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

Meryl Bothua, Leila Hassani, Marie Jubault, Philippe Suignard *
EDF R&D
7, boulevard Gaspard Monge 91120 Palaiseau
prenom.nom@edf.fr

RÉSUMÉ

Ce papier présente la participation d'EDF R&D à la campagne d'évaluation DEFT 2023. Notre équipe a participé à la tâche de réponse automatique à des questions à choix multiples issus d'annales d'examens en pharmacie en français. Le corpus utilisé est FrenchMedMCQA. Nous avons testé des Large Language Models pour générer des réponses.

ABSTRACT

EDF RD Participation to DEFT 2023

This paper describes the participation of EDF R&D at DEFT 2023. Our team worked on the Multiple Choice Questions Answering task proposed. The corpus was FrenchMedMCQA and is composed of questions from pharmacy exam annals. We used Large Language Models to predict the answers.

MOTS-CLÉS : Gros modèles de langue, Bloom, BloomZ, ChatGPT, Questions à choix multiples, pharmacie.

KEYWORDS: Large Language Models, Bloom, BloomZ, ChatGPT, MCQA, pharmacy.

1 Introduction

L'objectif du défi DEFT 2023 est de répondre automatiquement à des questionnaires à choix multiples issus d'annales d'examens de pharmacie. Le corpus utilisé, FrenchMedMCQA (Labrak *et al.*, 2022), se compose de questions fermées en français provenant d'annales d'examens de pharmacie. Chaque question contient : un identifiant, la question posée, cinq réponses possibles et l'ensemble des réponse(s) correcte(s).

À la différence des défis des années passées, il est impossible de répondre aux questions posées sans avoir accès à des connaissances extérieures. Plusieurs possibilités existent comme utiliser des bases de données spécialisées dans le domaine concerné (comme PubMed, par exemple), utiliser le Web comme source de données ou bien utiliser les Larges Modèles de Langues (LLM en anglais) construits sur de grosses quantités de données et qui ont ainsi « acquis » un certain nombre de connaissances (Wei *et al.*, 2022).

∗. Cités par ordre alphabétique

En participant à ce défi, nous avons choisi de tester plusieurs "Larges Modèles de Langue" comme GPT3 et Bloom, ainsi que leur sur-couche ChatGPT et BloomZ.

2 ChatGPT

L'arrivée de ChatGPT (OpenAI, 2021) a provoqué une onde choc assez impressionnante dans le monde du TAL. Cette technologie est très prometteuse et offre de belles perspectives pour la résolution des tâches classiques du TAL mais également pour de nouvelles. Nous en voulons pour preuve les nombreuses solutions concurrentes qui commencent à émerger comme « Open Assistant [1] ». Plusieurs articles utilisent déjà cette technologie pour répondre à des questions à choix multiples tels que (Kung *et al.*, 2023) et (Liévin *et al.*, 2022). Il nous a semblé intéressant d'utiliser le cadre de ce défi pour tester cette technologie en mode *zero-shot*, c'est-à-dire sans entraînement particulier et en demandant directement à ChatGPT quelles étaient les bonnes réponses parmi les cinq réponses possibles.

2.1 Aspect technique

ChatGPT se présente comme un modèle que l'on peut interroger avec différents paramètres. Deux distinctions principales existent : le mode "*chat*" et le mode "*completion*". Dans le mode "*completion*", on lui indique un début de phrase qu'il va venir compléter ainsi que le nombre de mots que doit fournir ChatGPT, ce qui est un peu la "fonctionnalité de base" d'un LLM. Dans le mode "*chat*", la réponse sera plus complète et prendra plus de temps. C'est ce mode que nous avons utilisé.

Le modèle utilisé est `gpt-3.5-turbo`, avec 0 comme valeur de température. Cette valeur est comprise entre 0 et 2. Plus la valeur est proche de 0 et plus la réponse sera précise, plus elle est proche de 2 et plus la réponse sera "créative".

