Accepted Manuscript

Submission Date: 2024-09-25
Accepted Date: 2025-01-18
Accepted Manuscript online: 2025-01-29

Methods of Information in Medicine

Leveraging Guideline-Based Clinical Decision Support Systems with Large Language Models: A Case Study with Breast Cancer

Solène DELOURME, Akram REDJDAL, Jacques BOUAUD, Brigitte SEROUSSI.

Affiliations below.

DOI: 10.1055/a-2528-4299

Please cite this article as: DELOURME S, REDJDAL A, BOUAUD J et al. Leveraging Guideline-Based Clinical Decision Support Systems with Large Language Models: A Case Study with Breast Cancer. Methods of Information in Medicine 2025. doi: 10.1055/a-2528-4299

Conflict of Interest: The authors declare that they have no conflict of interest.

Abstract:
Background: Multidisciplinary tumor boards (MTBs) have been established in most countries to allow experts collaboratively determine the best treatment decisions for cancer patients. However, MTBs often face challenges such as case overload, which can compromise MTB decision quality. Clinical decision support systems (CDSSs) have been introduced to assist clinicians in this process. Despite their potential, CDSSs are still underutilized in routine practice. The emergence of large language models (LLMs), such as ChatGPT, offers new opportunities to improve the efficiency and usability of traditional clinical decision support systems (CDSSs).

Objectives: OncoDoc2 is a guideline-based CDSS developed using a documentary approach and applied to breast cancer management. This study aims to evaluate the potential of LLMs, used as question-answering (QA) systems, to improve the usability of OncoDoc2 across different prompt engineering techniques (PETs).

Methods: Data extracted from breast cancer patient summaries (BCPSs), together with questions formulated by OncoDoc2, were used to create prompts for various LLMs, and several PETs were designed and tested. Using a sample of 200 randomized BCPSs, LLMs and PETs were initially compared on their responses to OncoDoc2 questions using classic metrics (accuracy, precision, recall, and F1 score). Best performing LLMs and PETs were further assessed by comparing the therapeutic recommendations generated by OncoDoc2, based on LLM inputs, to those provided by MTB clinicians using OncoDoc2. Finally, the best performing method was validated using a new sample of 30 randomized BCPSs.

Results: The combination of Mistral and OpenChat models under the enhanced zero-shot PET showed the best performance as a question-answering system. This approach gets a precision of 60.16%, a recall of 54.18%, an F1 Score of 56.59%, and an accuracy of 75.57% on the validation set of 30 BCPSs. However, this approach yielded poor results as a CDSS, with only 16.67% of the recommendations generated by OncoDoc2 based on LLM inputs matching the gold standard.

Conclusions: All the criteria in the OncoDoc2 decision tree are crucial for capturing the uniqueness of each patient. Any deviation from a criterion alters the recommendations generated. Despite a good accuracy rate of 75.57% was achieved, LLMs still face challenges in reliably understanding complex medical contexts and be effective as CDSSs.

Corresponding Author:
Master of Science, Level 2 Biomedical Informatics Solène DELOURME, Sorbonne Universite, Paris, France, solene.delourme@epita.fr

Affiliations:
Solène DELOURME, Sorbonne Universite, Paris, France
Solène DELOURME, LIMICS, Paris, France
Solène DELOURME, EPITA, Le Kremlin-Bicetre, France
[...]
Brigitte SEROUSSI, AP-HP, Paris, France

Thieme

Leveraging Guideline-Based Clinical Decision Support Systems with Large Language Models: A Case Study with Breast Cancer

Solène DELOURME [a,b], Akram REDJDAL[c,a,], Jacques BOUAUD [a] and Brigitte SEROUSSI [d,e]

[a] *Sorbonne Université, Université Sorbonne Paris Nord, INSERM, LIMICS, Paris, France*

[b] *Epita, Paris, France*

[c] *Laboratoire de Biomécanique Appliquée, Université Gustave-Eiffel, Aix-Marseille Université, Marseille, France*

[d] *Sorbonne Université, AP-HP, Tenon Hospital, Public Health Department, INSERM, Université Sorbonne Paris Nord, Limics 75006 Paris, France*

[e] *APREC, Paris, France*

Corresponding Author: Akram REDJDAL, Email: redjdalakram300@gmail.com, Mailing address: 8 avenue de ceinture 9400, Creteil, France.

Abstract

Background: *Multidisciplinary tumor boards (MTBs) have been established in most countries to allow experts collaboratively determine the best treatment decisions for cancer patients. However, MTBs often face challenges such as case overload, which can compromise MTB decision quality. Clinical decision support systems (CDSSs) have been introduced to assist clinicians in this process. Despite their potential, CDSSs are still underutilized in routine practice. The emergence of large language models (LLMs), such as ChatGPT, offers new opportunities to improve the efficiency and usability of traditional clinical decision support systems (CDSSs).*

Objectives: *OncoDoc2 is a guideline-based CDSS developed using a documentary approach and applied to breast cancer management. This study aims to evaluate the potential of LLMs, used as question-answering (QA) systems, to improve the usability of OncoDoc2 across different prompt engineering techniques (PETs).*

Methods: *Data extracted from breast cancer patient summaries (BCPSs), together with questions formulated by*

OncoDoc2, were used to create prompts for various LLMs, and several PETs were designed and tested. Using a sample of 200 randomized BCPSs, LLMs and PETs were initially compared on their responses to OncoDoc2 questions using classic metrics (accuracy, precision, recall, and F1 score). Best performing LLMs and PETs were further assessed by comparing the therapeutic recommendations generated by OncoDoc2, based on LLM inputs, to those provided by MTB clinicians using OncoDoc2. Finally, the best performing method was validated using a new sample of 30 randomized BCPSs.

Results: *The combination of Mistral and OpenChat models under the enhanced zero-shot PET showed the best performance as a question-answering system. This approach gets a precision of 60.16%, a recall of 54.18%, an F1 Score of 56.59%, and an accuracy of 75.57% on the validation set of 30 BCPSs. However, this approach yielded poor results as a CDSS, with only 16.67% of the recommendations generated by OncoDoc2 based on LLM inputs matching the gold standard.*

Conclusions: *All the criteria in the OncoDoc2 decision tree are crucial for capturing the uniqueness of each patient. Any deviation from a criterion alters the recommendations generated. Despite a good accuracy rate of 75.57% was achieved, LLMs still face challenges in reliably understanding complex medical contexts and be effective as CDSSs.*

Keywords

Clinical Decision Support Systems, Oncodoc2, Breast Cancer, Large Language Models, Question-Answering Systems

1 Introduction

The International Agency for Research on Cancer (IARC) predicts a 77% increase in cancer cases between 2022 and 2050[1]. As the incidence rises, the management of breast cancer is becoming increasingly complex. Multidisciplinary tumor board meetings (MTBs) have been established to support therapeutic decision-making for cancer patients. However, they are often confronted to clinical case overloads and limited discussion time, which affects the quality of care. Clinical decision support systems (CDSSs) have proven to be effective tools to assist clinicians in their decision-making process[2]. When based on clinical practice guidelines (CPGs), CDSSs provide patient-centered therapeutic recommendations aligned with guidelines. However, despite their potential to improve the compliance of MTB decisions with CPGs, many guideline-based CDSSs remain underutilized, and only few have successfully been integrated into routine clinical practices due to technical or usability barriers[3].

OncoDoc[4] and OncoDoc2[5] are decision support systems for the management of breast cancer, based on a documentary approach. Clinicians navigate through a knowledge base structured as a decision tree, answering questions to describe a patient personal and familial antecedents, general state, and tumor characteristics, to finally obtain patient-specific recommendations based on Cancer Est guidelines[5]. OncoDoc2, has been routinely used at the Tenon Hospital (Assistance Publique – Hôpitaux de Paris, Paris, France) between February 2007 and October 2009 showing a 91.7% compliance of Tenon MTB decisions with OncoDoc2 across about 2,000 decisions[5]. However, usability issues with OncoDoc2 have finally hindered its routine use as a CDSS.

The emergence of large language models (LLMs), such as OpenAI's ChatGPT in 2022, has opened new possibilities for enhancing CDSSs. LLMs are artificial intelligence (AI) algorithms based on deep neural network architectures like Transformers[6]. Trained on vast amounts of textual data, these models are able to generate text with varying degrees of coherence and contextual relevance. Despite their promise, LLMs currently face limitations in understanding highly specialized medical contexts, often generating outputs that look syntactically coherent but may lack of medical semantic accuracy. This reliability gap is a critical barrier to their application in clinical settings. In addition to accuracy, sustainability plays a key role in the selection and evaluation of LLMs, particularly in healthcare settings as computational efficiency can preserve the planet natural resources and reduce operational costs.

Nevertheless, the capabilities of LLMs for analyzing and synthesizing large volumes of textual data can be particularly relevant in oncology, where the volume of information is substantial, and the personalization of treatments is crucial[7]. Thus, LLMs could enhance CDSSs, if their outputs are carefully evaluated to ensure they provide medically sound recommendations. In the case of OncoDoc2, LLMs may offer the opportunity to automatically answer the CDSS questions based on patient records and summaries, thus automating the navigation through the decision tree, potentially increasing the system's usability. Yet, the extent to which LLMs can accurately interpret the specific medical context of each patient and provide reliable, patient-centered recommendations remains an open question.

To test this hypothesis, we developed and evaluated a Question-Answering (QA) system based on open-access LLMs to automate OncoDoc2 decision support process for breast cancer treatment. The aim was to generate best patient-specific therapeutic recommendations by integrating OncoDoc2 decision tree within the LLM reasoning process, based on breast cancer patient summaries (BCPSs), streamlining the decision-making process for MTB clinicians. Through this work, the aim was to use LLMs to improve an existing CDSS while addressing the following research questions:

1. Can LLMs accurately navigate OncoDoc2 decision tree based on patient data extracted from BCPSs?
2. How effective are different prompting techniques in enhancing LLM performance for clinical decision-making?
3. What are the limitations and future possibilities for integrating LLMs into guideline-based CDSSs to enhance the quality of care for breast cancer patients?

2 Materials

2.1 OncoDoc2

OncoDoc is a guideline-based clinical decision support system (CDSS) designed for the management of breast cancer patients[4]. Developed in a documentary paradigm, it allows users to interactively navigate its knowledge base structured as a decision tree (see Figure 1). By answering questions that characterize a patient's specific condition, users can obtain tailored therapeutic recommendations. The decision tree includes 69 clinical parameters (nodes), corresponding to decision variables. Arcs of the decision tree correspond to the values of the clinical parameters. The leaves of the tree give the treatment proposals recommended for the patient profiles represented by the paths through the decision tree, leading to the leaves. The following version of OncoDoc, OncoDoc2, has been extended to provide an enriched decision tree made of 2,305 possible paths, covering therapeutic decisions according to clinical practice guidelines (CPGs) for non-metastatic breast cancer. We used the set of 1,886 decisions made between February 2007 and October 2009, when OncoDoc2 was routinely used at the Tenon hospital (Assistance Publique – Hôpitaux de Paris, Paris, France). Decisions were summarized in an Excel spreadsheet of 1,886 rows. Each row represented the decision made for patient cases discussed during MTB meetings (see an example in Appendix 1) and described in 69 attributes. The XML representation of OncoDoc2 decision tree, consisting of the 69 clinical attributes and 2,305 paths, was used to make the LLM automatically navigate the decision tree until getting recommendations, that we compared with the recommendations obtained by MTB clinician navigations for the same patient cases, considered as the gold standard.

The study where OncoDoc2 has been used in the MTB of Tenon hospital was declared to the Comité Consultatif de Protection des Personnes en matière de Recherche Biomédicale (Institutional Review Board) of Saint-Antoine Hospital in Paris, as well as to the Commission Nationale de l'Informatique et des Libertés (French Data Protection Authority), ensuring compliance with all applicable legal and ethical standards. In line with institutional policies of Greater Paris University Hospitals (AP-HP), patients were informed that their health data could be reused for research purposes and were made aware of their right to object to such reuse.

2.2 Breast cancer patient summaries

Breat cancer patient summaries (BCPSs) are narrative natural language documents that summarize the clinical situation of a breast cancer patient, describe the reasoning process to establish the cancer diagnosis, and provide the collective decision made by MTB clinicians. BCPSs contain information about the patient, examination results, tumor characteristics, and the treatments already received. An example of a BCPS is shown in Figure 2. The information necessary to accurately describe a clinical case should exist in BCPSs. However, it happens that BCPSs are incomplete or inconsistent, making the task of extracting relevant information from BCPSs even more challenging. Besides, each BCPS is unique, making the overall information extraction process complex.

For this study, we used a set of 230 BCPSs randomly selected among the set of BCPSs for which the MTB clinician navigation through OncoDoc2 was available in the provided Excel file (see section 3.1). BCPSs were retrieved from the Tenon Hospital data warehouse (Assistance Publique – Hôpitaux de Paris, Paris, France). The dataset included cases treated between February 2007 and October 2009. Among the 230 BCPSs, 200 were randomly selected for the evaluation ($S^{Evaluation}$) and the 30 remaining BCPSs were used for the system validation ($S^{Validation}$). Among $S^{Evaluation}$, a sample of 20 randomly selected BCPSs were used for the model selection ($S^{Selection}$).

2.3 Large Language Models (LLMs)

Since BCPSs contain personal health data, we chose to use open-source large language models that may be processed locally to ensure that personal health data is not sent to external servers (as it is the case with ChatGPT), thereby guaranteeing the confidentiality and security of this sensitive data. As (i) Orca and Gemma have been trained on medical data[8-10], (ii), Openchat has been trained on raw unstructured data[12], (iii) Llama has proven its effectiveness as a question-answer system[9], and (iv) Mistral has shown good performance in real-world applications[11], we chose to work with Orca 7B[8] (Microsoft), Llama 7B[9] (Meta), Gemma 7B[10] (independent researchers), Mistral 7B[11], and Openchat 7B[12] (independent researchers). Additionally, we tested a combination of models to assess potential performance gains. We also tested Mixtral 8x7B[13] on the best prompt engineering technique studied to test the performance of a larger model.

3 Methods

Figure 3 illustrates the methodology we implemented to develop the Q/A system, working with BCPSs and OncoDoc2. The process involves retrieving data from BCPSs, creating prompts for LLMs, generating responses with OncoDoc2. We conducted a two-step evaluation by (i) comparing LLM and MTB clinician answers to characterize a clinical case, and (ii) comparing the treatment recommendations generated for the clinical case by LLMs' and MTB clinicians' use of OncoDoc2.

1. **Data retrieval**: The first step was to retrieve the navigations performed by clinicians during MTBs and to gather all the questions asked by OncoDoc2 in order to compare MTB clinician navigations with the navigations generated by LLMs. We also collected data from BCPSs (patient history, pathology data, operative information, etc.) to provide the necessary context for LLMs to navigate.

2. **Creation of instructions for LLMs:** With the retrieved information, we created the instructions for LLMs. Various prompt engineering techniques (PETs) have been used to guide the model towards the expected answers. The instructions, in French language, include the patient history, the question asked, and possible

answer options.

3. **Generation of LLM responses**: The instructions were provided to the various LLMs, and the generated texts were processed to get LLM responses to be used to evaluate the method (see Figure 4).

4. **Two-step evaluation of LLM performance:**

   a. First, for each Oncodoc2 question, we assessed the concordance of LLM responses with the responses given by MTB clinicians for the same questions (details in Section 3.3) considered as the gold standard (GS). We used accuracy, precision, recall, and F1 Score to evaluate LLM performance as Question/Answer systems.

   b. Then, we evaluated LLM ability to provide correct therapeutic recommendations to a given patient clinical case described by her BCPS by comparing the recommendations proposed by OncoDoc2 following the MTB clinician navigation to those proposed by OncoDoc2 following the LLM navigation for the same patient.

Each LLM was first downloaded and loaded onto the Ollama platform[14] to allow the models to be deployed locally without additional configuration. In this way, models' execution did not require an internet connection and security of health data was guaranteed. We then used the Ollama API to retrieve the responses generated by the different LLMs. Figure 4 resumes the pipeline developed.

