Tâches et systèmes de sélection automatique de réponses à des QCM dans le domaine médical : Présentation de la campagne DEFT 2024

Adrien Bazoge, Labrak Yanis, Richard Dufour, Benoît Favre, Mickaël Rouvier

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

Tâches et systèmes de sélection automatique de réponses à des QCM dans le domaine médical : Présentation de la campagne DEFT 2024

Adrien Bazoge[1, 4] Yanis Labrak[2, 5]
Richard Dufour[1, 2] Benoit Favre[3] Mickaël Rouvier[2]
(1) Laboratoire des Sciences du Numérique de Nantes (LS2N), Nantes Université, France
(2) Laboratoire Informatique d'Avignon (LIA), Avignon Université, France
(3) Aix Marseille Université, CNRS, LIS, Marseille, France
(4) Clinique des données, CHU de Nantes, Nantes, France
(5) Zenidoc, Marseille, France
prenom.nom@univ-avignon.fr, prenom.nom@univ-nantes.fr,
benoit.favre@lis-lab.fr,

RÉSUMÉ

L'édition 2024 du DÉfi Fouille de Textes (DEFT) met l'accent sur le développement de méthodes pour la sélection automatique de réponses pour des questions à choix multiples (QCM) en français. Les méthodes sont évaluées sur un nouveau sous-ensemble du corpus FrenchMedMCQA, comprenant 3 105 questions fermées avec cinq options chacune, provenant des archives d'examens de pharmacie. Dans la première tâche, les participants doivent se concentrer sur des petits modèles de langue (PML) avec moins de 3 milliards de paramètres et peuvent également utiliser les corpus spécifiques au domaine médical NACHOS et Wikipedia s'ils souhaitent appliquer des approches du type Retrieval-Augmented Generation (RAG). La second tâche lève la restriction sur la taille des modèles de langue. Les résultats, mesurés par l'Exact Match Ratio (EMR), varient de 1,68% à 11,74%, tandis que les performances selon le score de Hamming vont de 28,75% à 49,15% pour la première tâche. Parmi les approches proposées par les cinq équipes participantes, le meilleur système utilise une chaîne combinant un classifieur CamemBERT-bio pour identifier le type de question et un système RAG fondé sur Apollo 2B, affiné avec la méthode d'adaptation LoRA sur les données de l'année précédente.

ABSTRACT

Tasks and automatic response selection systems for MCQA in the medical domain : Presentation of the DEFT 2024 campaign.

The 2024 edition of the text mining challenge DÉfi Fouille de Textes (DEFT) is focused on developing methods for automatically selecting answers for multiple-choice questions (MCQs) in French. The methods are evaluated on a new subset of the FrenchMedMCQA corpus, which includes 3,105 closed questions with five options each, sourced from French pharmacy exam archives. The first task introduced by this edition limits participants to using small language models (SLMs) with fewer than 3 billion parameters. The second task removes this limit. Particpants can also use the NACHOS medical domain-specific corpus and Wikipedia if they wish to apply Retrieval-Augmented Generation (RAG) approaches. The results, measured by the Exact Match Ratio (EMR) metric, range from 1.68% to 11.74%, while the Hamming score performances range from 28.75% to 49.15% for the first task. Among the various approaches proposed by the five participating teams, the best system

Actes du Défi Fouille de Textes@TALN 2024, pages 1–10.

utilizes a pipeline combining a CamemBERT-bio classifier for identifying question type and an Apollo 2B-based RAG system, fine-tuned with the LoRA adaptation method on previous year data.

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

KEYWORDS: Multiple-choice question answering ; Medical domain ; Large Language Models ; SLM ; RAG ; TALN.

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.

L'édition 2024 du défi [1] est dans la continuité de l'édition 2023. Elle porte sur la mise en place d'approches permettant de répondre automatiquement à des questionnaires à choix multiples issus d'annales d'examens de pharmacie. Une des difficultés et originalités du corpus réside dans le fait que chaque question contient une inconnue sur le nombre de réponses associées, là où d'autres corpus (Jin *et al.*, 2021; Hendrycks *et al.*, 2021; Pal *et al.*, 2022a) attendent une seule réponse par question. Cette difficulté a permis aux équipes participantes d'explorer et de proposer de nouvelles approches pouvant s'écarter de celles actuellement proposées pour des tâches plus classiques en TALN.

