Tâches et systèmes de détection automatique des réponses correctes dans des QCMs liés au domaine médical : Présentation de la campagne DEFT 2023

Yanis Labrak, Adrien Bazoge, Béatrice Daille, Richard Dufour, Emmanuel Morin, Mickael Rouvier

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

Tâches et systèmes de détection automatique des réponses correctes dans des QCMs liés au domaine médical : Présentation de la campagne DEFT 2023

Yanis Labrak[1, 3] Adrien Bazoge[2] Béatrice Daille[2] Richard Dufour[2]
Emmanuel Morin[2] Mickael Rouvier[1]
(1) Laboratoire Informatique d'Avignon (LIA), Avignon Université, France
(2) Laboratoire des Sciences du Numérique de Nantes (LS2N), Nantes Université, France
(3) Zenidoc, France
{prenom.nom}@univ-avignon.fr, {prenom.nom}@univ-nantes.fr

RÉSUMÉ

L'édition 2023 du DÉfi Fouille de Textes (DEFT) s'est concentrée sur le développement de méthodes permettant de choisir automatiquement des réponses dans des questions à choix multiples (QCMs) en français. Les approches ont été évaluées sur le corpus FrenchMedMCQA, intégrant un ensemble de QCMs avec, pour chaque question, cinq réponses potentielles, dans le cadre d'annales d'examens de pharmacie. Deux tâches ont été proposées. La première consistait à identifier automatiquement l'ensemble des réponses correctes à une question. Les résultats obtenus, évalués selon la métrique de l'Exact Match Ratio (EMR), variaient de 9,97 % à 33,76 %, alors que les performances en termes de distance de Hamming s'échelonnaient de 24,93 à 52,94. La seconde tâche visait à identifier automatiquement le nombre exact de réponses correctes. Les résultats, quant à eux, étaient évalués d'une part avec la métrique de F1-Macro, variant de 13,26 % à 42,42 %, et la métrique *Accuracy*, allant de 47,43 % à 68,65 %. Parmi les approches variées proposées par les six équipes participantes à ce défi, le meilleur système s'est appuyé sur un modèle de langage large de type LLaMa affiné en utilisant la méthode d'adaptation LoRA.

ABSTRACT

Tasks and systems for automatic question-answering in the medical field : presentation of the DEFT 2023 campaign.

The 2023 edition of the text mining challenge *DÉfi Fouille de Textes* (DEFT) focused on the development of methods for automatically selecting answers in multiple-choice question (MCQ) in French. The approaches were evaluated on the FrenchMedMCQA corpus, which includes a set of MCQ with five potential answers for each question, based on pharmacy exam archives. Two tasks have been proposed. The first one aimed to automatically identify all the correct answers to a question. The obtained results, evaluated using the Exact Match Ratio (EMR) metric, ranged from 9.97% to 33.76%, while the performances in terms of Hamming score ranged from 24.93 to 52.94. The second task aimed to automatically identify the exact number of correct answers. The results, on the other hand, were evaluated using the F1-Macro metric, ranging from 13.26% to 42.42%, and the Accuracy metric, ranging from 47.43% to 68.65%. Among the various approaches proposed by the six participating teams in this challenge, the best system relied on a large language model of the LLaMa type, fine-tuned using the LoRA adaptation method.

MOTS-CLÉS : Question à choix multiples ; Domaine médical ; Modèle de langue large ; TALN.

Actes du Défi Fouille de Textes@TALN2023, pages 57–67.

KEYWORDS: Multiple Choice Question ; Medical field ; Large Language Model ; NLP.

1 Introduction

Le DÉfi Fouille de Textes (DEFT) est une campagne d'évaluation annuelle francophone qui permet à plusieurs équipes, souvent issues du monde académique et/ou industriel, de confronter des méthodes originales en traitement automatique du langage naturel (TALN) sur une ou plusieurs tâches régulièrement renouvelées.