Enfin, plusieurs façons d'interroger le modèle lui-même sont possibles dont cURL, pour *client URL request library*, une interface en ligne de commande pour requêter des ressources informatiques accessibles sur un réseau, le format d'échange de données étant JSON.

2.2 Formatage des données

Une des clés de l'utilisation de cette technologie est la manière de poser la question (*prompt engineering*). La manière de poser la question est ici inspirée par (Liévin *et al.*, 2022) : « Q : » suivi du texte de la question, puis par les 5 réponses possibles précédées de « A) », « B) », etc. avec « \n » venant séparer les différents champs, comme dans l'exemple suivant :

```
Q : Texte de la question \n A) Réponse 1 \n B) Réponse 2 \n C)
Réponse 3 \n D) Réponse 4 \n E) Réponse 5.
```

ChatGPT fournit une réponse qu'il faut ensuite analyser via des expressions régulières. Sur les 2171 questions du corpus d'entraînement, les réponses ont pour type :

— « `B) Réponse 2 \n C) Réponse 3` » (s'il a jugé que B et C étaient les bonnes réponses) (2063 cas sur 2171). Il faut donc récupérer les lettres majuscules A,B,C, D ou E suivi d'une parenthèse pour obtenir la réponse : « BC » ;

1. disponible à l'adresse : https ://open-assistant.io/

— « `toutes les (réponses|propositions|affirmations) sont (vraies|possibles|correctes|exactes)` ». (37 cas sur 2171). Dans ces cas-là, on génère la suite « ABCDE », puisque toutes les réponses sont justes ;
— « `Réponse : A et D sont exactes.\n \n Explication : ...` ». Dans ce cas, il faudrait extraire A et D, mais qui ne sont pas suivis d'une parenthèse. Le mot « Réponse » pouvant d'ailleurs être facultatif ;
— D'autres formes de réponses sont également possibles, mais sans qu'un lien puisse facilement être établi entre les réponses possibles et la réponse fournie (pas de lettre, ni de parenthèse). Dans ces deux derniers cas de figure (71 cas sur 2171), les expressions régulières retournent une chaîne de caractère vide, la réponse « ABCDE » est donc attribuée arbitrairement.

2.3 Résultats obtenus sur le corpus d'entraînement

Les scores obtenus par ChatGPT sur le corpus d'entraînement en *zero-shot*, sont de 29,13% en EMR (Exact Match Ratio) et 58,26% pour Hamming.

Quand on analyse le corpus d'entraînement, les réponses sont équitablement réparties. Les nombres de réponses A, B, C et D sont pratiquement égales. Par contre, le nombre de réponses à E est légèrement inférieur. Si on analyse maintenant les résultats obtenus par ChatGPT sur ce même corpus, on constate, à peu près la même répartition, avec un nombre de réponses à E encore inférieur. Le nombre moyen de réponses apportées par ChatGPT est très légèrement inférieur au nombre réel de réponses dans le corpus d'entraînement (2,22 réponses contre 2,37 réponses par question).