Les données d'évaluation proviennent du corpus FrenchMedMCQA (Labrak *et al.*, 2022) qui se compose de questions fermées en français issues d'annales d'examens de pharmacie en français. Pour l'édition 2024, un nouveau corpus d'évaluation a été collecté à partir d'une source de données différente et reprend le format du corpus FrenchMedMCQA. Nous ferons référence à ce nouveau corpus d'évaluation sous la dénomination ($DEFT_{2024}$). Le défi propose deux tâches aux participants :

1. **Tâche principale** : identifier automatiquement l'ensemble des réponses correctes parmi les cinq options possibles pour une question donnée. Les systèmes proposés pour cette tâche devront faire moins de 3 milliards de paramètres, tout en laissant la possibilité de puiser des informations à partir des bases de connaissance NACHOS (Labrak *et al.*, 2023) et Wikipedia, qui ont été mises à disposition des participants, afin par exemple de permettre d'appliquer des approches du type Retrieval-Augmented Generation (RAG).
2. **Tâche annexe** : cette tâche est identique à la tâche principale, à savoir identifier automatiquement l'ensemble des réponses correctes parmi les cinq options proposées pour une question donnée, mais sans aucune limite imposée sur la taille des modèles.

La campagne, lancée le 11 mars 2024, a permis l'accès aux données d'entraînement après la signature d'un accord par tous les membres des équipes participantes. La phase d'entraînement s'est déroulée sur presque deux mois, du 11 mars 2024 au 27 mai 2024, suivie de la phase de test du 27 mai au 31 mai 2024. Cinq équipes se sont inscrites et ont mené la campagne à son terme :

— *LIMICS* (Delourne *et al.*, 2024) : Équipe du Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS) de l'Université Sorbonne Paris-Nord.

1. https://deft2024.univ-avignon.fr

— *LIUM-CREN* (Okat *et al.*, 2024) : Équipe jointe entre les équipes du CREN et du LIUM de Le Mans Université, et le Centre Hospitalier du Mans.
— *CRIM* (Moubtahij *et al.*, 2024) : Équipe du Centre de Recherche Informatique de Montréal (CRIM).
— *ExBERT* (Charlot *et al.*, 2024) : Équipe du Laboratoire des sciences du numérique de Nantes (LS2N) et de la Faculté des Sciences (Master ATAL) de Nantes Université.
— *SPQRR* (Lejeune, 2024) : Équipe jointe entre l'équipe STIH du CERES de Sorbonne Université, le LIG de l'Université de Grenoble Alpes, le LORIA de l'Université de Lorraine et l'équipe DTIPG de la SNCF.

2 Corpus

Lors de ce défi, nous avons mis à disposition le corpus FrenchMedMCQA (Labrak *et al.*, 2022), ainsi qu'un nouveau corpus collecté spécialement pour DEFT 2024. Ces deux corpus de QCM portent sur le domaine médical et sont constitués à partir d'annales d'examens réels de pharmacie en français. Ils sont similaires à ceux que l'on retrouve dans d'autres langues, telles que l'anglais, avec les corpus MedMCQA (Pal *et al.*, 2022b) et SciQ (Welbl *et al.*, 2017). Chaque QCM contient cinq réponses potentielles, parmi lesquelles se trouvent une ou plusieurs réponses correctes, réalisées manuellement par des experts médicaux et utilisées lors d'examens de pharmacie.

Le corpus de l'édition 2023 de DEFT, FrenchMedMCQA, contient un ensemble de 3 105 QCMs. Il a été constitué en collectant des questions et leurs réponses associées à partir du site *Remede.org* [2]. Pour cette nouvelle édition, un nouveau corpus d'évaluation ($DEFT_{2024}$), contenant un ensemble de 477 QCMs, a été collecté. Ce corpus a été constitué à partir de questions et de leurs réponses associées provenant d'annales d'examens réels de pharmacie en français obtenues sur le site MedShake [3]. Le corpus $DEFT_{2024}$ suit les mêmes contraintes que le corpus FrenchMedMCQA, à savoir cinq réponses potentielles par question, et une ou plusieurs réponses correctes possibles.

Le Tableau 1 ci-dessous fournit la distribution du jeu de données FrenchMedMCQA selon son découpage pour l'apprentissage, le développement ainsi que du corpus d'évaluation collecté spécialement pour cette nouvelle édition. Le corpus de test ($DEFT_{2024}$) est constitué de 477 nouvelles questions. Le nombre de questions avec une réponse unique passe de 52% à 19%, permettant de tester la généralité des systèmes à ce changement de distribution.

