Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240486

MedFrenchmark, a Small Set for Benchmarking Generative LLMs in Medical French

Amandine QUERCIA[a,b,1], Jamil ZAGHIR[a,b], Christian LOVIS[a,b], and Christophe GAUDET-BLAVIGNAC[a,b]
[a]*Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland*
[b]*Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland*
ORCiD ID: Amandine Quercia https://orcid.org/0009-0004-3273-9288

Abstract. Generative Large Language Models (LLMs) have become ubiquitous in various fields, including healthcare and medicine. Consequently, there is growing interest in leveraging LLMs for medical applications, leading to the emergence of novel models daily. However, evaluation and benchmarking frameworks for LLMs are scarce, particularly those tailored for medical French. To address this gap, we introduce a minimal benchmark consisting of 114 open questions designed to assess the medical capabilities of LLMs in French. The proposed benchmark encompasses a wide range of medical domains, reflecting real-world clinical scenarios' complexity. A preliminary validation involved testing seven widely used LLMs with a parameter size of 7 billion. Results revealed significant variability in performance, emphasizing the importance of rigorous evaluation before deploying LLMs in medical settings. In conclusion, we present a novel and valuable resource for rapidly evaluating LLMs in medical French. By promoting greater accountability and standardization, this benchmark has the potential to enhance trustworthiness and utility in harnessing LLMs for medical applications.

Keywords. Generative AI, benchmark, NLP, LLM

1. Introduction

Large Language Models (LLMs) are a category of artificial intelligence algorithms that leverage deep learning methods and massive datasets to understand, synthesize, generate, and anticipate new textual content. They are already commonly used in fields such as finance, marketing, and many others but in the medical field, their usage is still at the research stage [1]. Improving the use of these LLMs in medicine could assist healthcare professionals by reducing their workload through faster access to knowledge and clinical decision support, ultimately reducing long-term professional burn-out [2]. The most well-known generative model specialized in dialogue is ChatGPT by OpenAI and studies have already been conducted to demonstrate the capability of this model for clinical decision support in radiology, showing the feasibility of using ChatGPT in radiological

[1] Corresponding Author: Amandine Quercia; E-mail: amandine.quercia@etu.unige.ch.

decision-making [3]. A study conducted in 2023, which created a benchmark of 31 questions about myopia, highlighted the potential of ChatGPT-4.0 to provide accurate and comprehensive responses to questions related to this topic [4]. Other models have also been studied to answer medical questions and showed "that medical answering capabilities (recall, reading comprehension and reasoning skills) improved with scale" [5]. Furthermore, new models are being released every day, making it difficult to ascertain their value, especially in medical French. Therefore, we have sought to establish a tool to quickly evaluate different generative models in medical French.

2. Methods

To create this benchmark, we decided to cover a wide range of specialties for a wide range of question types, to assess the breadth of skills of different algorithms and to determine if certain systems or categories of questions are more difficult to answer than others. Cardiology, gastroenterology, orthopedic, pneumology, and other subjectively less dense systems were grouped together to broaden the scope of knowledge to be evaluated without increasing the number of initially desired questions. The groups were assembled based on their similarities and the common questions they might have, so neurology was grouped with ENT (ear, nose, and throat), urology was grouped with nephrology, infectiology was grouped with immunology and hematology, dermatology was grouped with ophthalmology, endocrinology was grouped with gynecology, pharmacology was grouped with molecular system. As for certain broader topics chosen that could apply to different groups such as radiology, pediatrics, and oncology, they were included in a scattered manner across these groups, to avoid having too many categories.

For the question types, we aimed to mimic the abilities needed for physicians to reach the correct diagnosis. They need to be able to read a text while understanding the relevant elements related to the case, grasp the introduced concepts, engage in complex reasoning based on their medical knowledge to decide on the course of action leading to the diagnosis, and finally, they must be able to explain this entire process to patients. Therefore, questions were split into the following categories: word processing (including summarizing/finding important information/answering only the asked question/popularize), medical jargon, medical/clinical knowledge, trap spotting and complex reasoning. For each question a correct answer in French was written.

To show the utility of our benchmark, these questions were submitted to four small models each with approximately 7 billion parameters chosen for their open source availability: Llama2, Mistral, Meditron, and MedLlama2 [6–8]. Our study evaluated the performance of those models in delivering accurate responses to common medical questions. These models were tested at a temperature of T=0.8 and then T=0. The temperature is a value between 0 and 1, with 1 meaning the model will be more creative but verbose, while 0 will make the model more concise and precise. Therefore, T=0.8 is supposed to give us a more voluminous and creative response, whereas T=0 is the default setting that will provide the most precise response. A superset consisting of one question per system and two questions per task category was selected, to quickly evaluate models. Each question of the superset was posed to the LLMs, and their responses were independently graded by one medical student on a four-point accuracy scale (wrong, partially correct, correct, english answers).

