Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Evaluation et analyse des performances des grands modèles de langue sur des épreuves d'examen de médecine français

Adrien KUHNAST[1] Loïc VERLINGUE[1]
(1) Centre de Recherche en Cancérologie de Lyon, 28 rue Laennec, 69373 Lyon, France
adrien.kuhnast69@gmail.com, loic.verlingue@lyon.unicancer.fr

RESUME

Les grands modèles de langue (GMLs) ont démontré leur capacité à répondre correctement à des questions de médecine sur des bases anglaises. Or, leur paramétrage par apprentissage profond les soumet au biais linguistique et doivent ainsi être évalués dans la langue de l'utilisateur.

Nous avons évalué des GMLs sur 278 questions à choix multiples provenant d'examens de médecine (Lyon-Est 2024) de différentes spécialités et respectant les recommandations nationales.

Nos résultats montrent que les GMLs sont aussi bons que les étudiants mais qu'il existe d'importantes variations selon les spécialités. Améliorer la consigne en précisant de s'appuyer sur les recommandations françaises modifie significativement les notes obtenues ce qui démontre la nécessité d'éprouver les GMLs selon différents contextes géographiques et linguistiques. Nous avons également analysé le type d'erreur que font les GMLs ce qui ouvre la porte à des améliorations plus ciblées.

ABSTRACT

Evaluation and analysis of the performance of large language models on French medical exams

Large language models (LLMs) have demonstrated their ability to correctly answer medical questions on English benchmarks. However, as they are set up using deep learning, they are subject to linguistic bias and must therefore be evaluated in the user's language.

We evaluated LLMs on 278 multiple-choice questions from medical examinations (Lyon-Est 2024) in different specialities and in line with national recommendations.

Our results show that LLMs are as good as the students, but that there are significant variations between specialities. Improving the instructions by specifying that they should be based on French recommendations significantly modified the scores obtained, demonstrating the need to test LLMs in different geographical and linguistic contexts. We also analysed the type of errors made by LLMs, which opens the door to more targeted improvements.

Actes de l'atelier Traitement du langage médical à l'époque des LLMs 2025 (MLP-LLM), pages 14–24.

MOTS-CLES : Grand modèle de langue, biais linguistique, base de données française, type d'erreur.
KEYWORDS : Large language model, linguistic bias, french benchmark, error type.

1 Introduction

Le secteur de la santé manifeste un intérêt croissant pour les grands modèles de langue (GMLs). Leur performance doit pour cela être évaluée afin d'assurer leur efficacité et leur sécurité.

Les grands modèles de langue ont démontré leur capacité à répondre correctement à des questions de médecine mais principalement sur des bases de données anglaises (MedQA, MedMCQA, PubMedQA, ...) (Agerri et al., 2024 ; Natarajan et al., 2023). Ces modèles basés sur l'apprentissage sont très dépendants de leur base d'entraînement. Ainsi, le texte produit en sortie des modèles de langage les plus communs mettent en avant les sources anglaises. Or en France, la médecine et les bonnes pratiques médicales sont régies par des recommandations d'autorités nationales ou européennes. Cela permet de garder une cohérence vis à vis des ressources disponibles et des particularités des populations. Les grands modèles de langage, du fait de leur nature à contenir une quantité massive de paramètres entraînés non équitablement entre les différentes langues, sont soumis au biais linguistique et doivent alors être éprouvés dans chacune des langues de leur utilisation (Abadir & Chellappa et al., 2024 ; Liu et al., 2024). Une simple traduction d'une langue à une autre des réponses obtenues ne permet pas de les contextualiser aux recommandations médicales locales.