3.1 *Selection of LLMs for Question Answering*

In order to select the LLMs to work with, we tested five different LLMs: Mistral, Openchat, Orca2, Gemma, and Llama (cf. Section 2.3) on their ability to answer the OncoDoc2 questions and on their ecological impact. We used the sample of 20 BCPSs randomly selected ($S^{Selection}$) from the 200 used for model evaluation ($S^{Evaluation}$). For each BCPS, we used a set of questions from OncoDoc2 decision tree. The prompts were executed twice to verify models' reproducibility and the results were compared to the answers provided by MTB clinicians for each question.

We also considered the size of the five models to assess their overall ecological impact, and measured the carbon footprint of LLM computations on the 20 BCPSs of $S^{Selection}$, particularly when calling LLMs to generate responses to OncoDoc2 questions (we used a Python package called CarbonTracker[15]).

We finally selected the two models that demonstrated the highest performance while maintaining the lowest environmental impact, denoted BPM1 and BPM2, for the training and evaluation phases of the analysis.

3.2 *Prompt Engineering*

The following prompting engineering techniques (PETs) were applied to evaluate LLM performance:

- **Zero-Shot technique**[16] (see Appendix 2) involves direct instruction without providing any example to the model (e.g., “Cancer with breast tumor? Yes/No”).
- **Enhanced Zero-Shot technique** (see Appendix 3) refines the formulation of questions and answer options from OncoDoc2 decision tree without providing examples. For instance, the question: “Cancer with breast tumor? with option 1: Yes, and option 2: No” becomes: “Does the patient have a cancer with a breast tumor? with option 1: Yes, the patient has a cancer with a breast tumor and option 2: No, the patient does not have a cancer with a breast tumor.”
- **Zero-Shot Chain-of-Thought technique (Zero-Shot CoT)**[17] (see Appendix 4) encourages the model to follow a series of reasoning steps before providing an answer. This allows the model to structure more logically and coherently its reasoning process (e.g., 'Based on pathology data, determine if cancer is present, then decide if the tumor is a breast tumor, think step by step').

3.3 *Evaluation of LLMs used as Question-Answering Systems*

Using the two best-performing models (BPM1 and BPM2), we automated the method presented in section 3.1 (see Figure 5) for the 200 BCPSs of $S^{Evaluation}$. For each question, the answer of LLMs was compared to the answer given by MTB clinicians for the same question, and the same clinical patient case, at the moment the case was discussed. We evaluated the accuracy for all questions for all BCPSs. The results were categorized into three groups: accuracy above 80% (**high**), between 60% and 80% (**average**), and below 60% (**low**). Based on this evaluation, we kept the best-performing prompt technique for each question. The Enhanced Zero-Shot technique was only used when Zero-Shot accuracy was below 100%. For questions where the accuracy was low for both BPM1 and BPM2, we used the Zero-Shot CoT technique to attempt improvement. We also evaluated the Zero-Shot CoT technique across all BCPSs to assess its overall performance.

3.4 *Evaluation of LLMs used as Decision Support Systems*

Building on the previous evaluation, the top-performing LLMs and prompt engineering techniques were applied to the 200 BCPSs from S^{Evaluation} to assess how accurately the LLMs could generate appropriate therapeutic recommendations (as illustrated in Figure 6). We supplied the LLMs with BCPSs and the root question of OncoDoc2 decision tree. Based on the responses generated at each level from LLM inputs, the system navigated through OncoDoc2, progressing to the leaf level to obtain therapeutic recommendations. The evaluation focused on comparing the recommendations issued by the navigation performed by the LLM (denoted Recos^{LLM}) to the recommendations issued from the navigation performed by MTB clinicians considered as the gold standard (denoted Recos^{GS}). We made the difference between three main situations:

- When Recos^{LLM} = Recos^{GS}, then **Conf (Recos^{LLM}, Recos^{GS}) = identical**.
- When Recos^{LLM} ≠ Recos^{GS} but the care plans of both Recos^{LLM} and Recos^{GS} were made of the same treatment modalities (surgery, radiotherapy, hormone therapy, chemotherapy, or targeted therapies), AND were organized in the same order, we explored each proposed treatment modalities.
    - For surgery, we proposed a classification of surgery modalities (see Appendix 5) and an expert oncologist specified which surgery modalities could be considered as comparable.
    - For other modalities (radiotherapy, hormone therapy, chemotherapy, and targeted therapies), we considered that they were comparable as long as the modality in Recos^{LLM} was subsumed by the modality in Recos^{GS}.

    When all treatment modalities in the LLM proposal were *comparable* to those of the gold standard and proposed in the same order, then we considered the recommendations were comparable, i.e., **Conf (Recos^{LLM}, Recos^{GS}) = comparable.**
- When recommendations were neither identical nor comparable, then **Conf (Recos^{LLM}, Recos^{GS}) =**

different.

3.5 *Validation*

Validation was performed using the reserved 30 BCPSs of $S^{Validation}$. For each case, LLM inputs were compared to MTB-derived navigations. We conducted the validation of LLMs as question/answering systems and as decision support systems using a combination of BPM1 and BPM2 in an Enhanced Zero-Shot framework, where the models were used together while selecting the best-performing model for each question asked during the navigation throughout OncoDoc2 decision tree, based on their results on the evaluation dataset.

4 Results

4.1 Selection of LLMs

Figure 8 presents the performance of each of the five models with Zero-Shot prompt engineering technique (PET) to evaluate the best LLM on the sample of 20 BCPSs of $S^{Selection}$. Models are displayed by accuracy computed on a total of 220 questions. The number of correct answers is indicated on each bar. Results show that Mistral and OpenChat are the best performing models with an accuracy of [63.5% - 63.9%] and [69.9% - 70.3%] resp. Figure 9 illustrates the ecological impact of the five models on the same BCPSs. Again, Mistral and OpenChat demonstrated the lowest CO2 emissions, with Mistral's impact ranging from 0.48 gCO2eq to 0.54 gCO2eq and OpenChat ranging from 0.36 gCO2eq to 0.39 gCO2eq. Based on these results, **Mistral** (BPM1) and **OpenChat** (BPM2) were selected for the following evaluations in this study.

4.2 Comparison of prompt engineering techniques

We worked on the 200 randomly selected BCPSs of the evaluation sample $S^{Evaluation}$, corresponding to a total of 3,142 prompts concerning the 69 questions of OncoDoc2 decision tree. We analyzed various PETs on Mistral and OpenChat, the Zero-Shot, the Zero-Shot Chain-of-Thought, and the Enhanced Zero-Shot for Mistral 7B and OpenChat models. For the Enhanced Zero-Shot PET, which yielded the best results, we also used Mixtral 8x7B (a larger model with more parameters). Table 1 presents the results of the different PETs used on the three models, based on precision, recall, F1 Score, and accuracy metrics.

The simple Zero-Shot PET showed variable results, with OpenChat (achieving a precision of 72.22% and an accuracy of 69.95%), significantly outperforming Mistral (with 54.27% and 61.78%, resp.). The enhanced Zero-Shot PET enhanced models' performance: Mixtral 8x7B stands out with an accuracy of 77.08% (which may be explained by the fact it is the largest model), and OpenChat achieved the best overall performance with a precision of 73.54% and an accuracy of 72.47%. Among the evaluated techniques, Enhanced Zero-Shot consistently outperformed Zero-Shot and Zero-Shot CoT in precision, recall, and F1 Score, indicating its superiority to disambiguate questions and guide models. Zero-Shot CoT, while intended to encourage logical reasoning, underperformed in simpler questions due to hallucination effects, reducing accuracy.

4.3 Assessment of LLMs used as Question-Answering Systems

We worked on the same sample of 200 randomly selected BCPSs of $S^{Evaluation}$. We analyzed the distribution of all 69 questions used to navigate the OncoDoc2 decision tree with the enhanced zero-shot prompt engineering (PE), previously identified as the most effective technique based on the accuracy achieved during the training phase step (see Table 2).

A- LLM accuracy on OncoDoc2 questions

Figure 10 presents models' accuracy in answering OncoDoc2 questions with the percentages of correct answers, according to the three groups, high, average and low (ranges 60%, 80%, and 100%). This distribution illustrates the performance of Mistral and OpenChat models, with the enhanced Zero-Shot PET (larger versions of figures are presented in Appendix 6 and Appendix 7, and Zero-Shot CoT distributions for both models are presented in Appendixes 8 and 9).

The enhanced Zero-Shot PET results with Mistral are quite variable. For example, for the question about the presence of a breast tumor, the model achieves 97.49% of correct answers (194/200 BCPSs) but drops to 4.9% for the question about the existence of anthracyclin contraindications (3/61 BCPSs). OpenChat demonstrated higher overall accuracy compared to Mistral e.g., achieving 98.40% accuracy for anthracyclin contraindications (60/61 BCPSs). However, Mistral excelled in specific questions, with 27 questions (nodes of OncoDoc2 decision tree) achieving > 80% accuracy, highlighting its capacity to better handle nuanced scenarios than OpenChat, even though more than half of these 27 questions were actually rarely asked (14/27 questions were asked on less than 10 BCPSs with a performance above 80%).

B- Choosing the best model for each OncoDoc2 question

Table 3 presents the results of the enhanced Zero-Shot and Zero-Shot CoT PETs, combining both Mistral and OpenChat models, based on their performance on OncoDoc2 questions as previously obtained (See Figure 10). Questions were assigned to each model based on their respective performance in order to maximize the overall efficiency. Table 3 shows that both enhanced Zero-Shot and Zero-Shot CoT PETs achieved quite similar scores in terms of precision, recall, F1 Score, and accuracy.

4.4 Assessment of LLMs used as Decision Support Systems

We worked on $S^{Evaluation}$ to compare the recommendations generated by Mistral and OpenChat ($Recos^{LLM}$) to the recommendations obtained by MTB clinicians ($Recos^{GS}$) (see Section 3.4) and assess whether, based on LLM inputs, the provided recommendations were identical, comparable, or different from the GS. The results of this comparison are presented in Table 4. We compared two PETs:

- The Enhanced Zero-Shot PET using Mistral alone, OpenChat alone and the combination of both models,
- The Zero-Shot COT with the combination of Mistral and Openchat.

For the Enhanced Zero-Shot PET, the combination of Mistral and OpenChat achieved better results compared to using the models separately, with 17.91% identical recommendations, 19.40% comparable recommendations, and 62.68% different recommendations. In contrast, the Zero-Shot CoT PET with the combined models produced 7.46% identical recommendations, 14.92% comparable recommendations, and 79.10% of recommendations categorized as different.

4.5 Validation

The validation phase was conducted on the sample of 30 BCPSs, randomly selected for validation ($S^{Validation}$). We first evaluated the performance of the combination of Mistral and Openchat as a Q/A system using the Enhanced Zero-Shot PET on OncoDoc2 questions (as represented in Table 5). Then, we assessed LLM ability to navigate OncoDoc2 and generate recommendations (as represented in Table 6). The similarity in results of Table 5 and Table 6 confirms that LLMs performed reasonably well as question-answering systems, with a precision of

60.16%, recall of 54.18%, F1 Score of 56.59%, and accuracy of 75.57%. However, used as decision support systems, they provided poor results, with only 3.34% identical recommendations, 13.33% comparable, and 83.33% different recommendations. These results underscore the challenges of aligning LLM-driven navigations with MTB clinician practices, particularly in cases requiring nuanced contextual understanding.

5 Discussion

Many studies have evaluated the use of LLMs as CDSSs. A scoping review of 21 studies was recently conducted[18] focusing on studies that used LLMs as CDSSs, and found that the majority of studies (12/21) use LLMs to address clinical cases. The review showed that performance could vary depending on the wording of the questions and the source of the clinical cases. Studies using real patients (5/12) showed lower results (16% to 83%) compared to fictitious patients (58% to 98%). The use of fictitious data was mainly due to confidentiality concerns. Open-source models like Llama[9], which are recommended in the medical field for data security, were used by only three studies ([19], [20], [21]). Users were *favorable* to using ChatGPT as a CDSS in seven studies, *moderate* in six studies, *neutral* in four studies and *non-favorable* in four others. Despite varying performance, perceptions were generally positive, even for studies with average or low performance[18].

In this work, we wanted to validate the capability of open LLMs to augment an existing CDSS. We started the work with an evaluation to select the best performing models. The results of evaluating five LLMs on 20 BCPSs ($S^{Selection}$) showed that Mistral and OpenChat stood out for their performance and low ecological impact compared to the others. OpenChat demonstrated slightly higher accuracy with a range of 69.90% to 70.30% over the two evaluations conducted, compared to 63.50% to 63.90% for Mistral. In terms of carbon impact, OpenChat also showed lower results (0.36 gCO2eq - 0.39 gCO2eq) compared to Mistral (0.48 gCO2eq - 0.54 gCO2eq). While these criteria guided our selection, it is crucial to critically consider the trade-off between ecological efficiency and clinical accuracy. Our choice reflects a balance between these factors and aligns with Rillig et al. (2023)[22], which underscores the importance of energy efficiency to mitigate the ecological impact of LLMs.

Prompt engineering improved LLM performance, with the enhanced Zero-Shot outperforming both simple Zero-Shot and Zero-Shot CoT PETs. The success of enhanced Zero-Shot in disambiguating questions and guiding binary answers (e.g., "Yes, the patient has breast tumor cancer" and "No, the patient does not have breast tumor cancer") provided valuable insight into how contextual clarity can improve model responses. Applying the Zero-Shot CoT technique structured models' reasoning process more coherently for some complex questions. However, the decrease in performance with Zero-Shot CoT raised some concerns about potential hallucinations in PET strongly using reasoning. Our hypothesis is that there are many simple questions to which LLMs can easily respond (if the answer is in the context provided), and when adding step-by-step reasoning with Zero-Shot CoT, LLMs may start to hallucinate. For example, for the question on oestrogen receptors (OR), the percentage of correct answers when switching from enhanced Zero-Shot to Zero-Shot CoT, dropped from 73.70% to 21.10% for Mistral, and from 60% to 26.30% for OpenChat, suggesting that the added reasoning steps may introduce errors whereas the question is relatively simple. Although results with enhanced Zero-Shot are promising, they reflect specific conditions (when customizing the Zero-Shot prompts) tied to OncoDoc2 decision tree and BCPS dataset. Generalizability to other CDSSs or clinical contexts remains uncertain.

Despite OpenChat showed higher overall accuracy than Mistral (72.47% vs. 66.10%), Mistral outperformed OpenChat in specific questions, achieving a higher proportion of results with over 80% accuracy. This suggests that OpenChat is more consistent across different questions, while Mistral excels at handling specific cases. These results can be attributed to OpenChat being pre-trained with varied data and primarily using a Zero-Shot PET, allowing it to respond effectively even with unstructured information, such as the one retrieved in BCPSs [12]. Combining the two models allowed to leverage their respective performances. A similar approach was described by Yu et al. [19], where the authors compared treatment options proposed by four different models and retained the options chosen by at least two LLMs. They found that combining the treatment options of LLMs produced better results than each model individually. In our context, avoiding excessive resource consumption by combining the models based on their performance also showed promising results suggesting that intelligent model-switching is more efficient.

When evaluating LLMs as CDSSs, the combination of Mistral and OpenChat models in enhanced Zero-Shot showed poor results, with 62.68% of the recommendations being different from GS on the $S^{Evaluation}$ and 83.33% of the recommendations being different from GS on $S^{Validation}$. Indeed, when navigating OncoDoc2 decision tree, if the LLM answers one question wrong, this could change the path of the navigation, resulting in the description of a different patient profile leading to different recommendations. Moreover, during manual verification of the results, we found that there were some BCPSs for which the LLM did not provide the correct answer because the elements of response were not present in BCPSs. Therefore, it is important to differentiate errors due to a lack of context and those caused by a poor LLM understanding whereas the context was correctly given, in order to better identify and correct the issues.

Another specific problem may have negatively affected the performance of the model combination. Mistral may not respond if the question is not clearly contextualized, which is a problem since a lack of response leads to stopping the navigation and, consequently, to the absence of recommendations.

