Étude comparative de réponses humaines et de grands modèles de langue à des QCM en pharmacie

Ricardo Rodriguez[1] Stéphane Huet[1] Benoît Favre[2] Mickael Rouvier[1]
(1) LIA, Avignon Université, France
(2) LIS, Aix Marseille Université, France
prenom.nom@{univ-avignon.fr, lis-lab.fr}

RÉSUMÉ

Cet article propose d'étudier les réponses générées par plusieurs Grands Modèles de Langue à un ensemble de Questions à Choix Multiple en pharmacie. Ces réponses sont comparées aux réponses données par des étudiants, afin de comprendre quelles sont les questions difficiles pour les modèles par rapport aux humains et pour quelles raisons. Nous utilisons les logits internes des modèles pour construire des distributions de probabilité et analyser les caractéristiques principales qui déterminent la difficulté des questions via une approche statistique. Nous apportons aussi une extension du jeu de données FRENCHMEDMCQA avec des paires question-réponses en pharmacie, enrichies avec les réponses des étudiants, la ponctuation assignée aux réponses, les thématiques cliniques correspondantes et des annotations manuelles sur la structure et certains traits sémantiques des questions.

ABSTRACT

Comparative Analysis of Human and Large Language Model Performance in Pharmacology Multiple-Choice Questions.

In this article, we study the answers generated by a selection of Large Language Models to a set of Multiple Choice Questions in Pharmacology, and compare them to the answers provided by students, to understand which questions in this clinical domain are difficult for the models when compared to humans and why. We extract the internal logits to infer probability distributions and analyse the main features that determine the difficulty of questions using statistical methods. We also provide an extension to the FrenchMedMCQA dataset, with pairs of question-answers in pharmacology, enriched with student response rate, answer scoring, clinical topics and annotations on question structure and semantics..

MOTS-CLÉS : question à choix multiples, grands modèles de langue, frenchmedmcqa, médical, pharmacologie.

KEYWORDS: multiple choice question, large language models, frenchmedmcqa, clinical, pharmacology.

ARTICLE : **Contribution originelle.**

1 Introduction

Les Grands Modèles de Langue (Large Language Model - LLM) comme ChatGPT (OpenAI *et al.*, 2024) et Llama (Meta, 2024) ont marqué une avancée significative dans le domaine de la génération automatique de texte. Ces progrès ont eu un impact considérable sur de nombreuses tâches de Traitement Automatique du Langage Naturel (TALN) comme le résumé automatique, la réponse à des questions ou la traduction automatique.

Afin d'évaluer les capacités de ces modèles, de nombreux travaux ont été réalisés pour mesurer leurs performances sur divers référentiels d'évaluations (*benchmarks*). Au fil du temps, ces référentiels ont évolué, passant des tâches simples (la prédiction du mot suivant ou le calcul de la perplexité) à des tâches plus complexes (Liu *et al.*, 2023; Zhang *et al.*, 2024), telles que des examens médicaux (Pal *et al.*, 2022; Labrak *et al.*, 2022) ou juridiques (Guha *et al.*, 2023).

Les performances des modèles sur les référentiels d'évaluations, comme par exemple la précision prédictive (*accuracy*) ou la F-Mesure, font partie des métriques les plus communes pour évaluer les capacités des LLM. En revanche, elles n'apportent qu'une compréhension superficielle des systèmes, indiquant principalement si un modèle peut compléter une tâche donnée de manière efficace, telle que répondre à des questions, sans pour autant apporter une perception plus profonde du comportement sous-jacent des modèles, ce qui pourrait permettre de mieux comprendre les causes des erreurs ou les subtilités dans la dégradation des performances (Liang *et al.*, 2022; Ribeiro *et al.*, 2020).

Les difficultés rencontrées par les modèles quand il s'agit de répondre à des questions dépendent de plusieurs facteurs. D'une part, certaines questions sont intrinsèquement complexes, car elles requièrent des compétences telles qu'une connaissance approfondie du contexte, la capacité d'établir des liens entre différents sujets ou de mener des raisonnements en plusieurs étapes. D'autre part, la fréquence des concepts pendant l'entraînement joue un rôle clé : les LLMs ont plus de difficultés à traiter des termes rares ou à répondre à des questions très spécifiques.

Dans ce travail, nous proposons d'explorer pourquoi ces modèles échouent dans certaines conditions et dans quelle mesure ils rencontrent les mêmes obstacles que les humains quand il s'agit de répondre à des questions. Nous pensons qu'améliorer notre compréhension sur ces questions est essentiel pour pouvoir concevoir des modèles plus robustes et capables de raisonner de manière plus fiable dans des contextes variés.

Nous proposons d'utiliser le jeu de données FRENCHMEDMCQA (Labrak *et al.*, 2022) constitué de Questions à Choix Multiples (QCM). Ce jeu de données, qui contient des paires question-réponses du domaine public provenant d'annales du concours d'entrée aux formations d'internat en pharmacie, et pour lequel il est possible d'obtenir les taux de réponse des étudiants, nous permet de comparer les résultats de plusieurs modèles à ceux d'humains.

Les principales contributions de ce travail sont les suivantes :

— Nous proposons une analyse comparative originale entre les réponses humaines et celles générées par les LLM pour des questions à choix multiples dans le domaine de la pharmacologie en utilisant le jeu de données FRENCHMEDMCQA.
— Nous publions une version enrichie du jeu de données FRENCHMEDMCQA avec des annotations supplémentaires qui incluent les réponses agrégées des étudiants, un ensemble d'étiquettes incorporées manuellement par question (par exemple : la négation et l'identification de réponse fausse), des caractéristiques syntaxiques et les thèmes cliniques. Cette

nouvelle version du corpus est publiée en accès libre. [1]

— Nous proposons une analyse détaillée qui identifie les caractéristiques les plus importantes pour déterminer le degré de difficulté des questions tant pour les humains que pour les modèles.