Il existe pour l'instant deux bases de données publiques françaises qui traite de questions de médecine mais elles présentent des limites importantes :

- FrenchMedMCQA (Gourraud et al., 2023) : il s'agit de 3 105 questions-réponses provenant d'examen de pharmacie et non de médecine. La plupart des questions s'intéressent aux thérapeutiques médicamenteuses et le raisonnement médical y est peu représenté. De plus, les questions-réponses n'ont pas été relues ni triées ; des erreurs peuvent y persister.
- FrBMedQA (Kaddari & Bouchentouf, 2022) : Plus de 41 000 questions-réponses du domaine biomédical ont été généré automatiquement depuis Wikipedia. Ces questions ne s'intéressent pas aux problématiques rencontrés en pratique clinique sur lesquelles les étudiants en médecine sont évalués.
- MedExpQA (Agerri et al., 2024) : C'est une base de données de 622 questions-réponses de médecine multilingues. En réalité, les toutes questions proviennent des examens nationaux espagnols d'entrée à l'internat : CasiMedicos. Ces questions ont été ensuite traduites en anglais, français et italien. Elles ne sont donc pas adaptées aux recommandations locales.

Nous avons éprouvé pour la première fois plusieurs grands modèles de langue usuels à des questions à choix multiples d'examen de médecine français dans le but de déterminer leur précision. Nous avons aussi analysé le type d'erreur que font les modèles dans le but de les comparer sur des critères détaillés. L'objectif de cette étude est aussi pointer des risques de l'utilisation massive des GLMs grand public par les étudiants en médecine.

2 Méthode

2.1 Base de données de questions-réponses

Les questions-réponses proviennent des examens des étudiants de 5ème année de médecine de la Faculté Lyon-Est au cours de l'année 2024. L'étude porte sur l'ensemble des 278 questions-réponses tombées cette année. Les différentes épreuves parcourent une large diversité de spécialités médicales : gynécologie-obstétrique, médecine intensive et réanimation (MIR), hépato-gastro-entérologie (HGE), handicap et vieillissement, psychiatrie, oto-rhino-laryngologie (ORL), urologie, néphrologie et pédiatrie avec en moyenne 30 questions-réponses et une lecture critique d'article avec 15 questions-réponses. Ces questions sont rédigées par les professeurs et enseignants de la faculté dans le but d'évaluer les connaissances des étudiants et de les préparer à l'examen d'entrée à l'internat. Elles respectent les recommandations françaises et européennes et s'appuient sur les bonnes pratiques médicales. Les questions sont relues par d'autres enseignants et critiquées par les étudiants au cours de l'épreuve ce qui parfois amène à des annulations de question. Ici, une seule question a été annulée et a été exclue de cette étude. Ces relectures permettent d'assurer la bonne qualité des questions-réponses utilisées dans l'étude. Pour les questions les plus difficiles, les enseignants justifient leur réponse.

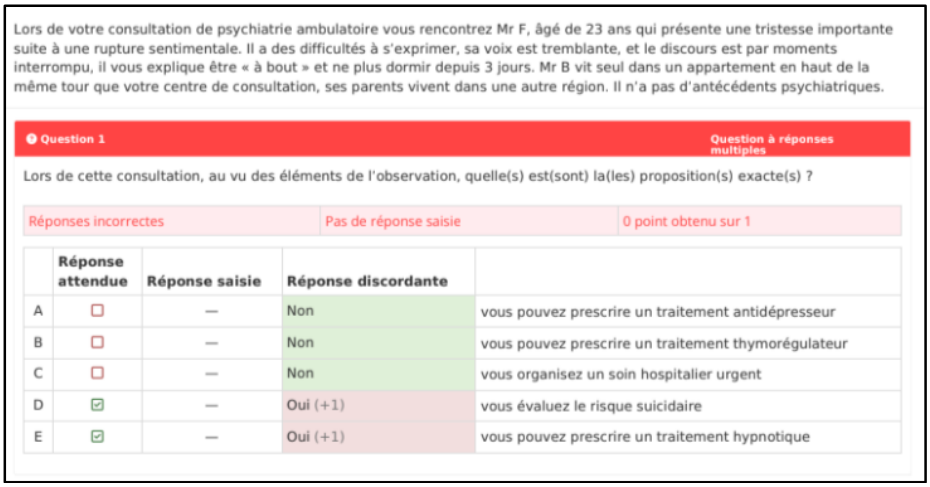
Les questions-réponses se présentent selon plusieurs modalités :

