LIMICS@DEFT'24 : Un mini-LLM peut-il tricher aux QCM de pharmacie en fouillant dans Wikipédia et NACHOS ?

Solène Delourme[1, *] Adam Remaki[1, *] Christel Gérardin[1] Pascal Vaillant[1]
Xavier Tannier[1] Brigitte Seroussi[1] Akram Redjdal[1]
(1) Sorbonne Université, Inserm, Université Sorbonne Paris-Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, Paris, France
solene.delourme@epita.fr, adam.remaki@etu.sorbonne-universite.fr, christel.gerardin@aphp.fr, vaillant@univ-paris13.fr, xavier.tannier@sorbonne-universite.fr, brigitte.seroussi@aphp.fr, akram.redjdal@sorbonne-universite.fr

*Both authors contributed equally

RÉSUMÉ

Ce papier explore deux approches pour répondre aux questions à choix multiples (QCM) de pharmacie du défi DEFT 2024 en utilisant des modèles de langue (LLMs) entraînés sur des données ouvertes avec moins de 3 milliards de paramètres. Les deux approches reposent sur l'architecture RAG (Retrieval Augmented Generation) pour combiner la récupération de contexte à partir de bases de connaissances externes (NACHOS et Wikipédia) avec la génération de réponses par le LLM Apollo-2B. La première approche traite directement les QCMs et génère les réponses en une seule étape, tandis que la seconde approche reformule les QCMs en questions binaires (Oui/Non) puis génère une réponse pour chaque question binaire. Cette dernière approche obtient un Exact Match Ratio de 14.7 et un Hamming Score de 51.6 sur le jeu de test, ce qui démontre le potentiel du RAG pour des tâches de Q/A sous de telles contraintes.

ABSTRACT

LIMICS@DEFT'24 : Can a mini-LLM cheat in pharmacy MCQs by searching Wikipédia and NACHOS

This paper investigates two approaches to tackle the DEFT 2024 pharmacy multiple-choice question (MCQ) answering challenge using language models (LLMs) trained on open data with less than 3 billion parameters. Both approaches rely on the Retrieval Augmented Generation (RAG) architecture to combine context retrieval from external knowledge bases (NACHOS and Wikipédia) with answer generation by the Apollo-2B LLM and a CamemBERT classifier. The first approach processes the MCQs directly and generates the answers in a single step, while the second approach first reformulates the MCQs into binary (Yes/No) questions and then generates an answer for each binary question. The latter approach obtains an Exact Match Ratio of 14.7 and a Hamming Score of 51.6 on the test set, highlighting the potential of RAG for Q/A tasks under such constraints.

MOTS-CLÉS : DEFT, LLM, RAG, prompt, LoRA, Apollo.

KEYWORDS: DEFT, LLM, RAG, prompt, LoRA, Apollo.

1 Introduction

Cette nouvelle édition 2024 du Défi Fouille de Textes (DEFT) propose, à l'instar de l'édition 2023, d'utiliser le corpus FrenchMedMCQA (Labrak *et al.*, 2022). Pour cette nouvelle édition, un nouveau corpus de test de 477 questions a été collecté. La tâche principale consiste à identifier les réponses correctes aux questions parmi les cinq options proposées. Toutefois, les systèmes proposés doivent respecter les contraintes suivantes :

1. Ne pas rechercher sur internet les originaux des données fournies.
2. Utiliser des modèles pré-entraînés dont les données d'entraînement sont ouvertes (i.e. ChatGPT, Mistral et autres modèles de ce type ne peuvent pas être utilisés).
3. Utiliser uniquement comme corpus additionnels NACHOS et Wikipédia.
4. Faire moins de 3 milliards de paramètres.

La tâche secondaire est identique à la tâche principale mais il n'y a aucune limite sur la taille des modèles. Avec l'ajout d'un nouveau corpus de test de 477 questions, nous avons décidé de répartir le corpus FrenchMedMCQA en deux sous-parties : le corpus d'entraînement (2171 questions, 70% du total) et le corpus de développement (934 questions, 30% du total).

L'édition DEFT 2023 (Favre, 2023) a mis en évidence la performance des grands modèles de langues (LLMs) tels que GPT ou Llama. Cependant, cette année, les participants doivent utiliser des systèmes ne dépassant pas 3 milliards de paramètres et dont les données de pré-entraînement sont ouvertes, ce qui exclut les grands LLMs de la tâche principale. L'objectif de cette année est d'utiliser des petits LLMs, parmi ces modèles on retrouve Appolo-2B (Wang *et al.*, 2024a), BLOOM-3B (Scao *et al.*, 2022) ou encore CroissantLLM (Faysse *et al.*, 2024) qui remplissent les exigences du défi. Ces modèles ont été entrainés sur des données multilingues incluant le français, ce qui est nécessaire pour le traitement du corpus du défi FrenchMedMCQA, ce qui nous a poussé à les évaluer. La contrainte sur le faible nombre de paramètres de notre système nous pousse également à utiliser une base de connaissance externe pour aider le système à répondre aux questions. Cette fouille en comparant les documents préalablement convertis en embeddings à l'aide d'un modèle de langue de type SentenceBERT (Reimers & Gurevych, 2019) qui est un modèle de type BERT affiné pour la comparaison de phrases.

2 Méthodes

2.1 Détection du type de question

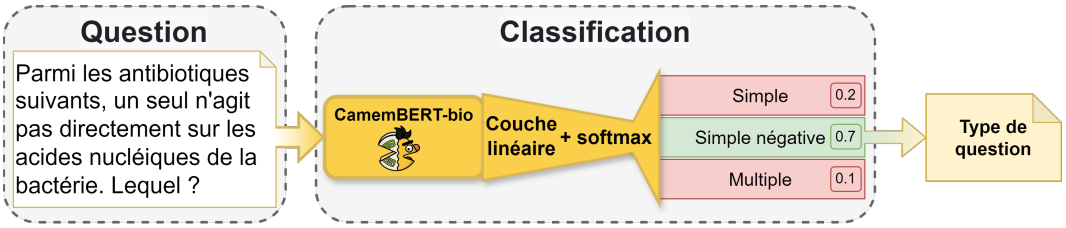


FIGURE 1 – Méthode de détection du type de question

La détection du type de question est cruciale pour le système, car elle guide la fouille de contexte et l'évaluation des réponses aux questions. Comme le montre la Figure 1, nous avons utilisé un modèle de classification basé sur ***CamemBERT-bio*** (Touchent *et al.*, 2023). Nous distinguons les questions en trois types : "simple", "multiple" et "simple négative", cette dernière catégorie faisant référence aux questions pour lesquelles le choix de réponse est négatif (cf. exemple de la Figure 1).

