exemples de données annotées. Nous sélectionnons deux exemples avec plusieurs entités cliniques extraites et un autre sans entités. Ce paramétrage permet de fournir une diversité sémantique d'entités à InstructGPT-3 (Figure 4).

Après plusieurs essais empiriques et analyses qualitatives, nous ajoutons un exemple sans entités cliniques pour éviter trop de faux positifs extraits par le modèle. Nous insérons dans les prompts un nombre réduit de mots clés associés à la définition de l'E3C des entités cliniques. De plus, nous ajoutons les tokens de *requête guidée* pour expliciter la structure de la réponse afin de faciliter le *parsing* de la sortie du LLM (Agrawal *et al.*, 2022).

5 Expériences

5.1 Jeu de données

Nous utilisons le corpus multilingue (anglais, basque, espagnol, français, italien) E3C (Magnini *et al.*, 2020) pour nos expériences, qui se compose de deux types d'annotations : temporelles et entités cliniques. Comme mentionné précédemment, notre objectif est d'extraire des entités cliniques en français. Dans E3C, l'une des tâches supplémentaires est de lier les entités extraites avec des entrées du métathésaurus de l'UMLS. Toutefois, nous nous concentrons sur l'extraction des entités cliniques.

Le jeu de données E3C est organisé en trois couches, chacune correspondant à un ensemble de données annotées d'une certaine manière :

— La première couche (appelée **layer 1**) consiste en une annotation manuelle complète ;
— La deuxième couche (**layer 2**) consiste en une annotation semi-automatique réalisée via une extraction par dictionnaire, qui contient des termes de l'UMLS et des termes extraits de la **layer 1**. Un sous-ensemble de cette couche (environ 10%) a été corrigé manuellement [1] (**layer 2 validate**) et est également utilisé séparément dans notre étude .
— La troisième couche (**layer 3**) est une couche non annotée que non-utilisée dans nos travaux.

Nous avons sélectionné le sous-ensemble français d'E3C comprenant des données annotées manuellement et automatiquement pour nos expériences de supervision faible. Cependant, le jeu de données a une quantité limitée de données dans ses différentes couches. Pour pallier ce problème, nous avons divisé le **layer 2** en utilisant une validation croisée à 5 partitions (folds).

5.2 Protocole Expérimental

Nous menons des expériences sur des tâches d'extraction d'entités cliniques. Pour toutes nos expériences, nous utilisons *camembert-base* (Martin *et al.*, 2019) comme modèle étudiant pour l'étape de distillation des connaissances via de la supervision faible. Nous testons nos modèles sur le **layer 1**. Les différents corpus d'entraînement utilisés dans les différentes configurations seront issus du **layer 2** et du **layer 2 validate**. Nous utilisons soit les annotations initiales, soit les annotations inférées par InstructGPT-3, soit encore un mélange des deux. Nous menons nos expériences en utilisant quatre configurations de jeux de données différentes pour l'affinage de *camembert-base* :

— **Configuration A** : nous utilisons le **layer 2** comme jeu d'entraînement et comparons deux modèles camemBERT entraînés respectivement sur des annotations d'extraction de dictionnaire et des annotations prédites par InstructGPT-3 ;
— **Configuration B** : nous utilisons le **layer 2 validate** comme jeu d'entraînement et comparons deux modèles camemBERT entraîné respectivement sur des annotations corrigées manuellement et des annotations prédites par InstructGPT-3 ;
— **Configuration C** : nous nous basons sur la **Configuration A** mais nous conservons les annotations de base du **layer 2 validate** pour les deux modèles. Ainsi, une petite partie (environ 10%) des annotations prédites par InstructGPT-3 et des annotations d'extraction de dictionnaire sont remplacées par des annotations manuelles ;
— **Configuration D** : nous utilisons le **layer 2** comme jeu d'entraînement mais mixons un ratio r d'annotations prédites par InstructGPT-3 avec $(1 - r)$ d'annotations prédites par InstructGPT-3. Dans ce contexte, nous testons et comparons des modèles entraînés pour divers valeurs de ratio r.

