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

Passe ta pharma d'abord !

Simon Meoni [1, 2, *] Rian Touchent [1, *] Éric de la Clergerie [1]
(1) Inria, Paris, France
(2) Arkhn, Paris, France
simon.meoni@arkhn.com, rian.touchent@inria.fr,
eric.de_la_clergerie@inria.fr

RÉSUMÉ

Nous présentons les 3 expériences menées par l'équipe ALMAnaCH - Arkhn et leurs résultats pour le DÉfi Fouille de Textes (DEFT) 2023. Les scores sont encourageants mais suggèrent surtout de nouveaux éléments à prendre en compte pour réussir ce défi. Nous avons exploré différentes approches avec des modèles de tailles variables et modélisé la tâche de différentes manières (classification multi-labels, implication textuelle, séquence à séquence). Nous n'avons pas observé des gains de performance significatifs. Nos expériences semblent montrer la nécessité de l'utilisation de bases de connaissances externes pour obtenir de bons résultats sur ce type de tâche.

ABSTRACT

Graduate Pharma First !

We present 3 experiments and results obtained by the ALMAnaCH - Arkhn team for the Text Mining Challenge (DEFT) 2023. We have explored various approaches with models of varying sizes and by modeling the task differently (multi-label classification, natural language inference, sequence-to-sequence). The results are encouraging but suggest new elements to consider to succeed in this challenge. We have not observed significant performance gains. Our experiments indicate the necessity of using external knowledge bases to achieve good results on this type of task.

MOTS-CLÉS : biomédical, pharmacologie, QCM, implication textuelle, affinage avec instruction, TAL.

KEYWORDS: biomedical, pharmacology, MCQA, textual entailment, prompt-tuning, NLP.

1 Introduction

Dans cet article, nous présentons notre participation au DÉfi Fouille de Textes (DEFT) 2023, une campagne d'évaluation francophone. L'objectif de ce défi est de développer des approches permettant de répondre automatiquement à des questionnaires à choix multiples issus d'annales d'examens de pharmacie.

*. Ces auteurs ont contribué de manière égale à ce travail

Actes du Défi Fouille de Textes@TALN2023, pages 68–76.

2 Corpus

FrenchMedMCQA (Labrak *et al.*, 2022) est composé de 3105 questions fermées, extraites d'annales françaises d'examens de pharmacie. Chaque question est associée à un identifiant et cinq réponses possibles. Le nombre de bonnes réponses par question varie entre 1 et 5. Le corpus est divisé en trois sous-ensembles : entraînement, développement et test. Le corpus d'entraînement représente 70% des questions, celui de développement 10%, et celui de test 20%.

