Can Generative AI Support Patients' & Caregivers' Informational Needs? Towards Task-Centric Evaluation Of AI Systems

Shreya Rajagopal[1], Jae Ho Sohn[2], Hari Subramonyam[3] and Shiwali Mohan[4,*]

[1]*University of Michigan, Ann Arbor*
[2]*University of California, San Francisco*
[3]*Stanford University, Stanford*
[4]*SRI International*

Abstract

Generative AI systems such as ChatGPT and Claude are built upon language models that are typically evaluated for accuracy on curated benchmark datasets. Such evaluation paradigms measure predictive and reasoning capabilities of language models but do not assess if they can provide information that is useful to people. In this paper, we take some initial steps in developing an evaluation paradigm that centers human understanding and decision-making. We study the utility of generative AI systems in supporting people in a concrete task - making sense of clinical reports and imagery in order to make a clinical decision. We conducted a formative need-finding study in which participants discussed chest computed tomography (CT) scans and associated radiology reports of a fictitious close relative with a cardiothoracic radiologist. Using thematic analysis of the conversation between participants and medical experts, we identified commonly occurring themes across interactions, including clarifying medical terminology, locating the problems mentioned in the report in the scanned image, understanding disease prognosis, discussing the next diagnostic steps, and comparing treatment options. Based on these themes, we evaluated two state-of-the-art generative AI systems against the radiologist's responses. Our results reveal variability in the quality of responses generated by the models across various themes. We highlight the importance of patient-facing generative AI systems to accommodate a diverse range of conversational themes, catering to the real-world informational needs of patients.

Keywords

Generative AI Systems, Conversational AI, Task-centric Evaluation, AI System Evaluation, Patient-centric Healthcare,

1. Introduction

Conversational systems built upon generative AI technology are general-purpose tools that enable people to access information through natural conversations [1]. Healthcare is primed to have significant benefits with reliable, robust and trustworthy adoption of such systems, assisting different medical interactions/conversations involving patients, doctors, clinicians, and healthcare administrators. Purportedly, generative AI systems have medical commonsense at par with doctors as evidenced by their high scores in standard medical knowledge evaluation paradigms like the United States Medical Licensing Examination (USMLE) [2]. Despite impressive results on benchmark medical and clinical datasets, the case for their adoption in real world clinical scenarios is still unclear. While several clinical researchers and practitioners have expressed excitement about their potential in healthcare [3, 4], several critical questions must be answered before their adoption can begin. These questions include what role generative AI systems can play in a healthcare system, how their behavior should be modulated, and how their performance can be evaluated. Answering these questions requires going beyond benchmark datasets prevalent in AI & ML research to understand the nature of real world problems and characteristics of desired solutions.

In this paper, we study a real world clinical usecase for generative AI systems. Patients (along with their caregivers) who are better educated about their disease have better outcomes, especially in serious illnesses such as cancer [5, 6, 7, 8]. However, they face complex information and learning challenges when navigating healthcare systems. They not only must learn what clinical terminology means but also what their medical reports convey, how the severity of their disease is diagnosed, what their treatment options are etc. This often leads to anxiety and stress not only from the disease but also from the complexity of understanding and managing it. To alleviate these challenges, healthcare systems have invested significantly in patient education resources. Meaningful and empathetic interactions with healthcare providers has a significant, positive impact on a patient's journey through the healthcare system [9]. While arguably the most useful, it also takes the experts' time away from their other critical responsibilities, adding to their time burden. On the other hand, patients often desire additional informational support and augment their understanding with generic and unverified information from search engines [10] and forums [11]. Integrating generative AI systems in patient-facing tools requires that generative AI systems be designed to teach and explain diagnosis, and prognosis; weigh risks, benefits, and costs of treatments; and understand complex human emotional and social contexts. Lack of adequate considerations can lead to harmful consequences for patients, including misinterpretation of generated output, over-reliance on AI, and poor disease management.

We examine the scenario where patients and caregivers attempt to understand their medical scans and reports by interacting with a clinical expert. In collaboration with a cardiothoracic radiologist at a major university hospital (and a co-author), we conducted a formative need-finding study with 5 participants. The study was framed as a conversational interview in which participants discussed chest computed tomography (CT) scans and associated radiology reports of a fictitious close relative with a radiologist. We limited the participants to a small number because the subject of the study is highly sensitive and could trigger anxiety in our participants. Additionally, each interview required $3hrs$ of the radiologist's time. We analyzed the recorded interviews with inductive thematic analysis to identify emerging themes and categories in the conversations. Next, we evaluated two state-of-art multi-modal generative AI systems ChatGPT-4o [12] and Claude 3.5 Sonnet [19] - against the radiologist's responses in the realistic interview task. Our findings highlight the strengths and weaknesses of generative AI models in addressing the information needs of patients and caregivers.

Our work situates and evaluates generative AI capabilities in a specific human context of seeking information to enhance understanding. This context enables us to identify the variety of information patients seek and study if AI systems are up to the challenge. Further, generative AI systems are typically evaluated on datasets curated by and for clinicians. These datasets contain questions and answers that are most interesting to clinicians, while performance measurement includes criteria such as diagnostic accuracy. Such an approach introduces an implicit bias in the design and analysis of these systems - they are designed to replace clinicians by mimicking the output of their problem-solving process, following the *AI-is-automation* philosophy. Instead, our research takes the *AI-is-augmentation* view and centers patients and caregivers. We explore if generative AI systems can support humans - patients and caregivers - in a specific task - making sense of the their reports and scans and accessing knowledge about their disease and treatment.

Advancing the science of human-centered, task-centered evaluation of generative AI systems, our paper makes the following contributions. It,

- Identifies an opportunity for adoption of generative AI system in clinical practice and specifies an AI task reflecting a real world problem;

- Characterizes the variability in types of information people seek when making sense of medical reports and scans;

- Proposes an evaluation approach for conversational systems quantitatively using semantic similarity as well as qualitatively, building upon the Gricean maxims of effective communication;

- Evaluates two state-of-art generative AI systems - ChatGPT-4o and Claude 3.5 Sonnet - on the proposed evaluation rubric.

2. Background

Recently, large language models (LLMs) like PaLM [13], GPTs [14, 15, 16], LLaMA [17, 18], and Claude [19] have significantly advanced the state-of-the-art in various natural language processing (NLP) tasks [20, 21]. This has inspired the development of many medical LLMs [22, 23, 24, 25, 26, 27] to assist the healthcare professionals and improve patient care [28, 29, 30]. Based on publicly available LLMs, specialized medical versions of the models like ChatDoctor [26], MedAlpacha [24], PMC-LLaMA [31], BenTsao [32], and Clinical Camel [25] have come up in the past years. Models like Med-PaLM [22] and Med-PaLM-2 [23] built on top of PaLM have achieved close to human expert scores in the United States Medical Licensing Examination (USMLE) [2]. Apart for only text-based models several multimodal models [33, 34, 35, 12] have come to prominence in the last year with the capability of processing both images, text and videos to generate responses. Similar to NLP models such multimodal models have promising application in the medical domain. They can be used as visual question answering systems for patient medical reports, medical report summarization and analysis tasks among many other possibilities. In this work, we evaluate Claude [19] and GPT-4V [12] in a patient-facing, clinical usecase.

GPT-4 model with vision capabilities (GPT-4V) [12] has been identified as the most advanced LLM thus far, particularly in its application to radiology domain, where it has been compared against the current state-of-the-art (SOTA) models [36]. The research, which referenced various sources, thoroughly examined GPT-4V across a broad spectrum of standard radiology text-based tasks, including MS-CXR-T[37] and CheXbert[38]. The findings revealed that GPT-4 either surpasses or matches the performance of specialized fine-tuned radiology models. A detailed investigation [12] of GPT-4V studied its utility in generating radiology reports. This investigation demonstrated that GPT-4V could accurately diagnose and recommend management based on X-ray images. The findings from this research indicate that GPT-4V holds significant promise as an AI assistant in radiology report generation. Claude [19] models are more recent but boast multi-modal reasoning capabilities at par to GPT and were the top-performing models on a recent evaluation on medical, multi-modal answer generation benchmarks [39].