# Réponses	FrenchMedMCQA			$DEFT_{2024}$	Total
	Apprentissage	Développement	Test	Test	
1	595	164	321	93	1 173
2	528	45	97	146	816
3	718	71	141	160	1 090
4	296	30	56	73	455
5	34	2	7	5	48
Total	2 171	312	622	477	3 582

TABLE 1 – Distribution du corpus FrenchMedMCQA selon son découpage en apprentissage, développement et test, ainsi que du corpus de test pour DEFT 2024.

2. http://www.remede.org/internat/pharmacie/qcm-internat.html
3. https://medshake.net/

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 du corpus FrenchMedMCQA 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 dans une question) et chaque réponse en contient 2 en moyenne (représentant 36 % des mots dans une réponse). Enfin, toujours en moyenne, un mot spécifique au domaine médical apparaît dans 2 questions et dans 8 réponses.

Pour le corpus d'évaluation $DEFT_{2024}$, ces métriques sont similaires au corpus FrenchMedMCQA. La longueur moyenne des questions est de 15,79 mots et la longueur moyenne des réponses est de 6,90 mots. Le vocabulaire de ce jeu d'évaluation compte 7 531 mots, dont 1 856 sont spécifiques au domaine médical (soit environ 24,64%). En moyenne, chaque question contient 2,57 mots spécifiques au domaine médical (représentant 15,28% des mots dans une question) et chaque réponse en contient 2,13 (représentant 35,05% des mots dans une réponse). Enfin, toujours en moyenne, un mot spécifique au domaine médical apparaît dans 0,64 question et 2,71 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.

Pour l'ensemble des tâches et pistes, les participants ont eu à leur disposition les données d'entraînement, de développement et de test de l'année précédente, comme décrit dans la Section 2. Les annotations du nouveau corpus de test ($DEFT_{2024}$), sur lequel toutes les équipes ont été évaluées, n'ont jamais été fournies durant la campagne d'évaluation. Cependant, elles ont été rendues disponibles librement à la fin de la campagne, permettant ainsi aux équipes d'évaluer les systèmes non soumis.

Enfin, nous avons fourni à l'ensemble des équipes un système état-de-l'art pour chacune des deux tâches. Ces premiers résultats permettaient aux équipes d'avoir un repère quant aux performances de leurs approches.

3 Tâches et métriques

3.1 Tâches

Lors de ce défi, deux tâches liées aux QCMs médicaux ont été proposées. L'objectif pour les deux tâches a consisté 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 les correctes. Pour l'ensemble des tâches, les systèmes des participants ne devaient pas rechercher sur internet les originaux des données fournies et devaient utiliser des modèles pré-entraînés dont les données d'entraînement sont connues (i.e. ChatGPT, Mistral et autres modèles de ce type ne pouvaient pas être utilisés).

3.2 Métriques

Contrairement à une tâche de classification classique où il est demandé d'associer une étiquette à un problème donné, la tâche de QCM peut impliquer une réponse partiellement correcte. Par exemple, si l'on doit retrouver deux réponses correctes parmi les cinq options disponibles pour une question ciblée, mais qu'un système automatique n'est capable que d'en retrouver une seule, alors la réponse est 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*) entre les sorties et la référence.

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

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.

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

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.

4 Résultats

Cinq équipes ont participé à la *tâche principale*, tandis qu'aucune n'a pris part à la *t*âche annexe. Dans les sections suivantes, nous présentons tous les résultats officiels de la campagne DEFT 2024, y

compris de brèves descriptions des systèmes proposés par chaque équipe ainsi que plusieurs systèmes de référence évalués par les organisateurs de l'événement.

4.1 Tâche principale

Les résultats décrits dans cette partie, constituent les performances officielles des systèmes proposés par les participants à la campagne d'évaluation DEFT 2024.

Les participants pouvaient chacun soumettre 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 des équipes est calculé en fonction de la métrique EMR. Notons que nous avons également intégré dans ce tableau les résultats de nos méthodes *baseline*, à savoir, deux affinages des modèles Apollo 0.5B et 2B sur les données d'entraînement de l'année précédente et en scénario d'inférence *zero-shot*.