The outcome of this study reinforces the conclusion that, even with improved instructions, current LLMs are not yet sufficiently reliable to be used as standalone CDSSs. These results are consistent with the conclusions obtained by the review on the use of LLMs as CDSSs[18]. Evaluations on real cases showed poor performance with ChatGPT, a much larger model than those used in this study, in the same breast cancer context ([23], [24], [25]). The three studies presented lower results (respectively 58.8%, 70%, and 16.05%) than those obtained here, which can be explained by the use of OncoDoc2, whereas the other studies examined LLM recommendations compared to those of experts.

For future work, exploring specialized models such as CancerLLM[26] or larger LLMs like Mixtral 8x7B could improve performance. However, using larger models comes at the cost of environmental impact, as they require more resources. Moreover, simply increasing model size is unlikely to resolve the challenges related to contextual understanding. A more promising direction lies in developing hybrid systems that leverage the complementary strengths of different models. For instance, OpenChat consistency in general question-answering could be combined with Mistral’s ability to handle nuanced cases through a model-switching framework. In addition to hybridization, integrating structured data from EHRs with LLMs’ natural language capabilities could be beneficial and help bridge information gaps. Similarly, incorporating Retrieval-Augmented Generation (RAG) techniques could dynamically enhance LLM outputs by linking them to external knowledge bases, enabling more contextually accurate and reliable recommendations. The findings of Pranab Sahoo et al.[27], which emphasize the divergent behavior of different LLMs to the same prompt, further support the need for model-specific prompt engineering. Optimizing Mistral prompts to prevent navigation failure in complex decision trees could lead to substantial gains in performance and usability.

6 Conclusion

While this study demonstrates the potential of LLMs to augment CDSSs like OncoDoc2, their current performance remains insufficient for routine clinical use as CDSSs. The majority of the recommendations provided by the models diverge significantly from the gold standard. However, continued research into hybrid models, improved prompt engineering techniques, and the integration of structured data offers promising pathways to enhance LLM performance, paving the way for their reliable application in clinical decision support.

Références

1. Global cancer burden growing, amidst mounting need for services. https://www.who.int/news/.
2. Muhiyaddin R, Abd-Alrazaq AA, Househ M, Alam T, Shah Z. The Impact of Clinical Decision Support Systems (CDSS) on Physicians: A Scoping Review. *Stud Health Technol Inform*. 2020;272:470-473. doi:10.3233/SHTI200597
3. Novikava N, Redjdal A, Bouaud J, Seroussi B. Clinical Decision Support Systems Applied to the Management of Breast Cancer Patients: A Scoping Review. *Stud Health Technol Inform*. 2023;305:353-356. doi:10.3233/SHTI230503
4. Séroussi B, Bouaud J, Antoine EC. ONCODOC: a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. *Artif Intell Med*. 2001;22(1):43-64. doi:10.1016/s0933-3657(00)00099-3
5. Séroussi B, Bouaud J, Gligorov J, Uzan S. Supporting multidisciplinary staff meetings for guideline-based breast cancer management: a study with OncoDoc2. *AMIA Annu Symp Proc*. 2007;2007:656-660.
6. Blacker SN, Kang M, Chakraborty I, et al. Utilizing Artificial Intelligence and Chat Generative Pretrained Transformer to Answer Questions About Clinical Scenarios in Neuroanesthesiology. *J Neurosurg Anesthesiol*. Published online December 19, 2023. doi:10.1097/ANA.0000000000000949
7. Luchini C, Pea A, Scarpa A. Artificial intelligence in oncology: current applications and future perspectives. *Br J Cancer*. 2022;126(1):4-9. doi:10.1038/s41416-021-01633-1
8. Mitra A, Del Corro L, Mahajan S, et al. Orca 2: Teaching Small Language Models How to Reason. Published

online November 21, 2023. doi:10.48550/arXiv.2311.11045

9. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. Published online July 19, 2023. doi:10.48550/arXiv.2307.09288

10. Gemma Team, Mesnard T, Hardin C, et al. Gemma: Open Models Based on Gemini Research and Technology. Published online April 16, 2024. doi:10.48550/arXiv.2403.08295

11. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. Published online October 10, 2023. doi:10.48550/arXiv.2310.06825

12. Wang G, Cheng S, Zhan X, Li X, Song S, Liu Y. OpenChat: Advancing Open-source Language Models with Mixed-Quality Data. Published online March 16, 2024. doi:10.48550/arXiv.2309.11235

13. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of Experts. Published online January 8, 2024. doi:10.48550/arXiv.2401.04088

14. Ollama. 2024. https://ollama.com

15. Carbone tracker, mlco2/codecarbon. https://github.com/mlco2/codecarbon, June 2024. Original-date: 2020-05-12T14:44:03Z.

16. Minaee S, Mikolov T, Nikzad N, et al. Large Language Models: A Survey. Published online February 20, 2024. doi:10.48550/arXiv.2402.06196

17. Jin F, Liu Y, Tan Y. Zero-Shot Chain-of-Thought Reasoning Guided by Evolutionary Algorithms in Large Language Models. Published online February 7, 2024. doi:10.48550/arXiv.2402.05376

18. Delourme S, Redjdal A, Bouaud J, Seroussi B. Measured Performance and Healthcare Professional Perception of Large Language Models Used as Clinical Decision Support Systems: A Scoping Review. *Stud Health Technol Inform*. 2024;316:841-845. doi:10.3233/SHTI240543

19. Yu P, Xu H, Hu X, Deng C. Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare (Basel)*. 2023;11(20):2776. doi:10.3390/healthcare11202776

20. Fisch U, Kliem P, Grzonka P, Sutter R. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. *BMJ Health Care Inform*. 2024;31(1):e100978.

doi:10.1136/bmjhci-2023-100978

21. Shiraishi M, Lee H, Kanayama K, Moriwaki Y, Okazaki M. Appropriateness of Artificial Intelligence Chatbots in Diabetic Foot Ulcer Management. *Int J Low Extrem Wounds*. Published online February 28, 2024:15347346241236811. doi:10.1177/15347346241236811

22. Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and Benefits of Large Language Models for the Environment. *Environ Sci Technol*. 2023;57(9):3464-3466. doi:10.1021/acs.est.3c01106

23. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. 2023;9(1):44. doi:10.1038/s41523-023-00557-8

24. Lukac S, Dayan D, Fink V, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch Gynecol Obstet*. 2023;308(6):1831-1844. doi:10.1007/s00404-023-07130-5

25. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in Senology-An Assessment of Concordance with Breast Cancer Tumor Board Decision Making. *J Pers Med*. 2023;13(10):1502. doi:10.3390/jpm13101502

26. Li M, Huang J, Yeung J, et al. CancerLLM: A Large Language Model in Cancer Domain. Published online September 1, 2024. doi:10.48550/arXiv.2406.10459

27. Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. Published online February 5, 2024. doi:10.48550/arXiv.2402.07927

Figure 1: OncoDoc2 decision tree and navigation via the OncoDoc2 user interface.

Figure 2: Building BCPS content from the different reports of a patient EHR.

Figure 3: Workflow illustrating the method implemented.

Figure 4: Ollama operational flowchart.

Figure 5: LLM used as a Q/A system based on OncoDoc2 decision tree.

Figure 6: LLM used as a CDSS based on OncoDoc2 decision tree.

Figure 7: Comparison of the leaves of LLM navigations with MTB clinician navigations, and evaluation of recommendation conformity.

Figure 8: Accuracy of the five models in Zero-Shot PET on $S^{Selection}$, first trial on the left and second trial on the right.

Figure 9: Total CO2 equivalent in grams for the five models on $S^{Selection}$, first trial on the left and second trial on the right.

Figure 10: Distributions of Mistral (top) and Openchat (bottom) models' accuracy to answer OncoDoc2 questions with the enhanced Zero-Shot PET.

Table 1: Results of the different PETs with Mistral, Mixtral 8x7B, and Openchat models on 3,142 prompts

PET	**Model**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Accuracy (%)**
Zero-Shot	*Mistral*	54.27	49.34	47.44	61.78
	Openchat	**72.22**	**70.73**	**71.30**	**69.95**
Enhanced Zero-Shot	*Mistral*	57.26	47.67	49.62	66.10
	Mixtral8x7B	61.69	58.06	59.45	**77.08**
	Openchat	**73.54**	**71.31**	**72.11**	72.47
Zero-Shot CoT	*Mistral*	54.80	44.27	45.70	59.83
	Openchat	55.74	49.13	51.60	**69.96**

Table 2: Accuracy percentages and number of questions for each accuracy category for Openchat and Mistral

models

Model	< 60%	Number	60%- 80%	Number	> 80%	Number	Total
Mistral	36.20	**25**	24.60	17	39.10	**27**	69
Openchat	33.33	23	36.23	**25**	30.43	21	69

Table 3 Results of the different PETs combining Mistral and Openchat models based on their performance on OncoDoc2 questions

PETs	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Enhanced Zero-Shot	59.05	54.40	56.35	77.05
Zero-Shot CoT	61.06	53.48	56.30	75.87

Table 4 Percentage of compliant recommendations produced by LLM navigation (Identical, Comparable, Different) as compared to those produced by MTB clinician navigations

PET	Model	Identical (%)	Comparable (%)	Different (%)
Enhanced Zero-Shot	*Mistral*	11.44	14.42	74.12
	Openchat	9.00	17.91	73.13
	Mistal&Openchat	**17.91**	**19.40**	**62.68**
Zero-Shot Cot	*Mistal&Openchat*	7.46	14.92	79.10

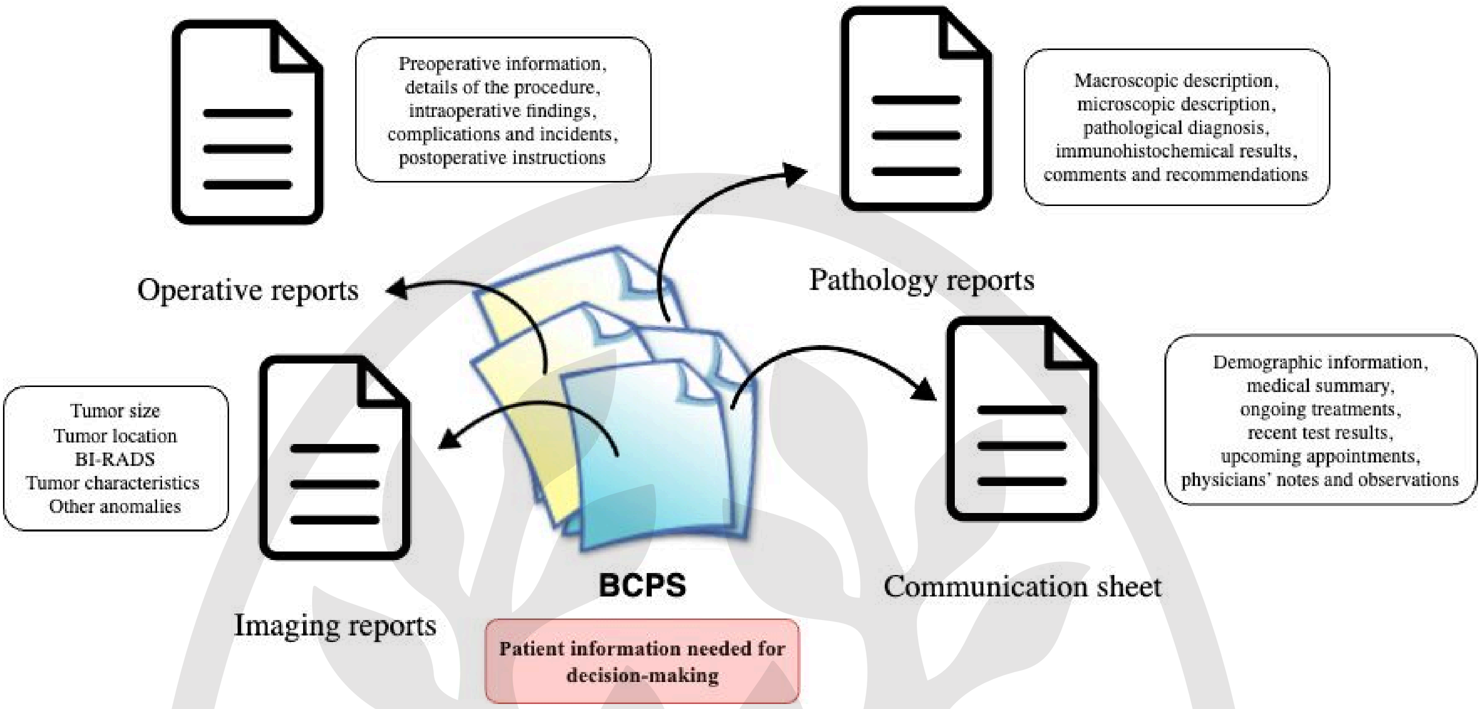
Table 5 Results of enhanced Zero-Shot combining Mistral and Openchat models based on their performance on the OncoDoc2 questions

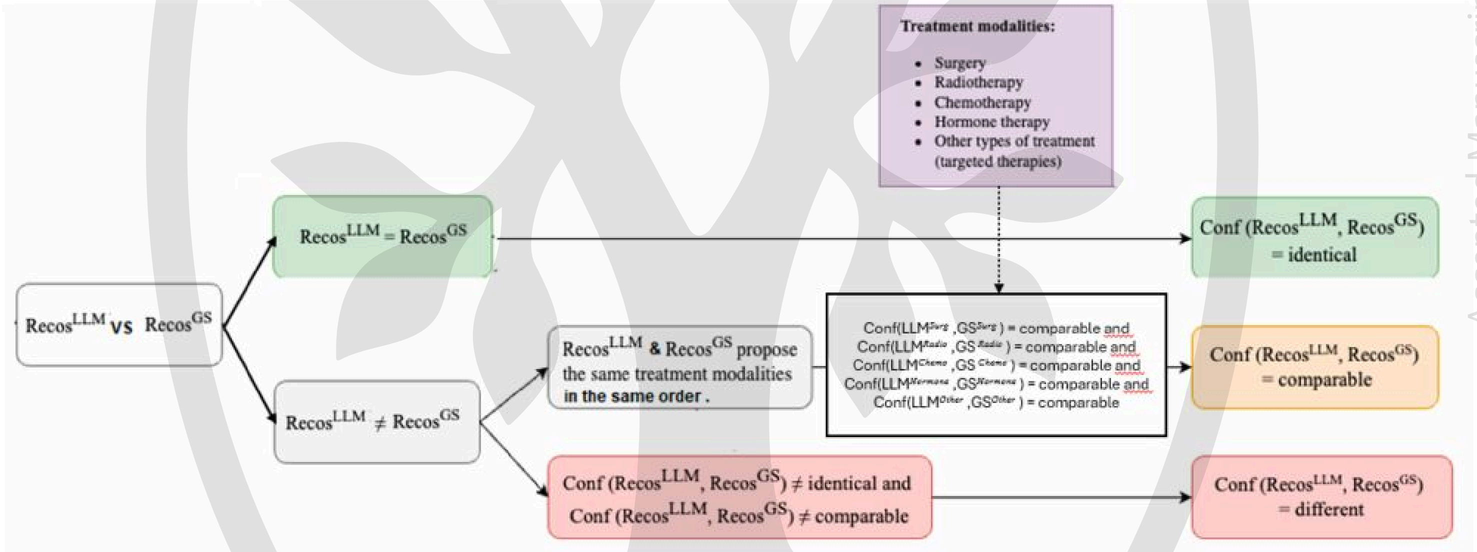
PETs	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)