Despite the growing interest in the application of Generative AI models in the medical and healthcare domains, there are a few key issues that need to be addressed for their reliable and effective usage in such a critical domain. Several recent surveys [20, 40, 41] on the application of LLMs for medicine highlight the best practices and challenges involved in their application and development. Effective evaluation of AI systems build around LLMs is one such challenge. Particularly in the medical domain, the current benchmarks and metrics often fail to evaluate LLMs' overall capabilities and emerging abilities. Current benchmarks such as USMLE MedQA [2] and MedMCQA [42] cover broad range of question-answering tasks but lack the information to evaluate on metrics like trustworthiness and helpfulness to different parties involved like patients, doctors and administrators [41]. It is imperative to prepare more domain and task specific evaluation protocols to adjudicate the properties of GenAI models. Singhal et al. [22] created a benchmark comprising the most commonly searched health queries to evaluate LLM responses. TruthfulQA [43] and HaluEval [44] attempt to evaluate truthfulness and factual accuracy, but not for any specific medical domain. The analyses reported in our work is the first attempt to reflect upon the conversational patterns in real-world patient-doctor interactions and evaluate strengths and weakness of generative AI system responses against those of human experts.

3. Preliminaries

The average patient's journey through the medical system in the United States is shown in Figure 1. Typically, a patient first enters the medical system and meets with a primary care physician (PCP) who collects their history, performs a physical evaluation, and orders further medical and radiology-based tests in order to inform a diagnosis. A radiologic technologist performs the radiology tests the referring PCP orders, and passes on these scans to the consulting radiologist. The radiologist analyses these scans and generates an associated radiology report. The referring physician receives these scans, reports,

and the results of other medical tests they had ordered, and combines this information to arrive at an appropriate diagnosis. They then meet with the patient again to discuss the diagnosis, and followup and treatment options. It is relatively rare for radiologists to directly meet with their patients to help them understand their scans and radiology reports [45, 46].
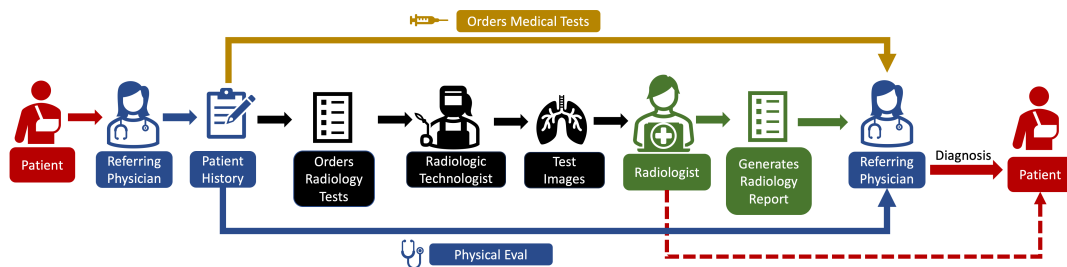
Figure 1: Patient journey through the medical system highlight various interactions they have with different healthcare experts.

A radiology report contains 3 major parts - Imaging techniques, Findings, and Impressions. Imaging techniques describe the specific imaging methodology used, and other parameters associated with the method. For example, in a chest CT, this section would describe the dimensional plane along with images that were collected (e.g. axial), the size of each voxel in the image, and whether or not intravenous contrast was administered. Findings describe whether or not any suspicious entities were found in the body parts scanned, and also describe their size, shape, and other appearance-based factors. Impressions are where the radiologists' expertise most comes into play, with the radiologist giving their opinions on what the findings likely mean in the diagnostic context of the tests ordered by the patient's physician. This could include likely diagnoses, the odds of each diagnosis, and suggested diagnostic follow-ups.

Since the radiology report is intended to aid the patient's physician in arriving at a diagnosis, and not the patient themselves, the report contains medical terminology that is not accessible to the average patient. However, the patient receives immediate access to their radiology report once it is generated, in their electronic health records. Since it is atypical and oftentimes difficult for a patient to schedule a one-on-one consultation with their consulting radiologist to understand their scan and the associated report, patients have to wait two to three weeks on average to meet with their referring physician instead. This can tend to create anxiety in patients faced with medical information they cannot understand. Often, patients resort to unstructured, unverified resources such as Reddit for further information.

An generative AI system can be useful during the period that the patient is waiting to meet with their PCP. The AI system can support a patient understand medical terminology used in the report and how the terms relate to the scan. It can help the patient understand the causes of their disease and its prognosis. It can support patients research various treatment options and weight their risks and benefits. Engaging with an intelligent information tool can prepare the patient (and their caregivers) for a productive meeting with their PCP.

4. Formative Study of Patient-Provider Interaction

We designed a formative study to closely resemble the medical context patients and caregivers find themselves in as they navigate the healthcare system (Section 3) - the timepoint when they have access to their scans and reports and are preparing to meet with the PCP. We formulated the study as an interview between a participant simulating the role of a caregiver with a radiologist. It recreated the realistic interaction between a close relative of a patient receiving some worrying medical information, and the radiologist who developed the report. The study was conducted virtually on Zoom, with the experimenter and participant meeting first, and the radiologist joining at a later stage in the task.

4.1. Study
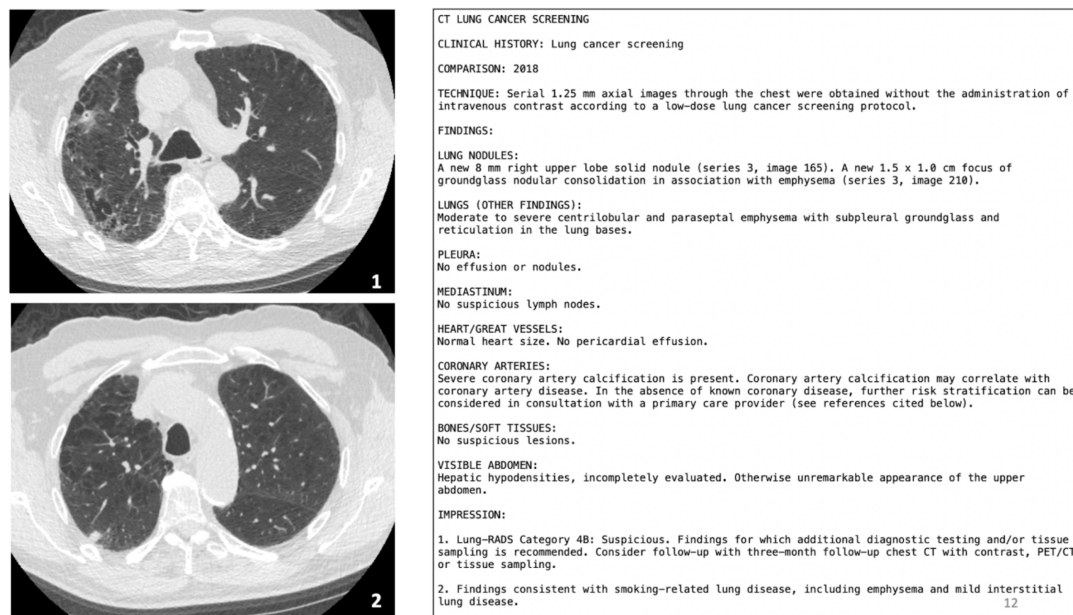
Figure 2: 2 axial slices of the lung CT and the radiology report presented to study participants
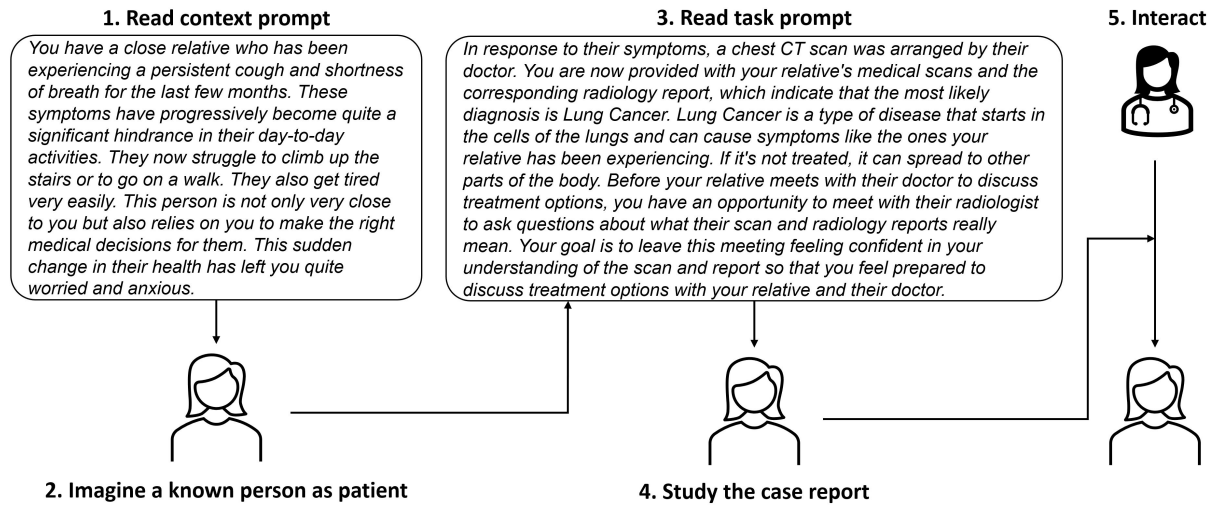
4.1.1. Materials