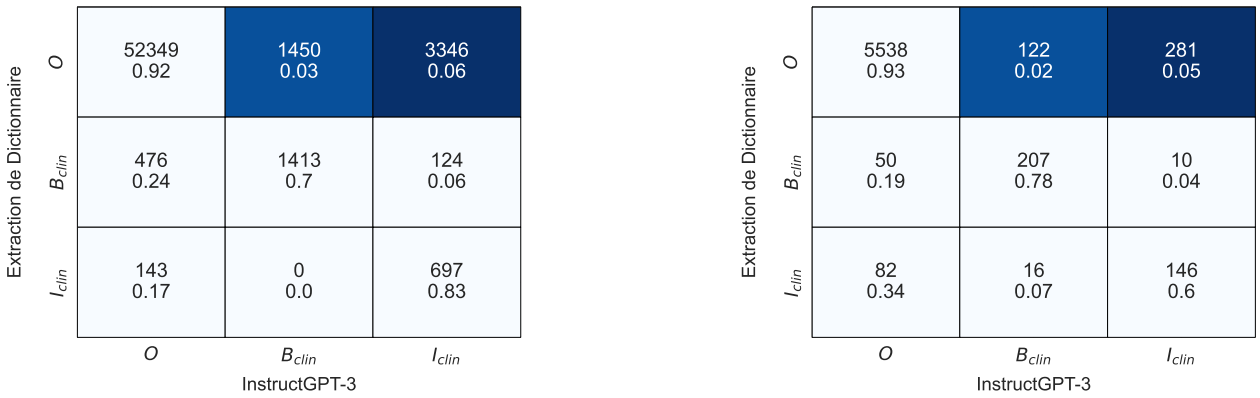
5.3 Résultats et discussion

Analyse des prédictions d'InstructGPT-3 Nous notons qu'InstructGPT-3 extrait presque deux fois plus d'entités (Table 1). Cette tendance est plus importante dans le **layer 2**, tandis que l'écart est réduit dans le **layer 2 validate** ; cela est certainement dû à la validation humaine sur cette couche. Cette différence pourrait s'expliquer par le fait qu'InstructGPT-3 n'a pas accès aux guidelines ; l'instruction reproduite en Figure 4 mentionne des "*disorders*", "*diseases*", ou "*symptoms*" à extraire :

1. ce sous-ensemble validée (**layer 2 validate**) n'a cependant pas été corrigé de la même manière que le **layer 1**

Layer	Méthodes	#Phrases	#Tokens	B_{clin}	I_{clin}
Layer 1	Dictionary Extraction	1109	28744	1258	1203
	InstructGPT-3			1398	2028
Layer 2	Dictionary Extraction	2389	59998	2013	840
	InstructGPT-3			2863	4167
Layer 2 Validate	Manual Extraction	293	6452	267	244
	InstructGPT-3			345	437

TABLE 1 – Nombre de tokens annotés par type d'annotation pour chaque couche



(a) **layer 2** (b) **layer 2 validate**

FIGURE 2 – Matrices de confusion pour le **layer 2** et le **layer 2 validate** montrant les relations entre les annotations par dictionnaire et via InstructGPT-3. Par exemple dans le **layer 2 validate**, l'intersection B_{clin} (en abscisse) à O (en ordonnée) indique qu'InstructGPT a identifié 122 tokens O comme B_{clin}en se basant sur l'extraction par dictionnaire.

ceci est moins restrictif que l'annotation de la guideline d'E3C. Les matrices de confusion (Figure 2) soulignent cette tendance. Néanmoins, les deux méthodes d'extraction ont presque étiqueté le même nombre de tokens en tant que O (0.93), ce qui confirme que l'annotation d'InstructGPT-3 est cohérente avec la tâche d'extraction. Nous notons que seulement 0.04 de B_{clin} sont annotés en tant que token I_{clin}, ce qui signifie qu'InstructGPT-3 discrimine bien avec la signification des tokens de début (B_{clin}) et des tokens internes (I_{clin}). Nous soulignons le fait est qu'InstructGPT-3 considère certains tokens O comme des tokens internes : les entités extraites sont plus longues que celles extraites par dictionnaire. D'où le plus grand nombre de tokens extrait par InstructGPT-3.