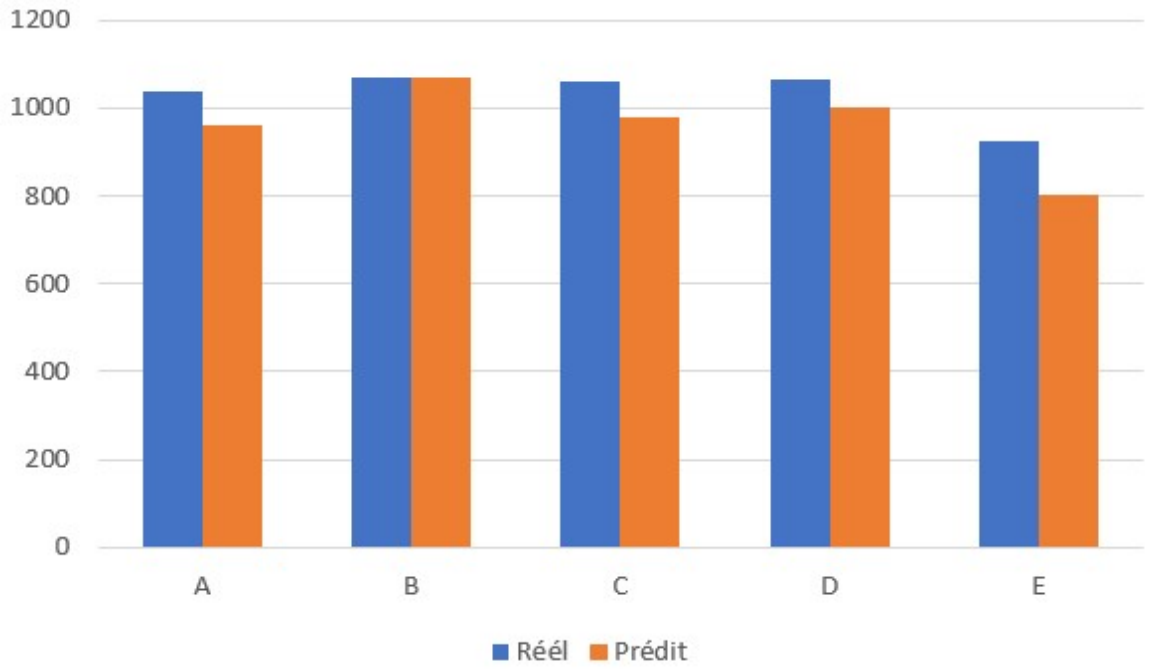

FIGURE 1 – Comparaison entre les réponses prédites par ChatGPT et les bonnes réponses

3 BloomZ

BloomZ (Workshop *et al.*, 2023) est un grand modèle de langue Open-Source et gratuit entraîné dans le cadre d'un projet international et collaboratif appelé BigScience, partiellement financé par le gouvernement français. Plus particulièrement, BloomZ est une itération de Bloom entraîné spécifiquement sur les tâches de compréhension d'instructions en *zero-shot*, c'est-à-dire sans exemple de résolution de la tâche attendue dans le prompt en entrée.

3.1 BloomZ 1.1B et 7.1B en inférence

Pour des questions de ressources matérielles, nous avons travaillé sur l'évaluation de BloomZ de 1.1 et 7.1 milliards de paramètres, et non pas sur le modèle le plus conséquent, qui atteint 176 milliards de paramètres et requiert des GPUs très performants. Le chargement du modèle 1.1B a été effectué à l'aide d'un GPU de 20Go, tandis que le modèle 7.1 milliards a dû être chargé à l'aide d'un GPU de 40Go.

Nous les avons testés directement en inférence avec du *zero-shot* et en *few-shot*. Les résultats sont, sans grande surprise, légèrement meilleurs avec le plus gros modèle.

Le tableau ci-dessous récapitule les résultats obtenus pour chaque *run*, à partir du Hamming Score (HS) et de l'Exact Match Ratio (EMR).

	1.1B	7.1B
Zero-Shot	HS : 0,53 EMR : 0,08	**HS : 0,57** **EMR : 0,1**
One-Shot	HS : 0,53 EMR : 0,08	HS : 0,56 EMR : 0,09
Two-Shot	HS : 0,5 EMR : 0,05	HS : 0,55 EMR : 0,09
Three-Shot	HS : 0,52 EMR : 0,07	HS : 0,55 EMR : 0,09

TABLE 1 – Résultats obtenus pour du ZSL et du FSL en inférence avec BloomZ 1.1 & 7.1B

Les meilleurs résultats pour le Hamming Score et l'EMR sont donc avec des prompts en *zero-shot*, bien que les différences de résultats sont parfois marginales. De plus, nous avons remarqué que pour le modèle 7.1B particulièrement, le choix des exemples envoyés au modèle a un impact plus ou moins fort. Par exemple, dans le cas du *one-shot learning* et avec le même prompt, les scores sont légèrement meilleurs si l'exemple envoyé au modèle est une question avec plusieurs bonnes réponses. Les résultats en *one-shot* avec une question qui n'accepte qu'une bonne réponse sont les suivantes :

* **BloomZ 1.1B** : 0,53 pour le Hamming Score, et 0,8 pour l'Exact Match Ratio.