Équipe	Run	EMR	Hamming	Classement
LIMICS	1 - APOLLIMICS	11.74	45.71	1
LIUM-CREN	1 - FT5L_DRAGON	10.69	47.97	-
	2 - FT5L_CARAGON	11.53	49.15	2
	3 - FT5_ROBOT	8.39	31.30	-
CRIM	1 - stablelm_16shot	2.51	29.68	-
	2 - stablelm_11shot_auto-judge	10.27	43.05	3
	3 - stablelm_11shot_bestof2	9.22	40.28	-
ExBERT	1 - ExBERT	2.73	45.85	-
	2 - REFT	4.40	30.77	4
SPQRR	1 - TTGV_Tom	2.94	38.07	-
	2 - TTGV_byfusion	4.19	26.97	5
	3 - TTGV_ollama_multilabel	1.68	28.75	-
Baseline	Apollo 0.5B + SFT + Zero-shot	2.73	35.11	-
	Apollo 2B + SFT + Zero-shot	1.05	30.41	-

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

Méthodes des participants Les participants ont utilisé des méthodes variées pour cette tâche principale. La majorité des équipes ont utilisé des grands modèles de langue (*Large Language Models* - LLMs), tout en respectant la limite à 3 milliards de paramètres imposée dans la tâche. Les trois premières équipes du classement (LIMICS, LIUM-CREN, CRIM) ont compensé cette limite de taille en s'aidant de connaissances externes avec des méthodes de génération augmentée par les résultats d'un moteur de recherche (*Retrieval-Augmented Generation* - RAG). L'équipe en tête du classement (LIMICS) a utilisé une chaîne combinant un classifieur CamemBERT-bio pour identifier le type de question, un système pour reformuler la question en binaire et un système RAG basé sur Apollo 2B (Wang *et al.*, 2024), affiné avec la méthode d'adaptation LoRA (Hu *et al.*, 2021) sur les données de DEFT 2023 pour répondre à la question reformulée. D'autres LLMs ont été explorés avec la méthode RAG par les autres participants, tels que Flan-T5 (Chung *et al.*, 2022) (LIUM-CREN, CRIM), Bloomz-3B (Muennighoff *et al.*, 2023) (CRIM) et stableLM-3b-4e1t (Wei *et al.*, 2023) (CRIM).

De même, les équipes ont employé différentes méthodes pour extraire du contexte dans les connais-

sances externes autorisées NACHOS et Wikipedia pour leur méthode de RAG, tels que BM25 (Trotman *et al.*, 2014) (CRIM), Sentence-CamemBERT-bio (Delourne *et al.*, 2024) (LIMICS), DrBERT (Labrak *et al.*, 2023) (LIUM-CREN), CamemBERT (Martin *et al.*, 2020) (LIUM-CREN) et TF-IDF (ExBERT).

Certaines équipes ont exploré des approches alternatives, avec ou sans l'utilisation de LLMs. L'équipe ExBERT enrichit les entrées d'un modèle de type BERT avec la méthode RAG décrite précédemment. Les représentations en découlant sont exploitées par un réseau de neurones prédisant un vecteur binaire correspondant aux 5 réponses possibles. Deux variantes ont été évaluées : un affinage ReFT (Wu *et al.*, 2024) de DrBERT et une combinaison *branch-train-mix* (Sukhbaatar *et al.*, 2024) de deux modèles : DrBERT et CamemBERT-bio (Touchent & de la Clergerie, 2024). L'équipe SQPRR, quant à elle, s'appuie sur une reformulation des QCM sous la forme d'assertions notamment au moyen d'expressions régulières et de règles définies manuellement. À partir de ces données transformées, trois approches ont été évaluées : 1) la vérification des assertions à partir d'un corpus additionnel par comparaison des plongements textuels par réseaux de neurones ; 2) similarité des n-grammes de caractères avec un corpus additionnel ; et 3) OpenLlama v2 (Geng & Liu, 2023) adapté sous la forme d'une classification multi-étiquette sans ajout de données externes.

Enfin, nous observons que tous les participants ont obtenu de meilleures performances que les modèles *baseline*, que ce soit sur la distance de Hamming ou la métrique EMR.

4.2 Tâche annexe

Sur la tâche annexe, aucune des équipes participantes n'a proposé de systèmes.

5 Conclusion

L'édition 2024 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 cinq 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 le corpus FrenchMedMCQA de DEFT 2023, et pouvaient également s'aider de connaissances externes avec les corpus NACHOS et Wikipedia. Les résultats obtenus sur les données de test de DEFT 2024 ont été fournis selon la métrique EMR, variant de 1,68% à 11,74%, alors que les performances en termes de distance de Hamming s'échelonnaient de 28,75% à 49,15%. Notre système *baseline* a obtenu 2,73% et 35,11% respectivement avec l'EMR et la distance de Hamming. L'utilisation de grands modèles de langue associés à des méthodes de RAG sur les corpus externes fournis s'est révélée la plus efficace.