- Les questions contiennent un énoncé et le plus souvent 5 propositions de réponse (de A à E) auxquelles il faut répondre vrai ou faux.
    - Les questions à réponse multiple (QRM) sont le type de question le plus fréquemment retrouvé.
    - Les questions à réponse unique (QRU) demandent à l'étudiant de sélectionner seulement l'item de réponse le plus pertinent.
- Les questions à réponse ouverte courte (QROC) nécessitent de répondre par un ou quelques mots les plus précis et qui correspondent le plus à la question posée. Ces questions ne donnent ainsi pas d'indice via les items de réponses proposés.
- Les questions à réponses précisées (QRP) proposent une liste entre 8 et 10 réponses possibles et dont le nombre de réponses est indiqué dans la consigne. Ces questions évaluent la capacité de l'étudiant à être exhaustif et ciblé à la fois.

Ces questions-réponses sont organisé en 2 ou 3 parties dans les épreuves :

- Les questions indépendantes (QI) qui n'ont pas de lien d'une question à une autre et correspondent souvent à des questions de cours.
- Les dossiers progressifs (DP) qui sont une suite de 6 à 8 questions logiques les unes avec les autres dont il n'est pas possible de revenir en arrière après avoir répondu à la question précédente. Ils présentent un cas clinique et demandent souvent une réflexion de la part de l'étudiant afin de répondre selon le contexte de l'ensemble du dossier. Une erreur au début du DP peut se répercuter sur l'ensemble du dossier.
- Les Key Feature Problems (KFP) sont une suite de 3 QRP. Ils explorent souvent une symptomatologie d'une maladie, la prescription d'examen complémentaire pertinent ou une prise en charge. Les questions suivantes ne demandent pas d'avoir répondu juste aux questions précédentes.

Nous travaillons actuellement avec le doyen de la Faculté de médecine Lyon-Est afin de rendre cette base de données disponible en ligne. Voici un exemple du type de question posée, ici la première question d'un dossier progressif de psychiatrie (Figure 1).



Lors de votre consultation de psychiatrie ambulatoire vous rencontrez Mr F, âgé de 23 ans qui présente une tristesse importante suite à une rupture sentimentale. Il a des difficultés à s'exprimer, sa voix est tremblante, et le discours est par moments interrompu, il vous explique être « à bout » et ne plus dormir depuis 3 jours. Mr B vit seul dans un appartement en haut de la même tour que votre centre de consultation, ses parents vivent dans une autre région. Il n'a pas d'antécédents psychiatriques.

Question 1 — Question à réponses multiples

Lors de cette consultation, au vu des éléments de l'observation, quelle(s) est(sont) la(les) proposition(s) exacte(s) ?

Réponses incorrectes | Pas de réponse saisie | 0 point obtenu sur 1

	Réponse attendue	Réponse saisie	Réponse discordante	
A	☐	—	Non	vous pouvez prescrire un traitement antidépresseur
B	☐	—	Non	vous pouvez prescrire un traitement thymorégulateur
C	☐	—	Non	vous organisez un soin hospitalier urgent
D	☑	—	Oui (+1)	vous évaluez le risque suicidaire
E	☑	—	Oui (+1)	vous pouvez prescrire un traitement hypnotique

FIGURE 1 : Première question à réponses multiple d'un dossier progressif de l'épreuve de psychiatrie. A titre d'exemple, aucune réponse n'a été saisie.

2.2 Grands modèles de langue

Les grands modèles de langue usuels ont été testé dans cette étude avec GPT-4o (OpenAI, version payante) et Gemini 2.0 (Google). Ils correspondent aux GMLs utilisés par les étudiants en médecine au cours de leur étude et pendant l'internat. Nous n'avons pas eu accès à EDN GPT, un GLM propriétaire augmenté avec l'inclusion des référentiels de l'externat en France.

Avant de débuter une épreuve, nous expliquons aux GLMs les modalités de l'épreuve et les types de questions posées. Nous avons demandé aux GLMs de ne répondre que par les lettres des réponses justes sans justifier les réponses par du texte. Nous avons précisé en entrée du modèle qu'il s'agissait d'un examen destiné aux étudiants en médecine français et que les réponses s'appuient sur les recommandations nationales et européennes (Figure 2).