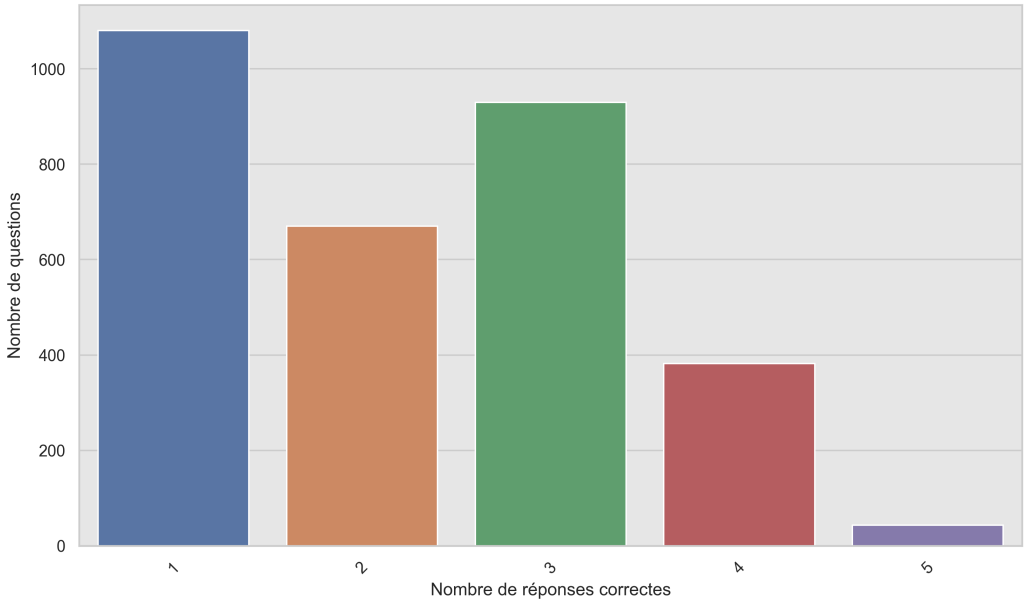

FIGURE 1 – Répartition du nombre de réponses correctes par question

Une exploration des données Fig. 1 montre un faible nombre de questions à 4 ou 5 réponses correctes. La majorité des questions n'ont qu'une seule bonne réponse, ce qui est souvent explicité dans la question elle-même. Le nombre de bonnes réponses n'est pas précisé dans la question quand il est supérieur à 1.

3 Description de la tâche

Le défi comporte deux tâches principales. La première tâche consiste à identifier automatiquement l'ensemble des réponses correctes parmi les cinq options proposées pour chaque question. L'évaluation de cette tâche est basée sur le taux de réponses parfaitement justes (*Exact Match Ratio*, EMR) et le taux de réponses justes parmi l'ensemble des réponses et références (*Hamming Score*). Le classement final des équipes se fait en fonction de l'Exact Match Ratio.

La deuxième tâche annexe consiste à estimer le nombre de réponses supposément justes pour chaque question, qui peut varier de 1 à 5. L'évaluation de cette tâche se fait à l'aide des métriques de précision et de score F1.

Seule la première tâche a été étudiée dans notre participation.

4 Méthode

4.1 Découpage du corpus et protocole expérimental

Run	Modèle	Taux d'apprentissage	Taille de lots	Accumulation de gradients	Epochs
run 1	`camembert-base`	$1e^{-5}$	8	2	25
run 2	`camembert-bio-base`	$4e^{-5}$	4	4	25
run 3	`flan-t5-xl`	$4e^{-5}$	1	16	10

TABLE 1 – Hyperparamètres retenus pour chaque run

Pour la run 1 et la run 2, nous avons exploré les modèles `camembert-base` (Martin *et al.*, 2020), `camembert-bio-base` (Touchent *et al.*, 2023) et `DrBERT-7GB` (Labrak *et al.*, 2023) et le taux d'apprentissage entre $1e^{-5}$ et $4e^{-5}$. Nous avons ensuite sélectionné les hyperparamètres (Table 1) donnant les meilleurs scores sur le jeu de développement en utilisant optuna (Akiba *et al.*, 2019)

Pour la run 3, nous avons exploré les modèles `t5-base`, `t5-large`, `flan-t5-xl`, `SciFive-base-Pubmed_PMC`, `mt5-base`.

Nous avons adapté le découpage initial du corpus à nos besoins pour l'ensemble de nos expériences comme suit :

— notre corpus d'entraînement contient 80% du jeu de données d'entraînement initial ;
— notre corpus de validation est le même que le jeu de données de développement initial ;
— notre corpus de test contient 20% du jeu de données d'entraînement initial ;
— notre corpus de soumission correspond au jeu de test initial ;

Lors de l'entraînement et pour chaque *epoch*, nous mesurons l'Exact Match Ratio sur le jeu de développement afin de sélectionner les meilleurs paramètres du modèle lors de la phase d'entraînement. Le corpus de test nous a permis d'évaluer nos approches avant la soumission. Nous sélectionnons le modèle ayant obtenu le meilleur score sur le corpus de test afin de l'utiliser pour la prédiction sur le corpus de soumission.

4.2 run 1 : Classification multi-labels

Notre première approche est basée sur de la classification multi-labels. La question et les 5 réponses correspondantes sont concaténées (Fig.2), séparées par un token spécifique, avant d'être encodées. La prédiction est ensuite réalisée avec une simple couche linéaire pour obtenir 5 logits, correspondants à la probabilité estimée par le réseau de neurones pour chaque question. La prédiction finale est obtenue en gardant les probabilités supérieures à 0.5, notre seuil, comme positifs.