We created a case report (shown in Figure 2) that comprised of 2 axial slices of the lung CT and a corresponding radiology report with findings. The 2 slices selected were the ones with evidence relevant for the findings in the report. The case report was developed using data collected for real cancer patients, modified and redacted for patient privacy purposes. The case was carefully selected such that it had some ambiguity both about visual evidence (i.e., there we two nodules that indicated malignancy) and the potential diagnosis (i.e., it wasn't certain that the paitent had cancer that this timepoint). These characteristics of the case enable us to understand how doctors communicate and manage uncertainty and how that might differ in Generative AI systems.

4.1.2. Protocol

Figure 3 shows the study protocol we implemented. The study progressed as follows (item numbers below correspond to stages of the study depicted in 3):

1. Each participant was provided with the context prompt containing some symptoms their relative has been experiencing, and their role in their relative's diagnostic journey.
2. The participants were asked to imagine a close relative as patients and describe them in terms their age, sex, and relationship with the participant. They were also asked what other health conditions the imagined relative had simultaneously been dealing with. This was done to aid the follow-up interaction with the radiologist, where responses to certain concerns would be contingent on this information about patient demographics and comorbidities.
3. The participants were provided with a task prompt that suggested that the most likely diagnosis was lung cancer, and described the participant's goals in their upcoming interaction with their relative's radiologist.
4. The participants studied the case report (Figure 2) shown on a screen and noted what they currently understood about their relative's diagnosis, and the questions and concerns they would like to discuss with the radiologist in their upcoming interaction. they were also prompted to remember their emotional concerns regarding their relative's well-being in addition to seeking information about

1. Read context prompt

You have a close relative who has been experiencing a persistent cough and shortness of breath for the last few months. These symptoms have progressively become quite a significant hindrance in their day-to-day activities. They now struggle to climb up the stairs or to go on a walk. They also get tired very easily. This person is not only very close to you but also relies on you to make the right medical decisions for them. This sudden change in their health has left you quite worried and anxious.

3. Read task prompt

In response to their symptoms, a chest CT scan was arranged by their doctor. You are now provided with your relative's medical scans and the corresponding radiology report, which indicate that the most likely diagnosis is Lung Cancer. Lung Cancer is a type of disease that starts in the cells of the lungs and can cause symptoms like the ones your relative has been experiencing. If it's not treated, it can spread to other parts of the body. Before your relative meets with their doctor to discuss treatment options, you have an opportunity to meet with their radiologist to ask questions about what their scan and radiology reports really mean. Your goal is to leave this meeting feeling confident in your understanding of the scan and report so that you feel prepared to discuss treatment options with your relative and their doctor.

5. Interact

2. Imagine a known person as patient

4. Study the case report

Figure 3: Protocol showing the context and task prompt as well as participant actions and interactions

the scans and reports. This step was designed to prepare the participant for meaningful interaction with the radiologist.

5. The radiologist was then invited into the virtual room, and following initial introductions, the experimenter turned their microphone and camera off to allow for a natural interaction between the radiologist and the participant. The radiologist began by briefly discussing their role in the patient's medical journey and introducing the patient to the problem at hand. The participants asked what questions came naturally to their mind as they looked at the case report. The scans and the report remained on the screen for the duration of the interaction, and the radiologist would often annotate the images on the screen to answer questions where referring to the scan was helpful.

4.1.3. Conducting the study

Our research company's internal review board conducted an *Expedited* review of our study material and protocol and approved it with the risk determined to be *minimal*. We recruited participants via an internal mailing list as well as through individual connections. During recruitment, we were careful to highlight the sensitive nature of the materials and the concern that participation in the study may trigger emotional response if the participants' close relative had gone through a similar medical experience. The email recruitment and the consent form prominently noted that the participant could request to stop the study at any point they desired without affecting their compensation. We recruited 5 participants and each study session took close to 2 hours. Each participant was paid $30 as an Amazon gift card. The radiologist participating in the study is a from a major hospital at a public university and is a co-author. The interaction between the participant and the radiologist was recorded.

4.2. Inductive Thematic Analysis

We transcribed the interaction sessions using Zoom's transcription service. Our final dataset consisted of 5 documents averaging 3049 ($R = 1642 - 4614$) words. We analyzed this data set with inductive thematic analysis [47]. We started with a single interview and went through the transcript one sentence at a time, assigning codes that captured the content of the sentence, or a set of related sentences. For each new interview we analyzed, we used existing codes generated based on previous interviews and assigned new codes only when there were no existing codes to capture the content of a specific statement or set of statements. Each time a new set of codes was added based on an interview, we revisited previously coded interviews to check if the new codes could be assigned to any statements in them. We found that we reached data saturation at 5 interviews with only 2 new codes being added in analyzing the final interview. We identified a set of 91 content-based codes across all transcribed

participant-radiologist interactions, and we grouped these codes into a set of 10 broad themes that best captured the relationship between the codes.

The primary set of content-based codes were grouped together into the following 10 themes. Each theme is described with some representative examples of quotations. These quotations consisting questions and answers are extracted and unmodified from the transcribed versions of the interview data, with the interviews these were drawn from indicated in parentheses. Quotations by the radiologist are indicated as (R-Interview Number).

Theme 1: Statistics of Lung Injury and Associated Damage to Other Organs comprised of content including the likelihood of the damage found in the lungs for a patient with specific demographics, the correlation of this lung damage with damage to other organs, and the likelihood of damage to other organs as well. For example, *"So usually, especially in this age group of my grandmother, emphysema can often lead to lung cancer, basically right?" (P03), "I only have one question about specifically the coronary artery part where it says there is severe coronary artery calcification present. Do you know if that is related to lung cancer at all?" (P02), "So it's very, very common for people, patients at age 80s to have coronary artery disease." (R-P03), "So now just having severe coronary artery calcification doesn't necessarily mean that your sister is going to have a heart attack, but it's certainly a statistical risk factor that we watch out for." (R-P02)*