Évaluation de la distillation des connaissances Nous comparons dans le Table 2 le modèle distillé via les annotations d'InstructGPT-3 (*Modèle Distillé*) et le modèle InstructGPT-3 sur le **layer 1**. Les deux modèles ont presque les mêmes performances, mais le modèle distillé a une meilleure précision, tandis qu'InstructGPT est plus performant en termes de rappel. Nous pouvons conclure que l'utilisation d'un modèle distillé est pertinent vis à vis de cette tâche.

Configuration A Si nous comparons le score F1 global (Table 3a), le modèle distillé via les annotations d'InstructGPT-3 (*Modèle Distillé*) est plus performant que le modèle (*Modèle par Supervision Faible*) formé avec la supervision faible par extraction par dictionnaire. Le *Modèle par Supervision*

Modèles	F1-Score	Précision	Rappel
Modèle Distillé	$0.74 \pm 6e^{-3}$	**0.72** $\pm$ 0.01	$0.78 \pm 6e^{-3}$
InstructGPT-3	0.74	0.70	**0.81**

TABLE 2 – Comparaison d'InstructGPT-3 avec un modèle distillé via de la supervision faible sur le **layer 2** d'E3C. Les scores sont calculés sur le **layer 1**.

Faible a un meilleur score F1 avec la classe B_{clin} , mais le rappel et le score F1 de I_{clin} sont relativement plus faibles.

Le *Modèle par Supervision Faible* a tendance à mieux extraire les termes cliniques composés d'un seul mot. Le *Modèle Distillé* a un meilleur rappel et reconnaît plus facilement les termes cliniques à plusieurs mots. Cependant, cette flexibilité est contrebalancée par la détection de faux positifs, ce qui fait baisser la précision.

Configuration B La quantité de tokens annotés est relativement faible par rapport au **layer 2** (Table 1). Cela nuit au résultat (Tableau 3b) du *Modèle par Supervision Faible* (0.57 avec le **layer 2 validate** contre 0.69 avec le **layer 2**). Contrairement au *Modèle Distillé*, où les performances sont relativement équivalentes par rapport au **layer 2**. Cependant, la stabilité du *Modèle Distillé* entre les 5 folds est affectée (0.01 avec le **layer 2 validate** vs. $6e^{-3}$ avec le **layer 2**). Le même problème de stabilité des folds est présent pour *Modèle par Supervision Faible*, ce qui peut être dû à la quantité relativement faible de tokens dans le jeu d'entraînement.

Configuration C Les résultats (Table 3c) montrent des performances légèrement meilleures pour les deux modèles. Nous remarquons que le *Modèle par Supervision Faible* a un meilleur rappel pour I_{clin} que dans la **Configuration A** (*Modèle par Supervision Faible* entraîné avec seulement des annotations de supervision faible) : respectivement 0.35 contre 0.24. Le *Modèle Distillé* présente une meilleure précision pour le B_{clin} et le I_{clin} par rapport au *Modèle Distillé* entraîné avec le **layer 2**. En conclusion, le mélange de jeux de données comportant une faible proportion d'annotations manuelles corrige la tendance du *Modèle par Supervision Faible* à ignorer les entités composées de plusieurs mots et celle du *Modèle Distillé* à inclure des mots supplémentaires aux entités présentes dans le corpus de test, notamment celles composées de peu de mots.

Configuration D Les résultats indiquent que l'utilisation qu'un ratio de 0.5 entre l'annotation prédite par InstructGPT-3 et l'annotation extraction par dictionnaire permet d'améliorer le F1-Score de 0.03 par rapport à l'utilisation d'un ensemble de données entièrement annoté avec l'annotation prédite par InstructGPT-3 (c'est-à-dire lorsque le rapport est de 1). En outre, l'incorporation d'une proportion d'annotations prédites par InstructGPT-3 au-dessus de 0.5 permet une meilleure stabilité entre les différentes folds. En conclusion, la combinaison de ces deux méthodes d'annotation disparates en termes de rappel et de précision permet d'obtenir un bon équilibre et d'augmenter le F1-Score.

