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

Flan-T5 avec ou sans contexte, telle est la question à choix multiples

Elias Okat[1] Hugo Brochelard[2, 3] Aghilas Sini[1]
Valérie Renault[3] Nathalie Camelin[1]
(1) Le Mans Université, LIUM, 72000 Le Mans, France
(2) CHM, 194 Av. Rubillard, 72037 Le Mans, France (3) Le Mans Université, CREN, 72000 Le Mans, France
{Elias.Okat.Etu, Hugo.Brochelard.Etu,
Aghilas.Sini,Valerie.Renault,Nathalie.Camelin}@univ-lemans.fr

RÉSUMÉ

Ce travail présente les systèmes développés par l'équipe LIUM-CREN pour l'atelier DEFT 2024. Nous avons participé à la tâche principale qui vise à inférer automatiquement les réponses correctes à des questions à choix multiples dans le domaine médical en utilisant le corpus FrenchMedMCQA. Nous avons soumis trois approches : (a) explorer l'espace de plongements afin de mettre en évidence les liens éventuels entre les questions et les réponses associées ; (b) utiliser la capacité de génération des modèles Text-To-Text tels que Flan-T5-Large pour générer les réponses correctes ; et (c) mettre en place une technique basique de Retrieval Augmented Generation (RAG) afin de fournir du contexte spécifique au modèle génératif Flan-T5-Large. Cet article vise à rapporter les résultats que nous avons obtenus et à étudier l'impact du contexte sur la capacité du Flan-T5 à inférer les réponses correctes.

ABSTRACT

Flan-T5 with or without context, that is the multiple-choice question

This work presents the systems developed by the LIUM-CREN team for the Défi Fouille de Textes (DEFT) 2024 workshop. We participated in the main task, which aims to infer the correct answers automatically for multiple-choice questions in the medical domain using the FrenchMedMCQA corpus. We submitted three approaches : (a) investigating the embedding space relative to the questions and answers ; (b) using the generation ability of Text-To-Text models such as Flan-T5-Large to generate the correct answers ; and (c) combining the Large Language Model (LLM) model with a basic Retrieval Augmented Generation (RAG) technique to build a "Context". This paper aims to report the results we obtained and investigate the impact of context on the ability of LLMs to predict the correct answers.

MOTS-CLÉS : Questionnaire à Choix Multiples (QCM), génération augmentée par extraction de données, grands modèles de langue, plongement.

KEYWORDS: Multiple choice questionnaire (MCQ), RAG, LLM, Embedding.

1 Introduction

Répondre à un QCM dans un domaine spécifique, tel que le médical, soulève des défis particuliers en raison de la complexité du lexique et des connaissances spécialisées requises. Les stratégies adoptées par les humains pour répondre aux QCMs varient selon les individus et les contextes. Parmi ces stratégies, on retrouve les connaissances préalables et d'étude (Ormrod, 2016), l'analyse des questions (Bruning et al., 2011), et la compréhension du format de QCM (Pauker, 2011), ainsi que bien d'autres

(Svinicki & McKeachie, 2014; Pauk & Owens, 2013; Tuckman & Kennedy, 2011; Biggs & Tang, 2011).

Quelles stratégies les machines mettent en œuvre pour répondre à un QCM ? Des études récentes ont mis en évidence les difficultés des techniques avancées d'apprentissage automatique telles que les Transformers (Raffel et al., 2020a; Zheng et al., 2024) à répondre à des QCMs. La campagne d'évaluation DEFT2023 (Labrak et al., 2023a) proposait d'explorer ces difficultés dans le domaine spécifique du médical. Plusieurs équipes participantes avaient utilisé des grands modèles de langue (LLM) (Favre, 2023; Blivet et al., 2023; Besnard et al., 2023) pour résoudre ce challenge. L'édition 2024 propose à nouveau le même défi, en ajoutant dans la tâche principale comme contraintes une limitation de taille du modèle utilisé et la transparence des données utilisées durant son apprentissage.

Nous avons eu accès aux données du corpus FrenchMedMCQA (Labrak et al., 2022). Ce corpus contient un ensemble de QCMs avec, pour chaque question, cinq propositions de réponses. Deux des principales difficultés résident dans le fait que le nombre de réponses correctes à sélectionner pour chaque question est inconnu, et que le vocabulaire utilisé est très spécifique au domaine pharmaceutique.