4.3 run 2 : NLI

Dans la seconde approche, nous modélisons le défi comme une tâche de prédiction d'implication textuelle. Pour chaque question, nous construisons 5 paires de question-réponse. Où chacune de ses paires de question-réponse correspondent à la question suivi d'une des 5 propositions de réponses

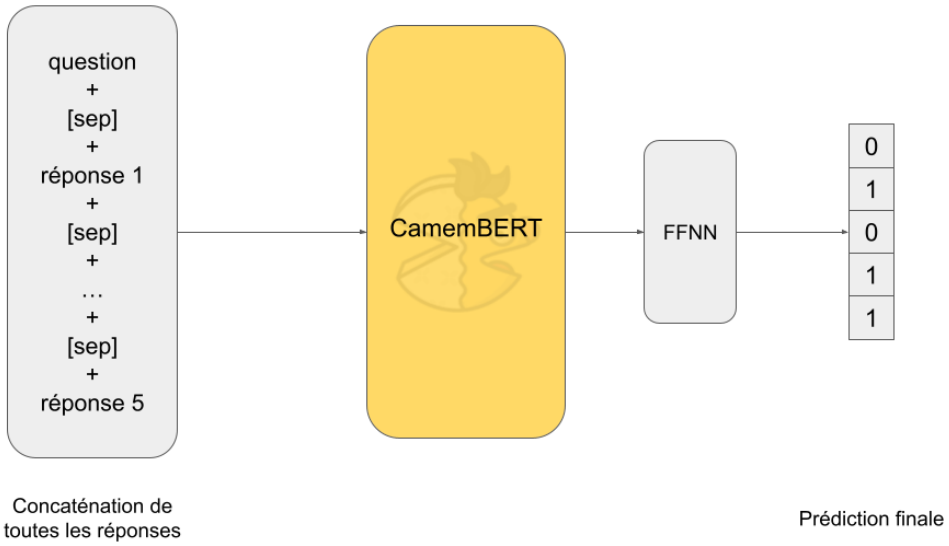

FIGURE 2 – Schéma de notre approche par classification multi-labels

(Fig.3). Chacune de ces paires est encodée par `camembert-base`, puis une couche linéaire va prédire si la question implique la réponse ou non, dans une tâche de classification binaire.

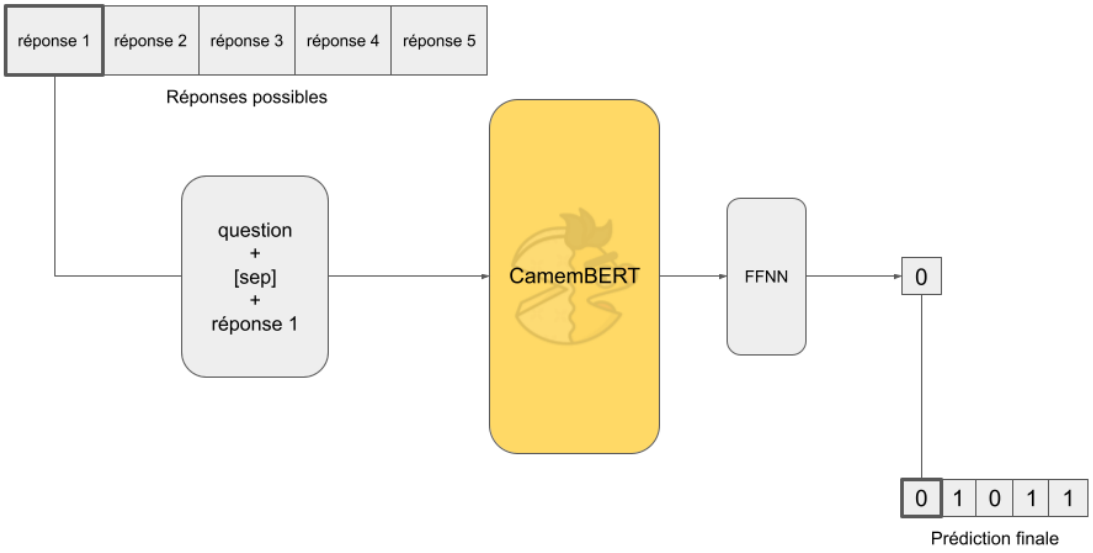

FIGURE 3 – Schéma de notre approche par NLI

La prédiction est alors réalisée 5 fois par question. Les 5 probabilités sont ensuite concaténées, et deviennent positives ou négatives en fonction du seuil, ce qui constitue la prédiction finale.