Pour cette édition 2023 du défi [1], nous avons proposé de travailler sur le corpus FrenchMedMCQA (Labrak *et al.*, 2022), intégrant un ensemble de QCMs en français issus d'annales d'examens de pharmacie. Une des difficultés, et originalités, du corpus, est que chaque question contient une inconnue sur le nombre de réponses associées, là où d'autres corpus attendent une seule réponse par question. Cette difficulté a permis aux équipes participantes d'explorer et de proposer des approches pouvant s'écarter de celles actuellement proposées pour des tâches plus classiques en TALN. Deux tâches ont été proposées aux participants pour cette édition :

1. Sélectionner automatiquement le sous-ensemble de réponses correctes parmi l'ensemble proposées pour une question donnée en s'aidant, ou non, de connaissances externes au corpus fourni.
2. Identifier, pour chaque question, le nombre exact de réponses correctes.

La campagne a été lancée le 27 février 2023. L'accès aux données d'entraînement était possible après signature d'un accord par tous les membres de l'équipe participante. La phase d'entraînement s'est déroulée sur pratiquement deux mois (27 février 2023 au 23 avril 2023). La phase de test s'est déroulée du 24 avril au 7 mai 2023. Six équipes se sont inscrites, et sont allées jusqu'au terme de la campagne :

— *ALMANACH-ARKHN* (Meoni *et al.*, 2023) : Équipe jointe entre l'entreprise Arkhn et l'INRIA.
— *LIS* (Favre, 2023) : LIS (Aix-Marseille Université).
— *LIUM-IRISA* (Besnard *et al.*, 2023) : Équipe jointe entre le LIUM (Le Mans Université) et l'IRISA (Université de Rennes).
— *SEQUOIA* (Bothua *et al.*, 2023) : Entreprise EDF R&D.
— *SPQR* (Bezançon *et al.*, 2023) : Équipe jointe entre le STIH (Sorbonne Université), le L3I (La Rochelle Université) et l'OBTIC (Sorbonne Université).
— *TTGV* (Blivet *et al.*, 2023) : Équipe jointe entre l'entreprise SNCF R&D, le LORIA (Université de Lorraine) et le LTCI (Telecom Paris).

2 Corpus

Le corpus FrenchMedMCQA (Labrak *et al.*, 2022) contient un ensemble de 3 105 QCMs en français portant sur le domaine médical, proche de ce que l'on retrouve dans d'autres langues telles que l'anglais avec le corpus MedMCQA (Pal *et al.*, 2022) ou SciQ (Welbl *et al.*, 2017). Ce corpus a été constitué en collectant des questions et leurs réponses associées à partir d'annales d'examens réels de pharmacie obtenus du site *Remede.org* [2]. Chaque QCM contient cinq réponses potentielles, parmi

1. https://deft2023.univ-avignon.fr
2. http://www.remede.org/internat/pharmacie/qcm-internat.html

lesquelles se trouve une ou plusieurs réponses correctes. Ces QCMs ont été réalisés manuellement par des experts médicaux et utilisés lors d'examens de pharmacie.

Le Tableau 1 fournit la distribution du jeu de données FrenchMedMCQA selon son découpage pour l'apprentissage, le développement et l'évaluation. Nous constatons que le corpus est composé de 1 080 questions ayant une réponse unique parmi les cinq potentielles ($\#Réponses = 1$), et de 2 025 questions avec de multiples réponses ($\#Réponses > 1$). Au total, le corpus contient 3 105 questions. Afin de garder un corpus équilibré, 70% des questions sont utilisées pour le corpus d'apprentissage, alors que 10 % ont été conservées pour le corpus de développement et 20% pour l'évaluation.

# Réponses	Apprentissage	Développement	Évaluation	Total
1	595	164	321	**1 080**
2	528	45	97	**670**
3	718	71	141	**930**
4	296	30	56	**382**
5	34	2	7	**43**
Total	**2 171**	**312**	**622**	**3 105**

TABLE 1 – Distribution du corpus FrenchMedMCQA selon son découpage en apprentissage, développement et test.