* **BloomZ 7.1B** : 0,55 pour le Hamming Score, et 0,8 pour l'Exact Match Ratio.

La différence n'est qu'infime dans le cas de l'EMR, mais plus significative dans le cas du HS.

3.2 BloomZ 1.1B avec apprentissage

Afin d'adapter le modèle BloomZ à la tâche de MCQA sur le domaine pharmaceutique, nous avons fait le choix d'outrepasser les méthodes classiques de *Fine-Tuning* afin d'expérimenter avec le *Prompt Tuning* (Lester *et al.*, 2021), un nouvelle méthode d'adaptation des grands modèles de langue qui ne requiert pas de modifier les paramètres du modèle de base.

Pour ce faire, nous avons utilisé la librairie PEFT (*Parameter-Efficient Fine-Tuning*) (Sourab Mangrulkar, 2022). Nous avons tenté le Prompt Tuning de BloomZ-1b1 seulement. Les résultats sont décevants : évalué sur des prompts en One-Shot, il n'améliore que très légèrement les résultats du Hamming Score (0.54 contre 0.53 pour le modèle 1b1 non entraîné), et n'améliore pas du tout l'Exact Match Ratio, qui reste de 0.08.

4 Résultats

Les résultats obtenus sur la tâche principale sont très satisfaisants : BloomZ obtient des scores équivalents aux méthodes présentées dans (Labrak *et al.*, 2022) et ChatGPT présente des scores environ 2 fois supérieurs, ce qui prouve sa qualité et justifie l'engouement qu'il suscite.

Système	Hamming	EMR
ChatGPT	**64,40**	**46,46**
BloomZ - run1	26,34	14,63
BloomZ - run2	35,90	15,27
BloomZ - run3	37,93	12,70

TABLE 2 – Résultats obtenus sur la tâche principale

Pour BloomZ, le *run1* correspond à un test avec du *zero-shot*, alors que les *run2* et *run3* correspondent à des tests en *one-shot* (avec des exemples différents).

Pour la tâche annexe (identifier le nombre de réponses supposément justes pour une question donnée), nous nous sommes contentés de calculer le nombre de réponses considérées comme étant justes par ChatGPT.

Système	Accuracy	F1-score macro
ChatGPT	65,92	44,36

TABLE 3 – Résultats obtenus sur la tâche annexe

5 Conclusion

La participation à la campagne DEFT 2023 nous a permis de tester les nouveaux *Large Language Models*, à la fois fermés via API, tel GPT3, et ouverts tels Bloom et BloomZ. Ces modèles, alliés aux mécanismes de *Prompt Engineering* sont très prometteurs pour le traitement des données textuelles au sein de EDF Commerce et d'autres entités du groupe EDF.

Références

KUNG T. H., CHEATHAM M., MEDENILLA A., SILLOS C., DE LEON L., ELEPAÑO C., MADRIAGA M., AGGABAO R., DIAZ-CANDIDO G., MANINGO J. *et al.* (2023). Performance of chatgpt on usmle : Potential for ai-assisted medical education using large language models. *PLoS digital health*, **2**(2), e0000198.

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.

LESTER B., AL-RFOU R. & CONSTANT N. (2021). The power of scale for parameter-efficient prompt tuning.

LIÉVIN V., HOTHER C. E. & WINTHER O. (2022). Can large language models reason about medical questions ? *arXiv preprint arXiv :2207.08143*.

OPENAI (2021). Gpt-3.5 language model. [online]. disponible sur https ://openai.com/gpt-3-5/.