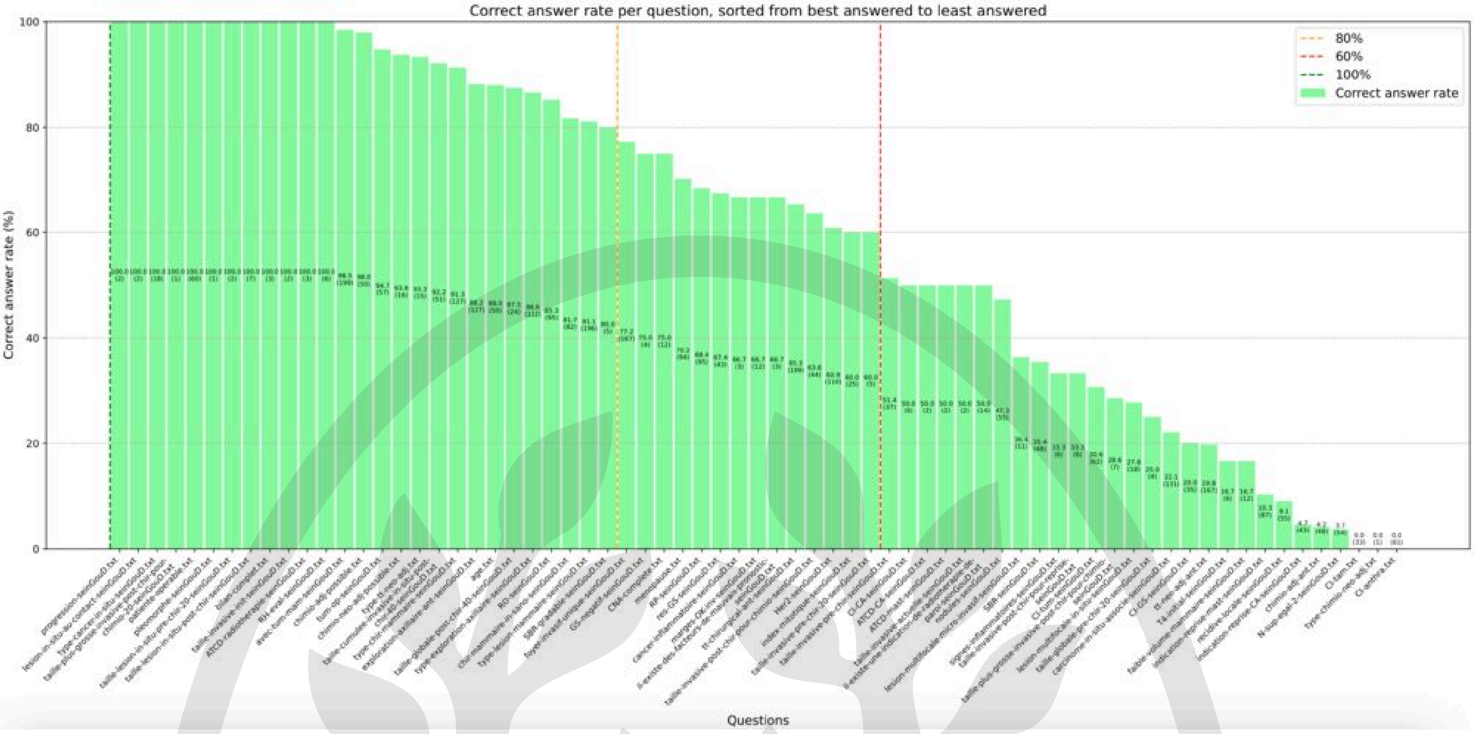
Enhanced Zero-Shot	60.16	54.18	56.59	75.57

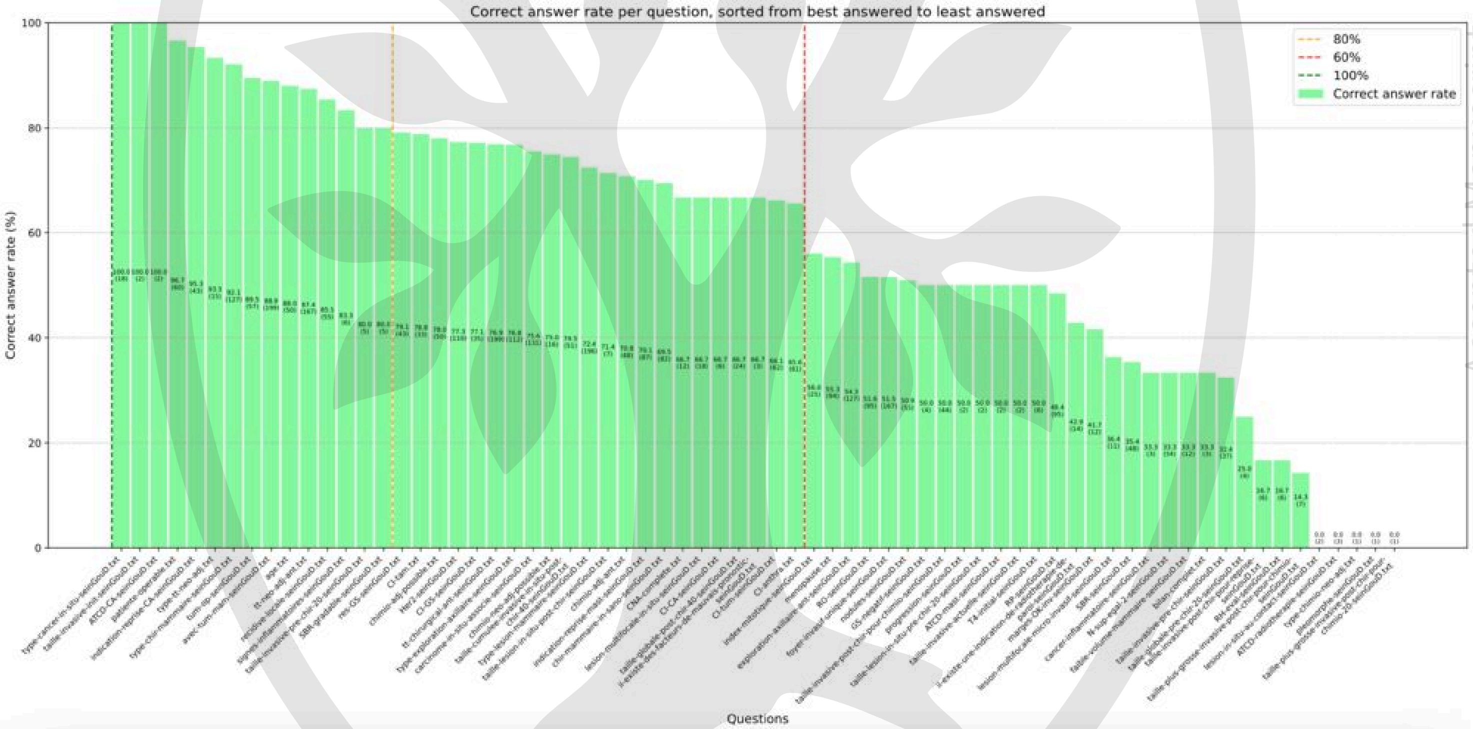
Table 6: Percentage of compliant recommendations produced by LLM navigations (Identical, Comparable, Different) compared to those issued by MTB clinician navigations.

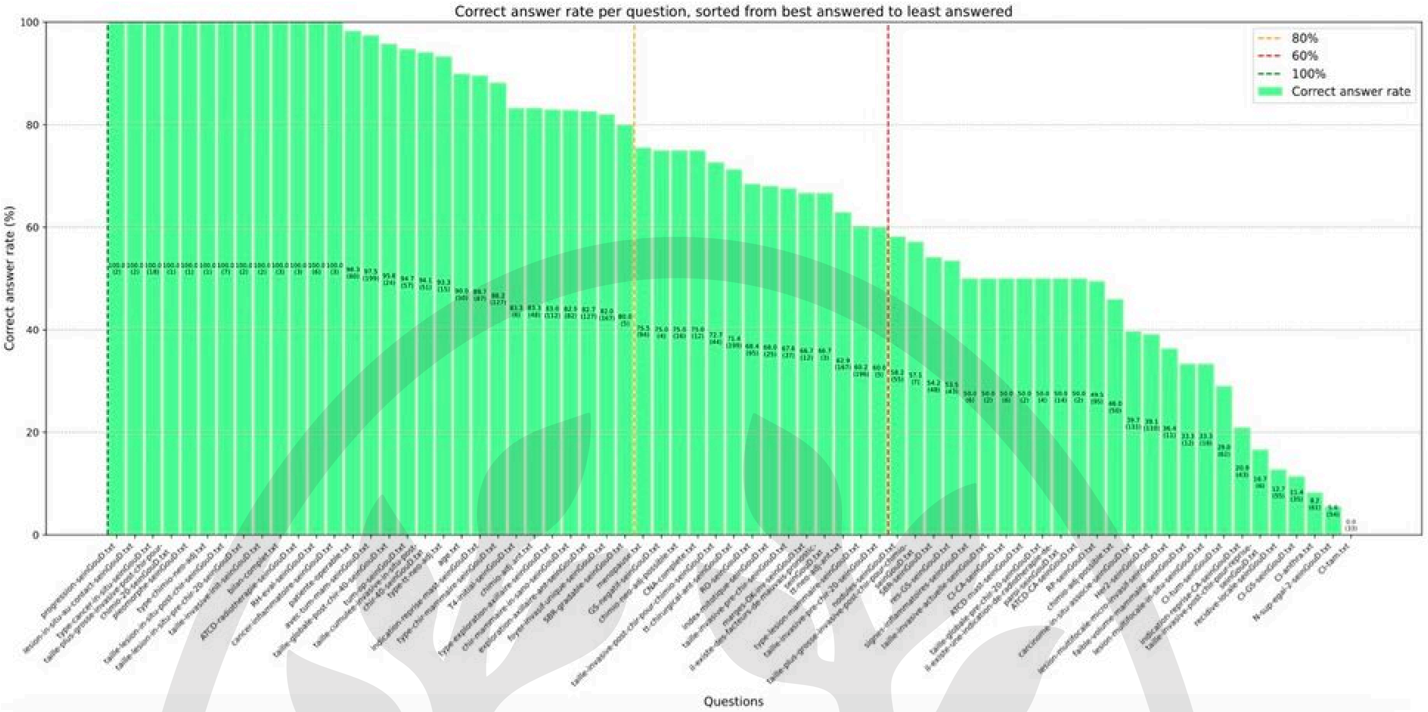
PETs	**Model**	**Identical (%)**	**Comparable (%)**	**Different (%)**
Enhanced Zero-Shot	*Mistral&Openchat*	3.34	13.33	83.33

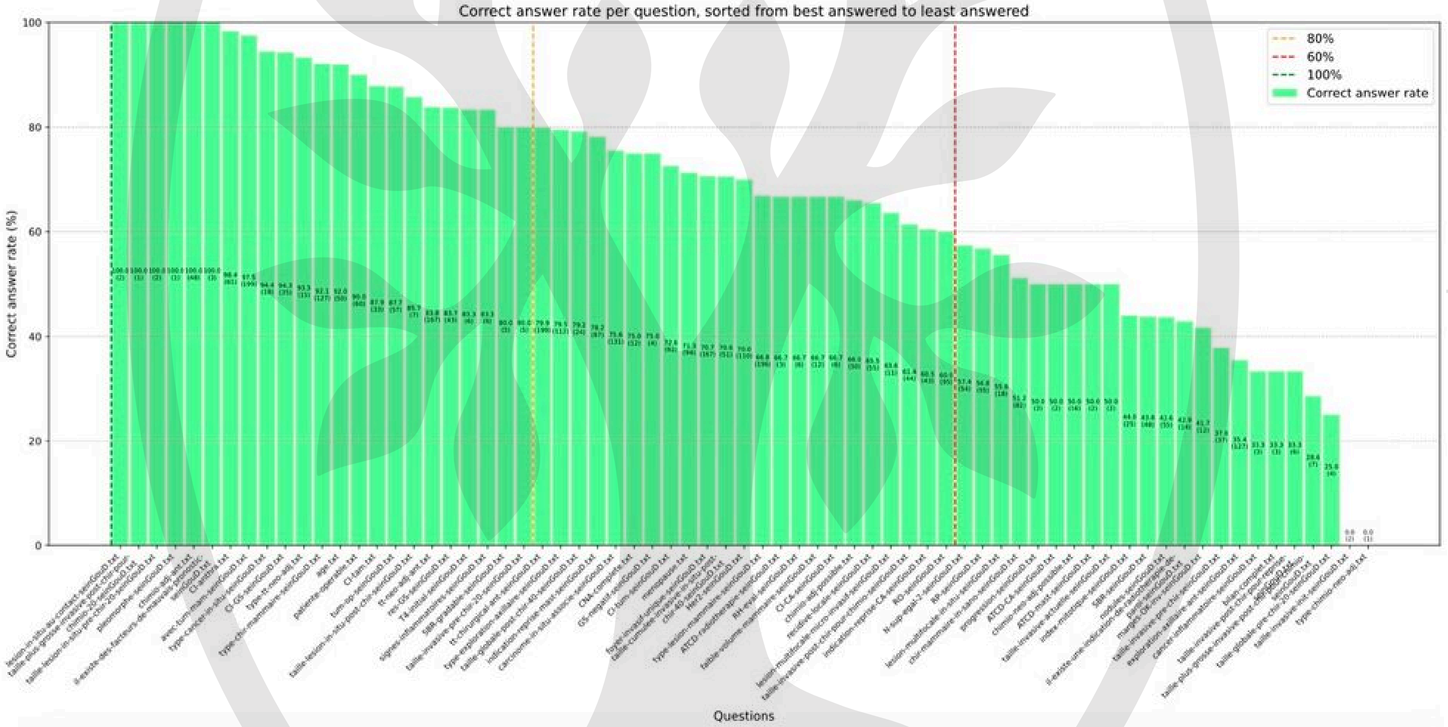

Preoperative information,
details of the procedure,
intraoperative findings,
complications and incidents,
postoperative instructions
Operative reports
Macroscopic description,
microscopic description,
pathological diagnosis,
immunohistochemical results,
comments and recommendations
Pathology reports
Tumor size
Tumor location
BI-RADS
Tumor characteristics
Other anomalies
Imaging reports
BCPS
Patient information needed for decision-making
Demographic information,
medical summary,
ongoing treatments,
recent test results,
upcoming appointments,
physicians' notes and observations
Communication sheet