Chaque instance du corpus comprend un identifiant, une question, cinq réponses potentielles (étiquetées dans le corpus de A à E), et la (ou les) réponse(s) correcte(s). La longueur moyenne des questions est de 14,17 mots et la longueur moyenne des réponses est de 6,44 mots. Le vocabulaire compte 13 000 mots, sachant que 3 800 d'entre-eux (soit environ 29 %) sont spécifiques au domaine médical. Dans le détail, en moyenne, chaque question contient 2,5 mots spécifiques au domaine médical (représentant 17 % des mots en moyenne dans une question) et chaque réponse en contient 2 (représentant 36 % des mots en moyenne dans une réponse). Enfin, toujours en moyenne, un mot spécifique au domaine médical ciblé apparaît dans 2 questions et dans 8 réponses. La Figure 1 donne un exemple d'une instance pour une question contenant plusieurs réponses correctes.


```
{
  "id": "6979d46501a3270436d37b98cf351439fbcbec8d5890d293dabfb8f85f723904",
  "question": "Cocher la (les) proposition(s) exacte(s) : Le métronidazole :",
  "answers": {
    "A": "Est un dérivé du pyrazole",
    "B": "Peut induire un effet antabuse",
    "C": "Peut être administré par voie parentérale intraveineuse",
    "D": "Peut être utilisé dans certaines parasitoses à protozoaires",
    "E": "Est inefficace dans les infections à germes anaérobies"
  },
  "correct_answers": ["B", "C", "D"],
  "nbr_correct_answers": 3,
}
```


Listing 1: Exemple d'une instance du corpus FrenchMedMCQA, comprenant un identifiant, une question, cinq réponses potentielles (étiquetées de A à E) et les réponses correctes.

Lors de ce défi, deux tâches liées aux QCMs médicaux ont été proposées. La tâche principale (voir Section 2.1.1) a consisté à choisir automatiquement, selon une question posée, l'ensemble des réponses correctes parmi cinq réponses possibles fournies. La tâche annexe (voir Section 2.1.2) a consisté à identifier automatiquement le nombre de réponses correctes. À ces deux tâches, nous avons proposé,

pour chacune d'entre-elles, deux pistes (voir Section 2.2) que les équipes pouvaient développer : 1) *recherche reproductible*, avec des approches et modèles dont les données d'entraînement ou de référence étaient connues et contrôlées, et 2) *aucune restriction*, laissant libre chaque participant de fournir des propositions sans contrainte de reproductibilité. À noter que la seconde piste de recherche n'apparaît qu'à des fins de recherche et ne compte pas dans le classement final des équipes à la campagne DEFT 2023.

Pour l'ensemble des tâches et pistes, les participants ont eu à leur disposition les données d'entraînement et de développement comme décrites dans la Section 2. Les annotations du corpus de test, sur lequel toutes les équipes ont été évaluées, n'a jamais été fourni durant la campagne d'évaluation, mais a été rendu disponible librement, tout comme les corpus d'entraînement et de développement, à la fin de la campagne.

Enfin, nous avons fourni à l'ensemble des équipes un système état-de-l'art (*baseline*) pour chacune des deux tâches. Ces systèmes ont été transmis sous la forme de recettes disponibles librement en ligne [3] que chaque participant pouvait entraîner. Ces premiers résultats permettaient aux équipes d'avoir un repère quant aux performances de leurs approches.

2.1 Tâches proposées

2.1.1 Tâche principale : Choix automatique des réponses correctes dans un QCM

Présentation La tâche principale consiste à identifier automatiquement la ou les bonne(s) réponse(s) parmi l'ensemble de réponses proposées. Les participants ont alors à leur disposition la question posée ainsi que les cinq réponses potentielles, leur système devant choisir celles qui sont correctes.