Synthèse entre les différentes configurations La Table 4 montre que la configuration D est celle qui a obtenu légèrement un meilleur F1-score (0.76). Le mélange d'annotations par extraction de dictionnaire et InstructGPT-3 obtenu avec $r = 0.5$ permet de réduire l'écart entre les Rappel

Annotations	extraction par dictionnaire	InstructGPT-3
F1-score	0.69 ± 0.04	**0.74** $\pm 6e^{-3}$
B_{clin}	F : **0.76** ± 0.02 P : **0.83** $\pm 2e^{-3}$ R : $0.70 \pm 4e^{-2}$	F : $0.73 \pm 8e^{-3}$ P : 0.68 ± 0.02 R : **0.80** ± 0.01
I_{clin}	F : 0.34 ± 0.10 P : **0.69** ± 0.01 R : 0.24 ± 0.08	F : **0.53** ± 0.01 P : 0.41 ± 0.01 R : **0.74** ± 0.01

(a) **Configuration A**

Annotations	extraction par dictionnaire	InstructGPT-3
F1-score	0.57 ± 0.02	**0.74** ± 0.01
B_{clin}	F : 0.72 ± 0.03 P : **0.65** ± 0.07 R : 0.82 ± 0.04	F : **0.74** ± 0.03 P : 0.64 ± 0.06 R : **0.87** ± 0.02
I_{clin}	F : 0.02 ± 0.04 P : 0.33 ± 0.30 R : 0.01 ± 0.02	F : **0.53** ± 0.01 P : **0.42** ± 0.03 R : **0.71** ± 0.03

(b) **Configuration B**

Annotations	extraction par dictionnaire	InstructGPT-3
F1-score	0.73 ± 0.01	**0.75** $\pm 5e^{-3}$
B_{clin}	F : **0.78** $\pm 7e^{-3}$ P : $0.84 \pm 4e^{-3}$ R : **0.72** ± 0.01	F : $0.77 \pm 5e^{-3}$ P : **0.85** ± 0.01 R : 0.71 ± 0.01
I_{clin}	F : 0.46 ± 0.03 P : **0.66** ± 0.04 R : 0.35 ± 0.04	F : **0.50** ± 0.01 P : 0.64 ± 0.02 R : **0.42** ± 0.02

(c) **Configuration C**

TABLE 3 – Performances obtenues en utilisant les deux paramétrages avec les deux jeux d'annotations. La ligne F1-score dénote les macro-F1-scores agrégés des labels O, B_{clin}, I_{clin}.

Configuration	F1-Score	Précision	Rappel
A	0.74 ± 0.01	0.69 ± 0.01	0.83 ± 0.01
B	0.74 ± 0.02	0.69 ± 0.03	**0.84** ± 0.01
C	0.75 ± 0.00	**0.82** ± 0.01	0.71 ± 0.01
D r=0.5	**0.76** ± 0.01	0.73 ± 0.02	0.81 ± 0.03

TABLE 4 – Performances pour les configurations décrites en section 5.2.

(0.81) et la Précision (0.73). Ceci nous permet d'affirmer que les deux méthodes peuvent être complémentaires : une combinaison des deux permet d'obtenir une diversité d'annotations plus importante que l'utilisation seule d'une des deux méthodes.

6 Limitations

Les limites de notre étude est la taille petite du corpus de test, ce qui peut avoir un impact sur la généralisation de nos résultats. Nous avons limité notre travail à l'extraction d'entités cliniques ; dans des travaux futurs, nous expérimenterons d'autres tâches en utilisant la couche de temporalité E3C pour couvrir une tâche de reconnaissance d'entités nommées et d'extraction de relations.

Enfin, les guidelines E3C ont été conçues pour l'extraction d'entités cliniques associables aux concepts de l'UMLS. Après la première étape d'annotation manuelle, une partie de l'étendue des entités a été modifiée pour correspondre le plus possible aux concepts sémantiques trouvés dans l'UMLS (Magnini *et al.*, 2020). Ces biais induisent des difficultés supplémentaires quant aux résultats de la prédiction des modèles sur le **layer 1**.