Ce papier est organisé de la manière suivante. Dans la section 2, nous présentons le jeu de données FRENCHMEDMCQA ainsi que les données supplémentaires qui ont été incorporés. Dans la section 3, nous présentons la méthodologie expérimentale utilisée dans notre étude. Nous poursuivons par la section 4 en faisant une comparaison détaillée des résultats humains et ceux des LLM, ainsi que de leur performance à travers les différents niveaux de difficulté. Finalement, nous apportons nos conclusions dans la section 5.

2 Jeu de données FrenchMedMCQA

Dans cette section nous décrivons le jeu de données FRENCHMEDMCQA et introduisons les éléments supplémentaires que nous avons collectés afin d'enrichir son contenu et utilisés comme base de notre analyse.

2.1 Description

Le jeu de données FRENCHMEDMCQA contient environ 3 000 questions à choix multiples (QCM) dans le domaine de la pharmacologie. Il regroupe des questions en français d'annales du concours d'internat en pharmacie, extraits du site web MedShake.net [2]. Ces données sont similaires à celles disponibles pour d'autres langues, telles que les jeux de données MEDMCQA (Pal *et al.*, 2022) et SCIQ (Welbl *et al.*, 2017).

Ce jeu de données a été choisi pour être d'une bonne qualité pour la tâche de répondre à des questions à choix multiples dans le médical et en français, et parce que la plateforme MedShake fournit les taux de réponse des étudiants, ce qui nous permet de faire une analyse comparant le comportement des systèmes face aux humains sur les mêmes données.

Chaque item comprend une question, cinq choix de réponse possibles identifiés par une lettre allant de « a » à « e » ainsi que la ou les réponses correctes.

2.2 Annotations supplémentaires

Sur MedShake.net, les étudiants peuvent s'exercer aux examens en ligne et d'évaluer leurs connaissances, obtenant un score final basé sur l'échelle de notation réelle utilisée lors de l'examen. La plateforme fournit également les réponses correctes et partiellement correctes parmi les choix possibles, le nombre de points attribués à chaque combinaison de choix, le nombre d'étudiants ayant

1. FrenchMedMCQA-extended sur le site HuggingFace : https://huggingface.co/datasets/uy-rrodriguez/FrenchMedMCQA-extended

2. « Annales QCM des concours d'internat en pharmacie » : https://www.medshake.net/pharmacie/concours-internat/annales/qcm/. Dernier accès : 2025-03-01.

A.

	Nb. Items	Réponses Étudiants	Nb. Choix Corrects				
			1	2	3	4	5
Train	2170	740	27%	24%	33%	14%	2%
Dev	312	601	52%	14%	23%	10%	1%
Test	622	650	52%	15%	23%	9%	1%

B.

	Négation		Composition		Intrus		Mode Phrase			Choix Explicites		
	non	oui	non	oui	non	oui	Q	I	A	S	M	U
Train	94%	6%	70%	30%	85%	15%	74%	14%	12%	56%	23%	21%
Dev	93%	7%	63%	37%	77%	23%	71%	22%	7%	47%	47%	6%
Test	94%	6%	65%	35%	78%	22%	67%	30%	3%	48%	43%	9%
Moyenne	94%	6%	69%	31%	82%	18%	72%	18%	10%	53%	30%	17%

TABLE 1 – (A) Nombre de questions et moyenne du nombre de réponses des étudiants par question dans le corpus FrenchMedMCQA, suivis par la distribution des questions selon le nombre de choix corrects (chaque question requiert entre 1 et 5 choix pour être répondue correctement). (B) Distribution des annotations manuelles ajoutées au corpus. Clarification des colonnes : **Mode de la Phrase :** Q=Question ; I=Instruction ; A=Affirmation. **Choix Explicites :** (Nombre explicite de choix) S=Unique (« Single ») ; M=Multiple ; U=Indéfini (« Undefined »).

répondu à chaque combinaison, le(s) thème(s) clinique(s) de la question [3], ainsi que l'année au cours de laquelle l'examen a eu lieu.

Le nombre de réponses des étudiants varie d'une question à l'autre, en dépendant des questions que les utilisateurs ont choisi de répondre. De plus, il est impossible d'identifier toutes les réponses d'un seul individu. En total il y a plus de 2,4 millions de réponses, avec une moyenne de 664 réponses par question.

Dans le cadre de ce travail, nous avons enrichi le jeu de données FRENCHMEDMCQA avec les réponses des étudiants, les thèmes cliniques et les années. Nous avons aussi étiqueté manuellement chaque question avec un ensemble de caractéristiques basées sur son contenu sémantique et sa formulation. Cette information additionnelle nous permet d'enrichir l'analyse statistique à suivre, notamment pour nous aider à identifier les caractéristiques dans ce corpus qui expliquent la difficulté des questions. Une ensemble de statistiques sur le corpus résultant est disponible dans la Table 1.

Les annotations manuelles décrivent les aspects suivants :

— **Négation** indique si la question contient un adverbe de négation.
— **Composition requise** indique quand la question est une phrase partielle et doit être combinée avec un ou plusieurs choix pour former une phrase correcte. *E.g. : « Le 'Crack' est une forme de : ». Choix : (a) héroïne ; (b) cocaïne.*
— **Identification de l'intrus** indique si la question demande à l'étudiant d'identifier la ou les choix qui ne respectent pas une certaine condition. *E.g. : « Quelle proposition ne correspond pas à norfloxacin ? ».*

3. Liste des thèmes cliniques : pharmacologie, physiologie, bactériologie, chimie analytique, toxicologie, hématologie, biochimie clinique, immunologie, santé publique, virologie, parasitologie, biophysique, épidémiologie, galénique, mycologie, pharmacocinétique, génétique, statistiques, enzymologie.