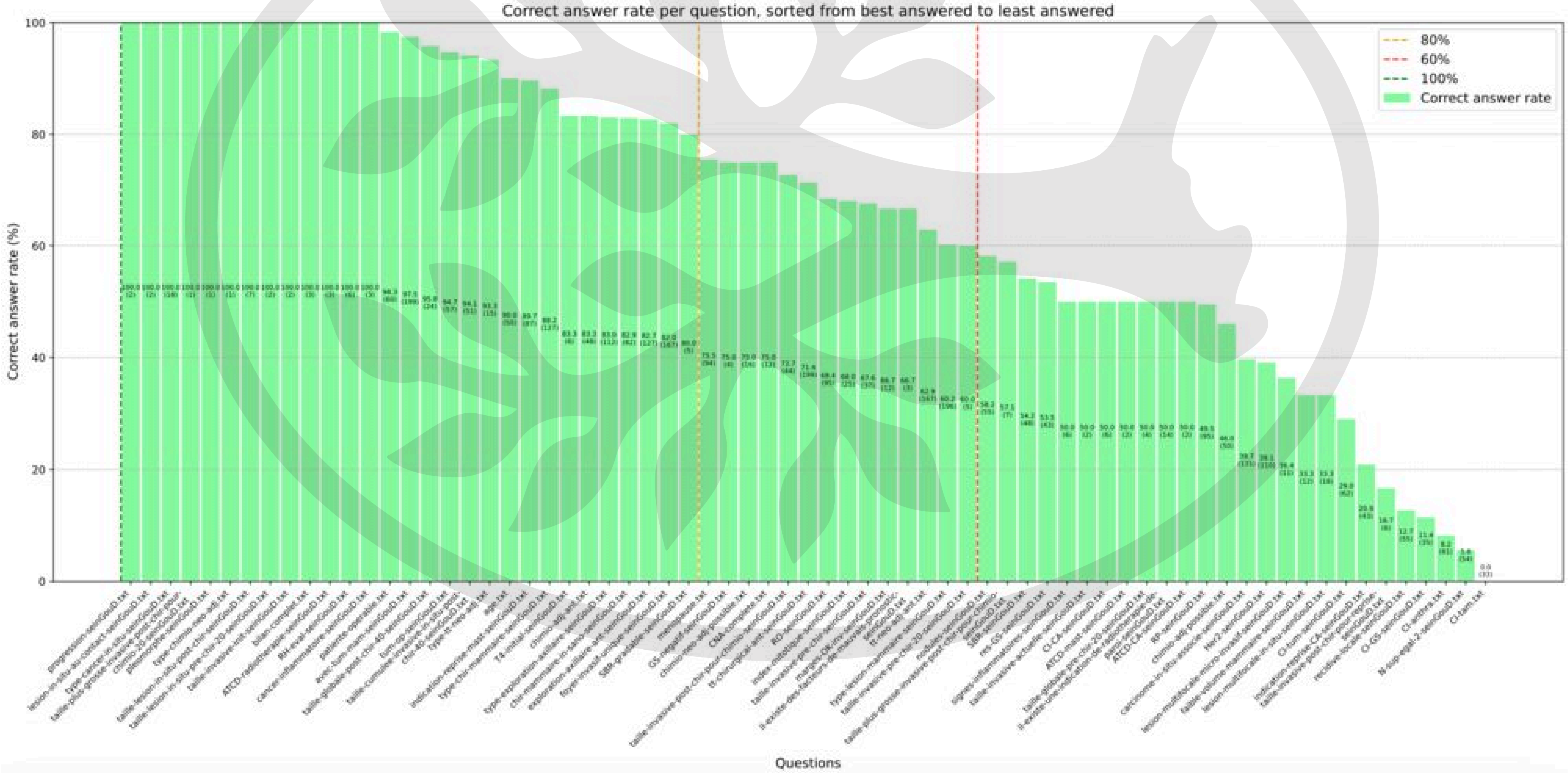

Treatment modalities:
Surgery
Radiotherapy
Chemotherapy
Hormone therapy
Other types of treatment (targeted therapies)
Recos^{LLM} vs Recos^{GS}
$\text{Recos}^{LLM} = \text{Recos}^{GS}$
$\text{Recos}^{LLM} \neq \text{Recos}^{GS}$
Recos^{LLM} & Recos^{GS} propose the same treatment modalities in the same order .
$\text{Conf}(\text{LLM}^{Surg}, \text{GS}^{Surg}) = \text{comparable and}$
$\text{Conf}(\text{LLM}^{Radio}, \text{GS}^{Radio}) = \text{comparable and}$
$\text{Conf}(\text{LLM}^{Chemo}, \text{GS}^{Chemo}) = \text{comparable and}$
$\text{Conf}(\text{LLM}^{Hormone}, \text{GS}^{Hormone}) = \text{comparable and}$
$\text{Conf}(\text{LLM}^{Other}, \text{GS}^{Other}) = \text{comparable}$
$\text{Conf}(\text{Recos}^{LLM}, \text{Recos}^{GS}) \neq$ identical and
$\text{Conf}(\text{Recos}^{LLM}, \text{Recos}^{GS}) \neq$ comparable
$\text{Conf}(\text{Recos}^{LLM}, \text{Recos}^{GS})$ = identical
$\text{Conf}(\text{Recos}^{LLM}, \text{Recos}^{GS})$ = comparable
$\text{Conf}(\text{Recos}^{LLM}, \text{Recos}^{GS})$ = different

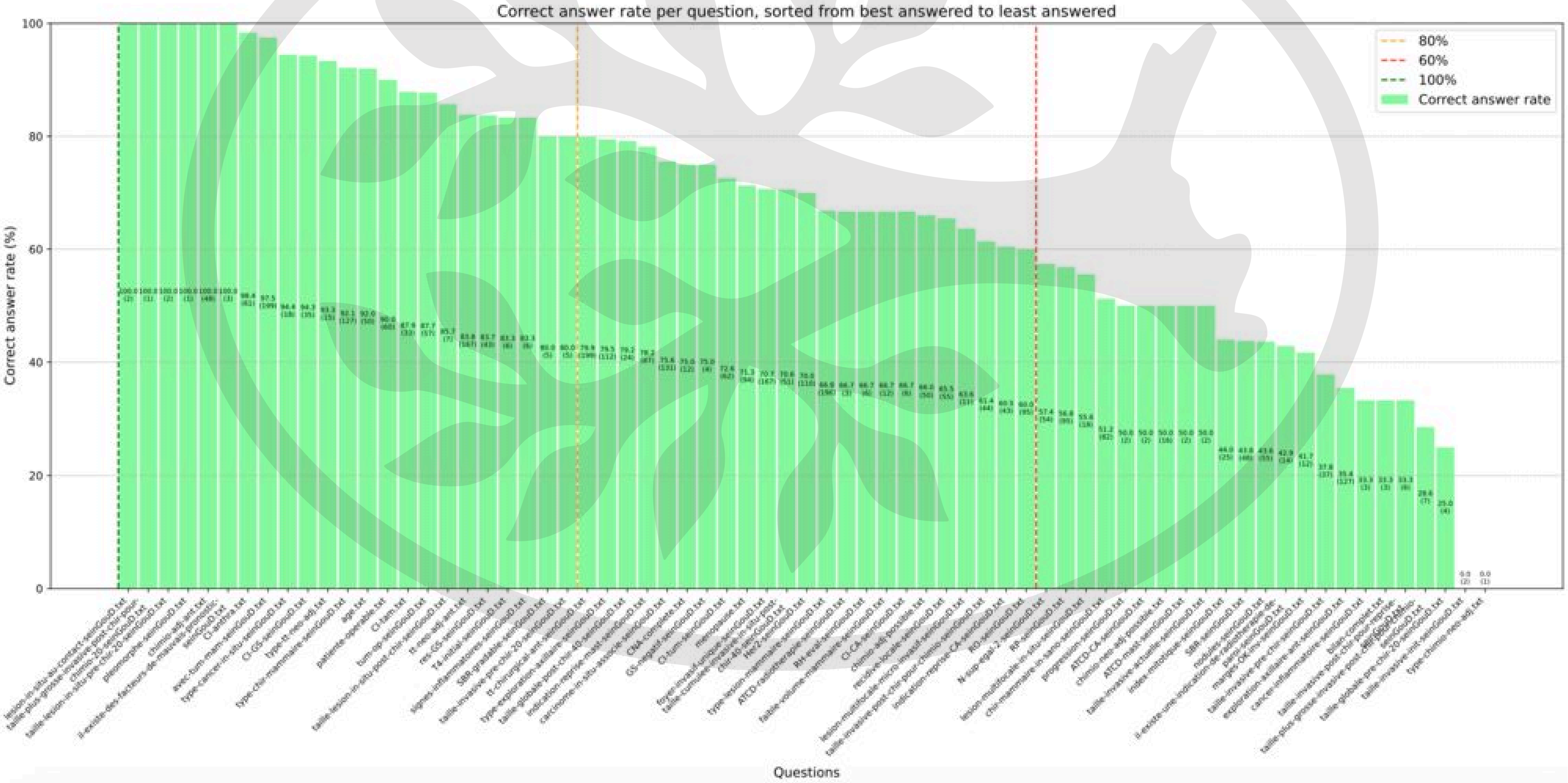

Correct answer rate per question, sorted from best answered to least answered
Correct answer rate (%)
0
20
40
60
80
100
80%
60%
100%
Correct answer rate
100.0 (2)
100.0 (2)
100.0 (18)
100.0 (1)
100.0 (60)
100.0 (1)
100.0 (2)
100.0 (7)
100.0 (3)
100.0 (2)
100.0 (3)
100.0 (6)
98.5 (199)
98.0 (50)
94.7 (57)
93.8 (16)
93.3 (15)
92.2 (51)
91.3 (127)
88.2 (127)
88.0 (50)
87.5 (24)
86.6 (112)
85.3 (95)
81.7 (82)
81.1 (196)
80.0 (5)
77.2 (167)
75.0 (4)
75.0 (12)
70.2 (84)
68.4 (95)
67.4 (43)
66.7 (3)
66.7 (12)
66.7 (3)
65.3 (199)
63.6 (44)
60.9 (110)
60.0 (25)
60.0 (5)
51.4 (37)
50.0 (6)
50.0 (2)
50.0 (2)
50.0 (2)
50.0 (14)
47.3 (55)
36.4 (11)
35.4 (48)
33.3 (6)
33.3 (6)
30.6 (62)
28.6 (7)
27.8 (18)
25.0 (4)
22.1 (131)
20.0 (35)
19.8 (167)
16.7 (6)
16.7 (12)
10.3 (87)
9.1 (55)
4.7 (43)
4.2 (48)
3.7 (54)
0.0 (33)
0.0 (1)
0.0 (61)
progression-seinGouD.txt
lesion-in-situ-au-contact-seinGouD.txt
type-cancer-in-situ-seinGouD.txt
taille-plus-grosse-invasive-post-chir-pour-chimio-20-seinGouD.txt
patiente-operable.txt
pleomorphe-seinGouD.txt
taille-lesion-in-situ-pre-chir-20-seinGouD.txt
taille-lesion-in-situ-post-chir-seinGouD.txt
bilan-complet.txt
taille-invasive-init-seinGouD.txt
ATCD-radiotherapie-seinGouD.txt
RH-eval-seinGouD.txt
avec-tum-mam-seinGouD.txt
chimio-adj-possible.txt
tum-op-seinGouD.txt
chimio-neo-adj-possible.txt
taille-cumulee-invasive-in-situ-post-chir-4D-seinGouD.txt
type-tt-neo-adj.txt
type-chir-mammaire-seinGouD.txt
exploration-axillaire-ant-seinGouD.txt
taille-globale-post-chir-4D-seinGouD.txt
age.txt
type-exploration-axillaire-seinGouD.txt
RO-seinGouD.txt
chir-mammaire-in-sano-seinGouD.txt
type-lesion-mammaire-seinGouD.txt
SBR-gradable-seinGouD.txt
foyer-invasif-unique-seinGouD.txt
GS-negatif-seinGouD.txt
CNA-complete.txt
menopause.txt
RP-seinGouD.txt
res-GS-seinGouD.txt
cancer-inflammatoire-seinGouD.txt
marges-OK-inv-seinGouD.txt
il-existe-des-facteurs-de-mauvais-pronostic-seinGouD.txt
tt-chirurgical-ant-seinGouD.txt
taille-invasive-post-chir-pour-chimio-seinGouD.txt
Her2-seinGouD.txt
index-mitotique-seinGouD.txt
taille-invasive-pre-chir-20-seinGouD.txt
taille-invasive-pre-chir-seinGouD.txt
CI-CA-seinGouD.txt
ATCD-CA-seinGouD.txt
ATCD-mast-seinGouD.txt
taille-invasive-actuelle-seinGouD.txt
il-existe-une-indication-de-radiotherapie-de-paroi-seinGouD.txt
nodules-seinGouD.txt
lesion-multifocale-micro-invasif-seinGouD.txt
SBR-seinGouD.txt
signes-inflammatoires-seinGouD.txt
taille-invasive-post-chir-pour-reprise-seinGouD.txt
CI-tum-seinGouD.txt
taille-plus-grosse-invasive-post-chir-pour-chimio-seinGouD.txt
lesion-multifocale-in-situ-seinGouD.txt
taille-globale-pre-chir-20-seinGouD.txt
carcinome-in-situ-associe-seinGouD.txt
CI-GS-seinGouD.txt
tt-neo-adj-ant.txt
T4-initial-seinGouD.txt
faible-volume-mammaire-seinGouD.txt
indication-reprise-mast-seinGouD.txt
recidive-locale-seinGouD.txt
indication-reprise-CA-seinGouD.txt
chimio-adj-ant.txt
N-sup-egal-2-seinGouD.txt
CI-tam.txt
type-chimio-neo-adj.txt
CI-anthra.txt
Questions

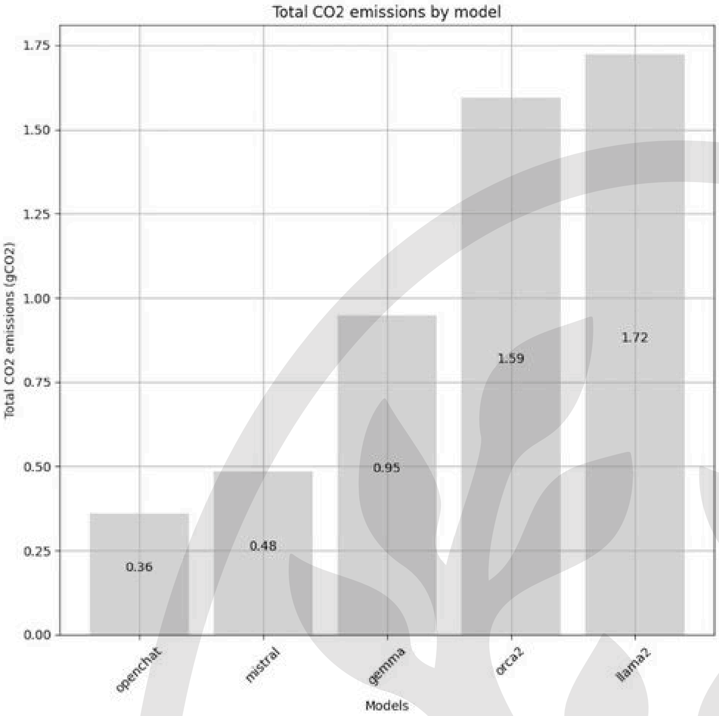

Correct answer rate per question, sorted from best answered to least answered
Correct answer rate (%)
0
20
40
60
80
100
80%
60%
100%
Correct answer rate
100.0 (18)
100.0 (2)
100.0 (2)
96.7 (60)
95.3 (43)
93.3 (15)
92.1 (127)
89.5 (57)
88.9 (199)
88.0 (50)
87.4 (167)
85.5 (55)
83.3 (6)
80.0 (5)
80.0 (5)
79.1 (43)
78.8 (33)
78.0 (50)
77.3 (110)
77.1 (35)
76.9 (199)
76.8 (112)
75.6 (131)
75.0 (16)
74.5 (51)
72.4 (196)
71.4 (7)
70.8 (48)
70.1 (87)
69.5 (82)
66.7 (12)
66.7 (18)
66.7 (6)
66.7 (24)
66.7 (3)
66.1 (62)
65.6 (61)
56.0 (25)
55.3 (94)
54.3 (127)
51.6 (95)
51.5 (167)
50.9 (51)
50.0 (4)
50.0 (44)
50.0 (2)
50.0 (2)
50.0 (2)
50.0 (2)
50.0 (6)
48.4 (95)
42.9 (14)
41.7 (12)
36.4 (11)
35.4 (48)
33.3 (3)
33.3 (54)
33.3 (12)
33.3 (3)
32.4 (37)
25.0 (4)
16.7 (6)
16.7 (6)
14.3 (7)
0.0 (2)
0.0 (3)
0.0 (1)
0.0 (1)
0.0 (1)
type-cancer-in-situ-seinGouD.txt
taille-invasive-init-seinGouD.txt
ATCD-CA-seinGouD.txt
patiente-operable.txt
indication-reprise-CA-seinGouD.txt
type-tt-neo-adj.txt
type-chir-mammaire-seinGouD.txt
tum-op-seinGouD.txt
avec-tum-mam-seinGouD.txt
age.txt
tt-neo-adj-ant.txt
recidive-locale-seinGouD.txt
signes-inflammatoires-seinGouD.txt
taille-invasive-pre-chir-20-seinGouD.txt
SBR-gradable-seinGouD.txt
res-GS-seinGouD.txt
CI-tam.txt
chimio-adj-possible.txt
Her2-seinGouD.txt
CI-GS-seinGouD.txt
tt-chirurgical-ant-seinGouD.txt
type-exploration-axillaire-seinGouD.txt
carcinome-in-situ-associe-seinGouD.txt
chimio-neo-adj-possible.txt
taille-cumulee-invasive-in-situ-post-chir-4D-seinGouD.txt
type-lesion-mammaire-seinGouD.txt
taille-lesion-in-situ-post-chir-seinGouD.txt
chimio-adj-ant.txt
indication-reprise-mast-seinGouD.txt
chir-mammaire-in-sano-seinGouD.txt
CNA-complete.txt
lesion-multifocale-in-situ-seinGouD.txt
CI-CA-seinGouD.txt
taille-globale-post-chir-4D-seinGouD.txt
il-existe-des-facteurs-de-mauvais-pronostic-seinGouD.txt
CI-tum-seinGouD.txt
CI-anthra.txt
index-mitotique-seinGouD.txt
menopause.txt
exploration-axillaire-ant-seinGouD.txt
RO-seinGouD.txt
foyer-invasif-unique-seinGouD.txt
nodules-seinGouD.txt
GS-negatif-seinGouD.txt
taille-invasive-post-chir-pour-chimio-seinGouD.txt
progression-seinGouD.txt
taille-lesion-in-situ-pre-chir-20-seinGouD.txt
ATCD-mast-seinGouD.txt
taille-invasive-actuelle-seinGouD.txt
T4-initial-seinGouD.txt
RP-seinGouD.txt
il-existe-une-indication-de-radiotherapie-de-paroi-seinGouD.txt
marges-OK-inv-seinGouD.txt
lesion-multifocale-micro-invasif-seinGouD.txt
SBR-seinGouD.txt
cancer-inflammatoire-seinGouD.txt
N-sup-egal-2-seinGouD.txt
faible-volume-mammaire-seinGouD.txt
bilan-complet.txt
taille-invasive-pre-chir-seinGouD.txt
taille-globale-pre-chir-20-seinGouD.txt
taille-invasive-post-chir-pour-reprise-seinGouD.txt
RH-eval-seinGouD.txt
taille-plus-grosse-invasive-post-chir-pour-chimio-seinGouD.txt
lesion-in-situ-au-contact-seinGouD.txt
ATCD-radiotherapie-seinGouD.txt
type-chimio-neo-adj.txt
pleomorphe-seinGouD.txt
taille-plus-grosse-invasive-post-chir-pour-chimio-20-seinGouD.txt
Questions