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, 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 2024 tient à remercier chaleureusement l'ensemble des équipes (CRIM, ExBERT, LIMICS, LIUM-CREN, SPQR) pour l'engagement et la qualité des systèmes proposés durant cette campagne d'évaluation. Le comité d'organisation tient également à remercier le comité scientifique de DEFT 2024 (Nathalie Camelin, Corinne Fredouille, Pierre-Antoine Gourraud, Natalia Grabar, Cyril Grouin, Pierre Jourlin, Solen Quiniou, Didier Schwab et Pierre Zweigenbaum).

L'organisation de cette campagne d'évaluation a pu être possible grâce au soutien de l'Agence Nationale de la Recherche (ANR) qui finance le projet ANR MALADES (ANR-23-IAS1-0005) ainsi que de l'entreprise Zenidoc.

Références

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

CHARLOT T., SISARITH E., STUCKY N., ILANGO R., GOUGET N., SEWRAJ H. & PILLET X. (2024). DÉfi fouille de texte 2024. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI Y., WANG X., DEHGHANI M., BRAHMA S., WEBSON A., GU S. S., DAI Z., SUZGUN M., CHEN X., CHOWDHERY A., CASTRO-ROS A., PELLAT M., ROBINSON K., VALTER D., 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.

DELOURNE S., REMAKI A., GÉRARDIN C., VAILLANT P., TANNIER X., SEROUSSI B. & REDJDAL A. (2024). Limics@deft'24 : Un mini-llm peut-il tricher aux qcm de pharmacie en fouillant dans wikipedia et nachos ? In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

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

GENG X. & LIU H. (2023). Openllama : An open reproduction of llama.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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.

JIN D., PAN E., OUFATTOLE N., WENG W.-H., FANG H. & SZOLOVITS P. (2021). What disease does this patient have ? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, **11**(14). DOI : 10.3390/app11146421.

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 A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16207–16221, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

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

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

LEJEUNE G. (2024). Spqr@deft2024. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language mode. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MOUBTAHIJ A., CUMMINGS C.-W., HANDAN A., GALY E. & CHARTON E. (2024). Participation du crim à deft 2024 : Utilisation de petits modèles de langue pour des qcms dans le domaine médical. 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., LE SCAO T., BARI M. S., SHEN S., YONG Z. X., SCHOELKOPF H., TANG X., RADEV D., AJI A. F., ALMUBARAK K., ALBANIE S., ALYAFEAI Z., WEBSON A., RAFF E. & RAFFEL C. (2023). Crosslingual generalization through multitask finetuning. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15991–16111, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.891.

OKAT E., BROCHELARD H., SINI A., RENAULT V. & CAMELIN N. (2024). Flan-t5 avec ou sans contexte, telle est la question à choix multiples. In *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*.

PAL A., UMAPATHI L. K. & SANKARASUBBU M. (2022a). Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In G. FLORES, G. H. CHEN, T. POLLARD, J. C. HO & T. NAUMANN, Éds., *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 de *Proceedings of Machine Learning Research*, p. 248–260 : PMLR.

PAL A., UMAPATHI L. K. & SANKARASUBBU M. (2022b). 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.

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

SUKHBAATAR S., GOLOVNEVA O., SHARMA V., XU H., LIN X. V., ROZIÈRE B., KAHN J., LI D., TAU YIH W., WESTON J. & LI X. (2024). Branch-train-mix : Mixing expert llms into a mixture-of-experts llm.

TOUCHENT R. & DE LA CLERGERIE É. (2024). CamemBERT-bio : Leveraging continual pre-training for cost-effective models on French biomedical data. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 2692–2701, Torino, Italia : ELRA and ICCL.

TROTMAN A., PUURULA A. & BURGESS B. (2014). Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, p. 58–65, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2682862.2682863.

WANG X., CHEN N., CHEN J., HU Y., WANG Y., WU X., GAO A., WAN X., LI H. & WANG B. (2024). Apollo : Lightweight multilingual medical llms towards democratizing medical ai to 6b people.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. & ZHOU D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

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.

WU Z., ARORA A., WANG Z., GEIGER A., JURAFSKY D., MANNING C. D. & POTTS C. (2024). Reft : Representation finetuning for language models.