— **Mode de la phrase** catégorise la « question » en tant que véritable question, instruction ou affirmation. *E.g. : Instruction : « Concernant le misoprostol, donner son mécanisme d'action. » ; Affirmation : « Une anémie s'observe généralement au cours des parasitoses suivantes : ».*
— **Nombre explicite de choix** indique si le nombre attendu de réponses est explicitement fourni (unique, multiple ou indéfini). *E.g. : Unique : « Une seule proposition est exacte. La sérotonine est la : » ; Multiple : « Lesquelles des propositions suivantes concernent IL-2 ? » ; Indéfini : « Que se passe-t-il pendant la systole ventriculaire ? ».*

Ce nouveau jeu de données appelé FRENCHMEDMCQA-EXTENDED est disponible en ligne. [1]

3 Méthodologie expérimentale

La section 3.1 présente les LLM utilisés dans notre étude ainsi que les approches d'affinement et d'inférence. Dans la section 3.2, nous décrivons les métriques employées pour évaluer ces modèles, puis nous proposons dans la section 3.3 une méthode pour évaluer la difficulté des questions. La section 3.4 fournit quelques statistiques sur les nouvelles annotations.

3.1 Sélection de modèles et affinement

Plusieurs LLM récents et disponibles sur le site de Hugging Face ont été utilisés dans notre analyse. Les modèles considérés sont à la fois des LLM généralistes états de l'art : Llama-3-8B et 70B (Touvron *et al.*, 2023), Mistral-7B (Jiang *et al.*, 2023), et des LLM spécialisés dans le domaine médical : BioMistral-7B (Labrak *et al.*, 2024) et Apollo-7B (Wang *et al.*, 2024). Ces LLM sont sélectionnés grâce à leur bons résultats dans le défi TALN-DEFT 2023 (Labrak *et al.*, 2023), pour avoir une référence de base.

Les modèles sont chargés en précision 4-bit pour ensuite être affinés sur l'ensemble d'apprentissage de FrenchMedMCQA en utilisant la méthode Low Rank Adaptation (LoRA) Hu *et al.* (2022), ce qui réduit le coût d'apprentissage et s'ajuste à notre infrastructure. Plus de détails sur les hyper-paramètres peuvent être trouvés dans l'Annexe B.

Deux formats de réponse attendue sont évalués, l'un comprenant le texte intégral des réponses et l'autre uniquement les lettres correspondant aux choix. Les instructions (*prompts*) sont présentées en Annexe A et comportent des descriptions de tâches simples et structurées.

Après l'affinage, nous évaluons la performance des modèles sur l'ensemble de test du jeu de données FRENCHMEDMCQA afin de mesurer à la fois les capacités de généralisation et l'efficacité globale.

Pour chaque LLM, l'algorithme d'inférence est exécuté quatre fois de manière indépendante et nous prenons en compte la moyenne des résultats pour déterminer la meilleure performance par modèle. Ces résultats se trouvent dans la Table 2.

3.2 Métriques

Contrairement à une tâche classique de classification de texte, les QCM peuvent permettre des réponses partiellement correctes. Par exemple, si une question invite à sélectionner deux options correctes et si une seule est choisie, alors la réponse est incomplète, mais peut tout de même être considéré comme acceptable. Dans cette étude, deux métriques proposées initialement dans (Labrak *et al.*, 2023), l'EMR et le score de Hamming, et une troisième originelle, le score de MedShake, sont utilisées afin de mesurer la proportion de réponses correctes (y compris incomplètes).

Dans les équations ci-dessous, N est le nombre de questions dans le jeu de données, y_i est l'ensemble de choix corrects pour la question numéro i, et $\hat{y}_i$ est l'ensemble de choix prédits pour la question numéro i.

— **Score de réponse exacte (EMR, *Exact Match Ratio*)** quantifie le nombre de questions pour lesquelles les réponses prédites correspondent exactement aux réponses attendues.

$$\text{EMR} = \frac{1}{N}\sum_{i=1}^{N}[y_i = \hat{y}_i]$$

— **Score de Hamming** est basé sur la distance de Hamming et mesure le nombre de correspondances entre les réponses prédites et celles attendues, en comparant la taille de leur intersection par rapport à celle de l'union.

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

— **Score de MedShake** est la métrique utilisée lors des examens. Elle attribue un score en fonction de la correspondance entre les réponses fournies et les réponses attendues. Une réponse entièrement correcte est notée 2,0 points. En présence d'une ou deux erreurs ou omissions, le score est réduit à 1,0 point et 0,4 respectivement. Toute autre configuration, incluant des réponses incorrectes ou incomplètes au-delà de ce seuil, est sanctionnée par un score nul (0 point).

3.3 Classification des questions par difficulté

Le niveau de difficulté d'une question correspond à la complexité qu'elle présente pour les répondants, en particulier dans le cadre d'une évaluation. Il reflète à quel point une question est susceptible de mettre en échec les étudiants, en fonction de la nature des connaissances et des raisonnements qu'elle mobilise.