Évaluation Contrairement à une tâche de classification classique où il est demandé d'associer une étiquette à un problème donné, notre tâche principale peu impliqué le fait d'avoir une réponse partiellement correcte. Par exemple, si nous devons retrouver deux réponses correctes parmi les cinq options disponibles pour une question ciblée, mais que notre système automatique n'est capable que d'en retrouve une seule, alors la réponse n'est pas juste, car incomplète. Il faut donc, dans ce cas, mettre en place une métrique permettant de prendre en compte la proportion de réponses justes, tout en pénalisant la/les réponse(s) incorrecte(s) dans le but d'éviter que les systèmes proposés répondent aux questions par l'ensemble du champ des possibilités. Dans cette optique, deux métriques différentes ont été utilisées, à savoir la correspondance exacte entre les réponses produites et la référence (*Exact Match Ratio*, EMR) ainsi que la Distance de Hamming (*Hamming Score*).

$$\text{Exact Match Ratio (EMR)} = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i = y_i]$$

où N est le nombre de questions, $\hat{y}_i$ est l'ensemble de réponses prédites pour la i-ième question, y_i est l'ensemble des bonnes réponses pour la i-ième question, et $[x]$ est une fonction indicatrice qui vaut 1 si x est vrai et 0 dans le cas contraire.

3. https://github.com/qanastek/DEFT-2023/tree/main/training_scripts

$$\text{Hamming Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}$$

où, N est le nombre de questions, y_i est l'ensemble des bonnes réponses pour la $i$ème question, $\hat{y}_i$ est l'ensemble des réponses prédites pour la $i$ème question, $|y_i \cap \hat{y}_i|$ est la taille de l'intersection des réponses vraies et prédites, et $|y_i \cup \hat{y}_i|$ est la taille de l'union des bonnes réponses et des réponses prédites.

Système *Baseline* Le problème de réponse automatique à une question peut être considéré dans notre cas comme étant un problème de classification multi-étiquettes. Nous avons donc proposé aux participants un système *baseline* s'appuyant sur l'affinage (*fine-tuning*) du modèle CamemBERT 138 GB OSCAR (Martin *et al.*, 2020), un modèle de langue générique pré-entraîné pour le français et fondé sur l'architecture RoBERTa (Liu *et al.*, 2019). Notons que la séquence d'entrée du modèle est composée de la question suivie des cinq réponses possibles, toutes séparées avec un token [SEP], selon le format suivant : [CLS] <question> [SEP] (A) <answer.a> [SEP] (B) <answer.b> [SEP] (C) <answer.c> [SEP] (D) <answer.d> [SEP] (E) <answer.e> [SEP] [EOS]. Pour ce qui est de la sortie, nous avons une couche de classification de dimension 5 suivie d'une SIGMOID et représentant, à partir d'un certain palier, soit l'absence ou la présence d'une classe, ici les lettres des réponses de A à E.

2.1.2 Tâche annexe : Nombre de réponses correctes dans un QCM

Présentation La tâche annexe à la campagne d'évaluation consiste à identifier automatiquement le nombre de réponses correctes, pour chaque question, parmi l'ensemble des réponses potentielles dans un QCM.

Évaluation Contrairement à la tâche principale, qui se retrouve très proche d'un problème multi-étiquettes, il s'agit ici de retrouver la valeur exacte correspondant au nombre de réponses correctes, cette valeur étant comprise entre 1 et 5 inclus. La tâche peut alors être vue comme un problème multi-classes (ici, 5 classes) : nous avons donc choisi d'évaluer les systèmes en termes de taux correct de classification (Accuracy) et de macro F-mesure (F1-Macro).

Système *Baseline* Le système *baseline* permet aux participants de résoudre cette tâche comme un problème de classification multi-classes, fondé, à l'instar de la tâche principale, sur l'affinage du modèle CamemBERT 138 GB OSCAR. La séquence d'entrée du modèle est exactement la même que pour la tâche principale (voir Section 2.1.1), mais ici, le modèle doit ne fournir, en sortie, qu'une seule et unique classe parmi les cinq qui sont possibles et qui représente le nombre de réponses correctes. Ici aussi, une fonction SIGMOID est appliquée sur le vecteur de sortie, mais nous choisissons par défaut la classe donnant le score le plus élevé.