Notre étude reprend certains éléments mis en avant par l'équipe LIUM-IRISA (Besnard et al., 2023), notamment la prise en compte des schémas de question et l'exploitation d'un contexte par les LLMs. Nous souhaitons aller plus loin en focalisant notre travail sur :

(a) **Proximité sémantique** : Analyser si la proximité entre la question et les réponses attendues dans l'espace de plongements est suffisante pour répondre efficacement.

(b) **Valeur ajoutée du contexte** : Évaluer la valeur ajoutée du contexte sur la performance des LLM dans la génération de sa réponse.

(c) **Pertinence du contexte** : Identifier les situations où l'ajout de contexte est le plus pertinent et améliore l'exactitude des réponses.

Nous commencerons par présenter nos analyses et traitements des données initiales issues du corpus. Ensuite, nous décrirons les approches méthodologiques utilisées par nos systèmes, avant de détailler les aspects expérimentaux et l'analyse des résultats obtenus. Enfin, nous terminerons par une discussion sur les perspectives et conclusions issues de cette campagne.

2 Corpus et schémas de questions

Dans le cadre de ce défi, nous avons eu accès au corpus FrenchMedMCQA (Labrak et al., 2022) qui est composé de 3105 questions fermées, ces questions extraites d'annales d'examens de pharmacie en français sont représentées par : un identifiant, la question en langage naturel, ses cinq options, l'ensemble de réponse(s) attendue(s), le nombre de réponses attendues ainsi que le type de question (*simple*/*multiple*). La distribution détaillée des jeux de données est décrite en Table 1.

Dans un premier temps, nous avons développé un système de classification des types de questions. Il a pour but de déterminer si une question attend une (*simple*) ou plusieurs (*multiple*) réponses. Cette information est importante, car elle nous permet d'adapter la stratégie de nos systèmes en fonction des exigences spécifiques de chaque question.

Suivant l'analyse de l'équipe LIUM-IRISA (Besnard et al., 2023), nous avons repris la caractéristique *correct*/*incorrect* définissant si la(les) réponse(s) attendue(s) correspond(ent) à l'information liée positivement à la question ou à son opposé. Par exemple, dans l'énoncé suivant "Laquelle des affirmations suivantes ne s'applique pas au cotrimoxazole ?", la réponse attendue est une réponse que

Dataset	Nb. questions	% questions simples	% questions incorrectes
Train	2171	27.41	14.51
Dev	312	52.56	23.7
Test 2023	622	51.60	21.5
Test 2024	477	19.49	5.03

TABLE 1 – Distribution des jeux de données des campagnes DEFT 2023 et 2024.

nous considérons dans la catégorie *incorrect*. Nous avons développé un second système qui permet de catégoriser les questions en *correct/incorrect*.

Ainsi, nous avons développé deux systèmes, basés sur des expressions régulières inspirées de (Besnard et al., 2023; Blivet et al., 2023) permettant de définir le **schéma d'une question** : ***simple/multiple*** et ***correct/incorrect***.

Notre système de classification *simple/multiple* obtient un F1-score de 88.7% sur le corpus de développement ainsi qu'un F1-score de 92.7% sur le corpus de test 2023. Les performances du système de classification *correct/incorrect* ne peuvent pas être calculées car la référence n'est pas disponible. Néanmoins, nous indiquons dans la dernière colonne de la Table 1 le pourcentage de questions *incorrectes* que nous détectons automatiquement afin de donner une idée de l'ordre de grandeur de cette caractéristique.

La définition du schéma de la question permet d'adater la stratégie de réponse de nos systèmes aux exigences spécifiques de chaque question et ainsi d'optimiser le processus de sélection des réponses.

3 Approches d'identification des réponses attendues

Nous avons envisagé deux approches afin de répondre à la tâche principale proposée dans DEFT2024. La première de ces approches repose sur l'exploitation d'espaces de plongements afin de tenir compte de la proximité sémantique entre les questions et les réponses. La seconde approche est basée sur la génération des réponses à l'aide de grands modèles de langue (LLM).

3.1 Procédure d'inférence des réponses à partir de l'espace de plongements

Notre première approche a consisté à utiliser des espaces de plongements dans l'objectif de projeter les représentations du contenu textuel des questions et des réponses dans un même espace.

Cette approche repose sur l'hypothèse de l'existence de corrélations entre les énoncés d'une question et de ses réponses.

Le principe au cœur de cette approche est de projeter chaque question et propositions de réponse associées dans un espace de plongements. Cela nous permet ensuite, en exploitant les proximités/éloignements, d'en déduire un ensemble de réponses à sélectionner. Pour cela, il suffit de créer des secteurs dans l'espace, comme schématisé dans la Figure 1, et de définir lequel répond à la question.