Je vais te poser des questions qui correspondent à un examen de médecine passé en France. Tu dois donc t'appuyer sur les recommandations médicales françaises actuelles.
Voici les modalités : il y a plusieurs dossiers progressifs et des sessions de questions individuelles. Les dossiers progressifs ont une logique intrinsèque que tu dois suivre tandis que les questions individuelles n'ont pas de liens les unes avec les autres.
Il y a plusieurs types de questions :

1. Question à réponse unique = 1 seule réponse juste ;
2. Question à réponses multiples = 1 ou plusieurs réponses justes ;
3. Question à réponse ouverte et courte = écrire 1 ou quelques mots qui répondent le mieux à la question posée ;
4. Question à réponses précisées = nombre de réponse attendue précisée.

J'aimerais que tu répondes que par les lettres qui correspondent aux items justes. Pas de justification. Es-tu prêt ?

FIGURE 2 : Instructions données aux GLMs pour leur présenter l'exercice. Une version de l'instruction ne comprend pas les indications spatiotemporelles.

2.3 Calcul de la note

Afin de pouvoir comparer les notes des GLMs à la note moyenne des étudiants en médecine de la promotion ayant passés ces examens, nous avons utilisé le même système de notation :

- pour les QRM :
  - 0 discordance entre les propositions vraies par le GLM et les réponses juste est notée 1/1
  - 1 discordance est notée 0,5/1
  - 2 discordances est notée 0,2/1
  - 3 discordances ou plus est notée 0/1
- pour les QRU
  - si la bonne réponse est donnée par le GLM, la note est 1/1, sinon 0/1.
- pour les QROC
  - si la réponse donnée par le GLM est identique à une des réponse juste, la note est 1/1 sinon 0/1
- pour les QRP
  - seule les bonnes réponses comptent, elles sont notées selon le pourcentage de réponse juste. Il n'est par ailleurs pas possible de cocher plus de réponses que le nombre précisé et aucun GLM n'a présenté ce cas de figure.

Il existe des items de réponses considérés comme inacceptable pour un étudiant en médecine et qui amène à une note de 0/1 en cas d'erreur, même si d'autres items de cette question étaient justes.

La note est rapportée sur 20 pour l'ensemble des épreuves des différentes spécialités. Une moyenne de l'ensemble des épreuves est calculée pour chaque GLM. Les moyennes sont comparées en valeur absolue. En effet, nous n'avons pas fait d'itération multiple d'un GLM sur une même question-réponse afin d'évaluer la variabilité de la réponse.

2.4 Analyse du type d'erreur

Nous avons voulu analyser le type d'erreur que pouvaient faire les GLMs. Pour cela nous avons classé les questions selon les compétences qu'elles évaluent. Ce classement s'appuie sur une observation personnelle du premier auteur Adrien KUHNAST qui, au cours de ses études de médecine, a su comprendre ses erreurs et les analyser. Cette analyse s'appuie également sur des ressources bibliographiques qui décrivent une classification de question similaire (Ganguly et al., 2024 ; Natarajan et al., 2025).

Les erreurs des GLMs sur les réponses-réponses peuvent être classées ainsi :

Erreur	Définition
De connaissance	correspond à des notions apprises “par cœur” par les étudiants. Elles sont décrites textuellement dans les supports de cours
De jugement	correspond au raisonnement clinique que les étudiants acquièrent dans leur cours et en stage
De précision	lorsque la question demande à être large ou précis sur les réponse. Il n’a pas forcément de réponse fausse mais l’énoncé indique un adjectif de précision comme “quelles sont les réponses vraies possibles” ou “quelles sont les réponses vraies les plus probables”
D’interprétation	correspond à une mauvaise interprétation du GLM de données biologiques numériques ou de coupe d’imagerie
En chaîne	lorsqu’une réponse fausse au début d’un DP se répercute sur l’ensemble du dossier
Du correcteur	lorsque nous avons un doute sur la véracité de la correction. Ces questions ont été exclues de l’étude

TABLE 1 : Classification des types d’erreur possible selon la question posée

3 Résultats

GPT-4o et Gemini 2.0 ont obtenu une note moyenne sur l’ensemble des épreuves de respectivement 15,4/20 et 15,6 ce qui est inférieur aux étudiants en médecine de cette promotion qui ont obtenu en moyenne 16/20. Ces résultats montrent que les GLMs ont été presque aussi bon que les étudiants en médecine sur cette épreuve.