4.4 run 3 : Instruction-seq2seq

Pour cette run, nous nous sommes inspirés des travaux de Wang *et al.* (2022). Dans ce papier, les auteurs convertissent des tâches d'extraction d'entités nommées en une tache de séquence à séquence afin de les adapter à des modèles de type encodeur-décodeur tel que `t5` (Raffel *et al.*, 2020). Le but est de fournir au modèle une instruction en entrée afin d'avoir en sortie les résultats désirés comme illustrés sur la figure 4. La phase d'entraînement est un affinage classique sur la tâche. Quant à la phase d'évaluation, elle consiste à structurer la sortie textuelle de `t5` en une sortie interprétable pour un script d'évaluation donné ou un cas d'usage précis. Pour les besoins de la tâche, nous avons adapté cette technique en utilisant seulement une seule tâche principale. D'autre part, Nous avons essayé différents modèles tels que `t5`, `t5-large` et `flan-t5-xl` (Chung *et al.*, 2022). Pour la soumission nous avons sélectionné le meilleur, à savoir `flan-t5-xl`.

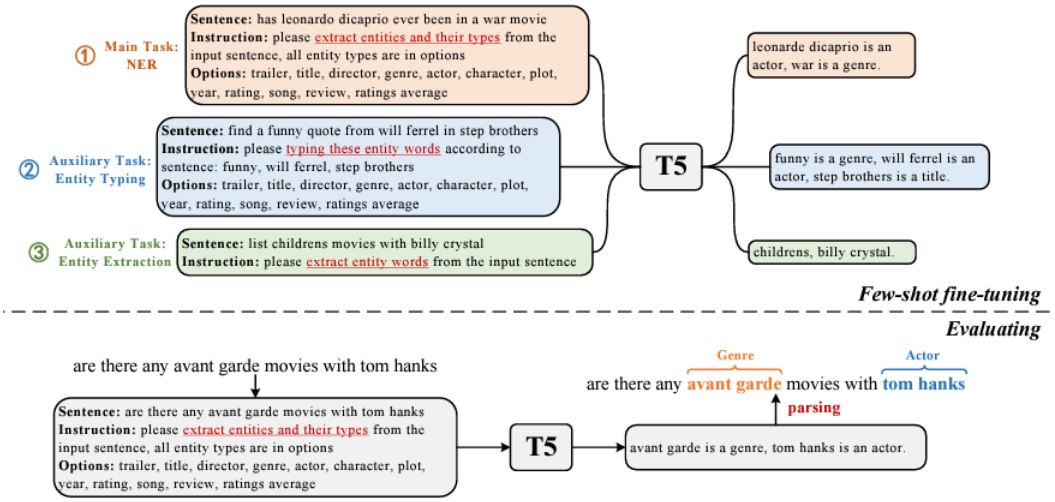

FIGURE 4 – Représentation du *framework* InstructionNER de Wang *et al.* (2022)

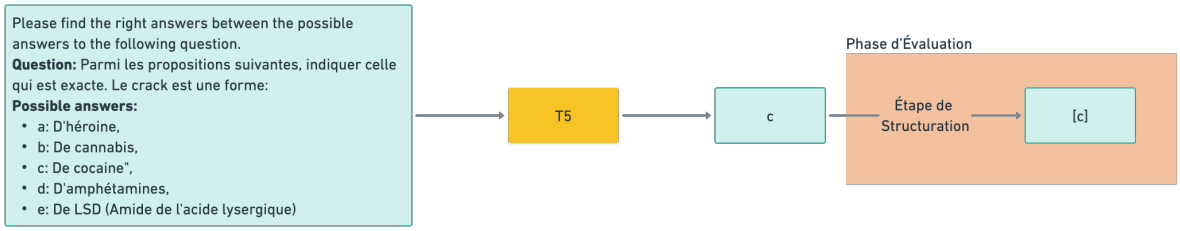

FIGURE 5 – Schéma de notre méthode par instruction sur un modèle de séquence à séquence

5 Résultats

En termes de taux de réponses parfaitement justes (Exact Match Ratio - EMR), le système NLI s'est légèrement démarqué des autres, bien que les différences ne soient pas significatives. En effet, les scores EMR des trois systèmes se situent dans le même ordre de grandeur et sont relativement faibles (Table 2). Nous n'avons pas réussi à obtenir des systèmes assez fiables pour cette tâche. Nous

Nom du système	Hamming	EMR
multilabel-classification	33.27	12.22
nli	33.67	14.15
instruction-seq2seq	35.96	13.67

TABLE 2 – Comparaison des performances entre les différents systèmes

pensons que c'est du à un manque d'accès à des connaissances précises dans certains cas de figures (par exemple sur des valeurs numériques). L'utilisation de connaissances externes aurait pu être une option.