Les corpus d'entraînement et d'évaluation ont été annotés manuellement pour déterminer les questions de type "simple", "multiple" et "simple-négative". Ensuite, le modèle a été entraîné et évalué en utilisant des techniques standards. Étant donné le déséquilibre des classes ("multiple" 72.8%, "simple" 16.8%, "simple-négative" 10.4% comme on peut le voir sur la Figure 7 en annexe), les poids des classes ont été intégrés dans une fonction de perte personnalisée pour gérer cette distribution inégale. Sur le jeu de validation, nous avons obtenu une précision de 0.95, un rappel de 0.96 et un F1-score de 0.95. Cette méthode permet de classer correctement les questions, facilitant ainsi les étapes suivantes du pipeline.

2.2 Fouille du contexte

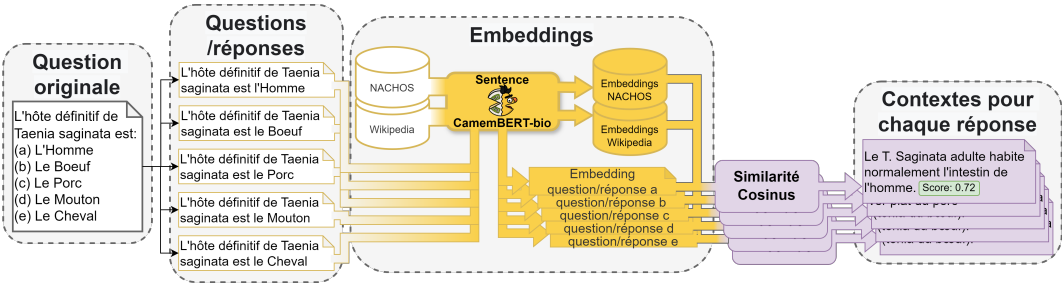


FIGURE 2 – Méthode de récupération du contexte

Comme décrit dans la Figure 2, nous avons suivi les étapes suivantes pour récupérer le contexte d'une question :

Choisir la base de connaissance DEFT 2024 nous propose d'utiliser les corpus Wikipédia et NACHOS (Labrak *et al.*, 2023). Comme il s'agit d'un QCM en français sur le thème médical, nous nous sommes restreints aux pages Wikipédia françaises dans la catégorie *Portail :Sciences/Articles liés* (202 933 pages) et l'intégralité du corpus NACHOS (1 698 029 documents).

Choisir le modèle d'embedding Pour la création d'embedding, nous avons affiné le modèle *CamemBERT-bio* sur la base de données STS Benchmark en français avec l'algorithme d'affinage de *SentenceBERT* (Reimers & Gurevych, 2019). *CamemBERT-bio* (Touchent *et al.*, 2023) est un modèle de langue de type BERT pré-entraîné sur du texte médical en Français. Le STS Benchmark (May, 2021) est une collection de 8630 paires de phrases extraites de légendes d'images, de titres d'actualités et de forums. À chaque paire de phrases est associée un score de similarité entre 0 et 5. Ces données sont disponibles dans 10 langues dont le français grâce à une traduction de l'anglais par les auteurs du jeu de données. De plus, nous avons utilisé une méthode d'augmentation des données (Thakur *et al.*, 2020) en utilisant le modèle *CrossEncoder CamemBERT-large*. Cette méthode consiste à former de nouvelles combinaisons de paires de phrases à partir du jeu de données et de leur attribuer un score de similarité *silver standard* prédit par le modèle *CrossEncoder CamemBERT-large* affiné sur le jeu

de donnée qui est plus performant mais trop lent pour être utilisé dans notre système de fouille. Le jeu de donnée a été augmenté de 31 149 paires de phrases, le modèle a été entraîné sur 10 époques avec une taille de *batch* de 16 et une optimisation par la méthode AdamW.

Convertir la question en embedding En première approche, nous avons converti en embedding l'intégralité de la question et des réponses. Cependant, dans le cas des questions de type "multiple", nous avons remarqué que les phrases de contexte en sortie du système étaient le plus souvent relatives à une seule réponse. Ainsi, comme le montre la Figure 2, nous avons accolé séparément chaque réponse à la question correspondante pour couvrir toutes les réponses de la question. les 5 options de réponses sont donc converties séparément en embedding de taille 768 à l'aide du modèle *Sentence CamemBERT-bio*.

Convertir la base de connaissance en embedding Nous avons divisé NACHOS en 2 376 553 parties de 1000 lignes chacune et Wikipédia en 2 752 036 paragraphes. Afin de conserver l'information du titre au sein de chaque paragraphe, nous avons accolé le titre de la page Wikipédia au début de chaque paragraphe. Chaque partie a ensuite été tronquée aux 512 premiers tokens avant d'être convertie en embedding par le modèle *Sentence CamemBERT-bio* dont la fenêtre de contexte maximale est de 512 tokens.

Classer les contextes Pour chaque question/réponse, la fouille du contexte se fait en 2 étapes : (i) L'embedding de la question/réponse est comparé avec les embeddings de Wikipédia et NACHOS par similarité cosinus. Les 100 parties de Wikipédia et les 100 parties de NACHOS avec le meilleur score sont sélectionnées. Ces 200 parties sont divisées en phrases et chaque phrase est convertie en embedding par *Sentence CamemBERT-bio*. (ii) L'embedding de la question/réponse est comparé par similarité cosinus avec les embeddings de phrases sélectionnées à la première étape. Le score de similarité ainsi calculé permet de classer les phrases les plus similaires.

2.3 Premier système : Architecture RAG sur questions à choix multiples

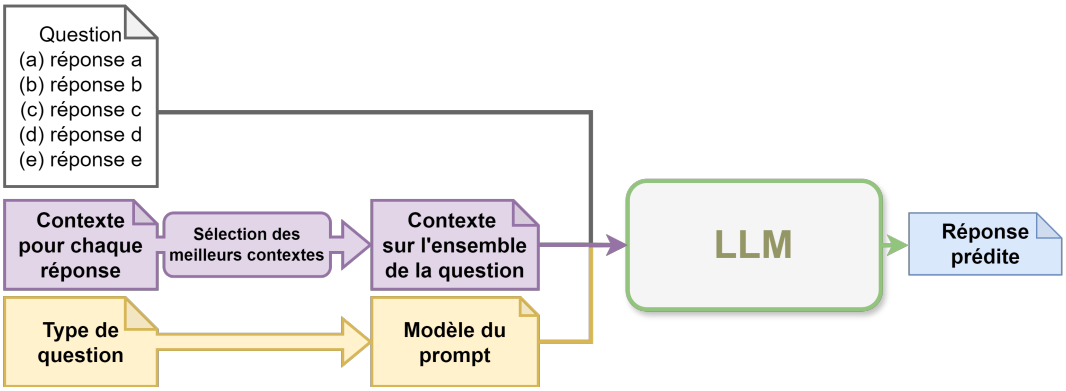


FIGURE 3 – Architecture RAG sur questions à choix multiples

À partir d'une question, nous avons préalablement prédit le type de la question (Section 2.1) et classé dans Wikipédia et NACHOS les phrases les plus similaires à chaque question/réponse (Section 2.2).