L'approche naïve pour déterminer la difficulté des questions consiste à utiliser le pourcentage d'étudiants qui ont donné la réponse correcte. Ainsi, si le pourcentage d'étudiants qui ont donné la réponse correcte est élevé, on considère la question comme facile, car la majorité parvient à identifier la bonne réponse. À l'inverse, si ce pourcentage est faible, la question est jugée difficile, indiquant que peu d'étudiants ont réussi à trouver la bonne réponse. Cependant, cette méthode ne prend pas en compte les réponses partielles ni le nombre de choix corrects, qui affectent les probabilités qu'une personne trouve la réponse correcte de manière aléatoire ; elle ne considère pas non plus les cas où les étudiants

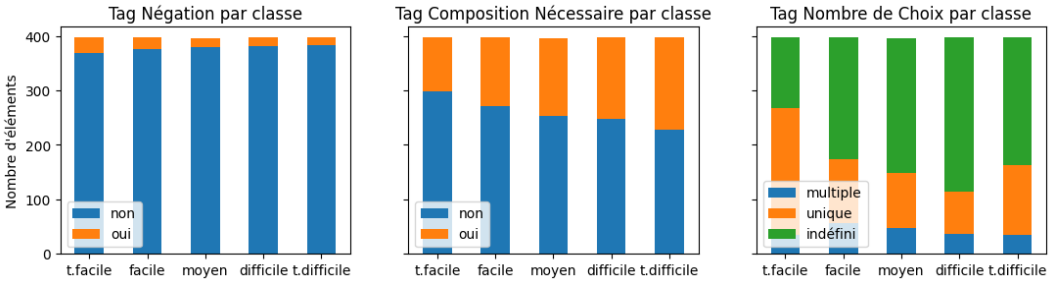

FIGURE 1 – Nombre de questions par classe et pour chaque étiquette annotée manuellement sur l'ensemble du corpus FrenchMedMCQA-extended.

hésitent parmi un sous-ensemble d'options. Afin de mieux intégrer ces facteurs, nous mesurons la difficulté des questions via une méthode basée sur l'entropie de Shannon (Shannon, 1948) :

$$H(P) = \frac{-\sum_{i=1}^{n} p_i log(p_i)}{log(n)}$$

où n est le nombre de toutes les combinaisons possibles des réponses et p_i est la proportion d'étudiants ayant choisi la réponse i. Ainsi, une valeur de 0 indique que la question est évidente pour les étudiants (tout le monde donne la même réponse), et une valeur de 1 signifie que la question est extrêmement difficile (le taux de réponse est équivalent à une sélection aléatoire).

Dans notre travail, H est calculé pour chaque question P dans le jeu de données et celui-ci est ensuite divisé en cinq quantiles de taille égale, qui correspondent à nos classes de difficulté : Très Facile, Facile, Moyenne, Difficile et Très Difficile.

3.4 Distribution des étiquettes annotées

La Figure 1 montre la répartition des étiquettes annotées manuellement sur l'ensemble du corpus. Nous pouvons constater que les classes de difficulté présentent une corrélation évidente avec les annotations. Par exemple, pour l'étiquette « Composition requise », le nombre de questions nécessitant une composition augmente avec le niveau de difficulté. En ce qui concerne « Nombre de choix », une proportion plus importante de questions indique explicitement de sélectionner une seule réponse dans les classes les plus faciles qu'au sein des classes les plus difficiles.

4 Résultats

Dans la section 4.1 nous décrivons les observations principales sur les réponses des LLM. Ensuite, dans la section 4.2 nous faisons une comparaison entre la performance du meilleur modèle face aux humains. Plus tard, dans la section 4.3 nous introduisons la stratégie pour construire une distribution de probabilité basée sur les logits internes du modèle, pour enfin l'utiliser dans la section 4.4 où nous analysons l'importance des caractéristiques via une régression linéaire.

4.1 Résultats des LLM et meilleur modèle

La Table 2 présente les résultats de l'ensemble des modèles. Les LLM sont évalués selon les trois métriques d'évaluation : MedShake, EMR (Exact Match Ratio) et Hamming.

Modèle	Médical	MedShake	EMR	Hamming	Instruction
Llama-3-8B	-	0,366	0,295	0,522	1
Llama-3-70B	-	0,189	0,138	0,381	1
Mistral-7B-v0.1	-	0,387	0,32	0,543	1
Mistral-7B-v0.3	-	0,391	0,318	0,539	1
BioMistral-7B	✓	0,289	0,224	0,475	2
Apollo-7B	✓	0,413	0,333	0,557	2
Llama-3-8B*	-	0,453	0,373	0,575	1
Llama-3-70B*	-	0,419	0,345	0,554	1
Mistral-7B-v0.3*	-	**0,491**	**0,418**	**0,626**	1
BioMistral-7B*	✓	0,404	0,326	0,551	2
Apollo-7B*	✓	0,417	0,339	0,575	2

TABLE 2 – Ce tableau résume les meilleurs résultats par modèle sur l'ensemble de test de FRENCH-MEDMCQA, affichant la moyenne sur toutes les classes de difficulté confondues. Les modèles affinés sont marqués avec *. Les modèles spécialisés dans le médical sont identifiés avec un check. L'instruction numéro « 1 » correspond au format simple en langage naturel, tandis que la numéro « 2 » correspond à la version plus structurée, comme décrit dans la section 3.1.