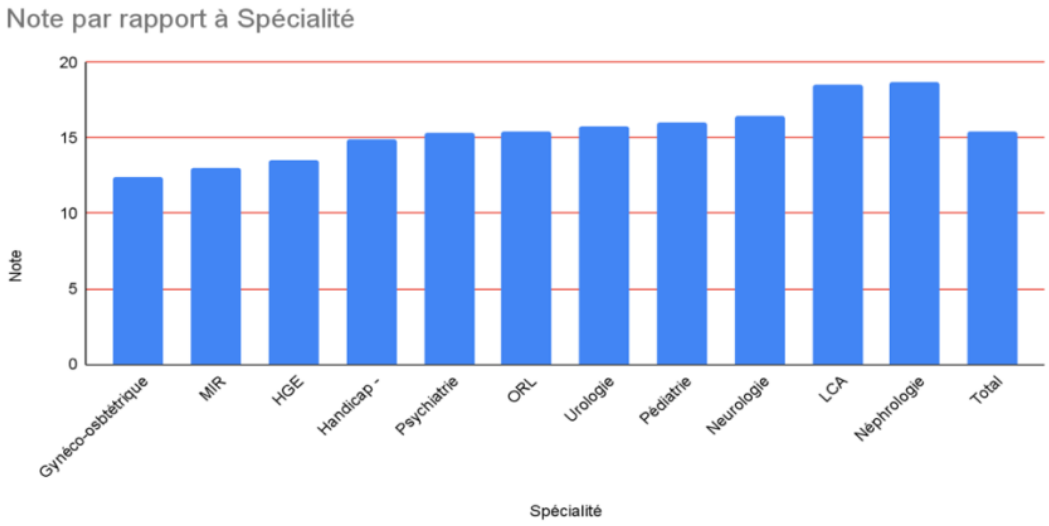


FIGURE 3 : Notes sur 20 de GPT-4o sur les différentes épreuves des examens de 5ème année médecine à Lyon-Est en 2024 et note moyenne sur l'ensemble des épreuves (Total)

Une note minimale de 10/20 est nécessaire pour chaque épreuve pour pouvoir valider la spécialité évaluée. En lecture critique d'article, GP4-4o a obtenu 18,55/20 là où les étudiants en médecine ont une moyenne de 13/20. Cette matière est redoutée par les étudiants en médecine car elle demande une lecture en anglais d'une étude clinique et de répondre à des questions sur les méthodes statistiques, sur les résultats et conclusion de l'étude en un temps imparti de 1h30. GPT-4o pourrait être un outil intéressant pour aider les étudiants dans cette matière. La note minimale qu'a obtenu GPT-4o est en gynécologie-obstétrique avec 12,4/20 et sa note maximale est en néphrologie avec 18,7/20 (Figure 3). La spécialité de gynécologie obstétrique s'appuie sur les recommandations nationales souvent différentes des recommandations anglo-saxonnes. Lorsque nous précisons à GPT-4o de s'appuyer sur des recommandations françaises, la note en gynécologie-obstétrique s'élève à 14,95/20 ce qui équivaut à la note obtenue par Gemini 2.0 (sans lui avoir précisé sur quelles recommandations s'appuyer) (Figure 4). La pédiatrie est également affectée par les variations des recommandations qui évoluent chaque année notamment concernant les vaccins obligatoires ou les maladies à déclaration obligatoire (Figure 5). Ces résultats montrent que les GLMs sont soumis aux biais linguistiques et temporels. Leur réponse est variable selon la nationalité sources qu'ils privilégient. Ils sont également figés dans le temps et ne prennent pas en considération les nouvelles recommandations. Cela démontre l'intérêt d'évaluer les GLMs dans la langue de l'utilisateur à un instant donné.