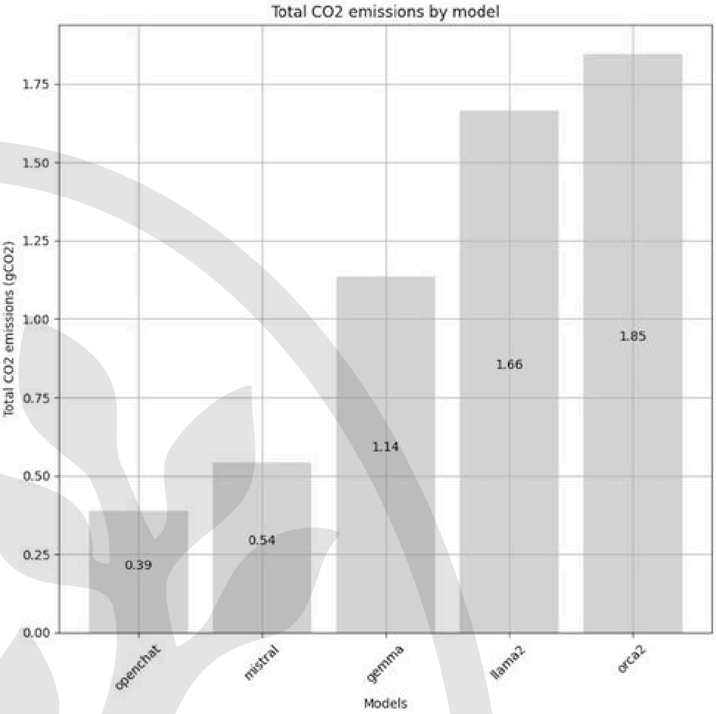

Correct answer rate per question, sorted from best answered to least answered
Correct answer rate (%)
Questions
0
20
40
60
80
100
80%
60%
100%
Correct answer rate

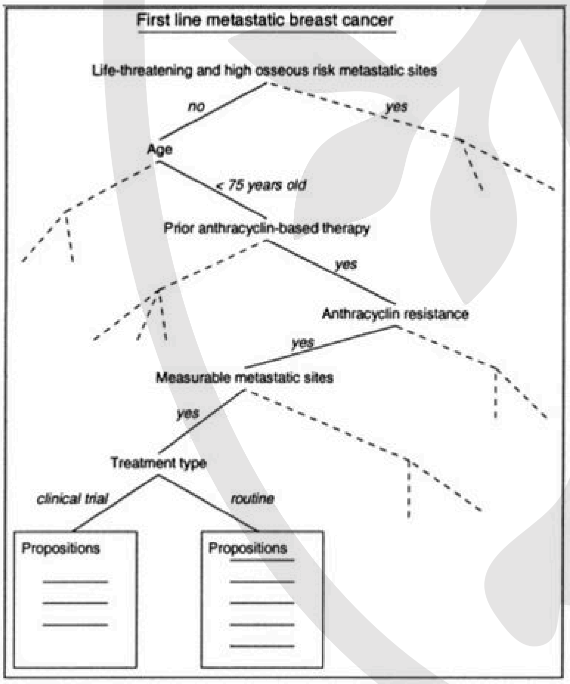

Correct answer rate per question, sorted from best answered to least answered
Correct answer rate (%)
Questions
0
20
40
60
80
100
80%
60%
100%
Correct answer rate

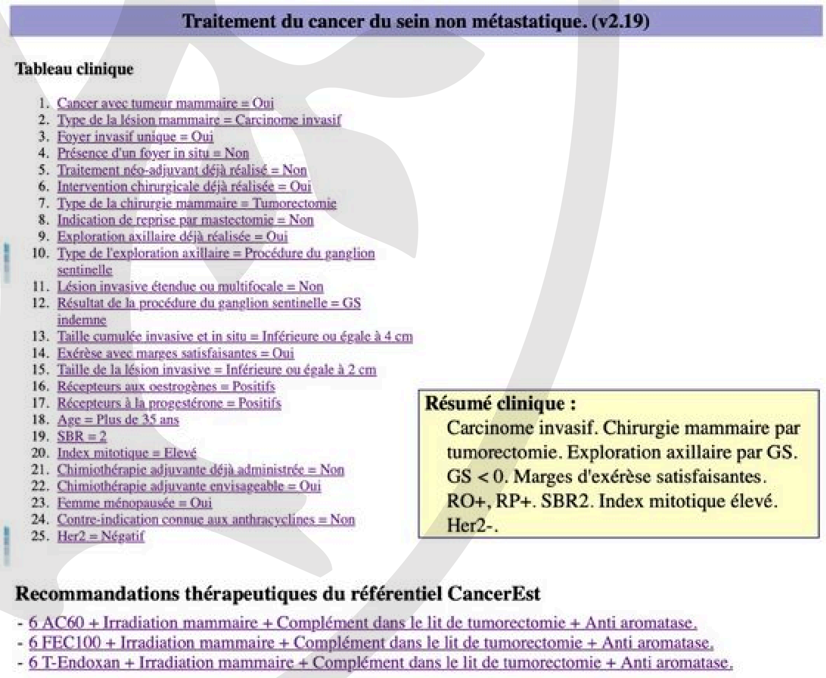

Correct answer rate per question, sorted from best answered to least answered
Correct answer rate (%)
0
20
40
60
80
100
Questions
80%
60%
100%
Correct answer rate
100.0 (2)
100.0 (2)
100.0 (18)
100.0 (1)
100.0 (1)
100.0 (1)
100.0 (7)
100.0 (2)
100.0 (2)
100.0 (3)
100.0 (3)
100.0 (6)
100.0 (3)
98.3 (60)
97.5 (199)
95.8 (24)
94.7 (57)
94.1 (51)
93.3 (15)
90.0 (50)
89.7 (87)
88.2 (127)
83.3 (6)
83.3 (48)
83.0 (112)
82.9 (82)
82.7 (127)
82.0 (167)
80.0 (5)
75.5 (94)
75.0 (4)
75.0 (16)
75.0 (12)
72.7 (44)
71.4 (199)
68.4 (95)
68.0 (25)
67.6 (37)
66.7 (12)
66.7 (3)
62.9 (167)
60.2 (196)
60.0 (5)
58.2 (55)
57.1 (7)
54.2 (48)
53.5 (43)
50.0 (6)
50.0 (2)
50.0 (6)
50.0 (2)
50.0 (4)
50.0 (14)
50.0 (2)
49.5 (95)
46.0 (50)
39.7 (131)
39.1 (110)
36.4 (11)
33.3 (12)
33.3 (18)
29.0 (62)
20.9 (43)
16.7 (6)
12.7 (55)
11.4 (35)
8.2 (61)
5.6 (54)
0.0 (33)
progression-seinGouD.txt
lesion-in-situ-au-contact-seinGouD.txt
type-cancer-in-situ-seinGouD.txt
taille-plus-grosse-invasive-post-chir-pour-
chimio-20-seinGouD.txt
pleomorphe-seinGouD.txt
type-chimio-neo-adj.txt
taille-lesion-in-situ-post-chir-seinGouD.txt
taille-lesion-in-situ-pre-chir-20-seinGouD.txt
taille-invasive-init-seinGouD.txt
bilan-complet.txt
ATCD-radiotherapie-seinGouD.txt
RH-eval-seinGouD.txt
cancer-inflammatoire-seinGouD.txt
patiente-operable.txt
avec-tum-mam-seinGouD.txt
taille-globale-post-chir-40-seinGouD.txt
tum-op-seinGouD.txt
taille-cumulee-invasive-in-situ-post-
chir-40-seinGouD.txt
type-tt-neo-adj.txt
age.txt
indication-reprise-mast-seinGouD.txt
type-chir-mammaire-seinGouD.txt
T4-initial-seinGouD.txt
chimio-adj-ant.txt
type-exploration-axillaire-seinGouD.txt
chir-mammaire-in-sano-seinGouD.txt
exploration-axillaire-ant-seinGouD.txt
foyer-invasif-unique-seinGouD.txt
SBR-gradable-seinGouD.txt
menopause.txt
GS-negatif-seinGouD.txt
taille-invasive-post-chir-pour-chimio-seinGouD.txt
chimio-neo-adj-possible.txt
CNA-complete.txt
tt-chirurgical-ant-seinGouD.txt
RO-seinGouD.txt
index-mitotique-seinGouD.txt
taille-invasive-pre-chir-seinGouD.txt
marges-OK-inv-seinGouD.txt
il-existe-des-facteurs-de-mauvais-pronostic-
seinGouD.txt
tt-neo-adj-ant.txt
type-lesion-mammaire-seinGouD.txt
taille-invasive-pre-chir-20-seinGouD.txt
nodules-seinGouD.txt
taille-plus-grosse-invasive-post-chir-pour-chimio-
seinGouD.txt
SBR-seinGouD.txt
res-GS-seinGouD.txt
signes-inflammatoires-seinGouD.txt
taille-invasive-actuelle-seinGouD.txt
CI-CA-seinGouD.txt
ATCD-mast-seinGouD.txt
taille-globale-pre-chir-20-seinGouD.txt
il-existe-une-indication-de-radiotherapie-de-
paroi-seinGouD.txt
ATCD-CA-seinGouD.txt
RP-seinGouD.txt
chimio-adj-possible.txt
carcinome-in-situ-associe-seinGouD.txt
Her2-seinGouD.txt
lesion-multifocale-micro-invasif-seinGouD.txt
taille-volume-mammaire-seinGouD.txt
lesion-multifocale-in-situ-seinGouD.txt
CI-tum-seinGouD.txt
indication-reprise-CA-seinGouD.txt
taille-invasive-post-chir-pour-reprise-
seinGouD.txt
recidive-locale-seinGouD.txt
CI-GS-seinGouD.txt
CI-anthra.txt
N-sup-egal-2-seinGouD.txt
CI-tam.txt


Correct answer rate per question, sorted from best answered to least answered
Correct answer rate (%)
0
20
40
60
80
100
Questions
80%
60%
100%
Correct answer rate
100.0 (2)
100.0 (1)
100.0 (2)
100.0 (1)
100.0 (48)
100.0 (3)
98.4 (61)
97.5 (199)
94.4 (18)
94.3 (35)
93.3 (15)
92.1 (127)
92.0 (50)
90.0 (60)
87.9 (33)
87.7 (57)
85.7 (7)
83.8 (167)
83.7 (43)
83.3 (6)
83.3 (6)
80.0 (5)
80.0 (5)
79.9 (199)
79.5 (112)
79.2 (24)
78.2 (87)
75.6 (131)
75.0 (12)
75.0 (4)
72.6 (62)
71.3 (94)
70.7 (167)
70.6 (51)
70.0 (110)
66.9 (196)
66.7 (3)
66.7 (6)
66.7 (12)
66.7 (9)
66.0 (50)
65.5 (55)
63.6 (11)
61.4 (44)
60.5 (43)
60.0 (95)
57.4 (54)
56.8 (95)
55.6 (18)
51.2 (82)
50.0 (2)
50.0 (2)
50.0 (16)
50.0 (2)
50.0 (2)
44.0 (25)
43.8 (48)
43.6 (55)
42.9 (14)
41.7 (12)
37.8 (37)
35.4 (127)
33.3 (3)
33.3 (3)
33.3 (6)
28.6 (7)
25.0 (4)
0.0 (2)
0.0 (1)
lesion-in-situ-au-contact-seinGouD.txt
taille-plus-grosse-invasive-post-chir-pour-
chimio-20-seinGouD.txt
taille-lesion-in-situ-pre-chir-20-seinGouD.txt
pleomorphe-seinGouD.txt
chimio-adj-ant.txt
il-existe-des-facteurs-de-mauvais-pronostic-
seinGouD.txt
CI-anthra.txt
avec-tum-mam-seinGouD.txt
type-cancer-in-situ-seinGouD.txt
CI-GS-seinGouD.txt
type-tt-neo-adj.txt
type-chir-mammaire-seinGouD.txt
age.txt
patiente-operable.txt
CI-tam.txt
tum-op-seinGouD.txt
taille-lesion-in-situ-post-chir-seinGouD.txt
tt-neo-adj-ant.txt
res-GS-seinGouD.txt
T4-initial-seinGouD.txt
signes-inflammatoires-seinGouD.txt
SBR-gradable-seinGouD.txt
taille-invasive-pre-chir-20-seinGouD.txt
tt-chirurgical-ant-seinGouD.txt
type-exploration-axillaire-seinGouD.txt
taille-globale-post-chir-40-seinGouD.txt
indication-reprise-mast-seinGouD.txt
carcinome-in-situ-associe-seinGouD.txt
CNA-complete.txt
GS-negatif-seinGouD.txt
CI-tum-seinGouD.txt
menopause.txt
foyer-invasif-unique-seinGouD.txt
taille-cumulee-invasive-in-situ-post-
chir-40-seinGouD.txt
Her2-seinGouD.txt
type-lesion-mammaire-seinGouD.txt
ATCD-radiotherapie-seinGouD.txt
RH-eval-seinGouD.txt
taille-volume-mammaire-seinGouD.txt
CI-CA-seinGouD.txt
chimio-adj-possible.txt
recidive-locale-seinGouD.txt
lesion-multifocale-micro-invasif-seinGouD.txt
taille-invasive-post-chir-pour-chimio-seinGouD.txt
indication-reprise-CA-seinGouD.txt
RO-seinGouD.txt
N-sup-egal-1-seinGouD.txt
RP-seinGouD.txt
lesion-multifocale-in-situ-seinGouD.txt
chir-mammaire-in-sano-seinGouD.txt
progression-seinGouD.txt
ATCD-CA-seinGouD.txt
chimio-neo-adj-possible.txt
ATCD-mast-seinGouD.txt
taille-invasive-actuelle-seinGouD.txt
index-mitotique-seinGouD.txt
SBR-seinGouD.txt
il-existe-une-indication-de-radiotherapie-de-
nodules-seinGouD.txt
paroi-seinGouD.txt
marges-OK-inv-seinGouD.txt
taille-invasive-pre-chir-seinGouD.txt
exploration-axillaire-ant-seinGouD.txt
cancer-inflammatoire-seinGouD.txt
bilan-complet.txt
taille-invasive-post-chir-pour-reprise-
seinGouD.txt
taille-plus-grosse-invasive-post-chir-pour-chimio-seinGouD.txt
taille-globale-pre-chir-20-seinGouD.txt
taille-invasive-init-seinGouD.txt
type-chimio-neo-adj.txt