3. Results

3.1. Superset

The 10 questions superset is described in Table 1. Question 9 was shortened to improve readability as it is a long description.

Table 1. list of the 10 questions belonging to the superset. Question 9 was shortened to improve readability as it is a long description.

Questions	Category task	Medical category
1. De quelle maladie la drépanocytose protège ?	Medical/clinical knowledge	Infectiology/ immunology/ hematology
2. Patient de 84 ans, connu pour une HTA et un diabète. A été opéré en 2018 pour une PTH à droite. Il fume 3 paquets par jour depuis 50 ans, il boit occasionnellement, il a deux chats depuis peu de temps, il est père de 3 enfants dont un qui travaille à la police et un autre qui a une mucoviscidose. Il a travaillé quarante ans dans une usine de textile avant de devenir comptable. Il se présente aux urgences avec une dyspnée et une perte de poids. Sa FC est à 90 BPM, il est apyrétique, et sa créatinine est à 0,7. Peux-tu me donner seulement les éléments importants qui m'aideront à poser mon diagnostic ?	Word processing	Pneumology
3. Dans quel ordre doit-on effectuer l'examen clinique d'un problème digestif entre l'Inspection, la palpation et l'auscultation ?	Medical/clinical knowledge	Gastroenterology
4. Selon la règle de Pulaski et Wallace, quel pourcentage du corps est atteint si une brulure touche les deux bras complets ?	Trap spotting	Dermatology/ophtalmology
5. Quel est le nom des fibres de collagène se trouvant dans la couche externe du périoste ?	Medical jargon	Orthopedic
6. Après la lésion du nerf abducens, le patient aura-t-il un strabisme divergent ?	Complex reasoningy	Neurology/ENT
7. Combien de couches lipidiques ont les mitochondries ?	Trap spotting	Pharmacology/ molecular system
8. Est-ce que l'albumine passe la barrière de filtration glomérulaire ?	Complex reasoning	Urology/nephrology
9. Peux-tu me résumer facilement en deux phrases maximum ce texte : Les complications à long terme de ce type d'opération incluent les obstructions du tunnel atrial, plus fréquentes dans l'opération de Mustard. Selon le placement du patch intra-auriculaire, … …Cette complication survient le plus souvent dans la deuxième et la troisième décade de vie. Les traitements médicamenteux sont souvent décevants et il faut avoir recours à la transplantation cardiaque ou au réentraînement du ventricule gauche afin de transformer le switch atrial en switch artériel.	Word processing	Cardiology
10. Comment appelle-t-on quand les menstruations sont très éloignées entre elles ?	Medical jargon	Gynecology/endocrinology

The evaluation of the 7B models results on the superset was made on a scale from 0 to 2 (0 = wrong, 1 = partially correct, 2 = correct) and EN when the language of the answer was English (and was therefore not evaluated). The questions were evaluated by

a single last year medical student. The superset covers each system and each task category with the aim of being representative. The questions were submitted to the LLMs only once without language specification to mimic real life situation.

The model that best meets the expectations is the medllama2 T=0, with three good answers and the ones that less meets the expectations are Llama2 T=0.8, Mistral T=0, Meditron T=0.8 and Medllama2 T=0.8, with only one correct answer (Table 2). Based on the superset, the category task that was the easiest to answer was complex reasoning, and the hardest category task was medical jargon with no correct answers.

Table 2. Evaluation of each question of the superset (T = temperature, 0 = wrong, 1 = partially correct, 2 = correct, EN = answer in English)

	ChatGPT	**Llama2**		**Mistral**		**Meditron**		**MedLlama2**	
	Default	**T=0.8**	**T=0**	**T=0.8**	**T=0**	**T=0.8**	**T=0**	**T=0.8**	**T=0**
1	1	0	2	2	0	0	0	EN	2
2	0	EN	EN	EN	EN	0	0	EN	EN
3	0	0	0	EN	EN	0	0	EN	EN
4	0	EN	EN	EN	EN	0	0	EN	EN
5	0	EN	0	EN	EN	0	0	0	0
6	0	EN	EN	EN	0	0	2	0	0
7	0	0	2	EN	EN	0	0	0	0
8	0	2	0	2	2	2	2	2	1
9	2	EN	EN	EN	EN	0	0	EN	1
10	0	0	0	EN	EN	0	0	EN	0