Theme 2: Role of Members of the Medical Team included description of a typical patient's journey through the medical system, and how different members of the medical team had different roles in diagnosing and designing a treatment plan for the patient. This included subjects like the radiologist's role, the role of the referring physician, the oncologist, and the multidisciplinary tumor board. E.g., *"For diagnosis of cancer we have these multiple experts talking to each other and determining what is the best next course of action based on our combination of expertise." (R-P03), "It would be important to talk to a board-certified medical oncologist to determine and discuss all the pros and cons of various different regimens, some of which are more toxic, and some of which are much less toxic and well tolerated." (R-P04), "So that would be something that would absolutely need to be discussed with a primary care doctor and likely with a cardiologist to ensure that your sister's risk of having a heart attack in the future is well controlled and treated." (R-P01), "With all these imaging data we actually have a very regular multidisciplinary team of doctors including myself, as well as thoracic surgeons, radiation oncologists, pathologists, pulmonologists, and oncologists all gathered together on a weekly basis to discuss the cases because there's not only diagnostic uncertainty but there are also treatment options that vary in their profile and some things are more appropriate than others."(R-P1)*

Theme 3: Locating Issues mentioned in the Radiology Report to the CT Scan included discussions of how the participant could locate a specific term from the radiology report on the CT scan. This information was also brought up by the radiologist himself in answering a question where pointing to the image might be helpful. The radiologist annotated the on-screen scan to demarcate parts of the scan. For example, *"If you look at the scans in image number 1 and 2, you can see that there are kind of dark bubbly spots and I can use the zoom technology to annotate for you. I'll just show an example area here. Other example areas like here are kind of dark bubbly spots that are present in the lungs and that represents areas of just ballooned out Airways as a result of destruction." (R-P02), "So how do I find a nodule here?" (P04), "If I can just sort of highlight that area here. That and that. Do you see these round sort of white dots, white sort of areas? These are the areas that I'm concerned might be because of lung cancer." (R-P02), "How would I differentiate between the scan of a healthy lung versus the one which is in stage one, stage 2, or stage 3? So, looking at it, what should I look for?" (P04)*

Theme 4: Selecting a Treatment Plan included a range of discussions comparing and contrasting the treatment plans available in terms of their success rate, degree of discomfort to the patient, and the costs involved. Examples are as follows – *"As long as she's in relatively and reasonably good health*

to walk around, typically surgery is an option and the exact evaluation of which I defer to my thoracic surgery colleagues to do the evaluation." (R-P03), "If the cancer has already spread throughout the body, then chemotherapy may be the only option available to shrink the cancer and seek care." (R-P04), "Between chemotherapy or radiation, which has the least side effects, because when I see some patients going through the treatment, the way they suffer, sometimes even they themselves feel that they would rather die than go through the treatment." (P04), "What treatment option would also say is like more expensive than others and vice versa?"(P03)

Theme 5: Understanding Medical Terminology from Radiology Report dealt understanding the content of the radiology report. For example, *"Got it, I know you mentioned emphysema. What exactly does that mean? Is that just like lung, you know, like problems with, like the lung or like lung being cancerous, is that but that stands for?" (P03), " For some terminology that I wasn't entirely clear on, like the pleura and then mediastinum, nothing special was mentioned, but can you just briefly mention what those mean and what you were looking for there?" (P05), "Should I be also worried about these Hepatic hyperdensities in the visible abdomen area?" (P1), "The one main finding is the lung RADS category 4B that's mentioned in the impression section and what that is referring to is based on the two lung nodules, and these are sort of tiny, sometimes very tiny lesions that are found in the lungs, many of which can be a consequence of having infection in the past or just air pollution or other things."(R-P02)*

Theme 6: Diagnostic Follow-up Plan included content related to the next diagnostic steps to be taken following the initial CT scan – discussions on the different kinds of follow-up options available and what would be the most appropriate in this given situation. E.g, – *"Tissue sampling means biopsy – so we insert a needle either through the skin or using a bronchoscope and then take a surgical sample of these lesions and then take a look at these them under the microscope to tell for sure." (R-P03), "In the case of PET CT, it's a metabolic imaging, so it highlights areas of high metabolic activity, which often corresponds to areas of cancer." (R-P03), "Would these scans happen pretty sequentially in just like a couple of weeks? Or would we have months between these scans or how long I guess would this last?" (P03), "So it seems like there is definitely a suspicion of lung cancer. What would you recommend to be the next steps? A chest CT with contrast has been recommended from the results. But how is that different from this and what would that reveal? (P03)"*

Theme 7: Risk Factors and Lifestyle Changes included content related to different kinds of risk factors involved in lung cancer, and discussions on whether certain lifestyle factors might help curb the progression of the disease. For example, *"Exercise has been shown to be helpful in numerous studies to reduce the risk of cardiovascular disease and would certainly be recommended assuming it's not just overdone and there are no conditions that cause issues with exercising." (R-P1), " Most commonly in the United States, it's because of smoking, but it can also be due to air pollution or some other conditions that may result in the destruction of that part of the lungs." (R-P02), "If smoking is continuing, making sure to stop smoking is important. Also if there's some sort of pollutant or other thing, if there's any way to really just keep your lungs healthy from here on, it will be important to make sure that it doesn't progress further and cause more shortness of breath." (R-P1), "Are there some foods that kind of help with these things? Or in general, is weight loss better?" (P1)*

Theme 8: Disease Progression and Prognosis included content about the way the disease was progressing, and what the patient could expect in the future. For example, *"I was mostly concerned about the health risks going forward and whether there was any way we can kind of stop the cancer from happening in some way." (P1), "If you were to follow this module, say in three months, and re-scan if this actually was truly a cancer, then it would grow in a pattern and growth rate that's typical of a lung cancer. Whereas if this represents something else and something else could be, say, the focus of infection or inflammation, these tend to either go away or increase or change in a pattern and growth rate that's not consistent with cancer." (R-P03), "Assuming that this nodule turns out to be cancer, it would be stage one A*

or one B, which is a really good stage to be in because I don't see any nodal metastasis." (R-P04), "Can they shrink, or can you make them disappear, such that in the follow-up scan, you will have a cleaner scan? Is this possible?" (P05)

Theme 9: Diagnostic Confidence and CT Scan Limitations   dealt with discussions on how confident the radiologist was about the findings and the general limitations of the presented scans. Examples include, *"Having said that, there is some diagnostic uncertainty with these that can be more definitively addressed by doing a follow-up CT scan or doing some additional imaging modalities or doing a biopsy." (R-P1), "I will add one caveat, we only scan the chest. We didn't scan the brain. We didn't scan the abdomen. So sometimes the cancers surprise us and say, without having any adjacent lymph, node metastasis, we see the first metastasis in the brain."(R-P04), "I personally would probably classify the prob of these being cancer somewhere between 20 to 50 percent."(R-P1), "I just want to mention again that there is some diagnostic uncertainty at this point. This could be cancer. This could also be an infectious process or something unrelated."(R-P1)*

Theme 10: Alleviating Patient Anxiety   included content that specifically dealt with managing participants' emotional responses to the diagnoses. Examples of quotations from this theme are as follows – *"Now, I don't want to have you get so worried right away immediately, because even though they raise suspicion and possibility of lung cancer. It is true many of them also don't turn out to be cancer." (R-P04), "It's very, very common for people, patients at age 80s to have coronary artery disease. So I wouldn't necessarily be so shocked about it and so deeply concerned about it."(R-P03), "And so in general, if I were to discuss this with my dad and he asked me how concerned should I be in general at this stage you know. What should I tell him, based on these findings?"(P05), "First of all, concern, and anxiety is never really a good thing in the medical diagnostic and treatment process. But that's very understandable with just receiving the impression that there is a suspicious lesion. Now, we don't know for sure that this is cancer. But this may turn out to be lung cancer, in which case there are some further steps. But what I usually tell patients is that for now work with us in the medical system and we will go through excellent, standard steps of care."(R-P05)*

4.3. Frequency of Themes across Radiologist-Patient Interactions

Figure 4 depicts the proportion of various themes in each participant's interaction with the radiologist. Although there are variations in how much each participant's conversation incorporates each theme, there appear to be several consistencies across participants. The following are a few noteworthy observations:

- Largely, all themes seem to be represented across the radiologist interactions of all 5 participants. The exceptions are Participant 2, who is missing the *Alleviating Anxiety* theme, and Participant 5, who is missing the *Statistics of Lung Injury and Associated Damage to Other Organs theme.* This is a good indication that we have identified a comprehensive, well-rounded set of themes that will likely suffice to describe interactions with any new participants.
- The *Understanding Medical Terminology from Radiology Report* and *Selecting a Treatment Plan* themes constitute the largest portion of the interactions. The emphasis on understanding terminology suggests that individuals are the most keen to move past the complex medical terminology and want to understand what the report means in accessible language. The emphasis on selecting treatment plans likely suggests that participants in the task were able to incorporate feelings of empathy for their imagined relative's condition, and are extremely interested in understanding treatment options and performing a cost-benefit analysis.
- *Alleviating Anxiety* and *Statistics of Lung Injury and Associated Damage to Other Organs* themes constitute the smallest portions of the interactions. This is interesting because participants seem minimally interested in statistical facts about the likelihood of different aspects of the disease, and also in emotionally soothing statements by the radiologist that only serve the purpose of alleviating anxiety.
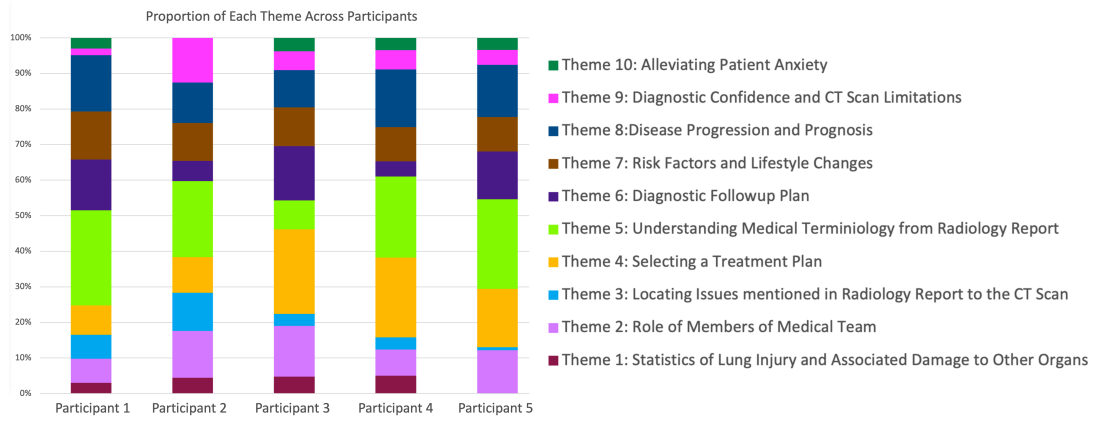
Figure 4: Stacked bar chart describing the proportion of each theme in each participant's interaction with the radiologist, with the portion of the conversation occupied by each theme on the y-axis, and participants along the x-axis.

- *Disease Progression and Prognosis, Risk Factors and Lifestyle Changes, Diagnostic Followup Plan* and *Role of Members of the Medical Team* all comprise fairly similar proportions of the interactions across participants. There seems to be a significant level of curiosity regarding what caused the disease and what the next steps are. There also seems to be considerable discussion on the roles of different members of the medical team, since participants often ask questions that are beyond the scope of the radiologist's role.

- *Locating Issues Mentioned in the Radiology Report in the CT Scan* comprises a surprisingly small proportion of all interactions. While participants are keen to understand what the report means, and what treatment options their relative will have, they seem to have little interest in understanding how the disease presents in the scans.

5. Generative AI Evaluation

Successful adoption of generative AI systems in patient-facing applications requires that they must respond reliably, factually, and meaningfully to questions asked by patients. In this section, we evaluate two state-of-art, commercially-deployed generative AI (GenAI) systems : Claude 3.5 Sonnet [19] and ChatGPT 4o [48]. Both of these systems are built upon language models that are top-scores in standardized multi-task understanding benchmarks such as MMLU-Pro [49]. MMLU-Pro is a question-answering dataset containing 12032 questions and multiple choice answers from health, psychology, and biology domains in addition to other fields. Our goal is to explore if these highly successful models can address questions humans ask naturally as they perform a sensemaking task.

5.1. Study

5.1.1. Conversational MultiModal Dataset

We, first, built a small multimodal dataset of question-answer pairs from the formative interview study and corresponding thematic analysis 4. Our dataset represents 5 themes;

1. Theme 3: locating issues mentioned in the radiology report to the CT scan,
2. Theme 4: selecting a treatment plan,
3. Theme 5: understanding medical terminology from radiology reports,
4. Theme 6: diagnostic follow-up plan, and
5. Theme 7: risk factors and lifestyle changes.

We chose these 5 themes because they allowed for the clearest formulation of meaningful question-answer pairs. The remaining themes were largely observed in parts of the radiologist's explanations to answers to questions from the selected themes. For each theme, we selected 5 questions that were

representative of the diversity in the questions study participants asked. We paired each question with the long-form answer that the radiologist provided during the interview study.

We structured the evaluation dataset as a set of 25 visual question-answering problems. Each problem consisted of (1) two images from radiology scans (in Figure 2), (2) corresponding radiology report capturing expert impressions (in Figure 2), and (3) a question asked by the patient. This setup requires the GenAI system under study to perform visual reasoning on the CT scan images and integrate that evidence with analytical understanding of the report to generate long-form answers.

5.1.2. Protocol

We accessed both CLAUDE-3.5-SONNET and CHATGPT-4o through their web interfaces. Generative AI systems are malleable and their behavior can be modulated through system instructions and various prompting techniques. Several of these prompting techniques (few-shot prompting, chain-of-thought reasoning) require the human to have expertise in interacting with generative AI systems and an understanding of their behavior. Patients are unlikely to have this expertise. Consequently, we studied two configuration that are most natural in our setting. In the first configuration (/config1), the system is provided with all background material (two scans and the report) and asked a question. This configuration reflects how a patient might use the publicly available GenAI systems. In the second configuration (/config2), we added a system prompt to guide AI's behavior. The system prompt provides the context in which the question is asked and includes who is asking the question (*who*: a patient's caregiver), what the questioner is looking for (*why*: they want to understand the scans and reports), and how the response should be generated (*how*: simple, non-clinical language, and a succinct answer). The configuration instructs GenAI systems that the questioner doesn't have a background in clinical practice, biasing them to use colloquial language and descriptions. This configuration reflects the guidance a GenAI system might receive from a healthcare system administrator as the system is deployed.

Both GenAI systems were tested under the two configurations on all 25 evaluation questions and their answers were recorded. After the system answered a question, the conversation history was reset. This was done to ensure independence in response generation. GenAI systems are autoregressive and generate answers by sampling from a token distribution based on which tokens were generated previously in the conversation. Resetting the history ensures that incorrect information in earlier response do not bias response generation for the question under study.

We complied the questions, expert answers, and responses generated by both configurations of the two GenAI systems understudy in an evaluation dataset[1].

5.2. Analysis

5.2.1. Empirical Findings

Table 1 shows the average length of expert responses as compared to those produced by the GenAI systems under both experimental conditions. Natively, both CLAUDE-3.5-SONNET and CHATGPT-4o tend to be verbose and produce significantly longer responses. However, when they are configured with the system prompt encouraging them to produce answers that are succinct, the response length reduces reliably for both.

Next, we evaluate if the GenAI system responses are similar to how a radiologist responds. Evaluating similarity of free-form text is a challenge in itself. Classical measures such as ROUGE score [50] rely on overlap in the words used in two pieces of text. However, difference in stylistic choices may result in lower scores even when the two texts have similar meaning.