2.2 Pistes développées

Pour cette campagne d'évaluation, nous avons proposé d'ouvrir deux pistes aux équipes : *Recherche reproductible* et *Aucune restriction*. Chacune des deux tâches, présentées précédemment dans la Section 2.1, peuvent s'inscrire dans ces deux pistes. Dans la piste *Recherche reproductible*, seuls les systèmes qui respectent les deux conditions suivantes sont acceptés : 1) ne pas rechercher sur Internet les originaux des données fournies, et 2) utiliser des modèles pré-entraînés dont les données d'entraînement sont connues. Pour la piste *Aucune restriction*, tous les systèmes sont acceptés sans limite de recherche, que ce soit au niveau des données collectées ou des modèles pré-entraînés utilisés.

Le classement des équipes se fait uniquement sur les sorties des systèmes déposés dans la piste *Recherche reproductible*, où un système doit obligatoirement être déposé. Dans cette piste, les participants sont autorisés à soumettre jusqu'à trois sorties de système (*run*) par tâche. En revanche, la participation à la piste *Aucune restriction* n'étant pas obligatoire, et difficile à contrôler, le classement des équipes dans la campagne d'évaluation n'intégrera pas les systèmes proposés dans cette piste. Notons également que le nombre de soumissions autorisées n'y a pas été limité.

3 Résultats

Dans la piste *Recherche reproductible*, six équipes ont participé à la tâche principale et trois équipes à la tâche annexe. Dans le cadre de la piste *Aucune restriction*, deux équipes ont participé aux deux tâches proposées.

Dans les sections suivantes, nous présentons, pour chacune des deux tâches, tout d'abord les résultats officiels de la campagne DEFT 2023, intégrant une description succincte des systèmes proposés par chaque équipe, correspondant à la piste *Recherche reproductible* (voir Section 3.1), alors que les systèmes correspondant à la piste *Aucune restriction* sont décrits dans la Section 3.2.

3.1 Piste : Recherche reproductible

Les résultats décrits dans cette partie, pour la tâche principale (Section 3.1.1) et la tâche annexe (Section 3.1.2), constituent les performances officielles des systèmes proposés par les participants à la campagne d'évaluation DEFT 2023.

3.1.1 Tâche principale

Les six participants ont chacun soumis trois fichiers de prédictions. Le Tableau 2 présente les résultats en termes de distance de Hamming et EMR obtenus par chaque équipe pour chaque fichier de prédiction (*Run*). Le classement de chaque équipe selon la métrique ciblée est fourni. Notons que nous avons également intégré dans ce tableau les résultats de notre méthode *baseline* (Section 2.1.1) et ceux de la *classe majoritaire*.

Méthodes des participants Les participants ont utilisé des méthodes variées pour cette tâche principale. Nous observons cependant un intérêt commun pour les grands modèles de langue (*Large*

Équipe	Run	Hamming	Classement	EMR	Classement
LIS	1	52,94	1	33,76	1
	2	47,43	-	27,81	-
	3	35,93	-	17,85	-
LIUM-IRISA	1	43,24	2	22,19	3
	2	37,24	-	18,65	-
	3	35,47	-	18,49	-
TTGV	1	41,54	3	23,95	2
	2	39,15	-	11,58	-
	3	37,22	-	15,43	-
SEQUOIA	1	26,34	-	14,63	-
	2	35,90	-	15,27	4
	3	37,93	4	12,70	-
ALMANACH-ARKHN	1	33,27	-	12,22	-
	2	33,67	-	14,15	5
	3	35,96	5	13,67	-
SPQR	1	24,93	6	8,52	-
	2	22,38	-	9,32	-
	3	23,94	-	9,97	6
Baseline	-	36,24	-	16,55	-
Classe majoritaire	-	23,93	-	13,67	-

TABLE 2 – Résultats et classement des équipes participantes pour la tâche principale dans la piste *Recherche reproductible*.