Notre système ROBOT (Rules on Bottom Or Top) permet de définir deux secteurs selon les règles suivantes :

— **Top** : ce secteur contient les réponses ayant une similarité supérieure ou égale à un seuil α.
— **Bottom** : ce secteur contient les réponses ayant une similarité inférieure ou égale au seuil α.

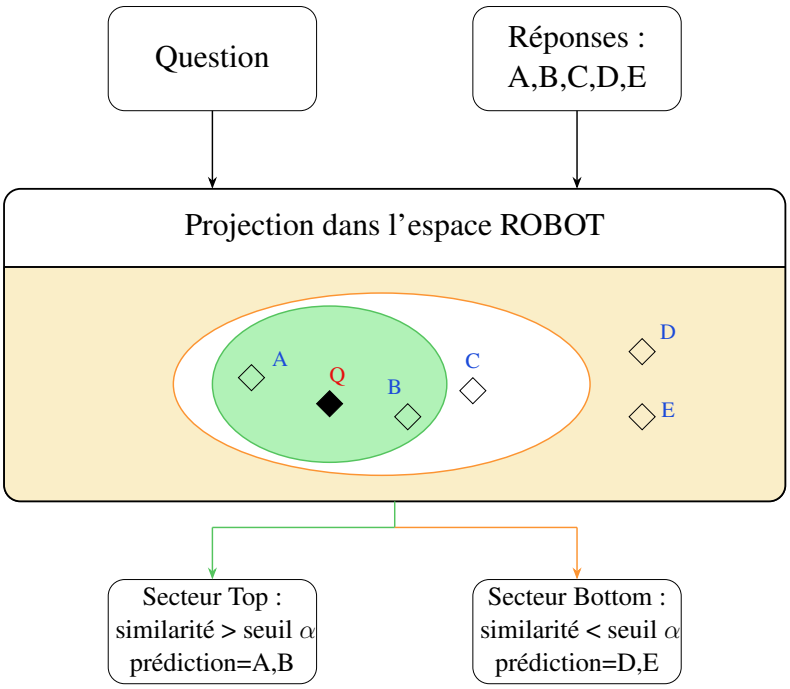

FIGURE 1 – Architecture du système ROBOT avec seuil α. Le secteur Top est en vert et le secteur Bottom est en jaune.

Des règles dépendant du schéma de chaque question permettent ensuite de définir le secteur choisi pour admettre les réponses. Les réponses qui n'appartiennent ni au secteur Top, ni au Bottom ne seront jamais choisies comme réponses attendues. Elles font partie d'une zone pour laquelle le système ne peut se prononcer avec certitude.

3.2 Génération de réponses à l'aide de grands modèles de langue

La seconde approche que nous avons explorée est une approche générative. Ce choix à été motivé par le fait que les grands modèles de langue (LLM) ont démontré d'impressionnantes capacités sur une grande variété de tâches en traitement du langage naturel (Raffel et al., 2020b; Chung et al., 2024). Cependant, ces modèles manquent souvent de connaissances dans des domaines spécifiques. Afin de pallier ce problème, nous avons étudié dans quelle mesure l'apport et l'exploitation d'éléments de contexte provenant de sources spécialisées permettraient d'améliorer les performances d'un LLM en s'appuyant sur le mécanisme de RAG (Gao et al., 2024) pour générer du contexte.

La Figure 2 décrit l'architecture des systèmes génératifs enrichie par RAG. Chaque question et les propositions de réponses associées sont d'abord vectorisés par un modèle de plongements. Par ailleurs, une base de données composée de nombreux documents (dont certains comportant des informations spécifiques au domaine médical) est constituée. Ensuite, un moteur de similarité compare les plongements des questions avec ceux de la base de connaissance afin d'y identifier les passages de texte pertinents. Ces derniers servent alors de contexte supplémentaire pour le modèle de langue qui génère les réponses attendues.

Le principe fondamental du RAG est de tirer parti de bases de connaissances afin de fournir des éléments de contextes pertinents aux modèles de langue. Nous espérons ainsi améliorer leur capacité à générer des réponses précises et appropriées à notre tâche.