SOURAB MANGRULKAR, SYLVAIN GUGGER L. D. Y. B. S. P. (2022). Peft : State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

WEI J., TAY Y., BOMMASANI R., RAFFEL C., ZOPH B., BORGEAUD S., YOGATAMA D., BOSMA M., ZHOU D., METZLER D. *et al.* (2022). Emergent abilities of large language models. *arXiv preprint arXiv :2206.07682*.

WORKSHOP B., :, SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N., DEL MORAL A. V., RUWASE O., BAWDEN R., BEKMAN S., MCMILLAN-MAJOR A., BELTAGY I., NGUYEN H., SAULNIER L., TAN S., SUAREZ P. O., SANH V., LAURENÇON H., JERNITE Y., LAUNAY J., MITCHELL M., RAFFEL C., GOKASLAN A., SIMHI A., SOROA A., AJI A. F., ALFASSY A., ROGERS A., NITZAV A. K., XU C., MOU C., EMEZUE C., KLAMM C., LEONG C., VAN STRIEN D., ADELANI D. I., RADEV D., PONFERRADA E. G., LEVKOVIZH E., KIM E., NATAN E. B., TONI F. D., DUPONT G., KRUSZEWSKI G., PISTILLI G., ELSAHAR H., BENYAMINA H., TRAN H., YU I., ABDULMUMIN I., JOHNSON I., GONZALEZ-DIOS I., DE LA ROSA J., CHIM J., DODGE J., ZHU J., CHANG J., FROHBERG J., TOBING J., BHATTACHARJEE J., ALMUBARAK K., CHEN K., LO K., WERRA L. V., WEBER L., PHAN L., ALLAL L. B., TANGUY L., DEY M., MUÑOZ M. R., MASOUD M., GRANDURY M., ŠAŠKO M., HUANG M., COAVOUX M., SINGH M., JIANG M. T.-J., VU M. C., JAUHAR M. A., GHALEB M., SUBRAMANI N., KASSNER N., KHAMIS N., NGUYEN O., ESPEJEL O., DE GIBERT O., VILLEGAS P., HENDERSON P., COLOMBO P., AMUOK P., LHOEST Q., HARLIMAN R., BOMMASANI R., LÓPEZ R. L., RIBEIRO R., OSEI S., PYYSALO S., NAGEL S., BOSE S., MUHAMMAD S. H., SHARMA S., LONGPRE S., NIKPOOR S., SILBERBERG S., PAI S., ZINK S., TORRENT T. T., SCHICK T., THRUSH T., DANCHEV V., NIKOULINA V., LAIPPALA V., LEPERCQ V., PRABHU V., ALYAFEAI Z., TALAT Z., RAJA A., HEINZERLING B., SI C., TAŞAR D. E., SALESKY E., MIELKE S. J., LEE W. Y., SHARMA A., SANTILLI A., CHAFFIN A., STIEGLER A., DATTA D., SZCZECHLA E., CHHABLANI G., WANG H., PANDEY H., STROBELT H., FRIES J. A., ROZEN J., GAO L., SUTAWIKA L., BARI M. S., AL-SHAIBANI M. S., MANICA M., NAYAK N., TEEHAN R., ALBANIE S., SHEN S., BEN-DAVID S., BACH S. H., KIM T., BERS T., FEVRY T., NEERAJ T., THAKKER U., RAUNAK