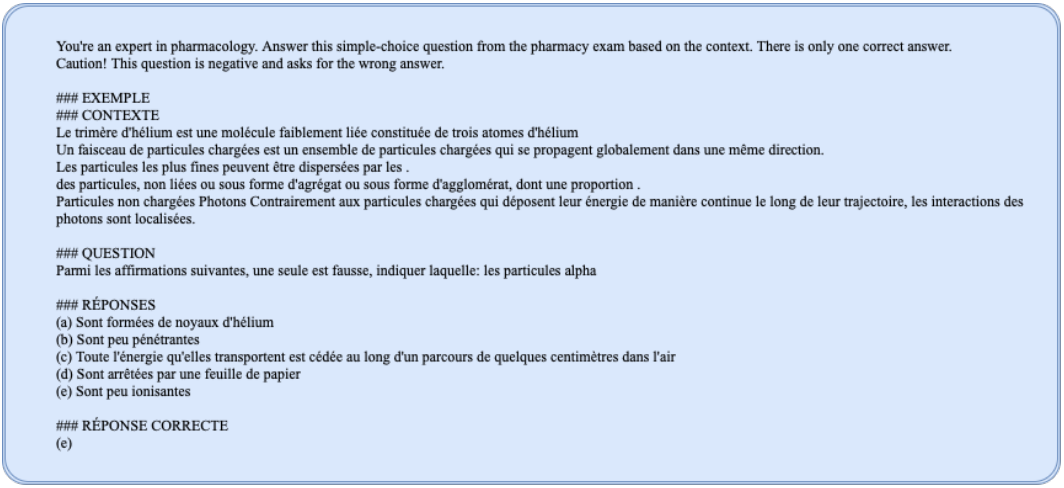
Comme le montre la Figure 3, notre premier système de génération augmentée de récupération (retrieval-augmented generation, RAG) consiste à fournir au LLM, la question accompagnée d'un prompt spécifique au type de question et d'une sélection des meilleurs contextes.

Sélectionner les meilleurs contextes Pour chaque question/réponse, nous avons gardé uniquement les 3 phrases les plus similaires à condition qu'elles aient un score de similarité supérieures à 0.7 pour une question de type "simple" ou "multiple" et 0.65 pour une question de type "simple-négative". Nous supposons qu'il faut plus de contexte pour éliminer les mauvaises réponses dans une question "simple-négative" que pour simplement trouver la ou les bonnes réponses dans une question "simple" ou "multiple". De plus, nous nous sommes assurés de supprimer les phrases qui sont trop proches à l'aide d'un modèle TF-IDF avec un seuil de similarité de 0.9. Au maximum, si pour chaque réponse, les 3 phrases les plus similaires ont un score supérieur au seuil et ne sont pas trop proches alors le contexte de la question sera formé de 15 phrases.

2.3.1 *Few-shot learning*

Pour augmenter l'efficacité du modèle, nous avons adopté une approche de *few-shot learning*. Cela consiste à fournir une série d'exemples au modèle avec les réponses afin qu'il apprenne à générer des réponses plus fidèles au format attendu. Pour cette tâche, nous avons utilisé le modèle *Apollo-2B* (Wang *et al.*, 2024a). Nous avons conçu des prompts spécifiques pour les trois types de questions : "simple", "simple-négative" et "multiple".

Lors de nos premiers tests, le *zero-shot* n'a pas donné de bons résultats, surtout pour le format des réponses. Nous avons donc fourni des exemples au LLM. Bien qu'il y ait des données en français dans le corpus d'entraînement du modèle *Apollo*, la majorité est en anglais (Wang *et al.*, 2024b). Le contexte ayant été extrait de données en français, nous gardons le contexte en français, la question et les options en français mais nous donnons des instructions en anglais pour améliorer le prompt.


You're an expert in pharmacology. Answer this simple-choice question from the pharmacy exam based on the context. There is only one correct answer.
Caution! This question is negative and asks for the wrong answer.

EXEMPLE
CONTEXTE
Le trimère d'hélium est une molécule faiblement liée constituée de trois atomes d'hélium
Un faisceau de particules chargées est un ensemble de particules chargées qui se propagent globalement dans une même direction.
Les particules les plus fines peuvent être dispersées par les .
des particules, non liées ou sous forme d'agrégat ou sous forme d'agglomérat, dont une proportion .
Particules non chargées Photons Contrairement aux particules chargées qui déposent leur énergie de manière continue le long de leur trajectoire, les interactions des photons sont localisées.

QUESTION
Parmi les affirmations suivantes, une seule est fausse, indiquer laquelle: les particules alpha

RÉPONSES
(a) Sont formées de noyaux d'hélium
(b) Sont peu pénétrantes
(c) Toute l'énergie qu'elles transportent est cédée au long d'un parcours de quelques centimètres dans l'air
(d) Sont arrêtées par une feuille de papier
(e) Sont peu ionisantes

RÉPONSE CORRECTE
(e)

FIGURE 4 – Prompt choix simple négative

Pour les questions attendant une unique bonne réponse, les questions de type "simple" et "simple-

négative", nous avons fourni un seul exemple. Pour les questions de type "multiple", attendant plus d'une réponse correcte, nous avons fourni des exemples illustrant toutes les possibilités de réponses correctes, à savoir deux, trois, quatre ou cinq. Cela permet au modèle de mieux comprendre la structure des réponses attendues.La Figure 4 donne un exemple de prompt pour les questions de type "simple-négative". Un exemple de prompt pour les questions de type "simple" et "multiple" se trouve en annexe (Figure 8 et Figure 9).

Après la préparation des prompts, nous avons inséré la sélection des meilleurs contextes, suivi de la question et des options de réponses possibles. La génération est effectuée sur un maximum de 32 nouveaux tokens, une température de 0.001 et des valeurs top-p et top-k de 0.90 et 40 respectivement.