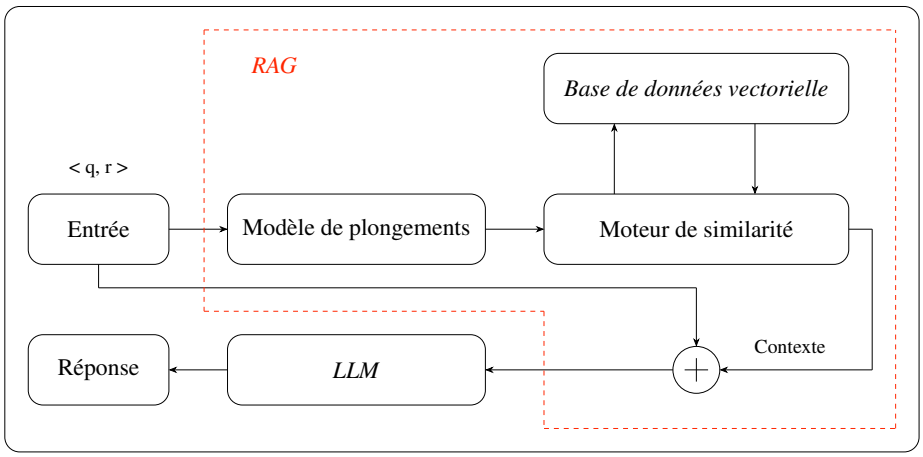

FIGURE 2 – Architecture générale des approches génératives enrichie par le mécanisme de RAG

4 Procédures expérimentales

Notre équipe a participé à la tâche principale de la campagne d'évalution DEFT 2024, consistant à identifier automatiquement l'ensemble des réponses correctes parmi les cinq proposées pour une question donnée. Les systèmes pour cette tâche avaient comme contraintes de faire moins de 3 milliards de paramètres, de ne pas rechercher sur Internet les originaux des données fournies, et de ne pas utiliser de corpus additionnels autres que NACHOS (Labrak et al., 2023b) et Wikipédia. Enfin, l'utilisation de modèles pré-entraînés était limitée à ceux dont les données d'entraînement sont connues.

4.1 Descriptif des différents modèles de langue mis en œuvre

Au cours de nos expérimentations, nous avons exploité divers modèles de langue : Flan-T5, sentence-camemBERT et DrBERT.

Nous avons choisi d'utiliser la version *Flan-T5-Large* [1] (780 millions de paramètres) du modèle Flan-T5 (Chung et al., 2022) comme modèle de **génération de réponse**. Les modèles Flan-T5 ont la particularité d'être très versatile, grâce à un entraînement par Google sur plus de 1000 tâches (traduction, complétion de texte, questions/réponses).

Par ailleurs, nous avons utilisé deux modèles de plongements pré-entraînés de type Bert pour représenter les questions et les réponses et **obtenir un contexte pertinent**. Sentence-CamemBERT (Martin et al., 2020) est un modèle adapté à la langue française et optimisé pour les segments de phrases (Nils Reimers, 2019) [2]. Sa capacité, à raisonner au niveau des phrases, nous a semblé particulièrement appropriée dans le cadre de l'ajout de contexte au LLM. Le modèle *DrBERT-7GB* [3] (Labrak et al., 2023b) a également été utilisé dans nos systèmes et pré-entraîné sur le corpus de données médicales en français NACHOS. Ce modèle a donc une compréhension approfondie des terminologies propres au domaine médical. Cette caractéristique en fait un choix pertinent pour enrichir le LLM avec du contexte.

1. https://huggingface.co/google/flan-t5-large
2. https://huggingface.co/dangvantuan/sentence-camembert-large
3. https://huggingface.co/Dr-BERT/DrBERT-7GB

4.2 Procédure d'affinage du modèle Flan-T5

Afin de spécialiser le LLM qui permet de générer les réponses à notre QCM, *Flan-T5-Large* a été affiné sur le corpus d'entraînement DEFT. Pour cela, un prompt a été créé contenant :

— un pré-prompt indiquant le schéma de la question.
— l'énoncé en langage naturel de la question et des réponses possibles.
— un contexte, qui peut être vide ou complété par un système de RAG (voir section 3.2).
— la liste des lettres des réponses attendues.

Le modèle *Flan-T5-Large* a été affiné sur 10 *époques*, avec un *batch* d'une taille de 4, un *taux d'apprentissage* de $1e^{-5}$, un *weight decay* de 0.01, et en utilisant la valeur 42 comme graine de génération pour l'aléatoire. Le modèle a été entraîné en utilisant 3 cartes "Quadro RTX 8000".