3.2. General tendencies of ChatGPT

In Table 3 we can see that the category task that showed the biggest number of good answers (42,9%) from ChatGPT is spotting a trap/finding missing information. And the category task that had the biggest number of wrong answers (55%) is medical jargon. General trends also observed a tendency to not only answer the question, and difficulty in defining important elements for making a diagnosis, as we can see in the question number 2 where it was asked to note the relevant elements that could help make a diagnosis. The different models all responded that the fact that one of the sons is a police officer is a relevant element to consider in the diagnostic process, although it is not medically relevant.

Table 3. Percentage of right, almost right and wrong answers by ChatGPT on each of the 114 questions (includes the superset) based on the category task.

	Right	Partially right	Wrong
Word processing	20%	45%	35%

Medical/clinical knowledge	28,6%	39.3%	32,1%
Spotting a trap	42,9%	19%	38,1%
Medical jargon	35%	10%	55%
Complex reasoning	36%	20%	44%

4. Discussion and Conclusions

The results of the superset evaluation showed that the tested models had a problem with answering to medical jargon questions. Moreover, 33/90 answer were in English and therefore out of scope of this paper. This shows that those models are limited in their handling of French language, perhaps due to the limited availability of resources in French in medical literature. LLMs being language-agnostic, meaning that when writing in French, the answer of the model will not necessarily be in the same language, in our case, an answer in English is irrelevant and must be considered wrong.

Considering the relevance of the questions, these questions were written by a single person. To address this subjectivity, the benchmark would benefit from peer review. Moreover, since it was created in a French-speaking hospital in Switzerland, it would benefit from a validation by a broader panel of specialists from various French-speaking hospitals, including hospitals in Belgium, France, and Luxembourg. Finally, the coverage of the benchmark could be improved by increasing the quantity of questions, adding systems, and defining more precise categories.

In conclusion, a new 114 questions benchmark is proposed for Medical French evaluation of Generative LLMs, and a superset of 10 questions is defined for quick capability overview. Four LLMs and ChatGPT-3.5-turbo were evaluated against this superset showing limited performance with a strong bias toward English language. The benchmark is available on request by contacting the corresponding author.

References

[1] Thirunavukarasu AJ, et al. Large language models in medicine. Nat Med [En ligne]. 17 juil 2023 [cité le 4 juin 2024]. Disponible : https://doi.org/10.1038/s41591-023-02448-8

[2] Moreno AC, Bitterman DS. Toward Clinical-Grade Evaluation of Large Language Models. Int J Radiat Oncol Biol Phys [En ligne]. Mars 2024 [cité le 4 juin 2024];118(4):916-20.

[3] Rao A, Dreyer KJ, Succi MD. Reply to Letter to the Editor : Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. J Am Coll Radiol [En ligne]. Sep 2023 [cité le 4 juin 2024]. Disponible : https://doi.org/10.1016/j.jacr.2023.08.026

[4] Lim ZW, Pushpanathan K, Yew SM, Lai Y, Sun CH, Lam JS, Chen DZ, Goh JH, Tan MC, Sheng B, Cheng CY, Koh VT, Tham YC. Benchmarking large language models' performances for myopia care : a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. eBioMedicine [En ligne]. Sep 2023 [cité le 4 juin 2024];95:104770. Disponible : https://doi.org/10.1016/j.ebiom.2023.104770

[5] Singhal K, Azizi S, et al. Large language models encode clinical knowledge. Nature [En ligne]. 12 juil 2023 [cité le 4 juin 2024]. Disponible : https://doi.org/10.1038/s41586-023-06291-2

[6] Touvron H, et al. arXiv.org [En ligne]. Llama 2 : Open Foundation and Fine-Tuned Chat Models ; 2023 [cité le 4 juin 2024]. Disponible : http://arxiv.org/abs/2307.09288

[7] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D de las, et al. arXiv.org [En ligne]. Mistral 7B ; 2023 [cité le 4 juin 2024]. Disponible : http://arxiv.org/abs/2310.06825

[8] Chen Z, et al. arXiv.org [En ligne]. MEDITRON-70B : Scaling Medical Pretraining for Large Language Models ; 2023 [cité le 4 juin 2024]. Disponible : http://arxiv.org/abs/2311.16079