Nous observons que les modèles affinées surpassent systématiquement leurs homologues non-affinés, quelle que soit la configuration du modèle. Par exemple, Mistral-7B-v0.3 atteint un score MedShake de 0,391 sans affinement, contre 0,491 après affinement. Ce modèle obtient d'ailleurs les meilleures performances globales sur l'ensemble des métriques (MedShake : 0,491, EMR : 0,418, Hamming : 0,626), ce qui illustre la robustesse du modèle une fois adapté à la tâche. En vue des résultats, ce modèle est choisi pour l'analyse de l'importance des caractéristiques dans la section 4.4

Parmi les modèles spécialisés dans le domaine médical, Apollo-7B obtient les meilleurs résultats parmi les modèles non-affinés (MedShake : 0,413, EMR : 0,333, Hamming : 0,557). Cependant, après affinement, ses performances demeurent inférieures à celles de Mistral-7B. Il est également à noter que les modèles biomédicaux affinés restent en retrait par rapport aux modèles généralistes affinés, malgré leur spécialisation.

4.2 Évaluation des humains par rapport aux LLM

La Table 3 présente une comparaison détaillée des performances entre le modèle Mistral-7B-v0.3 (après affinement) et les humains, en fonction des cinq niveaux croissants de difficulté des questions et sur les trois métriques d'évaluation : MedShake, EMR (Exact Match Ratio) et Hamming. Les scores correspondent à la moyenne de quatre exécutions indépendantes pour le modèle, tandis que les performances humaines sont dérivées des taux de réussite obtenus de la plateforme MedShake.net, comme décrit dans la section 2.2..

De manière générale, les performances décroissent de manière cohérente avec l'augmentation du niveau de difficulté, tant pour les humains que pour le modèle. Cette tendance reflète la validité de la

$\text{Modèle}_{\text{Métrique}}$	T. Facile	Facile	Moyenne	Difficile	T. Difficile	Total
Mistral-7B$_{MedShake}$	0,780	0,598	0,453	0,289	0,336	0,492
Mistral-7B$_{EMR}$	0,750	0,524	0,355	0,196	0,262	0,418
Mistral-7B$_{Hamming}$	0,814	0,706	0,637	0,506	0,467	0,626
Étudiants$_{MedShake}$	0,891	0,724	0,565	0,438	0,352	0,594
Étudiants$_{EMR}$	0,872	0,642	0,452	0,324	0,295	0,517
Étudiants$_{Hamming}$	0,903	0,775	0,667	0,579	0,460	0,677

TABLE 3 – Comparaison des performances entre le modèle affiné Mistral-7B-v0.3 et les humains, selon différents niveaux de difficulté et métriques d'évaluation. Les scores présentés correspondent à la moyenne de quatre exécutions (*runs*) indépendantes. Les scores Étudiants proviennent des taux de réussite dans l'ensemble de test.

catégorisation proposée et confirme que la difficulté perçue par les humains est globalement partagée par le modèle.

Sur l'ensemble des niveaux de difficulté, les humains surpassent Mistral-7B sur toutes les métriques. Par exemple, pour les questions de niveau « Très facile », les humains atteignent un score MedShake de 0,891, contre 0,780 pour le modèle. Cet écart reste présent à mesure que la difficulté augmente. Pour les questions « Difficiles », les scores EMR chutent à 0,324 pour les humains et 0,196 pour le modèle, illustrant les limites du modèle face à des questions complexes nécessitant une compréhension fine ou une déduction poussée.

Il est intéressant de noter que, pour la catégorie « Très Difficile » le modèle obtient des meilleurs résultats par rapport à la catégorie « Difficile », et que la différence avec les humains est réduite au minimum, avec les étudiants obtenant un score EMR seulement 3.3% par-dessus le modèle (0.295 contre 0.262). De manière générale, ces résultats surlignent la grande difficulté présentée par ces questions, qui pour la plupart requièrent la sélection de plusieurs choix pour être considérées correctes. Néanmoins, nous n'avons pas approfondi dans les raisons qui rendraient les questions dans cette classe légèrement moins difficiles pour Mistral, mais cela pourrait être un axe d'exploration à futur.

Enfin, la métrique Hamming, qui évalue la proximité partielle entre les réponses attendues et celles produites, montre un écart réduit entre les performances des LLM et celles des humains. Cela suggère que, même en l'absence de réponses entièrement correctes, le modèle parvient à identifier une partie des éléments pertinents.

Ces résultats soulignent à la fois les capacités actuelles des LLM dans des tâches spécialisées, et les marges d'amélioration nécessaires pour atteindre, voire dépasser, les performances humaines sur des tâches de spécialité.

4.3 Distribution de probabilité à partir du modèle

Pour comparer les réponses humaines à celles des LLM, étant donné que les réponses humaines sont exprimées en taux d'étudiants par combinaison de choix possibles, nous avons décidé de dériver une valeur comparable pour les modèles en faisant une conversion des logits internes en distribution de probabilité.

Nous faisons une extrapolation des probabilités apprises par le modèle via l'extraction des probabilités

au niveau des *tokens* (application de *softmax* sur les logits internes) pour des séquences suivant le même format d'instruction vu pendant l'affinage (comprenant l'introduction à la tâche, la question et les options de réponse).

Pour une séquence S, la log-probabilité $log\_prob(S)$ est calculée comme la somme des logarithmes des probabilités de chaque token. Nous avons évalué diverses stratégies pour calculer log_prob afin de prédire la réponse que le modèle sélectionnerait : en utilisant l'intégralité de la séquence, uniquement le segment de la réponse ou seulement les lettres correspondant aux choix, et nous avons ensuite comparé leurs performances à celles obtenues via l'inférence.

Nous avons observé quelques subtilités intéressantes entre ces stratégies et les différentes architectures. Pour Llama-3, la sortie sélectionnée en utilisant uniquement les probabilités sur les lettres correspondant aux choix est cohérente avec la génération classique fournie par le LLM ; cela montre que le modèle semble accorder beaucoup d'importance à la lettre. Le modèle Mistral-7B affiné pour prédire le texte complet des réponses (lettre et contenu) présente un comportement différent : il obtient un score élevé lorsque l'ensemble du segment de réponse est considéré, mais très faible lorsque seules les lettres des choix sont conservées. Ce comportement est corrigé quand le modèle est affiné pour ne prédire que la lettre des choix.