4.3 Exploration de la similarité sémantique question/réponse

Le système *ROBOT* utilise la similarité cosinus pour calculer la proximité entre les énoncés des questions et des réponses dans l'espace vectoriel défini par le modèle Flan-T5-Large affiné sans contexte. Le seuil α a été choisi empiriquement basé sur les tiers des scores de similarités.

Pour choisir un secteur et ainsi sélectionner les réponses attendues, deux règles sont appliquées :

(a) Pour une question simple, si l'un des secteurs contient plusieurs réponses, on privilégie l'autre par défaut.
(b) Pour les questions multiples, si l'un des secteur contient une seule réponse, on privilégie également l'autre par défaut.

Ensuite, la question est évaluée pour déterminer si elle vise à sélectionner la ou les réponses correctes ou incorrectes :

(a) Si la question recherche la ou les réponses correctes, le secteur *Top* sera choisi.
(b) Si la question recherche la ou les réponses incorrectes, le secteur *Bottom* sera choisi.

Cette approche permet une classification efficace des réponses en fonction de leur pertinence par rapport à la question posée, utilisant la similarité cosinus comme métrique principale pour mesurer la proximité sémantique.

4.4 Enrichissement de la génération de réponses par RAG

Nous définissons deux systèmes, *CARAGON* et *DRAGON*, reposant sur l'architecture définie en section 3.2. Le corpus Wikipedia a été utilisé comme base de connaissances et transformé en base de données vectorielle. *Flan-T5-Large* est le LLM utilisé pour générer les réponses. Lors de l'affinement du LLM, le contexte est obtenu par Camembert pour *CARAGON* (CAmembert RAG ON) et DrBert pour *DRAGON* (Drbert RAG ON). La Figure 3 donne un exemple de prompt fourni aux modèles lors de leur affinement pour une question au schéma *simple/correct*.

Le système CARAGON utilise *sentence-camembert-large* comme modèle de plongements, le système DRAGON repose lui sur l'utilisation du modèle *DrBert-7G*. Les processus de création de la base de données vectorielle et de recherche sont réalisés par la bibliothèque FAISS (Douze et al., 2024) à l'aide de LangChain (Chase, 2022). FAISS permet une indexation et une recherche rapide parmi des millions de vecteurs, optimisant ainsi le temps de réponse et la précision du système.

L'affinage a permis d'entraîner le modèle sur des données médicales, mais également de contraindre le modèle à respecter une syntaxe de génération spécifique. Le RAG a été utilisé à la fois lors de

Preprompt : Sélectionnez la réponse correcte.
Question : Citer parmi les composés suivants celui qui est cancérogène chez l'homme :
(A) Benzène ; (B) Chlorure de méthylène ; (C) Plomb ; (D) Monoxyde de carbone ; (E) Toluène
Contexte : Dans les conditions usuelles, le benzène est un liquide incolore, d'odeur caractéristique, volatil, très inflammable et cancérogène.
Réponse attendu : (A)

FIGURE 3 – Exemple de prompt utilisé lors de l'affinement de Flan-T5

l'inférence des deux systèmes, mais également pour l'entraînement du modèle *Flan-T5-Large*.

5 Résultats et analyse

La Table 2 rapporte l'ensemble des résultats de nos trois systèmes obtenus sur les jeux de données de test 2023 et 2024. Les résultats grisés sont ceux obtenus à la campagne officielle DEFT 2024.

Système	Corpus	HAMMING	EMR	EMR	
				Simple	Multiple
ROBOT	Test 2023	28.56	12.70	19.00	2.65
	Test 2024	31.30	8.39	20.72	5.18
DRAGON	Test 2023	40.11	17.04	27.10	6.31
	Test 2024	47.97	10.69	23.65	7.55
CARAGON	Test 2023	43.66	20.41	33.02	6.97
	Test 2024	49.15	11.53	23.65	8.59

TABLE 2 – Résultats 2023 et 2024

Dans un premier temps, nous observons que le système ROBOT, basé uniquement sur l'encodeur du LLM pour évaluer la proximité sémantique question/réponse obtient de moins bons résultats que les systèmes DRAGON et CARAGON, basés sur le LLM avec génération et RAG.

Concernant les systèmes avec génération de réponse, le système CARAGON donne de meilleurs scores sur les deux métriques, EMR et HAMMING sur les deux corpus de test. Les résultats obtenus par ce système reposent sur les facteurs suivants :

- La taille des segments a été optimisée, contrairement à DRAGON, où la taille des segments est restée fixe tout au long des expériences, ce qui peut être désavantageux dans certains cas.
- Contrairement à DRAGON, CARAGON utilise une représentation au niveau de la phrase (groupe de mots). La fonction d'agrégation a été optimisée pour obtenir les meilleures représentations, tandis que dans DRAGON, nous nous sommes limités à une agrégation moyenne (mean pooling), qui peut ne pas être adaptée.