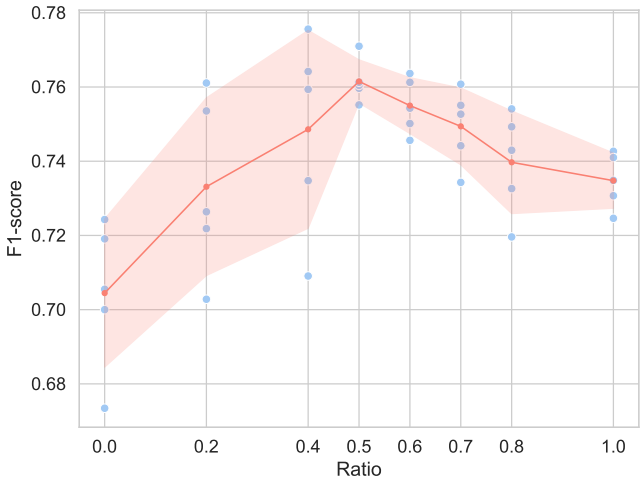


FIGURE 3 – La figure représente un graphique avec le F1-Score en ordonnée et le ratio r d'annotations par dictionnaire et d'annotations via InstructGPT-3 en abscisse comme précisé dans la **Configuration D** (Section 5.2). $r = 0$ indique l'utilisation exclusive d'annotations de dictionnaire, et $r = 1$ d'annotations de InstructGPT-3 exclusivement. Chaque point bleu représente une expérience différente avec un ensemble de données différent. Un point orange représente la moyenne des expériences pour r donné. La bande en orange clair représente l'écart-type pour r donné.

7 Conclusion

Nos résultats montrent que l'approche de distillation de connaissances avec InstructGPT-3 surpasse l'approche d'extraction de dictionnaire pour extraire les entités cliniques. Nous avons montré que mélanger ces approches pour construire un ensemble de données d'entraînement apporte de la diversité aux annotations et améliore les performances du modèle entraîné.

L'approche de faible supervision avec les LLMs est prometteuse pour créer un ensemble de données d'entraînement. Cela réduit le coût d'annotation et, en même temps, concentre l'annotation manuelle sur l'ensemble de test, qui est l'un enjeux majeurs des domaines à enjeux élevés comme la santé.

Dans les travaux futurs, nous aimerions tester cette solution sur plusieurs langues proposées dans E3C et étendre notre travail à d'autres tâches. Nous voulons également combiner plusieurs prédictions de différents grands modèles de langage pour obtenir plus de diversité dans les annotations.

De plus, nous avons l'intention d'étudier des techniques plus avancées pour combiner les différentes annotations en incorporant des mesures de confiance provenant des différentes prédictions ou bien des mesures de performance (telles que le rappel et la précision) pour décider quel type d'annotations (B_{clin} ou I_{clin}) conserver pour chaque méthode de prédiction.

Enfin, l'adaptation de CoT ou de connaissances générées (Wei *et al.*, 2022; Cobbe *et al.*, 2021) pour l'extraction d'entités cliniques pourrait être bénéfique pour améliorer la précision des LLMs. Nous pourrions créer un prompt où les différentes étapes d'annotation sont présentées à travers différents exemples. À chaque étape d'annotation, nous décrivons une instruction précise et son résultat. Par exemple, nous pouvons rédiger dans une instruction, les trois étapes d'annotation E3C pour encourager le LLM à être plus proche des guidelines d'identification de termes cliniques.

Références

AGRAWAL M., HEGSELMANN S., LANG H., KIM Y. & SONTAG D. (2022). Large Language Models are Few-Shot Clinical Information Extractors.

ALSENTZER E., MURPHY J. R., BOAG W., WENG W.-H., JIN D., NAUMANN T. & MCDERMOTT M. B. A. (2019). Publicly Available Clinical BERT Embeddings. DOI : 10.48550/arxiv.1904.03323.

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D. M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & OPENAI D. A. (2020). Language Models are Few-Shot Learners.

COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R., HESSE C. & SCHULMAN J. (2021). Training Verifiers to Solve Math Word Problems. DOI : 10.48550/arxiv.2110.14168.