Language Models - LLMs) de la part de la majorité des équipes. L'équipe en tête du classement (LIS) a utilisé pour ses soumissions le grand modèle de langue LLaMa (Touvron *et al.*, 2023) affiné sur les données d'entraînement de DEFT 2023 en utilisant la méthode d'adaptation LoRA (Hu *et al.*, 2021) (méthode d'adaptation à faible rang, *Low Rank Adaptation*) permettant de réduire le coût machine d'un tel affinage et de le rendre réalisable sur le matériel utilisé par l'équipe (NVIDIA A100 80 GB). Les différentes soumissions de l'équipe correspondent aux variantes de différentes tailles de ce modèle : 13 milliards (13B), 30 milliards (30B) et 65 milliards (65B) de paramètres. D'autres LLMs ont été explorés par les équipes participantes, tels que BloomZ (Muennighoff *et al.*, 2022) (SEQUOIA, TTGV, LIS), Flan-T5 (Chung *et al.*, 2022) (LIUM-IRISA, ALMANACH-ARKHN, LIS) et Vicuna 13B (Chiang *et al.*, 2023) (TTGV).

Des modèles de langue pré-entraînés pour le français et fondés sur RoBERTa ont aussi été utilisés, en particulier des modèles spécialisés dans le domaine médical, tels que DrBERT (Labrak *et al.*, 2023) (LIUM-IRISA, TTGV) et Camembert-BIO (Touchent *et al.*, 2023) (ALMANACH-ARKHN). À noter que ces derniers restent compétitifs, puisque les équipes LIUM-IRISA et TTGV obtiennent alternativement la deuxième et troisième place du classement selon la métrique étudiée.

Certaines équipes ont aussi exploré des approches différentes de celles maintenant classiques liées à l'affinage de modèles de langue. L'équipe SPQR s'est par exemple appuyée sur l'utilisation d'un corpus externe pour *vérifier* les réponses données dans les QCMs. Après plusieurs traitements des données, dont l'extraction de mots-clés médicaux, leurs systèmes intégraient différentes mesures de similarité afin de rapprocher les données des QCMs et des ressources externes biomédicales. L'équipe LIUM-IRISA a, quant à elle, proposé un système de fouille dans une base de connaissances, avec des approches TF.IDF et données Wikipédia, voire un méta-système intégrant plusieurs sources de connaissances (Flan-T5, DrBERT, etc.). Enfin, l'équipe TTGV a exploré un très grand nombre d'approches différentes, parmi lesquelles nous pouvons citer des méthodes par expressions régulières, modélisation par topics, régression logistique, approches multi-classes et multi-étiquettes, etc.

Enfin, nous observons que le modèle *baseline* fourni aux équipes participantes, se placerait à la cinquième position considérant la distance de Hamming, et à la quatrième position selon la métrique EMR.

3.1.2 Tâche annexe

Sur la tâche annexe, trois équipes ont proposé des systèmes originaux. Le Tableau 3 présente les résultats en termes de taux correct de classification (*Accuracy*) et F1-Macro, ainsi que le classement associé pour ces métriques. À l'instar de la tâche principale, nous intégrons dans ce tableau les résultats de notre *baseline* et ceux de la classe majoritaire.

Équipe	Run	F1-Macro	Classement	Accuracy	Classement
LIS	1	42,42	1	68,65	1
	2	35,26	-	65,92	-
	3	34,52	-	65,11	-
TTGV	1	27,98	-	62,54	2
	2	13,26	-	19,13	-
	3	31,51	2	60,45	-
SPQR	1	22,99	3	43,57	-
	2	15,29	-	47,43	3
	3	21,05	-	46,78	-
Baseline	-	28,79	-	67,04	-
Classe majoritaire	-	13,62	-	51,61	-

TABLE 3 – Résultats et classement des équipes participantes pour la tâche annexe dans la piste *Recherche reproductible*.

Méthode des participants Ainsi, l'équipe tête du classement (LIS) a dérivé ses sorties du système LLaMA de la tâche principale pour la tâche annexe.