To compare meaning-level similarity between expert and GenAI responses, we use the universal sentence encoder (USE) that encodes text into high dimensional vectors [51]. The model is trained and optimized for greater-than-word length text, such as sentences, phrases or short paragraphs making

---

[1]Raw dataset is available at https://docs.google.com/spreadsheets/d/e/2PACX-1vTZB2sw1JadKKc1TXF8dzF-WYL7cOFta6E61_IIdyVvR1ncDzm0nceFpA3JFPwZ5acWVjDM8hpYB9Ki/pubhtml

Table 1

A table noting the average length of responses from a radiologist and variations of GenAI models: Claude and ChatGPT

Response type	Avg. words (std. dev.)
Radiologist	146.2 (82.68)
Claude/config1	281.0 (35.85)
Claude/config2	167.88 (79.22)
ChatGPT/config1	178.36 (53.42)
ChatGPT/config2	134.12 (45.77)

it suitable for our purposes. It is trained on a variety of data sources and a variety of tasks with the aim of dynamically accommodating a wide variety of natural language understanding tasks. The input is variable length English text and the output is a 512 dimensional vector. Once the responses are converted into high-dimensional vectors, their cosine similarity ranges from 0 to 1 indicating how close they are in the semantic space, 1 when the responses are the same. In other words, the cosine similarity indicates the degree to which a GenAI system's response is aligned with an expert.

Figure 5 shows how similar the responses generated by the experimental configurations of GenAI systems are to those of a human expert. The scale goes from dark blue to dark red (dissimilar to similar). We make the following observations.

1. There is a large variability in measured similarity. This variability indicates that generative AI systems are not in full alignment with the expert radiologist. In a majority of the cases, their responses are different from that of a radiologist scoring less than 0.75 on cosine similarity.

2. Response similarity varies between the themes. Highest similarity are observed in **Theme 5: Understanding Medical Terminology in the Radiology Report**. This is not surprising given that clinical textbooks are included in the datasets LLMs are trained over. Lowest similarity responses are observed in **Theme 3: Locating Issues Mentioned in Radiology Report to the CT scan**. We investigate the reason for this observation in our qualitative analysis (4.2).

3. Configuration 2 that encourages the system to produce short text accessible to non-clinicians appears to reduce the similarity measure. Mostly, the impact is small. However in some cases (e.g, q3, q14 Claude; q21, q22 ChatGPT) it is especially pronounced. Reduction in similarity indicates that GenAI system do not know how to focus the generative process on relevant information. They can remove information that the radiologist considers important to communicate.

5.2.2. Qualitative Investigation

Similarity scores reported earlier indicate that there are differences between how a radiologist responds to patient queries and how a GenAI system approaches answer generation. However, the scores themselves don't explain why or how these answers are different. In this section, we qualitatively analyze the difference between the responses provided by the radiologist from those generated by the GenAI systems.

Drawing inspiration from the Griean maxims [52] of effective conversation, we identify the following dimensions of evaluation.

1. *Correctness* follows the Gricean maxim of quality, where one tries to be truthful, and does not give information that is false or that is not supported by evidence. We engaged an expert evaluator (a radiologist and a co-author on this paper) to judge if the responses produced by the GenAI systems were correct. We adopted a particularly strict criterion for correctness - whenever the expert identified even a single non-factual sentence in the response, it was marked to be incorrect. Such strict evaluation is necessary for critical applications of generative AI models that have been known to hallucinate information which may be dangerous in our usecase.

2. *Relevance and informativeness* follow the Gricean maxim of relation (where one tries to be relevant, and says things that are pertinent to the discussion) and quantity (where one tries
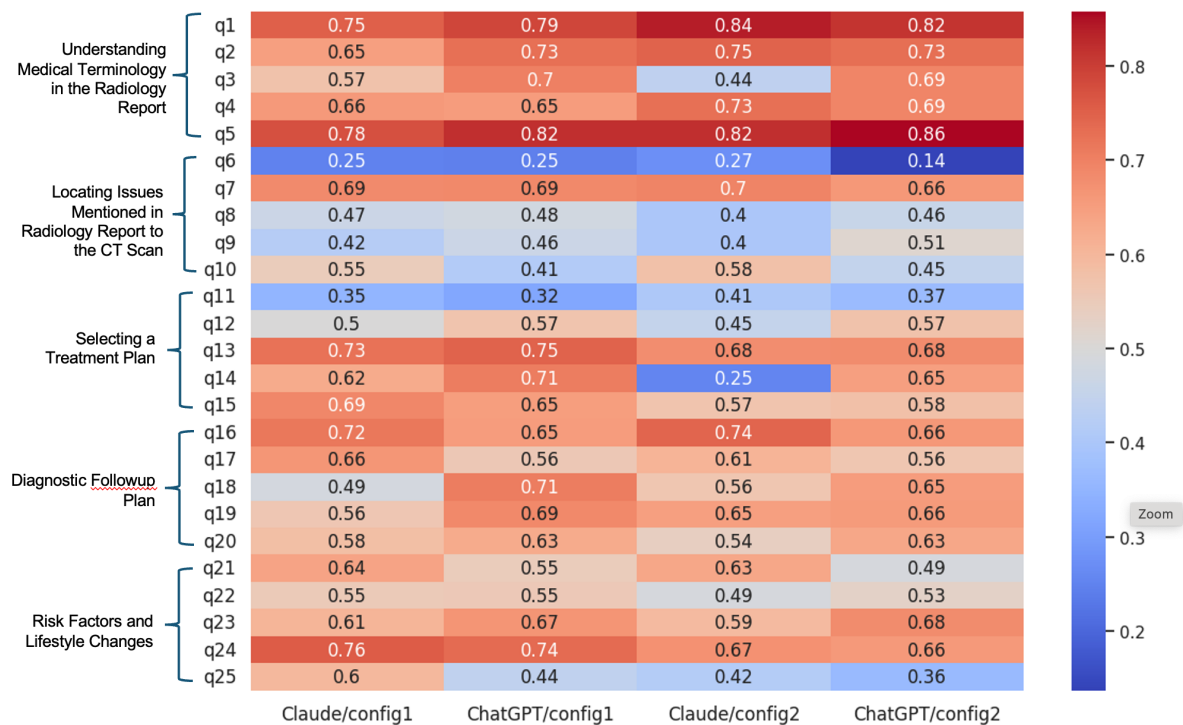
Figure 5: A heatmap plotting the semantic similarity of responses generated by Claude and ChatGPT under two experimental configurations to an expert's response.

	Question	Are these nodules, then are they present in in healthy individuals as well? Or is that specifically the cancer that's manifesting itself. Or what's the definition of that?
	Radiologist	So nodules are very widely present in most adult humans, I would say about one in 2 adults have pulmonary nodules, and many of them, I would say a vast majority of them do not turn out to be cancer. However, some of these nodules occasionally turn out to be cancer. And we, as radiologists, review various different factors, including how the nodule is looking and shaped in these ct scans. How big are the nodules. How fast is it growing over time? And where is the nodule located? These are some of the imaging features that we look-into in addition to patient age, gender, and any other clinical history of other types of cancer. These holistically go into our understanding of how suspicious are these nodules for potential lung cancer.
	Claude	To answer your question, let me clarify a few points about lung nodules: Definition: Lung nodules are small, round or oval-shaped growths in the lung. They appear as white spots on chest CT scans. Occurrence in healthy individuals: Lung nodules can occur in both healthy individuals and those with lung disease. Many lung nodules are benign (non-cancerous), especially in younger people without a history of smoking. New nodules in this case: The report mentions "A new 8 mm right upper lobe solid nodule". This is significant because: It's new (wasn't present in the 2018 comparison scan) It's solid (completely opaque on CT, which can be more concerning than non-solid nodules) It's 8 mm in size (nodules larger than 6 mm are generally considered more suspicious) Cancer risk: While not all nodules are cancerous, new solid nodules in the context of a lung cancer screening program (especially in individuals with risk factors like smoking history) are considered suspicious until proven otherwise. Classification: This finding, along with the "new 1.5 x 1.0 cm focus of groundglass nodular consolidation", has led to the Lung-RADS Category 4B classification, which indicates a high suspicion for malignancy. Next steps: The report recommends additional diagnostic testing, which may include follow-up CT, PET/CT, or tissue sampling to determine if these nodules are indeed cancerous. In summary, while lung nodules can be present in healthy individuals, new solid nodules of this size in the context of a lung cancer screening are considered suspicious and require further evaluation. They're not definitively cancer at this point, but they represent findings that need close follow-up and potentially more invasive testing to rule out malignancy.
	ChatGPT	In the context of your report, a "nodule" refers to a small, abnormal growth in the lung tissue. Nodules can appear in both healthy individuals and those with other conditions, such as infections, inflammation, or benign growths. However, nodules are often monitored closely in individuals undergoing lung cancer screening, as they can sometimes be an early sign of lung cancer. In your case, the report mentions a new 8 mm solid nodule in the right upper lobe, which is being closely evaluated. The fact that it is labeled under "Lung-RADS Category 4B: Suspicious" suggests that this particular nodule is considered concerning enough to warrant further diagnostic testing or a biopsy to rule out or confirm malignancy. The key point is that while lung nodules can occur in healthy individuals, some (like this one) may raise concern for cancer and require additional follow-up.