Nous avons finalement retenu la probabilité des lettres correspondant aux choix, plus stable dans nos expériences, comme une mesure suffisamment fiable pour construire une distribution de probabilité.

4.4 Importance des caractéristiques

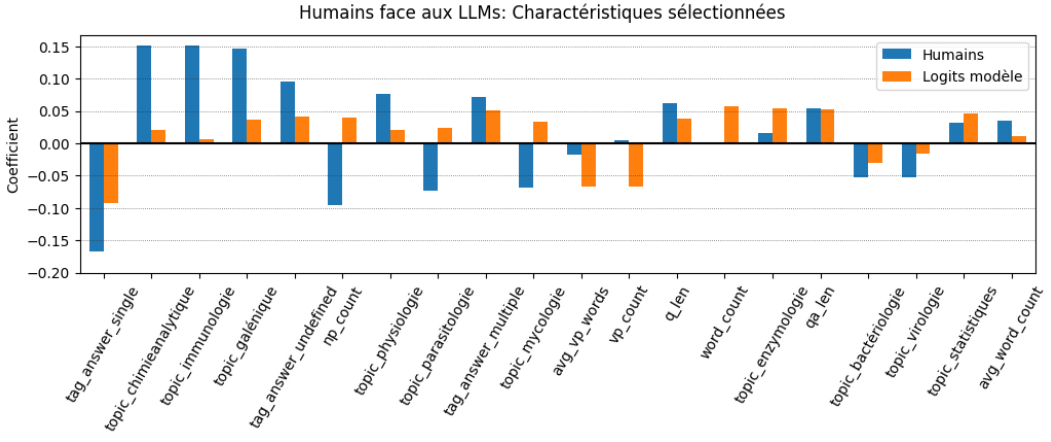

FIGURE 2 – Comparaison de coefficients obtenus via des régressions linéaires sur les résultats humains et pour le modèle Mistral-7B. Seules les caractéristiques avec les plus grands coefficients en valeur absolue sont affichées.

Notre objectif est d'identifier les caractéristiques qui déterminent le mieux le niveau de difficulté des questions, tant pour les humains que pour les LLM. Pour ce faire, nous avons entraîné un modèle de régression linéaire en utilisant une validation croisée à cinq plis avec une régularisation Ridge (L2), sur un jeu de données constitué de la fusion des ensembles d'apprentissage et de test (environ 1 600 éléments).

Il est important de clarifier que notre objectif n'est pas de développer le meilleur prédicteur de la difficulté d'une question. Ceci implique que des caractéristiques sémantiques telles les « embeddings » à partir des questions, souvent utilisés dans d'autres travaux de prédiction de difficulté comme par exemple Yaneva *et al.* (2024), ne sont pas incluses dans notre travail afin de ne considérer que des caractéristiques explicables et comparables aux humains. Cette décision peut réduire la capacité du modèle de régression à prédire la difficulté perçue par les LLM, ce qui est un limitation acceptable dans notre analyse.

Pour les données humaines, nous avons entraîné un modèle linéaire à prédire le taux de réponses correctes des étudiants. L'ensemble des caractéristiques provient des annotations présentées dans la section 2.2 ainsi que de plusieurs caractéristiques syntaxiques : la longueur de la question ; la somme des longueurs des réponses ; la longueur totale ; le nombre de mots dans la question ; le nombre moyen de mots par phrase ; la profondeur moyenne de l'arbre ; le nombre moyen de mots des groupes nominaux, prépositionnels et verbaux ; et le nombre total de ces groupes.

Pour les LLM, nous avons suivi la même approche en utilisant un modèle linéaire entraîné sur la distribution de probabilité basée sur les logits internes du modèle.

En raison de la taille limitée des données, nous avons sélectionné les caractéristiques uniquement en fonction de la magnitude de leurs coefficients plutôt que de nous appuyer sur des p-tests, lesquels se sont révélés instables d'une exécution à l'autre.

La Figure 2 montre la comparaison des coefficients des caractéristiques entre les humains et Mistral-7B. Dans l'ensemble, nous observons que la plupart des caractéristiques ont une relation similaire avec la difficulté des questions, tant pour les humains que pour les modèles, même quand la valeur des coefficients diffère. Notamment, les annotations *tag_answer_single* et *tag_answer_undefined*, qui décrivent respectivement quand la question indique de manière explicite qu'un seul choix est possible ou quand rien n'est indiqué, influencent de manière similaire les difficultés perçues par les humains et par le LLM. D'autre part, certains sujets, tels que l'*immunologie* et la *chimie analytique*, se révèlent être des indicateurs plus pertinents de la difficulté pour les humains, tandis que des indices syntaxiques, tels que le nombre moyen de mots dans les groupes verbaux (*avg_vp_words*), constituent de meilleurs prédicteurs pour le LLM. Ces observations suggèrent que les humains et les LLM évaluent généralement la difficulté des questions de manière équivalente, puisque la plupart des caractéristiques se rapportent de manière similaire à cette difficulté.

5 Conclusions

Cet article présente une analyse comparative des réponses aux QCM données par les humains et les LLM, révélant que les facteurs rendant une question difficile semblent être les mêmes pour les deux catégories. En effet, toutes deux rencontrent des difficultés face aux questions jugées difficiles. Par ailleurs, l'analyse de l'importance des caractéristiques montre que la plupart d'entre elles présentent une corrélation similaire avec le niveau de difficulté des questions. Notre étude expérimentale n'ayant été faite qu'avec un échantillon réduit de LLM et sur un seul corpus médical dans le domaine de la pharmacie uniquement en français, des expériences complémentaires devraient être menées pour étudier la généralisation des conclusions.