2.3.2 Affinage supervisé

Nous avons affiné le LLM sur les questions du corpus d'entraînement avec la méthode d'adaptation à faible rang (Low Rank Adaptation, LoRA) (Hu *et al.*, 2021). En effet, l'affinage des LLM nécessite beaucoup de mémoire GPU, mais LoRA permet de contourner cette contrainte en gelant le modèle original et en ajoutant des matrices de faible rang obtenues par le produit de deux petites matrices. Cela réduit le nombre de nouveaux paramètres et la mémoire requise. Cela repose sur l'idée que tous les paramètres du modèle n'ont pas besoin d'être modifiés lors de l'affinage avec relativement peu d'exemples. Nous avons affiné le modèle sur 1 époque avec une taille de *batch* de 24, un taux d'apprentissage de $3 \cdot 10^{-4}$, un *dropout* de 0.1 et une optimisation par la méthode AdamW avec un *warmup* de 3 % des *batchs*. Pour la méthode LoRA, nous avons choisi de modifier seulement les matrices de paramètres de projection de la clé et la valeur du mécanisme d'attention (q_{proj} et v_{proj} dans le modèle) avec $r = 8$ (rang des matrices de paramètres) et $\alpha = 16$. En effet, la méthode fonctionne aussi bien en se restreignant seulement aux matrices q_{proj} et v_{proj} avec un rang faible (Hu *et al.*, 2021).

2.4 Second système : Architecture RAG sur les questions binaires reformulées

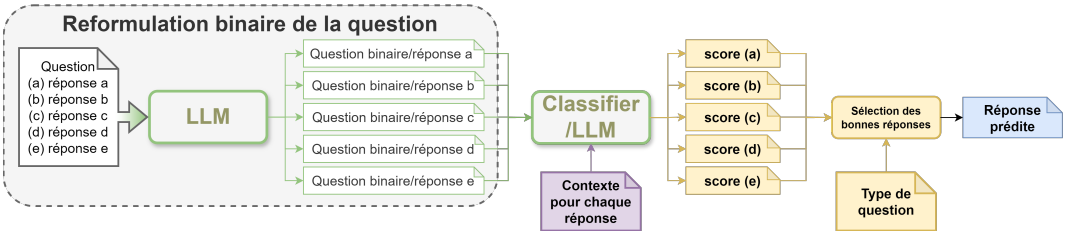


FIGURE 5 – Architecture RAG sur les questions binaires reformulées

Afin d'éviter de demander au LLM de répondre par plusieurs choix de réponses et pour faciliter la reconnaissance des réponses, nous avons développé une méthode se basant sur la reformulation des questions pour créer des questions fermées (avec seulement Oui/Non comme réponse possible). Ces questions sont appelées questions binaires.

Comme indiqué en Figure 5 nous commençons par la reformulation des questions via un LLM. Puis nous utilisons deux méthodes pour répondre aux questions : un classifieur binaire versus un LLM.

Reformulation du QCM en questions binaires Pour cette partie, nous avons utilisé le modèle *Apollo-2B*. Nous avons reformulé chaque question et chaque option de réponse individuellement en une question fermée à laquelle on peut seulement répondre par oui ou par non comme dans l'exemple présenté Figure 6. Cela permet de traiter chaque possibilité de réponse de manière isolée, facilitant ainsi l'évaluation de chaque réponse potentielle par le modèle. Comme illustré dans la Figure 10, nous avons spécifié au modèle de ne pas produire de questions sous forme négative afin de faciliter le post-traitement des réponses et d'identifier plus facilement la ou les réponse(s) correcte(s) de la question initiale selon son type.

Reformuler les questions via un LLM permet de simplifier le processus de réponse en transformant des questions à choix multiples, souvent complexes, en questions fermées (Oui/Non). Répondre directement à des questions à choix multiples nécessite une compréhension et un traitement approfondis de chaque option de réponse, ce qui peut être plus complexe et moins précis pour les petits LLMs.

```
### QUESTION
Parmi les affirmations suivantes, une seule est fausse, indiquer laquelle: les particules alpha

### OPTION
Sont formées de noyaux d'hélium

### QUESTION REFORMULÉE
Les particules alpha sont-elles formées de noyaux d'hélium ?
```

FIGURE 6 – Exemple question reformulée

Répondre aux questions binaires avec un classifier Pour répondre aux questions binaires reformulées, nous avons commencé par utiliser un modèle de classification basé sur *CamemBERT-bio* et *CamemBERT-base*. Chaque question est accompagnée du classement des phrases de contextes les plus similaires (voir Section 2.2). Les trois les plus similaires sont sélectionnés et concaténés pour former un contexte consolidé. Le modèle de classification est ensuite entraîné sur le corpus d'entraînement. Après l'entraînement, le modèle est utilisé pour prédire les réponses aux questions binarisées du jeu de test, en générant des probabilités pour les réponses "Oui" et "Non". Les prédictions finales sont déterminées en fonction des probabilités générées par le modèle, permettant une évaluation précise des réponses binaires.

Répondre aux questions binaires avec un LLM Nous avons utilisé pour cette méthode le modèle *Apollo-2B* (Wang *et al.*, 2024a). Les prompts utilisés sont similaires à ceux choisis pour les questions de type "simple" (voir Section 2.3.1) comme le montre la Figure 11 en annexe. Chaque prompt contient le contexte de la question, suivi de la question reformulée en question binaire finissant par Oui/Non. Les réponses générées par le modèle sont analysées pour identifier les "Oui" ou "Non" et les probabilités associées.

Post-traitement des réponses Les réponses finales sont déterminées en fonction du type des questions. Voici comment nous procédons :

— **Questions de type "simple"** : La réponse avec la probabilité la plus élevée est sélectionnée.
— **Question de type "simple-négative"** : La réponse avec la probabilité la plus faible est choisie.

— **Questions de type "multiple"** : Les réponses dont la probabilité de signifier un "Oui" dépasse un seuil de 0.5 sont retenues.

Ce post-traitement s'applique aussi bien au LLM qu'au classifieur, permettant la sélection des réponses finales en tenant compte des spécificités de chaque type de question.

3 Résultats

Le Tableau 1 présente les résultats du premier système. Ce système décrit dans la section 2.3 est basé sur *Sentence CamemBERT-bio* pour retrouver les contextes, *CamemBERT-bio* pour détecter le type des questions et un LLM pour générer les réponses. Pour la tâche principale, nous avons utilisé le LLM *Apollo 2B* pour générer les réponses avec trois méthodes différentes : (i) La méthode *few-shot learning*. (ii) l'affinage supervisé LoRA (iii) les deux méthodes. Pour répondre à la tâche annexe, nous avons utilisé le modèle *Apollo 7B* pour générer la réponse avec la méthode *few-shot learning*. Enfin, pour évaluer l'apport du RAG, nous présentons sur les dernières lignes du tableau, les performances du système sans l'architecture RAG (sans utilisation du contexte).