Nous remarquons que les résultats sur le corpus 2024 sont moins bons que ceux obtenus sur le corpus 2023. Nous pensons que cela est dû à la forte baisse de la proportion des questions simples (cf. Table 1 : 19,5% sur Test 2024 au lieu de 51,60 sur Test 2023), augmentant ainsi la proportion des questions multiples que nos systèmes ont du mal à traiter. Par exemple pour le corpus Test 2024, CARAGON obtient un EMR de 23,65% sur les questions simples tandis que les performances chutent drastiquement à 8,59% sur les questions multiples. En effet, nous relevons le manque de précision

mesuré en EMR de l'ensemble des algorithmes sur les questions à réponses multiples. Cela peut s'expliquer par le fait que l'algorithme doit inférer à la fois le nombre de réponses attendues et identifier les réponses attendues parmi celles proposées.

Néanmoins, les performances globales en terme de HAMMING suggèrent que nos systèmes gardent une capacité à identifier les réponses attendues, même si toutes ne sont pas toujours exactes en nombre et en labels. Afin d'avoir une première piste d'explication du comportement de nos systèmes, nous avons souhaité savoir si les systèmes avaient tendance à générer trop de réponses ou pas assez. La Figure 4 permet de visualiser la proportion des questions pour lesquelles le nombre de réponses attendues est égal/inférieur/supérieur à celui émis par nos système, calculé sur le corpus Test 2023.

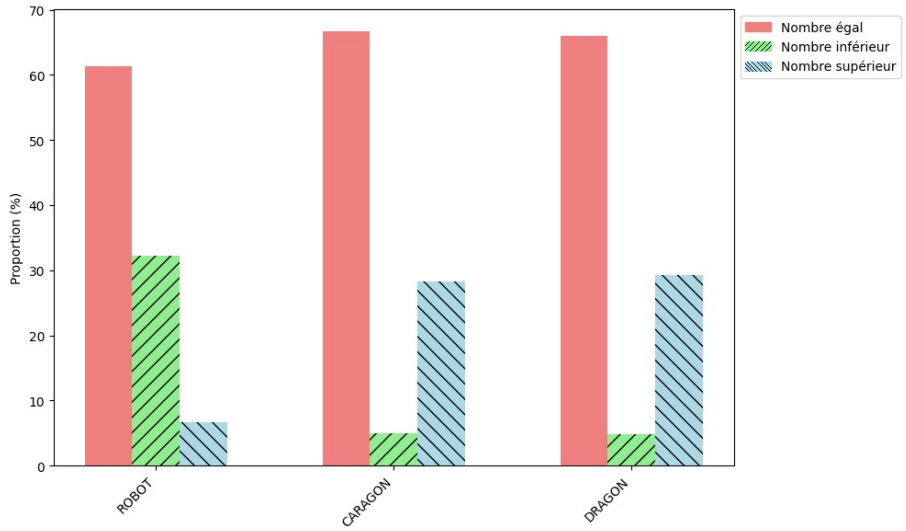

FIGURE 4 – Distribution du nombre de réponses émises par les systèmes par rapport à celui attendu : égal, insuffisant ou supérieur observé sur le corpus Test 2023.

Nous remarquons qu'une très large proportion de questions obtient le nombre de réponses attendu (plus de 60% pour tous les systèmes). Les scores d'EMR et HAMMING n'étant pas très élevés, cela nous laisse penser que la précision sur ces questions doit être faible. Par ailleurs, on observe des comportements différents entre ROBOT et les systèmes génératifs. Le nombre de réponses émises par ROBOT est souvent inférieur à celui attendu, tandis que pour les systèmes génératifs CARAGON et DRAGON c'est le contraire. En effet, dans une proportion de presque 30% les systèmes génératifs sont trop bavards. Il serait intéressant d'analyser l'origine de ce phénomène et de voir s'il est lié à l'utilisation du RAG.

Afin de mettre en évidence l'importance du RAG dans les performances de notre LLM, nous avons mené une étude d'ablation du RAG au niveau de l'affinement et de l'inférence. Les résultats sont présentés dans la Table 3.