Figure 6: Theme 5 question (id=q1) and responses by the expert and GenAI systems. Blue denotes sentences that contain a direct response to the question; green, relevant elaborations; orange, irrelevant elaboration; and purple, superfluous text.

to be as informative as one possibly can, and gives as much information as is needed, and no more). We characterized each sentence in the answers as: *direct response*, sentences that contain a clear answer to the question asked; *relevant elaboration*, sentences that elaborate the answer by providing additional, supporting information; *irrelevant elaboration*, sentences that contain information irrelevant to the question asked; and finally, *superfluous*, sentences with little information about the question. Figure 6 shows an example of such characterization where text in blue indicates direct response, green - relevant elaboration, orange - irrelevant elaboration, and purple - superfluous text. The characterization was done by authors who are not medical experts and appropriate judges to evaluate if the answers are relevant and informative.

Finding 1   On the correctness dimension, we found that out of 25 questions, CLAUDE/CONFIG1 was judged to be incorrect on 10 (40% error rate) while ChatGPT/CONFIG1 on 5 (20% error rate) leading the expert to reject their answers. In two instances the expert accepted the responses while noting that the response didn't answer the question but was correct. Errors manifested due to a variety of causes including: inablility to relate causal factors to diseases, incorrectly identifying risks and consequences, inability to distinguish generic textbook information from specifics of the case, and factual incorrectness due to hallucinations. The expert noted that the GenAI systems ramble on and often, include information that is not directly relevant to the question.

Finding 2   Our analysis of the relevance & informativeness dimension revealed several insights about how doctors address their patient's questions. The radiologist's responses consisted of sentences that directly answer the patient question followed by a relevant elaboration. An example is shown in Figure 6 where after explaining that the majority of nodules do not turn out to be cancer, the doctors describes the criteria that are used to evaluate nodules. Other elaborations included demonstrations, impact of the disease the patient can expect to see, what further evidence may manifest, how a diagnosis may be made, and how a treatment plan may be developed. This finding hints at the mixed-initiative nature of the patient-provider dialog where the role of the provider is not just to answer questions but also educate the patient so that they feel empowered to take informed decisions about their health. We didn't observe any other information categories in the radiologist's response.

Both CLAUDE-3.5-SONNET and ChatGPT-4o produced long answers with elaborations. An example is shown in Figure 6 where the relevant elaborations are shown in green and the irrelevant ones in orange. Explanation of how nodule's size or appearance influences the diagnosis of cancer is relevant for the patient. However, next steps are not pertinent to the question asked. Additionally, both generated some superfluous information shown in purple preface or summarize its actual response but doesn't add any useful information to the response. Production of irrelevant elaborations and superfluous information directly violates of desiderata of relevance & informativeness and the Gricean maxims of relation and quantity.

Finding 3   Both GenAI systems struggled with responding to questions in theme 3: *locating issues mentioned in the radiology report to the CT scan*, explaining the low similarity scores in Figure 5. An illustrative example is shown in Figure 7. To answer the question, the radiologist responded by first describing how to find relevant areas - nodules - in the scan and then, a demonstrative elaboration highlighting those in the scan to ground the caregiver's understanding.

The demonstrative elaboration was copiously missing in both CLAUDE/config1 and ChatGPT/config1 responses. They both generated generic guidance on how to locate the nodule in the scan and regurgitated language about nodules from the scan. It is to be expected. GenAI systems are not designed to guide the questioner's attention to visual evidence contributed to answer generation. Further, CLAUDE/config1 and CLAUDE/config2 incorrectly believe that the nodule cannot be localized in the scans. Our observations call into critique the assumed 'multi-modality' of GenAI system when these systems only consume visual information to guide response generation but cannot use the visual medium to communicate information.

ChatGPT/config2 generated an answer that is correct however, is entirely useless to the patient. It present banal/obvious information about how to localize the nodule. This finding indicates that we should revise our relevance and informativeness criteria to include banality.

FINDING 4 In several responses, the radiologist went beyond the exact question being asked and answered a higher level concern. This approach was particularly visible in *Theme 6: diagnostic follow-up plan*. An example is shown in Figure 8 where the patient enquires about expected expenses for various treatment options. The radiologist's response not only compares different treatment options in terms of how expensive they are, it also provides them with other tradeoffs that the patient didn't explicitly inquire about. Examples include observations such as radiation oncology requires the patient to come back several times and chemotherapy has side-effects that might require hospitalization. In contrast,

Question	So how do I? How do I find a nodule here?
Radiologist	So usually nodules are a little bit denser than the lung surrounding lungs, and in cts they come up as white, rounded sort of a lesion. So I just circle these 2 areas for you that are nodules. That I was concerned is suspicious for lung cancer. So these 2, and keep in mind the Ct scan is kind of shift it left and right. So what I circle there in your screen left is actually your brother's right lung where we see these nodules.
Claude Config1	To help you identify nodules in these CT images, let me explain what to look for: Appearance of nodules: Nodules typically appear as round or oval-shaped white spots in the lung tissue. They are denser (whiter) than the surrounding lung tissue, which appears darker. Size: The report mentions an 8 mm solid nodule and a 1.5 x 1.0 cm ground-glass nodule. These would be relatively small white spots on the image. Location: The 8 mm nodule is in the right upper lobe. The location of the ground-glass nodule isn't specified. Looking at the provided images: Image 1: This image shows a cross-section of the upper chest. The lungs are the dark areas on either side. I don't see any obvious nodules in this particular slice. Image 2: This is a lower cross-section of the chest. Again, the lungs are the dark areas. This image shows more prominent signs of emphysema (the very dark areas within the lung tissue). I don't see any clear nodules in this slice either. Important notes: The nodules mentioned in the report might be visible on different slices of the CT scan that aren't shown here. Identifying small nodules often requires examining multiple sequential images and can be challenging without specialized training. The report mentions specific image numbers (165 and 210) where the nodules are visible, but we don't have access to those particular images. If you're interested in seeing the nodules, it would be best to ask the radiologist or your doctor to point them out on the full set of CT images. They can show you exactly where these concerning areas are located.
Claude Config2	To help you locate the nodules mentioned in the report, I'll guide you through what to look for in the CT images: Right upper lobe nodule: This is a new solid nodule, 8 mm in size. Look for a small, round, white spot in the upper right part of the lung. It should be brighter (whiter) than the surrounding lung tissue. Focus on groundglass nodular consolidation: This is larger, measuring 1.5 x 1.0 cm. Look for a hazy, grayish area that's less defined than a solid nodule. It might look like a patchy, cloudy spot in the lung. In the CT images provided: Image 1 shows a cross-section of the chest. The lungs are the dark areas on either side. Image 2 is similar but at a different level of the chest. However, these images don't seem to show the specific slices where the nodules were found. The report mentions they're in "series 3, image 165" and "series 3, image 210", which we don't have here. To find these nodules, you'd need to look at those specific images with a radiologist or your doctor. They can point out exactly where these concerning areas are located.
ChatGPT Config1	To find the lung nodules mentioned in your scan, you can follow the findings described in the report: Right upper lobe nodule: There is an 8 mm solid nodule in the right upper lobe, which corresponds to series 3, image 165. Groundglass nodule: A 1.5 x 1.0 cm groundglass nodular consolidation is also noted, associated with emphysema, at series 3, image 210. You can use these references to locate the nodules in the corresponding series and images in your scan. Do you have access to the DICOM viewer for the images? That way, you can navigate to these specific slices.
ChatGPT Config2	To find a nodule in the scan based on the report, here's a guide: Location: There is an 8 mm solid nodule in the right upper lobe of the lung. It is located in series 3, image 165. Another groundglass nodule measuring 1.5 x 1.0 cm is seen in association with emphysema in series 3, image 210. Steps: Open the scan and navigate to series 3. Look for image 165 to find the 8 mm nodule in the right upper lobe. For the second nodule, go to image 210, where you'll find a groundglass consolidation linked to emphysema. Let me know if you need help with anything specific while viewing the images