CUSICK M., ADEKKANATTU P., CAMPION T. R., SHOLLE E. T., MYERS A., BANERJEE S., ALEXOPOULOS G., WANG Y. & PATHAK J. (2021). Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *Journal of Psychiatric Research*, **136**, 95–102. DOI : 10.1016/J.JPSYCHIRES.2021.01.052.

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), 140. DOI : 10.1186/s12911-017-0537-yï.

FRIES J. A., SEELAM N., ALTAY G., WEBER L., KANG M., DATTA D., SU R., GARDA S., WANG B., OTT S., SAMWALD M. & KUSA W. (2022). Dataset Debt in Biomedical Language Modeling. *Workshop on Challenges & Perspectives in Creating Large Language Models*, **5**, 137–145.

FRIES J. A., STEINBERG E., KHATTAR S., FLEMING S. L., POSADA J., CALLAHAN A. & SHAH N. H. (2021). Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*, **12**(1). DOI : 10.1038/s41467-021-22328-4.

GURURANGAN S., MARASOVÍCMARASOVÍC A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. (2020). Don't Stop Pretraining : Adapt Language Models to Domains and Tasks. *58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360.

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**. DOI : 10.1038/SDATA.2016.35.

KARAMANOLAKIS G., MUKHERJEE S., ZHENG G. & AWADALLAH A. H. (2021). Self-Training with Weak Supervision. p. 845–863. DOI : 10.18653/v1/2021.naacl-main.66.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). BioBERT : A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240. DOI : 10.1093/bioinformatics/btz682.

LEHMAN E., HERNANDEZ E., MAHAJAN D., WULFF J., SMITH M. J., ZIEGLER Z., NADLER D., SZOLOVITS P., JOHNSON A. & ALSENTZER E. (2023). Do We Still Need Clinical Language Models ?

LISON P., BARNES J. & HUBIN A. (2021). skweak : Weak Supervision Made Easy for NLP.

LISON P., BARNES J., HUBIN A. & TOUILEB S. (2020). Named Entity Recognition without Labelled Data : A Weak Supervision Approach. p. 1518–1533. DOI : 10.18653/V1/2020.ACL-MAIN.139.

LIU P., YUAN W., JIANG Z., HAYASHI H., NEUBIG G., FU J., YUAN W., JIANG Z., HAYASHI H., NEUBIG G. & FU J. (2023). Pre-train, Prompt, and Predict : A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, **55**(9), 1–35. DOI : 10.1145/3560815.

LU Q., DOU D. & NGUYEN T. H. (2022). ClinicalT5 : A Generative Language Model for Clinical Text.

LUO R., SUN L., XIA Y., QIN T., ZHANG S., POON H. & LIU T.-Y. (2022). BioGPT : Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in bioinformatics*, **23**(6). DOI : 10.1093/bib/bbac409.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2020). The E3C Project : Collection and Annotation of a Multilingual Corpus of Clinical Cases. *CEUR Workshop Proceedings*, **2769**. DOI : 10.4000/BOOKS.AACCADEMIA.8663.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE V., SEDDAH D. & SAGOT B. (2019). CamemBERT : a Tasty French Language Model. p. 7203–7219. DOI : 10.18653/v1/2020.acl-main.645.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). Training language models to follow instructions with human feedback. DOI : 10.48550/arxiv.2203.02155.

RAMSHAW L. A. & MARCUS M. P. (1995). Text Chunking using Transformation-Based Learning. p. 157–176. DOI : 10.48550/arxiv.cmp-lg/9505040.

ROSENBLOOM S. T., DENNY J. C., XU H., LORENZI N., STEAD W. W. & JOHNSON K. B. (2011). Data from clinical notes : A perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, **18**(2), 181–186. DOI : 10.1136/JAMIA.2010.007237.

SAINZ O., DE LACALLE O. L., LABAKA G., BARRENA A. & AGIRRE E. (2021). Label Verbalization and Entailment for Effective Zero- and Few-Shot Relation Extraction. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, p. 1199–1212. DOI : 10.48550/arxiv.2109.03659.

SCAO T. L., FAN A., AKIKI C., PAVLICK E. & ET AL. (2022). BLOOM : A 176B-Parameter Open-Access Multilingual Language Model. DOI : 10.48550/arxiv.2211.05100.