Les résultats obtenus sur les corpus Test 2023 et 2024 ne permettent pas d'affirmer que l'utilisation du RAG améliore les résultats. En effet, les résultats sur Test 2023 montrent que pour les questions simples, l'utilisation du RAG lors de l'inférence uniquement permet d'obtenir les meilleures performances tandis que pour les questions multiples, les meilleures performances sont observées sans utilisation du RAG à l'inférence. Malheureusement, ce même comportement n'est pas observé sur le corpus DEFT 2024 où globalement les meilleures performances sont observées lors de l'utilisation du RAG à l'affinement uniquement.

Configuration		Corpus	EMR		HAMMING	
Affinement	Inférence		Simple	Multiple	Simple	Multiple
✖	✖	Test 2023	28.34	9.96	29.49	54.87
		Test 2024	26.88	8.33	28.85	53.52
✖	✔	Test 2023	35.82	8.63	36.73	51.45
		Test 2024	24.73	6.77	26.52	51.01
✔	✖	Test 2023	28.66	8.30	29.54	55.12
		Test 2024	26.88	9.11	29.15	52.26
✔	✔	Test 2023	31.46	7.97	32.52	54.08
		Test 2024	25.80	8.85	28.17	52.00

TABLE 3 – Évaluation de l'impact du contexte selon la configuration (Affinement et/ou Inférence) en fonction du type de question pour le système CAmembert with RAG ON. Les meilleures performances obtenues sur Test 2024 sont colorées en bleu et celles sur Test 2023 en vert.

Il est à noter que la mesure de similarité, à savoir la similarité cosinus, utilisée pour identifier le contexte pertinent par le moteur de recherche dans le mécanisme de RAG, peut avoir contribué à la sous-exploitation des performances des modèles de plongements. Il est crucial de souligner l'absence de métriques permettant de mesurer la pertinence et l'adéquation du contexte sélectionné dans les deux systèmes considérés (CARAGON, DRAGON).

Les résultats obtenus montrent que l'intégration de systèmes de RAG basés sur des modèles de langue avancés comme CamemBert et DrBert est prometteuse. . Cependant, la performance en terme d'EMR indique qu'il reste des défis à surmonter pour améliorer l'exactitude des réponses.

Afin d'analyser nos points faibles, nous avons expérimenté une version ORACLE de notre système ROBOT. Cette version consiste à toujours choisir le bon nombre n de réponses dans l'espace de plongements. Les n réponses les plus similaires seront choisies dans le cas d'une question à schéma *correct*. Inversement, les n réponses les moins similaires seront choisies dans le cas d'une question à schéma *incorrect*. Sur le corpus Test 2024, ORACLE-ROBOT a obtenu un EMR de 14,46% et un HAMMING de 39,89%, soit 6,07 d'augmentation absolue en EMR et 8,59% en HAMMING par rapport aux résultats de ROBOT. Cette expérience montre qu'une de nos limites provient d'une faiblesse dans la détection du nombre de réponses à sélectionner. Cela paraît être une des raisons les plus plausibles quant à la différence observée entre score de HAMMING et EMR.

6 Perspectives d'amélioration

Les résultats obtenus et l'analyse réalisée mettent en lumière plusieurs pistes d'amélioration pour optimiser les performances des systèmes de réponse aux QCMs dans le domaine médical.

Tout d'abord, nous pensons qu'il serait intéressant d'enrichir les bases de connaissances avec des informations spécifiques au domaine pharmaceutique. Cela peut inclure l'intégration de données provenant de corpora adaptés au domaine médical, tels que des bases de données spécialisées et des publications scientifiques.

Ensuite, il semble nécessaire de développer des algorithmes plus sophistiqués pour améliorer la recherche de contexte pertinent dans la base de connaissances. L'utilisation de techniques avancées de traitement du langage naturel (NLP) et d'algorithmes de recherche contextuelle peut potentiellement

augmenter la pertinence du contexte sélectionné, améliorant ainsi les performances globales des systèmes.

Introduire de la flexibilité dans le choix des techniques de RAG ou du Non-RAG en fonction des types de questions et de leur complexité pourrait améliorer significativement les performances. Une approche adaptative permettrait de sélectionner la méthode la plus appropriée en fonction du schéma de chaque question. Nous pensons notamment à une approche intégrant des systèmes multi-agents. Comme proposé par (Puerto et al., 2023), les agents choisissent des approches différentes selon la typologie des questions dans une gestion plus dynamique et contextuelle des réponses.

Ces perspectives offrent un cadre stratégique pour l'amélioration continue de nos systèmes. En mettant en œuvre ces recommandations, nous espérons non seulement améliorer les scores EMR et HAMMING, mais aussi accroître la robustesse et la flexibilité de nos systèmes face à la diversité et à la complexité des questions du domaine médical.