Modèle	Few-shot	LoRA	EMR	Hamming
Apollo 2B	✓		13.8 [10.9, 17.2]	47.8 [45.0, 50.5]
Apollo 2B		✓	11.7 [9.0, 14.9]	44.8 [42.1, 47.6]
Apollo 2B	✓	✓	11.5 [8.8, 14.7]	45.5 [42.8, 48.4]
Apollo 7B	✓		21.4 [17.8, 25.2]	56.8 [53.8, 59.5]
Apollo 2B sans RAG	✓		10.7 [8.2, 13.6]	45.7 [43.1, 48.5]
Apollo 2B sans RAG		✓	11.1 [8.6, 14.3]	39.7 [37.0, 42.5]

TABLE 1 – Performance du premier système. un intervalle de confiance à 95% obetnu à partir d'une méthode de bootstraping est indiqué entre crochets.

Le Tableau 2 présente les résultats du second système. Ce système décrit dans la section 2.4 est basés sur *Apollo 2B* pour reformuler la question, *Sentence CamemBERT-bio* pour retrouver le contexte, *CamemBERT-bio* pour détecter le type de la question. Pour générer la réponse, nous avons expérimenté le LLM *Apollo 2B* avec La méthode *few-shot learning*, avec l'affinage supervisé LoRA ou avec les deux méthodes. Nous avons également expérimenté une classification binaire avec le modèle *CamemBERT-bio* ou *CamemBERT-base*.

Modèle	Few-shot	Affinage	EMR	Hamming
Apollo 2B	✓		**14.7 [11.7, 18.0]**	51.6 [48.9, 54.4]
Apollo 2B		✓	14.0 [11.1, 17.4]	46.9 [44.1, 49.9]
Apollo 2B	✓	✓	7.1 [5.0, 9.6]	52.0 [49.6, 54.4]
CamemBERT-bio		✓	8.2 [6.1, 10.9]	42.4 [39.6, 45.1]
CamemBERT-base		✓	9.4 [6.9, 12.4]	44.1 [41.4, 46.8]

TABLE 2 – Performance du second système avec reformulation des questions. Un intervalle de confiance à 95 % obtenu à partir d'une méthode de bootstraping est indiqué entre crochets.

En sommant les nombres de paramètres des modèles listés dans le Tableau 3 (à l'exception d'*Apollo-7B* qui est utilisé pour la tâche annexe), on obtient environ 2,8 milliards de paramètres ce qui respecte

bien la contrainte des 3 milliards de paramètres.

Modèle	Nombre de paramètres
CamemBERT-bio	110 655 493
CamemBERT-base	110 655 493
Sentence CamemBERT-bio	110 621 952
Apollo 2B	2 506 172 416
Apollo 7B	8 537 680 896

TABLE 3 – Nombre de paramètres des modèles utilisés

Lors de l'évaluation officielle, nous avions utilisé le premier système avec le modèle *Apollo 2B* pour générer les réponses et la méthode *few-shot learning* (cela correspond à la première ligne du Tableau 1). Nous avions obtenu un Hamming de 45.71 et un EMR de 11.74. Ce résultat est inférieur à celui présenté dans notre étude car au moment de l'évaluation officielle, l'algorithme de récupération du contexte Wikipédia utilisait seulement les titres des pages. Une meilleure optimisation mémoire nous a permis de résoudre ce problème et d'obtenir de meilleures performances globales.

4 Discussion & conclusion

En comparaison avec l'édition DEFT 2023 (Favre, 2023), où des grands modèles de langue comme *GPT* et *Llama* ont montrés des performances remarquables, cette édition a mis en lumière le potentiel des petits modèles entraînés sur des données ouvertes. Dans le cadre de la tâche principale, le modèle *Apollo 2B* avec *few-shot learning* sans l'architecture RAG a obtenu un EMR de 10.7 et un Hamming Score de 45.7. Le même modèle avec l'architecture RAG a atteint un EMR de 13.8 et un Hamming Score de 47.8. Enfin, la reformulation des questions en questions binaires, a permis d'améliorer encore les performances, avec un EMR de 14.7 pour le même modèle *Apollo 2B* utilisant le *few-shot learning*. Par ailleurs, sur la tâche secondaire, le modèle *Apollo 7B*, qui dépasse les contraintes de taille, a montré des performances supérieures avec un EMR de 21.4 et un Hamming Score de 56.8. Ce qui démontre que malgré les efforts effectués (reformulation des questions, utilisation de classifieurs ...etc), un modèle avec plus de paramètres arrive facilement à obtenir de meilleurs résultats.

Ces résultats montrent que l'utilisation de l'architecture RAG intégrant du contexte pertinent issu de bases de connaissances telles que NACHOS et Wikipédia améliore les performances sur la tâche principale. En effet, en regardant manuellement les réponses ou le LLM s'est trompé, il s'agissait souvent de questions où le contexte extrait ne fournissait pas les éléments nécessaires pour répondre à la question (cf. Figure 12 en Annexe). Les résultats montrent aussi que l'utilisation de techniques comme le *few-shot learning* ou la reforumlation des questions en questions binaires améliorent aussi les performances sur la tâche principale. Toutefois, le résultat sur la tâche secondaire montre que même si les méthodes utilisées (reformulation des questions, utilisation de classifieurs ...etc) semblent contribuer à augmenter légèrement les performances, cela n'égale pas l'utilisation d'un plus grand modèle de langue. L'affinage du modèle avec la méthode LoRA ne montre pas de gains de performances par rapport à la méthode *few-shot learning*.

En conclusion, cette étude a permis de montrer le potentiel de l'architecture RAG pour améliorer les résultats des petits modèles de langue dans des tâches de questions/réponses. Cette étude a permis également la publication du modèle *Sentence CamemBERT-bio* spécialisé dans la comparaison de

phrases en Français dans le domaine biomédical. Tout cela ouvre de nouvelles perspectives pour la recherche et les applications pratiques dans le domaine de la pharmacie.