First line metastatic breast cancer
Life-threatening and high osseous risk metastatic sites
no
yes
Age
< 75 years old
Prior anthracyclin-based therapy
yes
Anthracyclin resistance
yes
Measurable metastatic sites
yes
Treatment type
clinical trial
routine
Propositions
Propositions



Traitement du cancer du sein non métastatique. (v2.19)

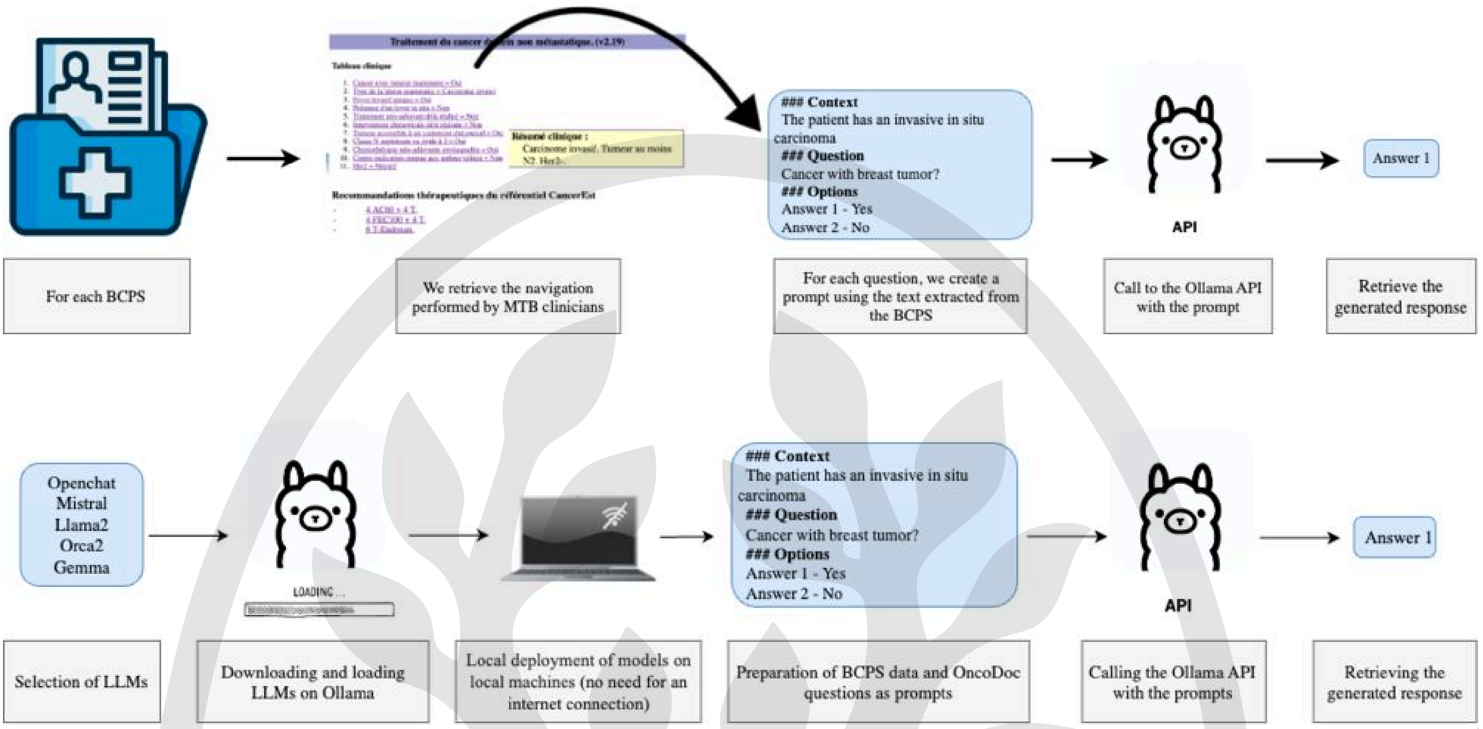
Tableau clinique

1. Cancer avec tumeur mammaire = Oui
2. Type de la lésion mammaire = Carcinome invasif
3. Foyer invasif unique = Oui
4. Présence d'un foyer in situ = Non
5. Traitement néo-adjuvant déjà réalisé = Non
6. Intervention chirurgicale déjà réalisée = Oui
7. Type de la chirurgie mammaire = Tumorectomie
8. Indication de reprise par mastectomie = Non
9. Exploration axillaire déjà réalisée = Oui
10. Type de l'exploration axillaire = Procédure du ganglion sentinelle
11. Lésion invasive étendue ou multifocale = Non
12. Résultat de la procédure du ganglion sentinelle = GS indemne
13. Taille cumulée invasive et in situ = Inférieure ou égale à 4 cm
14. Exérèse avec marges satisfaisantes = Oui
15. Taille de la lésion invasive = Inférieure ou égale à 2 cm
16. Récepteurs aux oestrogènes = Positifs
17. Récepteurs à la progestérone = Positifs
18. Age = Plus de 35 ans
19. SBR = 2
20. Index mitotique = Elevé
21. Chimiothérapie adjuvante déjà administrée = Non
22. Chimiothérapie adjuvante envisageable = Oui
23. Femme ménopausée = Oui
24. Contre-indication connue aux anthracyclines = Non
25. Her2 = Négatif

Résumé clinique :
Carcinome invasif. Chirurgie mammaire par tumorectomie. Exploration axillaire par GS. GS < 0. Marges d'exérèse satisfaisantes. RO+, RP+. SBR2. Index mitotique élevé. Her2-.

Recommandations thérapeutiques du référentiel CancerEst

- 6 AC60 + Irradiation mammaire + Complément dans le lit de tumorectomie + Anti aromatase.
- 6 FEC100 + Irradiation mammaire + Complément dans le lit de tumorectomie + Anti aromatase.
- 6 T-Endoxan + Irradiation mammaire + Complément dans le lit de tumorectomie + Anti aromatase.



Traitement du cancer du sein non métastatique. (v2.19)
Tableau clinique
Résumé clinique :
Carcinome invasif. Tumeur au moins N2. Her2-.
Recommandations thérapeutiques du référentiel CancerEst
4 AC60 + 4 T.
4 FEC100 + 4 T.
6 T-Endoxan.
API
API
LOADING...

For each BCPS

We retrieve the navigation performed by MTB clinicians

Context
The patient has an invasive in situ carcinoma
Question
Cancer with breast tumor?
Options
Answer 1 - Yes
Answer 2 - No

For each question, we create a prompt using the text extracted from the BCPS

Call to the Ollama API with the prompt

Answer 1

Retrieve the generated response

Openchat
Mistral
Llama2
Orca2
Gemma

Selection of LLMs

Downloading and loading LLMs on Ollama

Local deployment of models on local machines (no need for an internet connection)

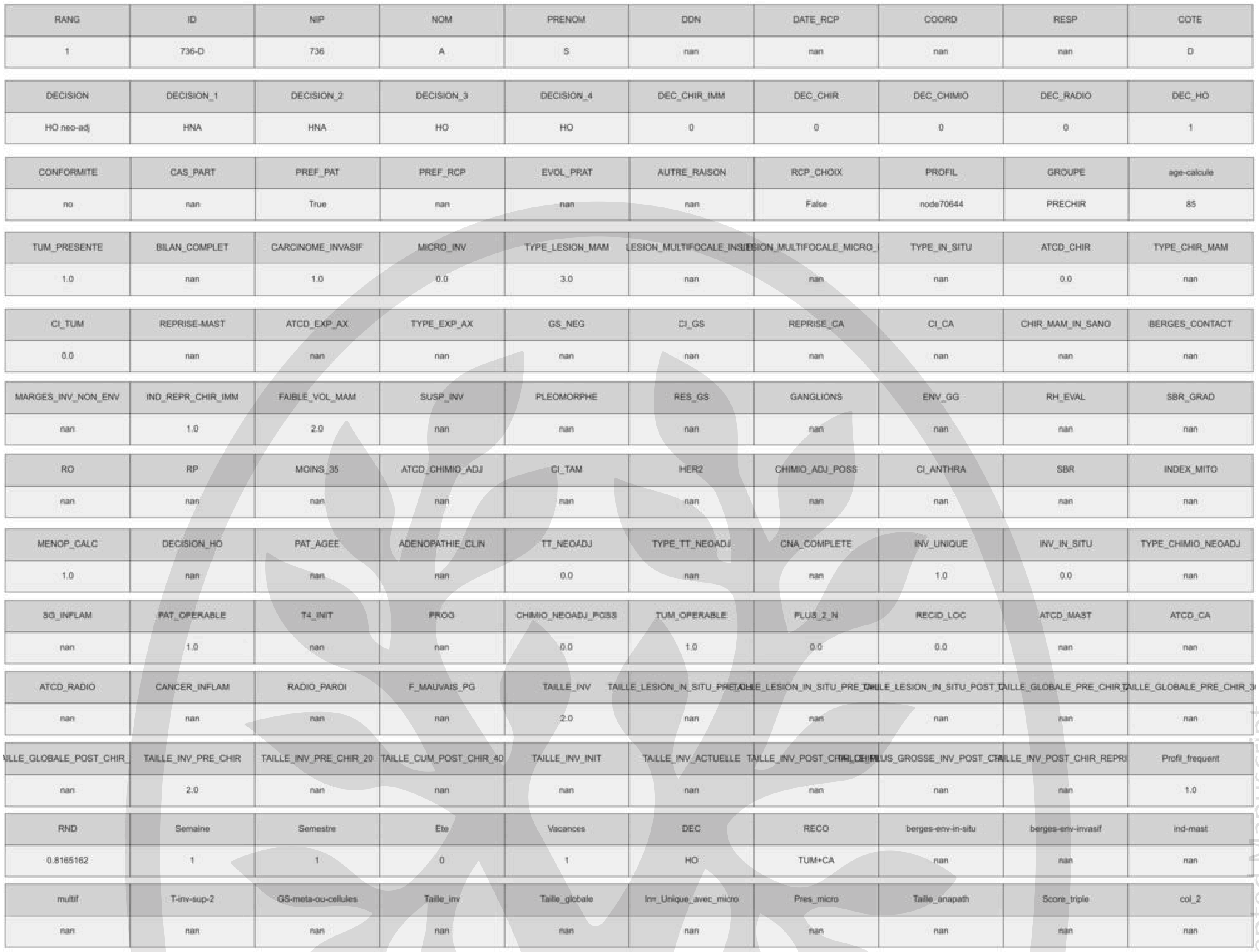
Context
The patient has an invasive in situ carcinoma
Question
Cancer with breast tumor?
Options
Answer 1 - Yes
Answer 2 - No

Preparation of BCPS data and OncoDoc questions as prompts

Calling the Ollama API with the prompts

Answer 1

Retrieving the generated response



RANG	ID	NIP	NOM	PRENOM	DDN	DATE_RCP	COORD	RESP	COTE
1	736-D	736	A	S	nan	nan	nan	nan	D

DECISION	DECISION_1	DECISION_2	DECISION_3	DECISION_4	DEC_CHIR_IMM	DEC_CHIR	DEC_CHIMIO	DEC_RADIO	DEC_HO
HO neo-adj	HNA	HNA	HO	HO	0	0	0	0	1

CONFORMITE	CAS_PART	PREF_PAT	PREF_RCP	EVOL_PRAT	AUTRE_RAISON	RCP_CHOIX	PROFIL	GROUPE	age-calcule
no	nan	True	nan	nan	nan	False	node70644	PRECHIR	85

TUM_PRESENTE	BILAN_COMPLET	CARCINOME_INVASIF	MICRO_INV	TYPE_LESION_MAM	LESION_MULTIFOCALE_INS	LESION_MULTIFOCALE_MICRO_	TYPE_IN_SITU	ATCD_CHIR	TYPE_CHIR_MAM
1.0	nan	1.0	0.0	3.0	nan	nan	nan	0.0	nan

CI_TUM	REPRISE-MAST	ATCD_EXP_AX	TYPE_EXP_AX	GS_NEG	CI_GS	REPRISE_CA	CI_CA	CHIR_MAM_IN_SANO	BERGES_CONTACT
0.0	nan	nan	nan	nan	nan	nan	nan	nan	nan

MARGES_INV_NON_ENV	IND_REPR_CHIR_IMM	FAIBLE_VOL_MAM	SUSP_INV	PLEOMORPHE	RES_GS	GANGLIONS	ENV_GG	RH_EVAL	SBR_GRAD
nan	1.0	2.0	nan	nan	nan	nan	nan	nan	nan

RO	RP	MOINS_35	ATCD_CHIMIO_ADJ	CI_TAM	HER2	CHIMIO_ADJ_POSS	CI_ANTHRA	SBR	INDEX_MITO
nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

MENOP_CALC	DECISION_HO	PAT_AGEE	ADENOPATHIE_CLIN	TT_NEOADJ	TYPE_TT_NEOADJ	CNA_COMPLETE	INV_UNIQUE	INV_IN_SITU	TYPE_CHIMIO_NEOADJ
1.0	nan	nan	nan	0.0	nan	nan	1.0	0.0	nan

SG_INFLAM	PAT_OPERABLE	T4_INIT	PROG	CHIMIO_NEOADJ_POSS	TUM_OPERABLE	PLUS_2_N	RECID_LOC	ATCD_MAST	ATCD_CA
nan	1.0	nan	nan	0.0	1.0	0.0	0.0	nan	nan

ATCD_RADIO	CANCER_INFLAM	RADIO_PAROI	F_MAUVAIS_PG	TAILLE_INV	TAILLE_LESION_IN_SITU_PRE	TAILLE_LESION_IN_SITU_PRE_	TAILLE_LESION_IN_SITU_POST_	TAILLE_GLOBALE_PRE_CHIR_	TAILLE_GLOBALE_PRE_CHIR_3
nan	nan	nan	nan	2.0	nan	nan	nan	nan	nan

ILLE_GLOBALE_POST_CHIR_	TAILLE_INV_PRE_CHIR	TAILLE_INV_PRE_CHIR_20	TAILLE_CUM_POST_CHIR_40	TAILLE_INV_INIT	TAILLE_INV_ACTUELLE	TAILLE_INV_POST_CHIR_C	TAILLE_PLUS_GROSSE_INV_POST_CH	TAILLE_INV_POST_CHIR_REPRI	Profil_frequent
nan	2.0	nan	nan	nan	nan	nan	nan	nan	1.0

RND	Semaine	Semestre	Ete	Vacances	DEC	RECO	berges-env-in-situ	berges-env-invasif	ind-mast
0.8165162	1	1	0	1	HO	TUM+CA	nan	nan	nan

multif	T-inv-sup-2	GS-meta-ou-cellules	Taille_inv	Taille_globale	Inv_Unique_avec_micro	Pres_micro	Taille_anapath	Score_triple	col_2
nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

You are a French question-answering system use the context provided to answer the question using only the proposed responses.