Remerciements

Nous tenons à remercier Pierre-Michel Bousquet pour son aide et ses conseils avisés tout au long de ce travail.

Ce travail a bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2024-A0161014871 et du soutien financier du projet ANR MALADES (ANR-23-IAS1-0005, https://anr-malades.github.io/).

Références

GUHA N., NYARKO J., HO D., RÉ C., CHILTON A., CHOHLAS-WOOD A., PETERS A., WALDON B., ROCKMORE D., ZAMBRANO D. *et al.* (2023). Legalbench : A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, **36**, 44123–44279.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2022). Lora : Low-rank adaptation of large language models. *ICLR*, **1**(2), 3.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b.

LABRAK Y., BAZOGE A., DAILLE B., DUFOUR R., MORIN E. & ROUVIER M. (2023). 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., MORIN E., GOURRAUD P.-A., ROUVIER M. & DUFOUR R. (2024). Biomistral : A collection of open-source pretrained large language models for medical domains.

LIANG P., BOMMASANI R., LEE T., TSIPRAS D., SOYLU D., YASUNAGA M., ZHANG Y., NARAYANAN D., WU Y., KUMAR A. *et al.* (2022). Holistic evaluation of language models. *arXiv preprint arXiv :2211.09110*.

LIU Y., FABBRI A. R., CHEN J., ZHAO Y., HAN S., JOTY S., LIU P., RADEV D., WU C.-S. & COHAN A. (2023). Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *arXiv preprint arXiv :2311.09184*.

META A. (2024). Introducing meta llama 3 : The most capable openly available llm to date. *Meta AI*, **2**(5), 6.

OPENAI, ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S., ANADKAT S., AVILA R., BABUSCHKIN I., BALAJI S., BALCOM V., BALTESCU P., BAO H., BAVARIAN M., BELGUM J., BELLO I., BERDINE J., BERNADETT-SHAPIRO G., BERNER C., BOGDONOFF L., BOIKO O., BOYD M., BRAKMAN

A.-L., BROCKMAN G., BROOKS T., BRUNDAGE M., BUTTON K., CAI T., CAMPBELL R., CANN A., CAREY B., CARLSON C., CARMICHAEL R., CHAN B., CHANG C., CHANTZIS F., CHEN D., CHEN S., CHEN R., CHEN J., CHEN M., CHESS B., CHO C., CHU C., CHUNG H. W., CUMMINGS D., CURRIER J., DAI Y., DECAREAUX C., DEGRY T., DEUTSCH N., DEVILLE D., DHAR A., DOHAN D., DOWLING S., DUNNING S., ECOFFET A., ELETI A., ELOUNDOU T., FARHI D., FEDUS L., FELIX N., FISHMAN S. P., FORTE J., FULFORD I., GAO L., GEORGES E., GIBSON C., GOEL V., GOGINENI T., GOH G., GONTIJO-LOPES R., GORDON J., GRAFSTEIN M., GRAY S., GREENE R., GROSS J., GU S. S., GUO Y., HALLACY C., HAN J., HARRIS J., HE Y., HEATON M., HEIDECKE J., HESSE C., HICKEY A., HICKEY W., HOESCHELE P., HOUGHTON B., HSU K., HU S., HU X., HUIZINGA J., JAIN S., JAIN S., JANG J., JIANG A., JIANG R., JIN H., JIN D., JOMOTO S., JONN B., JUN H., KAFTAN T., ŁUKASZ KAISER, KAMALI A., KANITSCHEIDER I., KESKAR N. S., KHAN T., KILPATRICK L., KIM J. W., KIM C., KIM Y., KIRCHNER J. H., KIROS J., KNIGHT M., KOKOTAJLO D., ŁUKASZ KONDRACIUK, KONDRICH A., KONSTANTINIDIS A., KOSIC K., KRUEGER G., KUO V., LAMPE M., LAN I., LEE T., LEIKE J., LEUNG J., LEVY D., LI C. M., LIM R., LIN M., LIN S., LITWIN M., LOPEZ T., LOWE R., LUE P., MAKANJU A., MALFACINI K., MANNING S., MARKOV T., MARKOVSKI Y., MARTIN B., MAYER K., MAYNE A., MCGREW B., MCKINNEY S. M., MCLEAVEY C., MCMILLAN P., MCNEIL J., MEDINA D., MEHTA A., MENICK J., METZ L., MISHCHENKO A., MISHKIN P., MONACO V., MORIKAWA E., MOSSING D., MU T., MURATI M., MURK O., MÉLY D., NAIR A., NAKANO R., NAYAK R., NEELAKANTAN A., NGO R., NOH H., OUYANG L., O'KEEFE C., PACHOCKI J., PAINO A., PALERMO J., PANTULIANO A., PARASCANDOLO G., PARISH J., PARPARITA E., PASSOS A., PAVLOV M., PENG A., PERELMAN A., DE AVILA BELBUTE PERES F., PETROV M., DE OLIVEIRA PINTO H. P., MICHAEL, POKORNY, POKRASS M., PONG V. H., POWELL T., POWER A., POWER B., PROEHL E., PURI R., RADFORD A., RAE J., RAMESH A., RAYMOND C., REAL F., RIMBACH K., ROSS C., ROTSTED B., ROUSSEZ H., RYDER N., SALTARELLI M., SANDERS T., SANTURKAR S., SASTRY G., SCHMIDT H., SCHNURR D., SCHULMAN J., SELSAM D., SHEPPARD K., SHERBAKOV T., SHIEH J., SHOKER S., SHYAM P., SIDOR S., SIGLER E., SIMENS M., SITKIN J., SLAMA K., SOHL I., SOKOLOWSKY B., SONG Y., STAUDACHER N., SUCH F. P., SUMMERS N., SUTSKEVER I., TANG J., TEZAK N., THOMPSON M. B., TILLET P., TOOTOONCHIAN A., TSENG E., TUGGLE P., TURLEY N., TWOREK J., URIBE J. F. C., VALLONE A., VIJAYVERGIYA A., VOSS C., WAINWRIGHT C., WANG J. J., WANG A., WANG B., WARD J., WEI J., WEINMANN C., WELIHINDA A., WELINDER P., WENG J., WENG L., WIETHOFF M., WILLNER D., WINTER C., WOLRICH S., WONG H., WORKMAN L., WU S., WU J., WU M., XIAO K., XU T., YOO S., YU K., YUAN Q., ZAREMBA W., ZELLERS R., ZHANG C., ZHANG M., ZHAO S., ZHENG T., ZHUANG J., ZHUK W. & ZOPH B. (2024). GPT-4 technical report.