Il est important de noter que même si cette étude montre des résultats encourageants, de nombreuses pistes n'ont pas pu être explorées dans le temps imparti au DEFT. Nous n'avons pas exploré de manière approfondie l'espace des hyperparamètres pour l'affinage LoRA (époques, rang r, facteur d'échelle α...etc), cela pourrait expliquer son inefficacité dans cette étude. En ce qui concerne le système de fouille du contexte, nous n'avons pas explorer les paramètres liés aux partitionnements des corpus additionnels (taille des partitions) ou à la sélection des meilleurs contextes (seuils des scores, nombre de phrases sélectionnées). Pour l'affinage du modèle *Sentence CamemBERT-bio* nous avons utilisé un jeu de données généraliste, l'utilisation de jeux de données dans le domaine biomédical comme CLISTER (Hiebel *et al.*, 2022) ou DEFT-2020 (Cardon *et al.*, 2020) pourrait améliorer les performances du modèle sur la tâche. Enfin, nous n'avons pas comparé notre modèle *Sentence CamemBERT-bio* à d'autres modèles de type BERT français comme *Sentence CamemBERT-large* ou multilingues comme *distiluse-base-multilingual-cased-v2* ou même à des modèles de sac de mots comme TF-IDF ou sa variante BM25 qui ont l'avantage de ne pas avoir de limite de contexte à 512 tokens. Nous n'avons pas évalué la méthode de détection du type de question à trois classes en la comparant avec une méthode plus simple à deux classes "simple" et "multiple".

Reproductibilité

Le code développé pour réaliser les expériences de cette étude est disponible sur GitHub à l'adresse suivante : https://github.com/Aremaki/deft_2024. Le modèle *Sentence CamemBERT-bio* décrit dans la Section 2.2, utilisé pour récupérer le contexte des questions, est disponible sur la plateforme HuggingFace via le lien suivant : https://huggingface.co/Aremaki/sentence-camembert-bio. Les données du DEFT 2024, avec les annotations manuelles des questions de type "simple-négative" décrites Section 2.1, sont disponibles sur HuggingFace via le lien suivant : https://huggingface.co/datasets/Aremaki/DEFT_2024.

Références

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

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation deft 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *DEFT 2020*.

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

FAVRE B. (2023). LIS@DEFT'23 : les LLMs peuvent-ils répondre à des QCM ? (a) oui ; (b) non ; (c) je ne sais pas. In A. BAZOGE, B. DAILLE, R. DUFOUR, Y. LABRAK, E. MORIN & M. ROUVIER, Éds., *Actes de CORIA-TALN 2023. Actes du Défi Fouille de Textes@TALN2023*, p. 46–56, Paris, France : ATALA.

FAYSSE M., FERNANDES P., GUERREIRO N. M., LOISON A., ALVES D. M., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P. H., CASADEMUNT A. B., YVON F., MARTINS A. F. T., VIAUD G., HUDELOT C. & COLOMBO P. (2024). CroissantLLM : A Truly Bilingual French-English Language Model. arXiv :2402.00786 [cs], DOI : 10.48550/arXiv.2402.00786.

HIEBEL N., FORT K., NÉVÉOL A. & FERRET O. (2022). Clister : Un corpus pour la similarité sémantique textuelle dans des cas cliniques en français (clister : A corpus for semantic textual similarity in french clinical narratives). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 287–296.

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 :2106.09685 [cs], DOI : 10.48550/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.

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.

MAY P. (2021). Machine translated multilingual sts benchmark dataset.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. arXiv :1908.10084 [cs], DOI : 10.48550/arXiv.1908.10084.

SCAO T., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A., YVON F., GALLÉ M., TOW J., RUSH A., BIDERMAN S., WEBSON A., AMMANAMANCHI P., WANG T., SAGOT B., MUENNIGHOFF N., MORAL A. & WOLF T. (2022). Bloom : A 176b-parameter open-access multilingual language model.

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

THAKUR N., REIMERS N., DAXENBERGER J. & GUREVYCH I. (2020). Augmented SBERT : data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *CoRR*, **abs/2010.08240**.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2023). Camembert-bio : a tasty french language model better for your health.

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

WANG X., CHEN N., CHEN J., HU Y., WANG Y., WU X., GAO A., WAN X., LI H. & WANG B. (2024b). Apollo : An Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. arXiv :2403.03640 [cs], DOI : 10.48550/arXiv.2403.03640.

5 Annexe

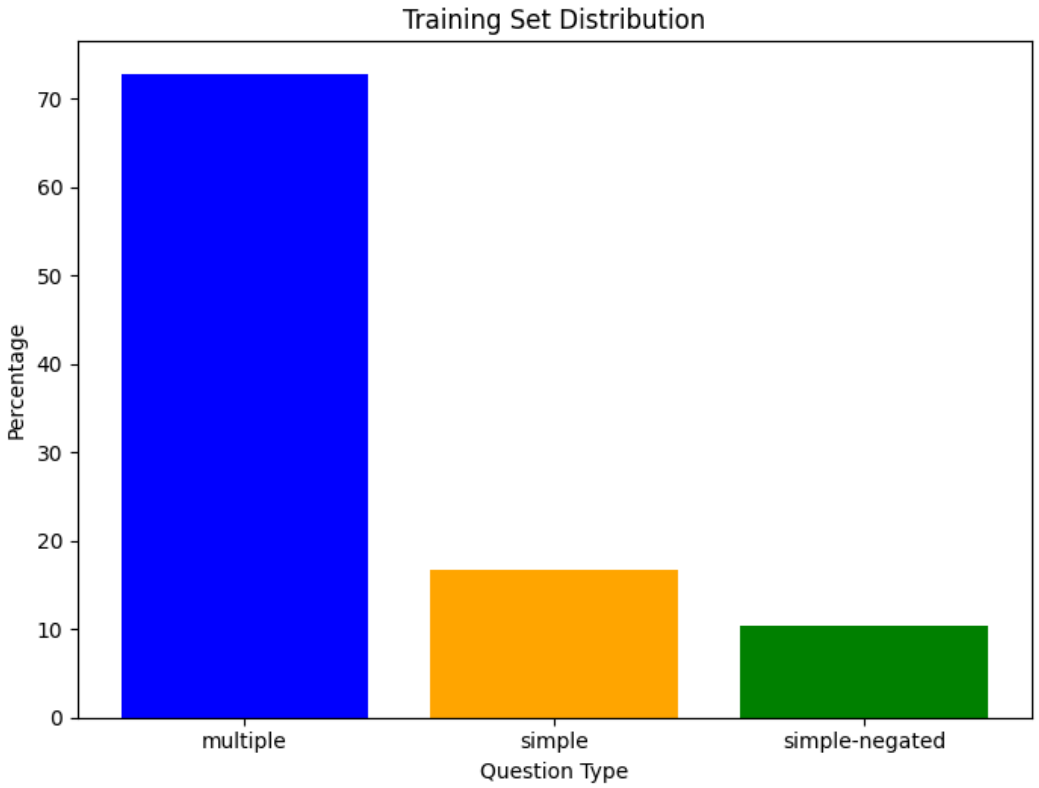


FIGURE 7 – Distribution des types de questions sur les jeu d'entraînements

You're an expert in pharmacology. Answer this simple-choice question from the pharmacy exam based on the context. There is only one correct answer.