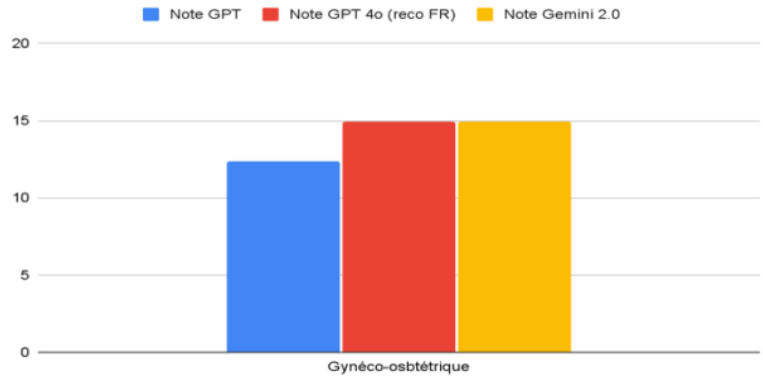


FIGURE 4 : Notes sur 20 de l'épreuve de gynécologie-obstétrique par GPT-4o, GPT-4o en lui précisant de s'appuyer sur les recommandations françaises et de Gemini 2.0

Quelle(s) proposition(s) est(sont) correcte(s) concernant la rougeole ?

Réponses incorrectes | Pas de réponse saisie | 0 point obtenu sur 1

	Réponse attendue	Réponse saisie	Réponse discordante	
A	☐	—	Non	il s'agit d'une infection cutanée bactérienne
B	☑	—	Oui (+1)	le signe de Köplick est pathognomonique
C	☐	—	Non	la confirmation paraclinique n'est pas obligatoire
D	☐	—	Non	le schéma vaccinal comporte 2 injections dans la 1ère année de vie
E	☑	—	Oui (+1)	la protection conférée par une seule dose de vaccin anti-rougeole est estimée à 90%

Commentaire de correction de la question

Il s'agit d'une infection virale à paramyxovirus : Morbillivirus. Le calendrier vaccinal recommande 2 injections à M12 et M16-18.

Il s'agit d'une maladie à Déclaration Obligatoire et une confirmation paraclinique doit être réalisée de manière systématique.

Intérêt d'une 2nde injection car seulement 90% des patients protégés avec une 1 injection et l'objectif de protection est une couverture > 95%. La France et l'Europe sont actuellement < 85% de protection.

FIGURE 5 : Question individuelle de pédiatrie. GPT-4o retient les items BCE or les réponses attendues sont BE. La rougeole n'est pas une maladie déclaration obligatoire dans tous les pays. En France, sa déclaration obligatoire nécessite une preuve le plus souvent faite sur analyse génétique d'un prélèvement nasopharyngé.

L'analyse du type d'erreur montre que GPT-4o commet principalement des erreurs de connaissances en ORL ou en gynécologie-obstétrique (Figure 6). Pour la gynécologie-obstétrique, cela peut s'expliquer par des recommandations variables selon les pays et même au sein de la France, plusieurs instances comme la Haute Autorité de Santé (HAS) et le Collège National des Gynécologues et Obstétriciens Français (CNGOF) émettent des recommandations contradictoires. Pour l'ORL, les questions font appel à des notions de cours très précises et parfois issus de recommandations récentes dont GPT-4o n'a sûrement pas connaissance. Pour répondre à ces questions, une analyse des justifications par les GLMs serait intéressante.

L'analyse du type d'erreur montre en revanche une forte proportion d'erreurs de jugement en psychiatrie (Figure 6). L'épreuve de psychiatrie demande une réflexion approfondie de la part des étudiants en médecine. Ils doivent analyser des symptômes dont le degré d'intensité et leur association les classent dans le cadre pathologique ou physiologique. Cette réflexion nécessite une bonne connaissance des cours théoriques ainsi qu'une expérience acquise en cours magistral et en stage. Cet empirisme est sûrement difficilement apprécié par les GLMs.