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

RIBEIRO M. T., WU T., GUESTRIN C. & SINGH S. (2020). Beyond accuracy : Behavioral testing of nlp models with checklist. *arXiv preprint arXiv :2005.04118*.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, **27**(3), 379–423.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). Llama : Open and efficient foundation language models.

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

people.

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

YANEVA V., NORTH K., BALDWIN P., REZAYI S., ZHOU Y., CHOUDHURY S. R., HARIK P., CLAUSER B. *et al.* (2024). Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, p. 470–482.

ZHANG T., LADHAK F., DURMUS E., LIANG P., MCKEOWN K. & HASHIMOTO T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, **12**, 39–57.

A Instructions

Cette annexe présente des exemples des deux instructions les plus efficaces dans nos expériences, avec des exemples (few-shots) ne contenant que la lettre des choix corrects au lieu du texte complet. Le Format 1 présente une introduction concise à la tâche en français, immédiatement suivie des exemples et de la question. Le Format 2 est plus structuré et propose des descriptions très concrètes de la tâche en anglais.

Format 1 : Chaque *shot* et la question finale ont un format identique et incluent une brève explication de la tâche en français.

Format 1

Ceci est une question de QCM de l'examen de pharmacie. Réponds avec la ou les lettres correspondant à la bonne réponse.

Parmi ces médicaments, quel(s) est (sont) celui (ceux) susceptible(s) d'augmenter la toxicité de la ciclosporine ?
(a) Kétoconazole.
(b) Rifampicine.
(c) Amphotéricine B.
(d) Josamycine.
(e) Diltiazem.
Réponse(s) : (a) (c) (d) (e)

Ceci est une question de QCM de l'examen de pharmacie. Réponds avec la ou les lettres correspondant à la bonne réponse.

Parmi les propositions suivantes, indiquer celle qui est exacte. Dans les conditions physiologiques, le pH le plus élevé est mesuré dans :
(a) Le suc gastrique.
(b) La bile vésiculaire.
(c) Le suc pancréatique.
(d) La salive.
(e) Les sécrétions intestinales.
Réponse(s) : (

Format 2 : Description de la tâche en anglais, suivie de trois sections distinctes marquées par les caractères spéciaux `"###"`. Chaque plan et la question finale ont un format identique.

Format 2

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction : We are giving you a scientific question and five answers options (associated to "a", "b", "c", "d", "e"). Your task is to find the correct answer(s) based on scientific facts, knowledge and reasoning. Don't generate anything other than one of the following characters : 'a b c d e'.

Input : Ceci est une question de QCM de l'examen de pharmacie. Réponds avec la ou les lettres correspondant à la bonne réponse. Parmi les phénomènes suivants, tous sont synchrones du premier bruit cardiaque sauf un, lequel ?
(a) Fermeture des valvules auriculo-ventriculaires.
(b) Remplissage ventriculaire rapide.
(c) Contraction ventriculaire isométrique.
(d) Contraction ventriculaire isotonique.
(e) Ouverture des valvules sigmoïdes.

Response : (b)

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction : We are giving you a scientific question and five answers options (associated to "a", "b", "c", "d", "e"). Your task is to find the correct answer(s) based on scientific facts, knowledge and reasoning. Don't generate anything other than one of the following characters : 'a b c d e'.

Input : Ceci est une question de QCM de l'examen de pharmacie. Réponds avec la ou les lettres correspondant à la bonne réponse. Parmi les propositions suivantes, indiquer celle qui est exacte. Dans les conditions physiologiques, le pH le plus élevé est mesuré dans :
(a) Le suc gastrique.
(b) La bile vésiculaire.
(c) Le suc pancréatique.
(d) La salive.
(e) Les sécrétions intestinales.

Response : (

B Hyper-paramètres et infrastructure

Hyper-paramètres

Epochs	1	Max seq. length	512
Batch size	4	Precision	bf16
Micro batch size	4	LoRA r	4
Gradient accumulation steps	1	LoRA alpha	16
Learning rate	$3e^{-4}$	LoRA dropout	0.05
Learning scheduler	Cosine	Quantisation	4-bit
Optimiser	Paged 32-bit AdamW	Quantisation compute dtype	float16
Weight decay	0.001	Quantisation data type	NF4
Max gradient norm	0.3		

Infrastructure

GPU	RTX 3090	VRAM	16GB

TABLE 4 – Paramètres d'affinement et infrastructure