EXEMPLE
CONTEXTE
Escherichia coli, en abrégée E. coli, est une bactérie intestinale des organismes à sang chaud1, Gram négatif, du genre Escherichia, en forme de bâtonnet appelée parfois colibacille. E. coli est une bactérie aéro-anaérobie. E. coli constitue, avec d'autres bactéries, 0,1% du microbiote intestinal2. Les bactéries Escherichia coli ont été observées pour la première fois dans des fèces de nourrissons en 1885 par l'allemand Theodore Escherich et ont été nommées en hommage à ce dernier en 1919 par Castellani et Chambert3, c'est un coliforme thermotolérant (fécal) généralement commensal. La plupart des souches sont inoffensives, voire bénéfiques pour l'être humain, produisant de la vitamine K4 ou empêchant la colonisation de l'intestin par des bactéries pathogènes, établissant alors une relation de type mutualiste. Cependant, certains sérotypes d'E. coli peuvent être pathogènes, entraînant alors des gastro-entérites, infections urinaires, méningites ou sepsis.

QUESTION
Parmi les propositions suivantes, quelle est celle qui s'applique à Escherichia coli?

RÉPONSES
(a) C'est une bactérie anaérobie stricte
(b) Il peut être responsable de méningites
(c) C'est une bactérie toujours immobile
(d) C'est une bactérie oxydase positive
(e) Il est toujours résistant aux aminopénicillines

RÉPONSE CORRECTE
(b)

FIGURE 8 – Prompt des questions de type "simple"

You're an expert in pharmacology. Answer this multiple-choice question from the pharmacy exam based on the context. There is more than one correct answer.

EXEMPLE
CONTEXTE
L'administration simultanée d'amitriptyline et d'IMAO peut entraîner un syndrome sérotoninergique (association de symptômes, dont notamment agitation, confusion, tremblements, myoclonie et hyperthermie).
Effets Lorsqu'une quantité suffisante d'anticholinergique est en circulation dans le corps, un toxidrome (intoxication) appelé syndrome anticholinergique aigu peut se produire.
Précautions d'emploi Convulsions Des cas de convulsions ont été rapportés en association avec un traitement par l'acide tranexamique.
En effet une augmentation du risque de bradycardies, d'hypotension et d'accident vasculaire incluant la mise en jeu du pronostic vital chez les patients ayant des facteurs de risque cardiovasculaire a été observée dans des essais thérapeutiques.

QUESTION
Concernant l'intoxication aiguë par l'amitriptyline, quelle(s) est (sont la (les) proposition(s) exacte(s) ?

RÉPONSES
(a) Un syndrome sérotoninergique est observé
(b) Un syndrome anticholinergique est observé
(c) Des convulsions peuvent être observées
(d) Le pronostic dépend des troubles cardiovasculaires
(e) La dose toxique est supérieure à 5 grammes chez l'adulte

RÉPONSE CORRECTE
(a),(b),(c),(d)

EXEMPLE
CONTEXTE
Le dosage de l'ostéocalcine.
La phosphatase alcaline spécifique osseuse et l'un des nouveaux marqueurs biologiques de la résorption osseuse, la pyridinoline (Pyr) urinaire qui reflète la dégradation du collagène osseux et cartilagineux cette série de 62 patients.
Elle se distingue des autres phosphatases acides de mammifères par sa masse moléculaire et sa résistance à l'acide tartrique.
Definition d'un site "acylophile" dans l'uridine diphosphate glucose-glycogène transférase.
L'effet de la calcitonine peut être suivi par la mesure de marqueurs appropriés du remodelage osseux tels que les phosphatases alcalines sériques ou l'hydroxyproline ou la déoxypyridinoline urinaires.

QUESTION
Parmi les marqueurs suivants, indiquer celui (ceux) qui reflète(nt) l'activité ostéoblastique.

RÉPONSES
(a) Ostéocalcine (sérique)
(b) Pyridinolines (urinaires)
(c) Phosphatase acide résistant à l'acide tartrique (sérique)
(d) Glycosides d'hydroxylysine (urinaires)
(e) Phosphatase alcaline osseuse (sérique)

RÉPONSE CORRECTE
(a),(e)

EXEMPLE
CONTEXTE
La chromatographie en phase inverse, ou plus précisément chromatographie de partage à polarité de phases inversée, est une méthode physico-chimique très utilisée en biochimie et visant à séparer les constituants d'un mélange en fonction de leur polarité.
Parmi les différentes phases stationnaires aujourd'hui utilisées en HPLC, on retrouve la phase inverse (RP) et la phase normale (NP), mais aucune de ces deux phases ne permet de séparer efficacement les molécules polaires.
Principe\nLa chromatographie d'absorption est une chromatographie avec une phase stationnaire solide avec des propriétés absorbantes et une phase mobile liquide (mélange de solvants possible).La chromatographie en phase inverse, ou plus précisément chromatographie de partage à polarité de phases inversée, est une méthode physico-chimique très utilisée en biochimie et visant à séparer les constituants d'un mélange en fonction de leur polarité.
Elle est utilisée comme méthode d'ionisation pour la spectrométrie de masse, préférentiellement couplée à une chromatographie en phase liquide.
La séparation est réalisée par CLHP en phase inversée sur une colonne de silice greffée C18 et une phase mobile.

QUESTION
Parmi les propositions suivantes, laquelle ou lesquelles est (sont) exacte(s)? La chromatographie à polarité de phases inversée

RÉPONSES
(a) Utilise une phase stationnaire polaire
(b) Utilise une phase mobile hydro-organique
(c) Peut être appliquée aux espèces ionisées
(d) Élu en premier les solutés les plus polaires
(e) Utilise une phase stationnaire en silice vierge

RÉPONSE CORRECTE
(b),(c),(d)

FIGURE 9 – Prompt des questions de type "multiple"

You're an expert in pharmacology. Answer this simple-choice question from the pharmacy exam based on the context. There is only one correct answer.

EXAMPLE
###CONTEXT
La demi- vie d' élimination de la clomipramine après administration intraveineuse est de 6,4 heures pour la clomipramine, et de 3,6 heures pour la déméthylclomipramine.

QUESTION
La clomipramine (ANAFRANIL®) a-t-elle une demi-vie d'élimination de 6h ? (Oui/Non)

ANSWER
Oui

EXAMPLE
CONTEXT
La morphine est un analgésique opioïde utilisé pour soulager la douleur sévère. Elle peut provoquer des effets secondaires tels que la somnolence, la constipation et des nausées.

QUESTION
La morphine peut-elle provoquer des nausées comme effet secondaire ? (Oui/Non)

ANSWER
Oui

EXAMPLE
CONTEXT
L'aspirine, également connue sous le nom d'acide acétylsalicylique, est un médicament couramment utilisé pour soulager la douleur légère à modérée, réduire la fièvre et diminuer l'inflammation.