En revanche, en considérant le score de Hamming, qui mesure la proportion de réponses correctes identifiées parmi les options proposées, le système instruction-seq2seq s'est avéré légèrement plus performant par rapport aux autres systèmes. Il a démontré une meilleure capacité à identifier les bonnes réponses parmi les options proposées. Afin d'améliorer ce système, nous aurions pu filtrer les réponses proposées par ce modèle à l'aide d'un modèle entraîné à cette ou à l'aide de techniques basées sur des graphes de connaissances.

La run 3 met en évidence un constat intéressant. Malgré l'utilisation de modèles plus larges tels que `flan-t5-xl`, nous n'avons pas observé d'augmentation significative des performances par rapport à l'utilisation de modèles de taille plus réduite. Ce qui signifie que la part de données pharmacologiques durant le pré-entraînement de `flant-t5-xl` n'est pas assez importante ou que les tâches de pré-entraînement ne sont pas assez adaptées à notre cas d'usage.

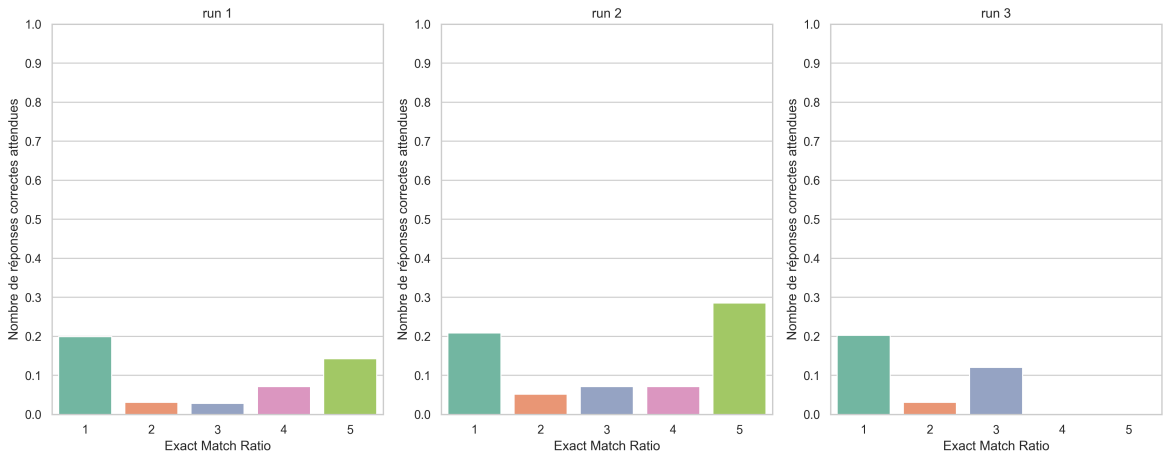

FIGURE 6 – EMR pour chaque nombre de bonnes réponses attendues

Dans la Fig. 6, on observe que la run 3 obtient un EMR de 0 si on ne prend en compte que les questions où entre 4 et 5 réponses sont attendues. Cela s'explique par une quasi absence de prédiction à 5 réponses de la part de `flan-t5-xl` (Fig.7). Il semblerait que T5 a amplifié le biais de la distribution du jeu d'entraînement, en se concentrant principalement sur des prédictions à 1 ou 3 réponses. C'est cependant aussi T5 qui obtient assez légèrement les meilleurs résultats sur ces deux catégories.

La run 2 semble moins sensible aux biais de la distribution du jeu d'entraînement. Le modèle de la run 2 semble moins frileux à prédire 5 réponses pour une question, ce qui lui permet d'obtenir un très bon score EMR pour les questions à 5 réponses (Fig.6). C'est également la run qui a le meilleur