V., TANG X., YONG Z.-X., SUN Z., BRODY S., URI Y., TOJARIEH H., ROBERTS A., CHUNG H. W., TAE J., PHANG J., PRESS O., LI C., NARAYANAN D., BOURFOUNE H., CASPER J., RASLEY J., RYABININ M., MISHRA M., ZHANG M., SHOEYBI M., PEYROUNETTE M., PATRY N., TAZI N., SANSEVIERO O., VON PLATEN P., CORNETTE P., LAVALLÉE P. F., LACROIX R., RAJBHANDARI S., GANDHI S., SMITH S., REQUENA S., PATIL S., DETTMERS T., BARUWA A., SINGH A., CHEVELEVA A., LIGOZAT A.-L., SUBRAMONIAN A., NÉVÉOL A., LOVERING C., GARRETTE D., TUNUGUNTLA D., REITER E., TAKTASHEVA E., VOLOSHINA E., BOGDANOV E., WINATA G. I., SCHOELKOPF H., KALO J.-C., NOVIKOVA J., FORDE J. Z., CLIVE J., KASAI J., KAWAMURA K., HAZAN L., CARPUAT M., CLINCIU M., KIM N., CHENG N., SERIKOV O., ANTVERG O., VAN DER WAL O., ZHANG R., ZHANG R., GEHRMANN S., MIRKIN S., PAIS S., SHAVRINA T., SCIALOM T., YUN T., LIMISIEWICZ T., RIESER V., PROTASOV V., MIKHAILOV V., PRUKSACHATKUN Y., BELINKOV Y., BAMBERGER Z., KASNER Z., RUEDA A., PESTANA A., FEIZPOUR A., KHAN A., FARANAK A., SANTOS A., HEVIA A., UNLDREAJ A., AGHAGOL A., ABDOLLAHI A., TAMMOUR A., HAJIHOSSEINI A., BEHROOZI B., AJIBADE B., SAXENA B., FERRANDIS C. M., CONTRACTOR D., LANSKY D., DAVID D., KIELA D., NGUYEN D. A., TAN E., BAYLOR E., OZOANI E., MIRZA F., ONONIWU F., REZANEJAD H., JONES H., BHATTACHARYA I., SOLAIMAN I., SEDENKO I., NEJADGHOLI I., PASSMORE J., SELTZER J., SANZ J. B., DUTRA L., SAMAGAIO M., ELBADRI M., MIESKES M., GERCHICK M., AKINLOLU M., MCKENNA M., QIU M., GHAURI M., BURYNOK M., ABRAR N., RAJANI N., ELKOTT N., FAHMY N., SAMUEL O., AN R., KROMANN R., HAO R., ALIZADEH S., SHUBBER S., WANG S., ROY S., VIGUIER S., LE T., OYEBADE T., LE T., YANG Y., NGUYEN Z., KASHYAP A. R., PALASCIANO A., CALLAHAN A., SHUKLA A., MIRANDA-ESCALADA A., SINGH A., BEILHARZ B., WANG B., BRITO C., ZHOU C., JAIN C., XU C., FOURRIER C., PERIÑÁN D. L., MOLANO D., YU D., MANJAVACAS E., BARTH F., FUHRIMANN F., ALTAY G., BAYRAK G., BURNS G., VRABEC H. U., BELLO I., DASH I., KANG J., GIORGI J., GOLDE J., POSADA J. D., SIVARAMAN K. R., BULCHANDANI L., LIU L., SHINZATO L., DE BYKHOVETZ M. H., TAKEUCHI M., PÀMIES M., CASTILLO M. A., NEZHURINA M., SÄNGER M., SAMWALD M., CULLAN M., WEINBERG M., WOLF M. D., MIHALJCIC M., LIU M., FREIDANK M., KANG M., SEELAM N., DAHLBERG N., BROAD N. M., MUELLNER N., FUNG P., HALLER P., CHANDRASEKHAR R., EISENBERG R., MARTIN R., CANALLI R., SU R., SU R., CAHYAWIJAYA S., GARDA S., DESHMUKH S. S., MISHRA S., KIBLAWI S., OTT S., SANG-AROONSIRI S., KUMAR S., SCHWETER S., BHARATI S., LAUD T., GIGANT T., KAINUMA T., KUSA W., LABRAK Y., BAJAJ Y. S., VENKATRAMAN Y., XU Y., XU Y., XU Y., TAN Z., XIE Z., YE Z., BRAS M., BELKADA Y. & WOLF T. (2023). Bloom : A 176b-parameter open-access multilingual language model.