L'équipe TTGV a testé plusieurs méthodes : celle ayant obtenu les meilleurs résultats a consisté à utiliser un modèle de langue pré-entraîné dans le domaine médical pour le français (DrBERT) affiné sur les données d'apprentissage de DEFT 2023 afin de résoudre le problème sous la forme de classification multi-classes.

Enfin, l'équipe SPQR s'est appuyée sur les résultats obtenus dans la tâche principale avec leur approche par similarité pour fournir le nombre de réponses correctes par question.

Nous observons qu'en termes de F1-macro, seule l'approche intégrant des LLMs surpasse largement les résultats de la *baseline* fournie aux participants. Notons cependant que cette observation n'est pas aussi franche en termes d'accuracy, où finalement le modèle CamemBERT affiné reste compétitif face à ces grands modèles.

3.2 Piste : Aucune restriction

La piste *Aucune restriction* n'était pas une piste obligatoire, celle-ci laissant la possibilité aux équipes participantes d'évaluer n'importe quelle approche. Contrairement à la piste *Recherche reproductible*, aucun classement n'a été fait dans cette piste, les résultats n'étant reportés qu'à titre informatif. Deux

équipes ont choisi de participer à la piste *Aucune restriction* dans les deux tâches (principale et annexe). Le Tableau 4 présente les résultats obtenus par chacune des équipes.

Équipe	Run	Tâche principal		Tâche annexe	
		Hamming	EMR	F1-Macro	Accuracy
LIS	1 - GPT-3.5-turbo	64,75	46,95	47,51	68,17
	2 - GPT-4	85,17	72,83	71,57	79,58
SEQUOIA	1 - GPT-3.5-turbo	64,40	46,46	44,36	65,92

TABLE 4 – Résultats et classement des équipes participantes pour la tâche principale dans la piste *Aucune restriction*.

Méthode des participants L'agent conversationnel privé ChatGPT (GPT-3.5-Turbo) [4] de l'entreprise OpenAI, a été utilisé par les deux équipes LIS et SEQUOIA. Les *Run 1* des deux équipes utilisent le modèle GPT-3.5 accessible par l'API de la firme, ce qui explique les résultats très proches (des petites différences de résultats pourraient être imputées à la manière d'interagir avec le modèle). Le *Run 2* de l'équipe LIS utilise, quant à lui, le modèle GPT-4 [5]. Il est intéressant de voir que les prédictions fondées sur le modèle GPT-4 obtiennent de meilleurs résultats que le système le plus performant de la piste *Recherche reproductible*. Toutefois, étant donné que nous ne disposons pas d'informations concernant la présence des données d'entraînement dans le corpus d'apprentissage, nous pouvons difficilement tirer des conclusions quant à cette performance.

4 Conclusion

L'édition 2023 du DÉfi Fouille de Textes (DEFT) s'est concentrée sur le développement de méthodes permettant de choisir les réponses dans des questions à choix multiples (QCMs) en français.

La première tâche a rassemblé six équipes et consistait à sélectionner automatiquement le sous-ensemble de réponses correctes parmi celles proposées pour une question donnée. Les équipes participantes avaient à leur disposition des données d'entraînement et de développement fournies avec le corpus FrenchMedMCQA, et pouvaient également s'aider de connaissances externes. Les résultats obtenus sur les données de test de ce corpus ont été fournis selon la métrique de l'Exact Match Ratio, variant de 9,97 % à 33,76 %, alors que les performances en termes de distance de Hamming s'échelonnaient de 24,93 à 52,94. Notre système *baseline* a obtenu 16,55 % et 36,24 respectivement avec l'EMR et la distance de Hamming. L'utilisation de grands modèles de langue comme LLaMa affiné sur les données mises à disposition en utilisant la méthode d'adaptation LoRA s'est révélée la plus efficace.