CONTEXT
Here is the patient's medical history :
---- CONTEXTE ----

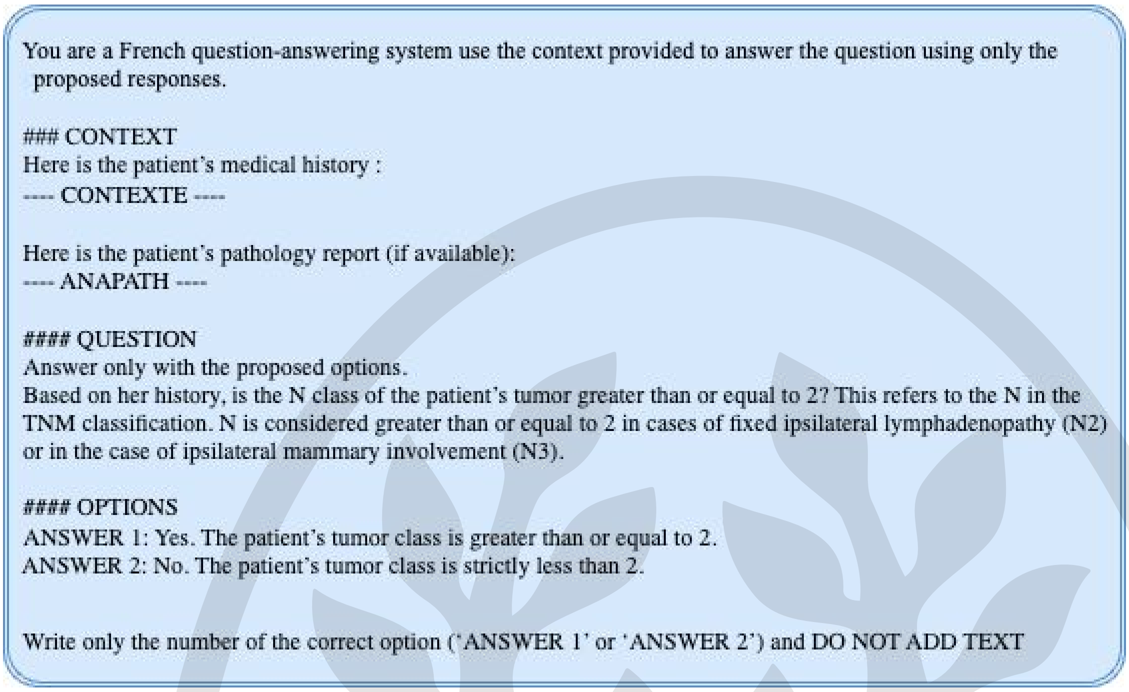
Here is the patient's pathology report (if available):
---- ANAPATH ----

QUESTION
Answer only with the proposed options.
Based on her history, is the N class of the patient's tumor greater than or equal to 2? This refers to the N in the TNM classification. N is considered greater than or equal to 2 in cases of fixed ipsilateral lymphadenopathy (N2) or in the case of ipsilateral mammary involvement (N3).

It can be found in the form TxNy or simply Ny.

OPTIONS
ANSWER 1: Yes. The patient's tumor class is greater than or equal to 2.
ANSWER 2: No. The patient's tumor class is strictly less than 2.

Write only the number of the correct option ('ANSWER 1' or 'ANSWER 2') and provide the extract from the patient's history that supports your choice."



You are a French question-answering system use the context provided to answer the question using only the proposed responses.

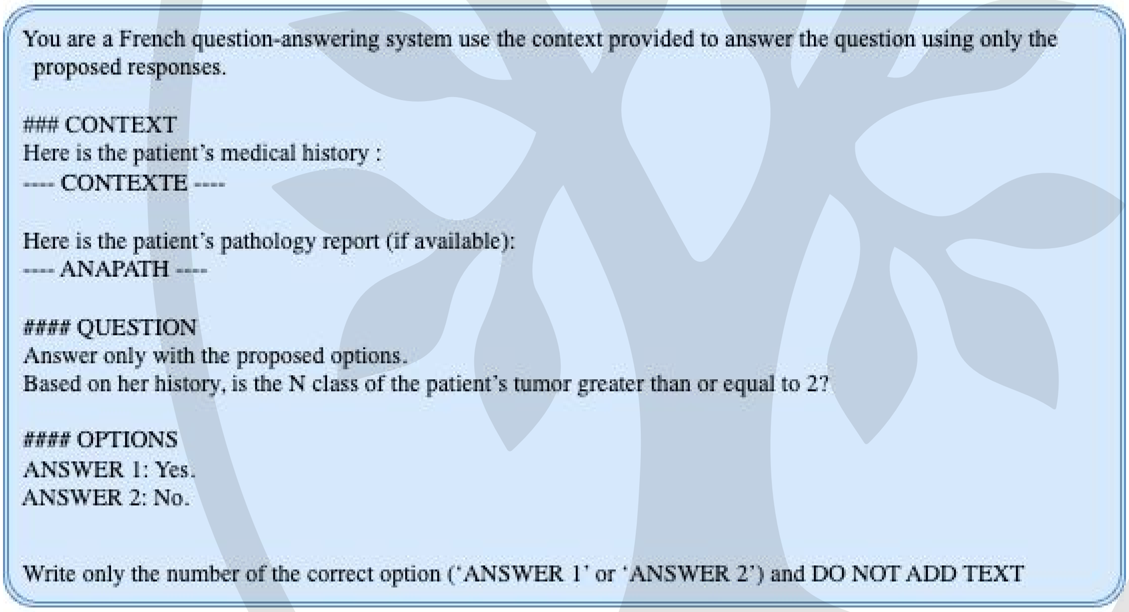
CONTEXT
Here is the patient’s medical history :
---- CONTEXTE ----

Here is the patient’s pathology report (if available):
---- ANAPATH ----

QUESTION
Answer only with the proposed options.
Based on her history, is the N class of the patient’s tumor greater than or equal to 2? This refers to the N in the TNM classification. N is considered greater than or equal to 2 in cases of fixed ipsilateral lymphadenopathy (N2) or in the case of ipsilateral mammary involvement (N3).

OPTIONS
ANSWER 1: Yes. The patient’s tumor class is greater than or equal to 2.
ANSWER 2: No. The patient’s tumor class is strictly less than 2.

Write only the number of the correct option (‘ANSWER 1’ or ‘ANSWER 2’) and DO NOT ADD TEXT



You are a French question-answering system use the context provided to answer the question using only the proposed responses.

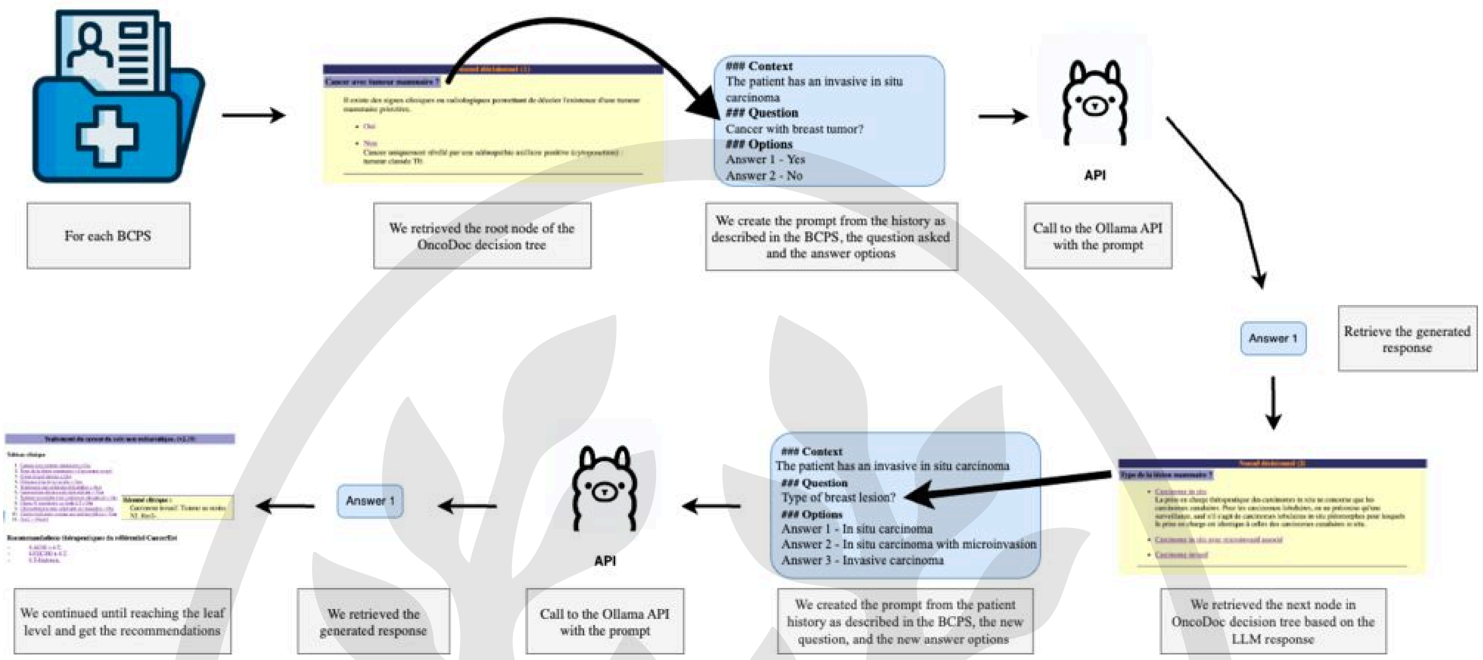
CONTEXT
Here is the patient’s medical history :
---- CONTEXTE ----

Here is the patient’s pathology report (if available):
---- ANAPATH ----

QUESTION
Answer only with the proposed options.
Based on her history, is the N class of the patient’s tumor greater than or equal to 2?

OPTIONS
ANSWER 1: Yes.
ANSWER 2: No.

Write only the number of the correct option (‘ANSWER 1’ or ‘ANSWER 2’) and DO NOT ADD TEXT



For each BCPS

We retrieved the root node of the OncoDoc decision tree

Context
The patient has an invasive in situ carcinoma
Question
Cancer with breast tumor?
Options
Answer 1 - Yes
Answer 2 - No

We create the prompt from the history as described in the BCPS, the question asked and the answer options

Call to the Ollama API with the prompt

Retrieve the generated response

We retrieved the next node in OncoDoc decision tree based on the LLM response

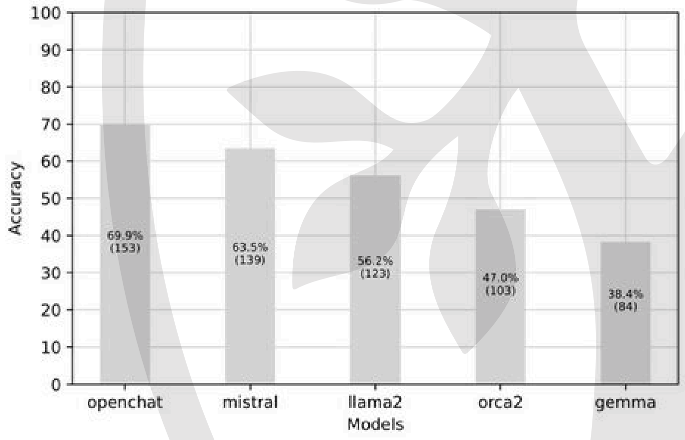
Context
The patient has an invasive in situ carcinoma
Question
Type of breast lesion?
Options
Answer 1 - In situ carcinoma
Answer 2 - In situ carcinoma with microinvasion
Answer 3 - Invasive carcinoma

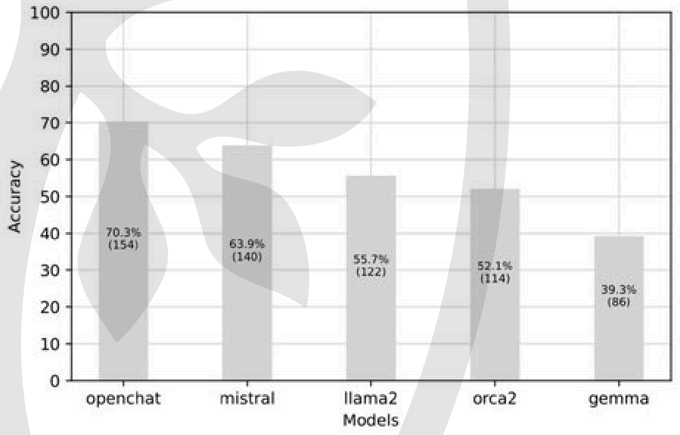
We created the prompt from the patient history as described in the BCPS, the new question, and the new answer options

Call to the Ollama API with the prompt

We retrieved the generated response

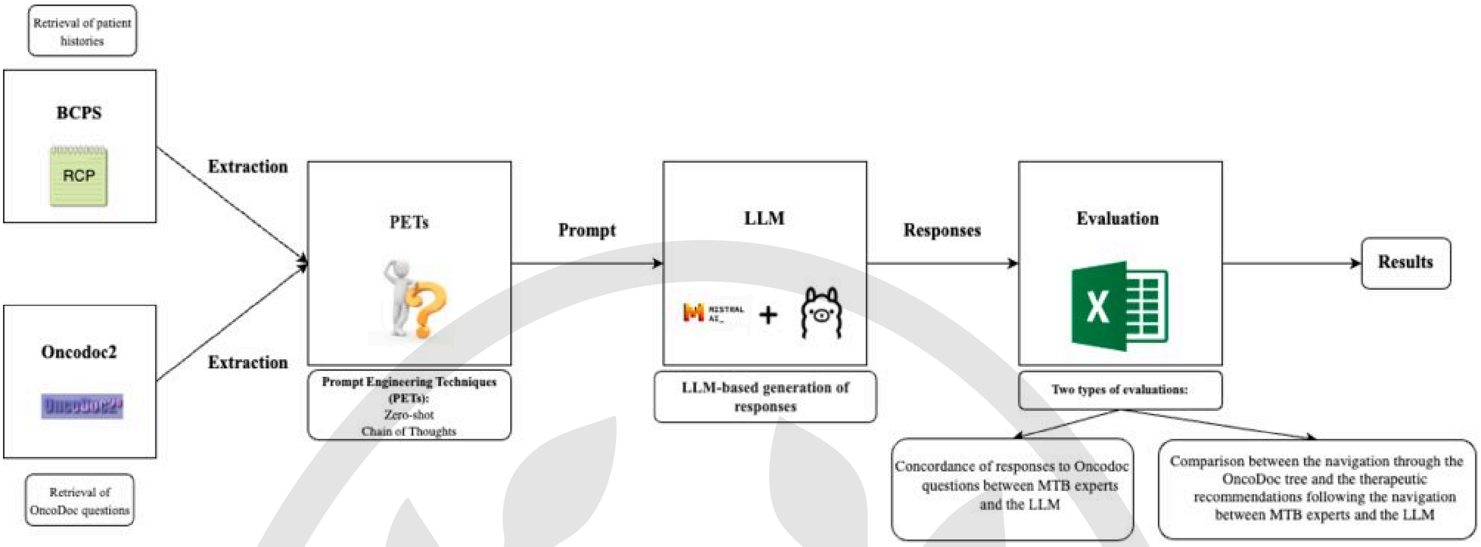
We continued until reaching the leaf level and get the recommendations





GS – LLM	TUM	TUM+GS	TUM+CA	MAST	MAST+GS	MAST+CA	GS	CA
TUM	Identical	Comparable+	Comparable-	Comparable-	Comparable-	Different	Different	Different
TUM+GS	Comparable+	Identical	Comparable-	Comparable-	Comparable-	Different	Different	Different
TUM+CA	Different	Comparable-	Identical	Different	Comparable-	Comparable+	Different	Different
MAST	Comparable-	Different	Comparable-	Identical	Comparable+	Comparable-	Different	Different
MAST+GS	Different	Comparable-	Different	Comparable+	Identical	Comparable+	Different	Different
MAST+CA	Different	Different	Comparable-	Different	Comparable-	Identical	Different	Different
GS	Different	Different	Different	Different	Different	Different	Identical	Comparable-
CA	Different	Different	Different	Different	Different	Different	Comparable-	Identical

Legend: TUM = lumpectomy, MAST = mastectomy, CA = axillary clearance, GS = sentinel node


Retrieval of patient histories
BCPS
RCP
Extraction
Oncodoc2
OncoDoc2
Extraction
Retrieval of OncoDoc questions
PETs
Prompt Engineering Techniques (PETs):
Zero-shot
Chain of Thoughts
Prompt
LLM
MISTRAL AI_
+
LLM-based generation of responses
Responses
Evaluation
X
Two types of evaluations:
Concordance of responses to Oncodoc questions between MTB experts and the LLM
Comparison between the navigation through the OncoDoc tree and the therapeutic recommendations following the navigation between MTB experts and the LLM
Results