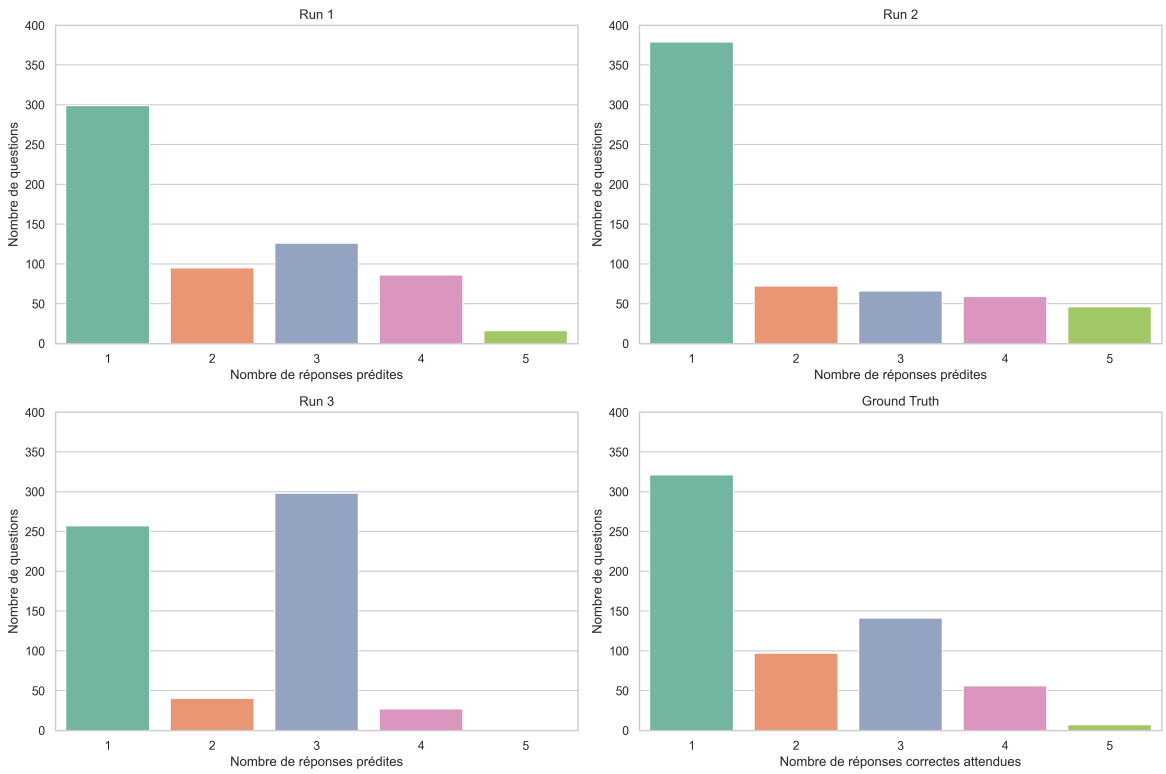

FIGURE 7 – Distribution du nombre de réponses prédites ou attendues

score EMR global. Cependant, on peut également déduire que ce modèle a du mal à prédire le bon nombre de réponses puisque la distribution finale du nombre de réponse prédite par question est assez différente de la distribution du jeu de test.

Enfin la run 1 reproduit correctement la distribution du nombre de réponses à prédire par question, elle a cependant un EMR global légèrement moins bon que la run 2. Il semblerait que le modèle de la run 1 soit meilleur pour prédire le nombre de réponse à prédire, mais plus faible pour trouver les bonnes réponses.

Avec ces résultats, on peut imaginer qu'une approche ensembliste pourrait permettre de rassembler les points forts de chaque run, qui semble souffrir de failles différentes. Il semblerait également que la seconde tâche aurait pu aider nos modèles. En effet, dans certaines de nos runs les modèles font de grandes erreurs dans le nombre de réponse à prédire avant même de trouver les bonnes réponses. Ainsi, jouer avec le seuil ou utiliser un modèle pour prédire le nombre de réponses correctes a priori pour permettre d'augmenter les performances.

6 Conclusion et perspectives

Notre participation au DÉfi Fouille de Textes (DEFT) 2023 nous donne des conclusions intéressantes. Les scores relativement faibles de nos systèmes, bien que d'approches différentes, semblent montrer qu'il est difficile d'obtenir de bonnes performances pour cette tâche sans utilisation de bases de connaissances externes. Ni les connaissances apprises à l'aide du jeu d'entraînement de FrenchMedMCQA, ni celles apprises lors du pré-entraînement de nos modèles ne semblent suffisantes. Les

jeux de pré-entraînement de nos modèles sont variés, avec des jeux spécialisés pour CamemBERT-bio et DrBERT, ou des jeux de très grande taille comme avec Flan-T5 (Chung *et al.*, 2022). Ces types de modèle montrent de bonnes performances sur un certain nombre de tâches biomédicales sans utilisation de bases de connaissances externes (Lehman *et al.*, 2023).