QUESTION
L'aspirine est-elle utilisée pour traiter les infections bactériennes ? (Oui/Non)

ANSWER
Non

EXAMPLE
CONTEXT
La metformine est un médicament utilisé pour traiter le diabète de type 2. Elle aide à contrôler la glycémie en diminuant la production de glucose par le foie et en augmentant la sensibilité à l'insuline.

QUESTION
La metformine est-elle utilisée pour traiter le diabète de type 2 ? (Oui/Non)

ANSWER
Oui

EXAMPLE
CONTEXT
Le clopidogrel (Plavix®) est un médicament antiplaquettaire utilisé pour prévenir les événements thromboemboliques tels que les crises cardiaques et les accidents vasculaires cérébraux.

QUESTION
Le clopidogrel est-il utilisé comme anticoagulant pour traiter les infections ? (Oui/Non)

ANSWER
Non

EXAMPLE
CONTEXT
Le propranolol est un bêta-bloquant utilisé pour traiter diverses conditions telles que l'hypertension artérielle, les migraines et l'anxiété.

QUESTION
Le propranolol est-il un antibiotique ? (Oui/Non)

ANSWER
Non

FIGURE 11 – Prompt des questions binaires

You're an expert in pharmacology. Answer this simple-choice question from the pharmacy exam based on the context. There is only one correct answer.

EXEMPLE
CONTEXTE
Escherichia coli, en abrégée E. coli, est une bactérie intestinale des organismes à sang chaud1, Gram négatif, du genre Escherichia, en forme de bâtonnet appelée parfois colibacille. E. coli est une bactérie aéro-anaérobie. E. coli constitue, avec d'autres bactéries, 0,1% du microbiote intestinal2. Les bactéries Escherichia coli ont été observées pour la première fois dans des fèces de nourrissons en 1885 par l'allemand Theodore Escherich et ont été nommées en hommage à ce dernier en 1919 par Castellani et Chambert3, c'est un coliforme thermotolérant (fécal) généralement commensal. La plupart des souches sont inoffensives, voire bénéfiques pour l'être humain, produisant de la vitamine K4 ou empêchant la colonisation de l'intestin par des bactéries pathogènes, établissant alors une relation de type mutualiste. Cependant, certains sérotypes d'E. coli peuvent être pathogènes, entraînant alors des gastro-entérites, infections urinaires, méningites ou sepsis.

QUESTION
Parmi les propositions suivantes, quelle est celle qui s'applique à Escherichia coli?

RÉPONSES
(a) C'est une bactérie anaérobie stricte
(b) Il peut être responsable de méningites
(c) C'est une bactérie toujours immobile
(d) C'est une bactérie oxydase positive
(e) Il est toujours résistant aux aminopénicillines

RÉPONSE CORRECTE
(b)

CONTEXTE
Ampérométrie L'ampérométrie est le terme désignant l'ensemble des techniques électrochimiques, dans lesquelles un courant est mesuré en fonction d'une variable indépendante, qui est généralement le temps ou le potentiel d'électrode.
D'autre part, la voltamétrie est également une sous-classe de l'ampérométrie, dans laquelle le courant est mesuré en faisant varier le potentiel appliqué à l'électrode.
La voltampérométrie cyclique (ou voltammétrie cyclique) est une technique électrochimique dans laquelle on enregistre la réponse en courant résultant d'une variation continue du potentiel de l'électrode de travail sur laquelle se produit la réaction électrochimique étudiée.
La mesure conductimétrique est une méthode d'électroanalyse qui permet de mesurer les propriétés conductrices d'une telle solution.
La voltampérométrie (ou voltammétrie) est une méthode d'électroanalyse basée sur la mesure du flux de courant résultant de la réduction ou de l'oxydation des composés tests présents en solution sous l'effet d'une variation contrôlée de la différence de potentiel entre deux électrodes spécifiques.

QUESTION
Une seule des perturbations suivantes ne concerne pas une hyperthyroïdie d'origine périphérique. Laquelle?

RÉPONSES
(a) Hypocholestérolémie
(b) Augmentation de la TSH plasmatique
(c) Présence d'une tachycardie
(d) Augmentation du métabolisme basal
(e) Augmentation du catabolisme protéique

RÉPONSE CORRECTE
(c)

VRAIE RÉPONSE
(b)

FIGURE 12 – Exemple de prompt avec la réponse générée par le LLM et la vraie réponse où le contexte fourni ne permet pas au LLM de répondre correctement à la question posée

You are a medical doctor answering real-world medical exam questions. Your role is to transform each question with the proposed answer into a yes/no affiramtive question. Do not use negation form.

EXAMPLE
QUESTION
L'adrénaline stimule tous ces processus métaboliques sauf un. Lequel?

Réponse
Glycolyse

Réponse reformulée
L'adrénaline stimule-t-elle la Glycolyse? (Oui/Non)

EXAMPLE
QUESTION
Quel(s) est (sont) le(s) mode(s) d'infestation possible par Toxoplasma gondii ?

Réponse
Greffe d'organe

Réponse reformulée
Le mode d'infestation possible par Toxoplasma gondii est-il la greffe d'organe? (Oui/Non)

EXAMPLE
QUESTION
Parmi les affirmations suivantes, une seule est fausse, indiquer laquelle: les particules alpha

Réponse
Sont formées de noyaux d'hélium

Réponse reformulée
les particules alpha sont-elles formées de noyaux d'hélium ? (Oui/Non)

EXAMPLE
QUESTION
Parmi les bactéries suivantes, une seule ne peut généralement pas être responsable d'une méningite aiguë, laquelle?

Réponse
Neisseria gonorrhoeae

Réponse reformulée
La bactérie Neisseria gonorrhoeae est-elle généralement responsable d'une méningite aiguë ? (Oui/Non)

EXAMPLE
QUESTION
Parmi les propositions suivantes, laquelle (lesquelles) est (sont) exacte(s)?Les modifications de structure permettant le passage du composé 1 au composé 2 entraînent:

Réponse
Un élargissement du spectre vers les cocci Gram+

Réponse reformulée
Les modifications de structure permettant le passage du composé 1 au composé 2 entraînent-elles un élargissement du spectre vers les cocci Gram+ ? (Oui/Non)

EXAMPLE
QUESTION
Parmi les techniques voltampérométriques, on trouve:

Réponse
La potentiométrie

Réponse reformulée
La potentiométrie fait-elle partie des techniques voltampérométriques ? (Oui/Non)

FIGURE 10 – Prompt pour la reformulation des questions