THOPPILAN R., DE FREITAS D., HALL J., SHAZEER N., KULSHRESHTHA A., CHENG H.-T., JIN A., BOS T., BAKER L., DU Y., LI Y., HUAIXIU H. L., ZHENG S., GHAFOURI A., MENEGALI M., HUANG Y., KRIKUN M., LEPIKHIN D., QIN J., CHEN D., XU Y., CHEN Z., ROBERTS A., BOSMA M., ZHAO V., ZHOU Y., CHANG C.-C., KRIVOKON I., RUSCH W., PICKETT M., SRINIVASAN P., MAN L., MEIER-HELLSTERN K., RINGEL M., TULSEE M., RENELITO D., SANTOS D., DUKE T., SORAKER J., ZEVENBERGEN B., PRABHAKARAN V., DIAZ M., HUTCHINSON B., OLSON K., MOLINA A., HOFFMAN-JOHN E., LEE J., AROYO L., RAJAKUMAR R., BUTRYNA A., LAMM M., KUZMINA V., FENTON J., COHEN A., BERNSTEIN R., KURZWEIL R., AGUERA-ARCAS B., CUI C., CROAK M., CHI E. & LE GOOGLE Q. (2022). LaMDA : Language Models for Dialog Applications.

WANG Y., SOHN S., LIU S., SHEN F., WANG L., ATKINSON E. J., AMIN S. & LIU H. (2019). A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, **19**(1), 1–13. DOI : 10.1186/S12911-018-0723-6/FIGURES/4.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. & ZHOU D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models.

YANG X., CHEN A., POURNEJATIAN N., SHIN H. C., SMITH K. E., PARISIEN C., COMPAS C., MARTIN C., COSTA A. B., FLORES M. G., ZHANG Y., MAGOC T., HARLE C. A., LIPORI G., MITCHELL D. A., HOGAN W. R., SHENKMAN E. A., BIAN J. & WU Y. (2022). A large language model for electronic health records. *npj Digital Medicine*, **5**(1). DOI : 10.1038/s41746-022-00742-2.

ZHANG J., HSIEH C.-Y., YU Y., ZHANG C. & RATNER A. (2022). A Survey on Programmatic Weak Supervision. DOI : 10.48550/arxiv.2202.05433.

ZHAO T. Z., WALLACE E., FENG S., KLEIN D. & SINGH S. (2021). Calibrate Before Use : Improving Few-Shot Performance of Language Models. DOI : 10.48550/arxiv.2102.09690.

A Appendice

```
Input: L'évolution était marquée deux mois plus tard, par
l'apparition de placards angiomateux au
niveau de l'avant bras droit, [...]
extract the exact match of disorders, diseases or
symptoms mentioned in the text or
return None if there is no clinical entity:

- "placards angiomateux"
- "lymphoedème"
- "lésions"

Input: De façon concomitante, le patient avait présenté une
altération progressive de l'état général,
une fièvre et des sueurs nocturnes
extract the exact match of disorders, diseases
or symptoms mentioned in the text or
return None if there is no clinical entity:

- "altération progressive de l' état général"
- "fièvre"
- "sueurs nocturnes"

Input: La vitesse de sédimentation était à 35mm à
la première heure, la protéine C réactive
était négative et
la Ferritinémie était à 900µg/l (soit à 4 fois la normale).
extract the exact match of disorders, diseases
or symptoms mentioned in the text or
return None if there is no clinical entity:

- "None"

Input: L'interrogatoire n'a retrouvé aucun antécédent
pathologique en particulier la notion d'éruption cutanée,
de troubles du transit, d'ictère,
d'épisode infectieux respiratoire ou de vaccination récente.
extract the exact match of disorders, diseases
or symptoms mentioned in the text or
return None if there is no clinical entity:
- "
```

FIGURE 4 – Un exemple du texte d'instruction utilisé dans notre expérience. Les exemples formatés sont présentés en bleu, tandis que l'exemple à prédire est présentés en orange. Les instructions sont présentées en violet, et la *requête guidée*, telle qu'utilisée dans Agrawal *et al.* (2022), sont présentées en vert.