En effet, les questions de ce jeu de données sont très spécifiques et basées sur de la connaissance. Il serait alors intéressant d'explorer par la suite des méthodes qui exploitent des bases de connaissances externes.

Une première méthode serait d'injecter dans les instructions que l'on donne à notre modèle de langue, un contexte pertinent vis-à-vis de la question. Ce contexte pourrait alors être extrait depuis un corpus biomédical dans laquelle on aurait fait une recherche sémantique ou une recherche par mots clés type MeSH, et ainsi donner les connaissances nécessaires pour répondre à la question (Noh & Kavuluru, 2018). Cette méthode n'assure cependant pas que le contexte contient les informations nécessaires pour répondre, mais seulement des phrases sémantiques proches de la question dans un corpus donné.

Une autre approche serait de reconnaître les différents types de question. En effet certaines sont basées sur du calcul numérique. On pourrait alors détecter le type de question et faire appel par la suite à un agent spécialisé pour ce problème. Schick *et al.* (2023) montrent qu'un modèle de langue génératif est capable d'identifier différents problèmes et d'utiliser d'autres agents pour y répondre. Cela demande cependant de trouver à l'avance les différents types de questions et de construire ou identifier un agent pertinent pour chacun d'entre eux.

Il est également possible d'utiliser une base de connaissances externe pour générer de nouvelles questions (Sileo *et al.*, 2023). Cela nous permettrait d'avoir un jeu d'entraînement plus conséquent d'une part, et d'autre part d'encoder des connaissances externes directement dans le jeu d'entraînement.

En conclusion, pour améliorer les performances de réponse automatique à des questionnaires à choix multiples en pharmacologie, il serait intéressant d'explorer des méthodes qui exploitent des bases de connaissances externes, en générant de nouvelles questions, en injectant du contexte pertinent dans les instructions données à un modèle génératif ou en construisant et en appelant des agents pertinents pour chacun des types de questions possibles. Ces pistes de recherche offrent des perspectives intéressantes pour de futurs travaux dans le domaine.

Références

AKIBA T., SANO S., YANASE T., OHTA T. & KOYAMA M. (2019). Optuna : A next-generation hyperparameter optimization framework. *CoRR*, **abs/1907.10902**.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI E., WANG X., DEHGHANI M., BRAHMA S., WEBSON A., GU S. S., DAI Z., SUZGUN M., CHEN X., CHOWDHERY A., NARANG S., MISHRA G., YU A., ZHAO V., HUANG Y., DAI A., YU H., PETROV S., CHI E. H., DEAN J., DEVLIN J., ROBERTS A., ZHOU D., LE Q. V. & WEI J. (2022). Scaling instruction-finetuned language models. DOI : 10.48550/ARXIV.2210.11416.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2022). FrenchMedMCQA : A French multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, p. 41–46, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

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.

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 ?

MARTIN L., MULLER B., JAVIER ORTIZ SUÁREZ P., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE E., SAGOT B. & SEDDAH D. (2020). Les modèles de langue contextuels Camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement. In C. BENZITOUN, C. BRAUD, L. HUBER, D. LANGLOIS, S. OUNI, S. POGODALLA & S. SCHNEIDER, Éds., *JEP-TALN-RECITAL 2020 - 33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 54–65, Nancy, France : ATALA. HAL : hal-02784755.

NOH J. & KAVULURU R. (2018). Document Retrieval for Biomedical Question Answering with Neural Sentence Matching. *Proc Int Conf Mach Learn Appl*, **2018**, 194–201.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

SCHICK T., DWIVEDI-YU J., DESSÌ R., RAILEANU R., LOMELI M., ZETTLEMOYER L., CANCEDDA N. & SCIALOM T. (2023). Toolformer : Language models can teach themselves to use tools.

SILEO D., UMA K. & MOENS M.-F. (2023). Generating multiple-choice questions for medical question answering with distractors and cue-masking.

TOUCHENT R., ROMARY L. & VILLEMONTE DE LA CLERGERIE E. (2023). CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. working paper or preprint.

WANG L., LI R., YAN Y., YAN Y., WANG S., WU W. & XU W. (2022). Instructionner : A multi-task instruction-based generative framework for few-shot ner.