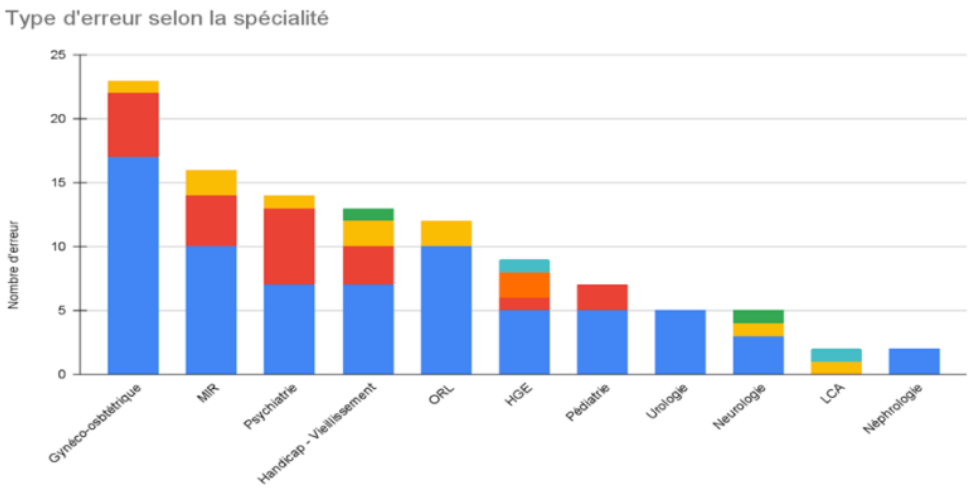


FIGURE 6 : Type d'erreur selon la spécialité par GPT-4o. La légende des couleurs ci-dessous :

- erreur de connaissance = "par coeur'
- erreur de jugement = raisonnement clinique
- erreur de précision = peu / trop précis par rapport à la question posée
- erreur d'interprétation = mauvaise interprétation d'un bilan biologique / d'une imagerie
- erreur en chaîne = une erreur au début d'un dossier progressif se répercute sur l'ensemble du dossier
- erreur du correcteur = doute sur la véracité de la correction

4 Discussion

L'utilisation pratique des GLMs par les étudiants en médecine, internes ou médecins font l'objet d'une évolution importante qui comporte des risques pour l'apprentissage de la médecine et, à terme, pour les patients (Blay et al., 2024). L'utilité et la sécurité de ces GLMs n'ont pas encore été évaluées en France sur des bases de données françaises.

Dans cette étude, nous proposons des résultats préliminaires d'évaluation de la capacité de GLMs comme GPT-4o ou Gemini 2.0 à répondre à des questions d'examen de médecine français. Ces deux GLMs sont des modèles propriétaires et nous avons peu d'information sur leur nombre de paramètre, leur configuration et leur donnée d'entraînement. Ils ont principalement été choisi car ce sont les deux agents conversationnels les plus utilisés par les étudiants en médecine en France actuellement. Ces résultats seront également confrontés à une base de données plus grande sur laquelle nous travaillons.

Les résultats de cette étude montrent que la note moyenne des GLMs sur l'ensemble des spécialités médicales est équivalente à celle des étudiants en médecine. Cependant, nous n'avons pas étudié la variabilité de la réponse des GLMs sur la même question posée. D'autres études ont montré que sur un nombre suffisamment important d'itération, le GLMs ne répondait pas toujours les mêmes items (Pal & Sankarasubbu, 2024).

On observe également que les variations géographiques des recommandations et bonnes pratiques médicales exercent une influence sur la réponse des grands modèles de langue. Une étude plus exhaustive en analysant les sources utilisées par les GLMs pour répondre aux questions serait intéressante.

L'analyse du type d'erreur montre que les GLMs ont des capacités variables selon les spécialités. Tandis que les erreurs liées aux différentes recommandations nationales sont facilement corrigeables, les erreurs de raisonnement clinique montre une limite importante des GLMs pour les questions faisant appel à l'empirisme médical. Ces analyses peuvent aussi permettre des améliorations plus ciblées des GLMs par fine-tuning ou par génération augmentée de récupération (RAG) par exemple.