Quant à la deuxième tâche, elle a rassemblé trois équipes et consistait à identifier, pour chaque question, le nombre exact de réponses correctes. Les résultats ont été fournis en utilisant la métrique de F1-Macro, variant de 13,26 % à 42,42 %, ainsi que la métrique *Accuracy*, allant de 47,43 % à 68,65 %. Notre système *baseline* a obtenu une F1-Macro de 28,79 % et une *Accuracy* de 67,04 %.

Cette nouvelle édition de DEFT se termine avec une grande variété de méthodes testées sur chacune des tâches proposées, et montre que l'utilisation de grands modèles de langue s'avère très efficace,

4. https://openai.com/product/chatgpt
5. https://openai.com/waitlist/gpt-4-api

alors même que certains de ces modèles ne sont pas adaptés au domaine traité (ici, le domaine médical).

Remerciements

Le comité d'organisation de DEFT 2023 tient à remercier chaleureusement l'ensemble des équipes (ALMANACH-ARKHN, LIS, LIUM-IRISA, SEQUOIA, SPQR, TTGV) pour l'engagement et la qualité des systèmes proposés durant cette campagne d'évaluation. Le comité d'organisation tient également à remercier Cyril Grouin, pour son aide précieuse à la mise en place de l'atelier, ainsi que le comité scientifique de DEFT 2023 (Nathalie Camelin, Liana Ermakova, Benoit Favre, Corinne Fredouille, Pierre-Antoine Gourraud, Natalia Grabar, Cyril Grouin, Pierre Jourlin, Fleur Mougin, Aurélie Névéol, Didier Schwab et Pierre Zweigenbaum).

Références

BESNARD C., ETTALEB M., RAYMOND C. & CAMELIN N. (2023). Qui de DrBERT, Wikipédia ou Flan-T5 s'y connaît le plus en questions médicales ? In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

BEZANÇON J., BOUBEHZIZ, TOUFIK ANX CHUTAUX C., ZINE O., ACENSIO L., BRIGLIA A., KOUDORO-PARFAIT C. & LEJEUNE G. (2023). SPQR@Deft2013 : Similarité Sorbonne Pour les Systèmes de Question Réponse. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

BLIVET A., DEGRUTÈRE S., GENDRON B., RENAULT A., SIOUFFI C., GAUDRAY-BOUJU V., CERISARA C., FLAMEIN H., GUIBON G., LABEAU M. & ROUSSEAU T. (2023). Participation de l'équipe TTGV à DEFT 2023 : Réponse automatique à des QCM issus d'examens en pharmacie. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

BOTHUA M., HASSANI L., JUBAULT M. & SUIGNARD P. (2023). Participation d'EDF R&D au défi DEFT 2023 : réponses automatiques à des questionnaires à choix multiples à l'aide de « Larges Modèles de Langue ». In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

CHIANG W.-L., LI Z., LIN Z., SHENG Y., WU Z., ZHANG H., ZHENG L., ZHUANG S., ZHUANG Y., GONZALEZ J. E., STOICA I. & XING E. P. (2023). Vicuna : An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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

FAVRE B. (2023). LIS@DEFT'23 : les LLMs peuvent-ils répondre à des QCM ? (a) oui ; (b) non ; (c) je ne sais pas. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2021). LoRa : Low-rank adaptation of large language models. *arXiv preprint arXiv :2106.09685*.

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. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper*, Toronto, Canada : Association for Computational Linguistics.

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*.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MEONI S., TOUCHENT R. & DE LA CLERGERIE (2023). Passe ta pharma d'abord ! In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

MUENNIGHOFF N., WANG T., SUTAWIKA L., ROBERTS A., BIDERMAN S., SCAO T. L., BARI M. S., SHEN S., YONG Z.-X., SCHOELKOPF H. *et al.* (2022). Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv :2211.01786*.

PAL A., UMAPATHI L. K. & SANKARASUBBU M. (2022). MedMCQA : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, p. 248–260 : PMLR.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. V. (2023). CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles*.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F. *et al.* (2023). Llama : Open and efficient foundation language models. *arXiv preprint arXiv :2302.13971*.

WELBL J., LIU N. F. & GARDNER M. (2017). Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 94–106, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4413.