7 Conclusion

Nos systèmes, bien que basés sur le même modèle affiné de Flan-T5-Large, montrent des performances différentes. Cela met en évidence les forces et les faiblesses spécifiques de chaque approche sur certaines particularités des jeux de données. Les architectures avec RAG étaient lourdes à mettre en place. Néanmoins, nous avons réussi à exploiter ces architectures et obtenus des réponses nous permettant de participer à la campagne DEFT 2024.

Il nous apparaît que l'espace de plongements seul ne suffit pas pour répondre avec précision aux questions à choix multiples. Notre analyse met en évidence la nécessité de continuer à affiner nos modèles et d'explorer de nouvelles approches pour améliorer la précision et l'exactitude des réponses. L'optimisation des processus de recherche de contexte, l'intégration de bases de connaissances enrichies, et l'expérimentation avec diverses architectures de RAG sont autant de pistes prometteuses pour l'avenir.

Remerciements

Cette étude a été réalisée dans le cadre du projet Perti'SAM (Réseau SAM - Comue Angers Le Mans) et de l'ANR AISSPER (ANR 19-CE23-0004). Nous tenons à remercier Madame Elsa Jouhanneau Treton Pharmacienne clinicienne praticien hospitalier au CH du Mans, portant le projet Perti'SAM. Ce travail a bénéficié d'un accès aux moyens de calcul de l'IDRIS grâce à l'allocation de ressources 2024-AD011015264R1 attribuée par GENCI.

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 18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, p. 1–10 : ATALA.

BIGGS J. B. & TANG C. (2011). Teaching for Quality Learning at University. Open University Press.

BLIVET A., DEGRUTÈRE S., GENDRON B., RENAULT A., SIOUFFI C., BOUJU V. G., CERISARA C., FLAMEIN H., GUIBON G., LABEAU M. ET AL. (2023). Participation de l'équipe ttgv à deft 2023˜ : Réponse automatique à des qcm issus d'examens en pharmacie. In 18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, p. 23–38 : ATALA.

BRUNING R. H., SCHRAW G. J. & NORBY M. M. (2011). Cognitive Psychology and Instruction. Pearson.

CHASE H. (2022). LangChain.

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.

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. (2024). Scaling instruction-finetuned language models. Journal of Machine Learning Research, **25**(70), 1–53.

DOUZE M., GUZHVA A., DENG C., JOHNSON J., SZILVASY G., MAZARÉ P.-E., LOMELI M., HOSSEINI L. & JÉGOU H. (2024). The faiss library.

FAVRE B. (2023). Lis@ deft'23 : les llms peuvent-ils répondre à des qcm ?(a) oui ;(b) non ;(c) je ne sais pas. In 18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, p. 46–56 : ATALA.

GAO Y., XIONG Y., GAO X., JIA K., PAN J., BI Y., DAI Y., SUN J., GUO Q., WANG M. & WANG H. (2024). Retrieval-augmented generation for large language models : A survey.

LABRAK Y., BAZOGE A., DAILLE B., DUFOUR R., MORIN E. & ROUVIER M. (2023a). Tâches et systèmes de détection automatique des réponses correctes dans des qcms liés au domaine médical : Présentation de la campagne deft 2023. In 18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, p. 57–67 : ATALA.

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. (2023b). Drbert : A robust pre-trained model in french for biomedical and clinical domains.

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.

NILS REIMERS I. G. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. https ://arxiv.org/abs/1908.10084.

ORMROD J. E. (2016). Human Learning. Pearson.

PAUK W. & OWENS R. J. Q. (2013). How to Study in College. Cengage Learning.

PAUKER J. D. (2011). Multiple choice questions : Tips and tricks. Medical Education Online.

PUERTO H., ŞAHIN G. & GUREVYCH I. (2023). MetaQA : Combining expert agents for multi-skill question answering. In A. VLACHOS & I. AUGENSTEIN, Éds., Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, p. 3566–3580, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.259.

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

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

SVINICKI M. D. & MCKEACHIE W. J. (2014). McKeachie's Teaching Tips. Cengage Learning.

TUCKMAN B. W. & KENNEDY G. J. (2011). The effect of motivation on test performance of first-year college students. Research in Higher Education, **52**, 361–373.

ZHENG C., ZHOU H., MENG F., ZHOU J. & HUANG M. (2024). Large language models are not robust multiple choice selectors.