Les examens d'entrée à l'internat sont une bonne méthode d'évaluation des GLMs pour la pratique quotidienne des médecins. Récemment, ces examens incluent des oraux sous la forme d'Examens Cliniques Objectifs Structurés (ECOS). La suite logique de cette étude est d'évaluer les GLMs sur ces ECOS. Pour les étudiants en médecine, cette modalité d'examen est un challenge majeur car depuis un énoncé et une consigne courte, ils doivent en 8 minutes aborder à l'oral spontanément les notions attendues. Cela diffère grandement des questions-réponses où les propositions de réponses sont sous les yeux de l'étudiant qui doit faire un choix de sélection des items vrais. Dans la préparation aux ECOS, nous avons vu émerger beaucoup d'outil de travail s'appuyant sur des agents conversationnels avec modèle de langage et du traitement automatique des langues naturelles.

Il est maintenant clair que les étudiants en médecine utilisent quotidiennement des GLMs comme Chat-GPT ou Gemini. Ils s'en servent comme un outil de travail pour mieux comprendre les mécanismes physiopathologiques, révéler les liens entre une maladie et ses symptômes, vérifier des informations ou des questions d'entrainement voire organiser leurs révisions et leur prodiguer des conseils d'apprentissages. L'utilisation des agents conversationnels et très rapide et permet de s'affranchir des longues minutes de recherches dans les référentiels de cours lorsque l'étudiant a une question précise. Le côté ludique de la génération de texte en temps réel et la clarté d'explication donne une forte crédibilité en apparence. Cette étude révèle quelques limites de ces agents conversationnels dans leur utilisation par des étudiants en médecine.

Cette étude s'inscrit également dans le développement des GLMs en pratique médicale (génération de compte rendus médicaux, agents conversationnels et assistants personnalisés, systèmes d'aide à la décision médicale, …). Cette étude montre l'importance d'évaluer ces systèmes dans la langue de leur utilisation ainsi que la mise à jour des bases de données avec les nouvelles recommandations.

5 Conclusion

Les GLMs évalués dans cette étude ont montré des performances similaires aux étudiants en médecine de Lyon sur leur épreuve de fin d'année. Cette étude montre également que les performances des GLMs sur des questions médicales varient selon la nationalité des recommandations. Ils doivent donc être évaluées dans la langue de leur utilisation. Il n'existe à ce jour aucune base de données françaises de questions-réponses qui traite précisément de questions médicales. Les questions-réponses utilisées dans cette étude sont un bon reflet des attendus d'un étudiant en médecine mais une base de données plus complète et représentative des examens d'entrée à l'internat serait plus appropriée. L'analyse de type d'erreur des GLMs permet d'identifier leur limite. Les GLMs les plus utilisés sont soumis à des biais liées à la surreprésentation de données d'entraînement anglo-saxonnes et figées dans le temps. Il serait intéressant de comparer les GLMs grand public avec des GLMs français dont on connait mieux le fonctionnement. Dans la perspective d'une utilisation des GLMs en pratique clinique, des GLMs locaux entraînés sur des données controlées pourraient permettre de conserver le secret médical et de s'affranchir des biais sus-cités.

6 Références

ALONSO I., ORONOZ M., & AGERRI R. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 2024, vol. 155, p. 102938.

SINGHAL K., AZIZI S., TU T., NATARAJAN V., *et al.* Large language models encode clinical knowledge. *Nature*, 2023, vol. 620, no 7972, p. 172-180.

SCHMIDGALL S., HARRIS C., ESSIEN I., ABAIR P. & CHELLAPPA R., *et al.*Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 2024, vol. 7, no 1, p. 295.

GUO Y., GUO M., SU J., LIU S.-S., *et al.* Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*, 2024.

ROY S., KHATUA A., GHOOCHANI F., GANGULY N., *et al.* Beyond accuracy: Investigating error types in GPT-4 responses to USMLE questions. In : *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2024. p. 1073-1082.

SINGHAL K., TU T., GOTTWEIS J., NATARAJAN V., *et al.* Toward expert-level medical question answering with large language models. *Nature Medicine*, 2025, p. 1-8.

VERLINGUE L., BOYER C., OLGIATI L., BLAY J.-Y., *et al.*Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice. *The Lancet Regional Health–Europe*, 2024, vol. 46.

PAL A. & SANKARASUBBU M. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In : *Proceedings of the 6th Clinical Natural Language Processing Workshop*. 2024. p. 21-46.