Figure 7: Theme 3 question (id=q7) and corresponding answers provided by the radiologist and GenAI systems

Question	I know there's a lot of more diagnostic studies that we need to do to make sure and understand if this is cancerous or not, but what treatment option would also say is like more expensive than others and vice versa.
Radiologist	Right. So to answer that question, there's one component we can tell you how much it is in general for patients that perhaps surgery in itself may be arguably the most expensive at one time point. But if you're doing a radiation oncology treatment in the grand scheme of the treatment, arguably you may have to come back for it several times or there may be some other additional things that may have to be done. And one option that I didn't mention is chemotherapy that also comes with some side effects that require possibility, for example hospitalization or some infection down the road. So a really hard for me to tell you exactly what the cost will be, but that's a very fair question to ask to the medical oncologist and the billing department. The other sort of end of the equation is that different insurances have different coverage level. And if your grandmother has a great insurance that covers all of these, it may be something close to $0.00 for a lot of these treatments. Whereas some other insurance may resist covering certain treatments, it's something that we really wanna be as transparent as possible just within the confines of the healthcare system that we work with
Claude Config1	You're right that more diagnostic studies are needed to confirm whether the findings are cancerous and to determine the best course of action. Regarding the costs of different treatment options, here's a general overview: Most expensive options (typically): Surgery: Especially if it involves extended hospital stays or complicated procedures. Newer targeted therapies and immunotherapies: These can be very costly due to their specialized nature. Moderately expensive: Radiation therapy: Costs can add up due to the need for multiple sessions, but it's often less expensive than surgery. Traditional chemotherapy: While still costly, it's often less expensive than newer targeted therapies. Less expensive (relatively): Some oral medications: Though some can be quite expensive, others might be more affordable. Palliative care: Focused on symptom management rather than curative treatment. Factors affecting cost: Duration of treatment Frequency of hospital visits Need for supportive medications Potential complications requiring additional care Geographic location and healthcare facility It's important to note: Costs can vary widely depending on individual circumstances, insurance coverage, and location. The most expensive option isn't always the most effective; treatment decisions should be based on medical factors first. Many hospitals and treatment centers offer financial counseling and assistance programs. Insurance coverage can significantly impact out-of-pocket costs. Remember, the best treatment isn't always the most or least expensive. It's the one that offers the best chance of success with manageable side effects, tailored to your specific situation. Always discuss both medical and financial concerns with your healthcare team and insurance provider.
ChatGPT Config1	Treatment options for lung cancer vary in cost depending on the type and stage of cancer, as well as the approach taken. Here are some general comparisons between expensive and more affordable treatments: More Expensive Treatment Options: Targeted Therapy: This involves drugs that specifically target cancerous cells' mutations, requiring genetic testing of the tumor. It's generally quite expensive but can be more effective for certain types of lung cancer. Immunotherapy: Drugs like pembrolizumab (Keytruda) can help the body's immune system fight the cancer but are costly and often used for advanced stages of lung cancer. Radiation Therapy: Although standard, newer technologies like stereotactic body radiotherapy (SBRT) can be expensive due to their precision in targeting tumors while sparing healthy tissue. Surgery: Advanced surgeries like video-assisted thoracic surgery (VATS) or robotic surgery, which are less invasive but more expensive, may be considered depending on the tumor's location and size. More Affordable Treatment Options: Chemotherapy: This is a standard and often less expensive option, though the cost can vary depending on the drugs used and the number of cycles required. Palliative Care: For late-stage lung cancer, focusing on symptom management (rather than curing the disease) can be a more affordable option. Given that your report suggests a Lung-RADS Category 4B (high suspicion of malignancy), the most appropriate next steps, such as a PET/CT scan or tissue biopsy, will help determine the exact type of treatment, potentially shifting the cost range depending on what is found.

Figure 8: Theme 6 question (id=q15) and corresponding answers provided by the radiologist and GenAI systems

both CLAUDE-3.5-SONNET and CHATGPT-4O focused on costs associated with treatment options, often elaborating to include less common or new therapies. In general we found the radiologist's responses were geared towards helping the patients make decisions whereas GenAI systems' responses were geared towards extensive enumeration of possibilities and definitions.

Finding 5 A meta-finding of our research is that the USE-based similarity metric isn't sufficiently selective for our tasks. Consider the example in Figure 7 where it is easy to see that the radiologist's responses in grounded in the CT scans and discusses localization of nodules. GenAI systems on the other hand produce responses containing different information. However, their similarity scores (Figure 5, line q7) are almost 0.7 indicating a good degree match. This observation indicates that USE supports a topic-level similarity but cannot compare the content as a human would, despite being one of the

standard measurement techniques in AI.

6. Discussion, Conclusions, and Future Work

In this paper, we study if current state-of-art in generative AI systems can support patients (and caregivers) as they attempt to understand their scans and reports. To structure our evaluation, we conducted a need-finding study of caregiver-radiologist interactions. We identified 10 themes of diverse information types that such interactions are typically comprised of. We used 4 themes to create an evaluation question-answer dataset and evaluated two state-of-art models: Claude-3.5-Sonnet and ChaptGPT-4o both quantitatively, by measuring response similarity with a human expert and qualitatively, by judging the responses on Gricean maxims of effective communication.

Our findings indicate that despite significant optimism about generative models in medical contexts, the current state-of-art doesn't sufficiently address patients' information needs. Surprisingly, we found that these models have high error rates (20% for ChatGPT-4o, 40% in Claude-3.5-Sonnet) in our evaluation set built on question asked in real interactions. This finding is concerning - patients are not experts of medicine and consequently, cannot identify misinformation in AI-generated responses. Any system that is deployed must guaranteed to provide factual information. The models were also deficient as conversational information systems and frequently produced irrelevant and superfluous elaborations violating Gricean maxims. This finding suggests that it is unlikely that patients will be able to use this system to access useful information.

Our findings also highlight a critical gap in the evaluation paradigm prominent in AI & ML and the challenges of accountable deployment of generative AI systems. In machine learning literature [53], these models are shown to have high accuracy on curated, task-agnostic datasets and claimed to have human-level performance. However, to be at par with human experts, the models not only have to answer questions but also operationalize this knowledge in practical, real-world tasks such as helping patients understand their illness and determine a treatment plan - a task doctors and healthcare worker routinely perform. Our work paves the way for task-specific evaluation protocols that can spur the generative AI movement in a beneficial direction.

There are several avenues for future work. The study described here is small (only 5 participants) and limited to a single case of radiological imaging. Consequently, the themes identified here do not represent an exhaustive set of informational needs patients have. In doing this arguably limited study, we formulated a concrete approach of evaluating GenAI systems. Building upon our learnings, we will expand the study to include a larger, heterogeneous population as well as cover a greater variation of medical contexts. A larger study will enable us to create a comprehensive evaluation dataset for GenAI systems, ensuring a rigorous task-centric evaluation of GenAI systems.

The second avenue is studying a wider range of GenAI system configurations. Generative AI systems are considerably malleable and their behavior can be modulated through various prompting strategies, supervised fine tuning, and reinforcement learning etc. This paper focuses on sketching out an evaluation paradigm by studying two simple configurations. In future, we will build and study methods that align GenAI system response to experts.

Third, prominent metrics such like BLEU [54], ROGUE [50], BERT Score [55], or USE [51] while commonly used to evaluate LLM responses, fall short of measuring if the response produced are human-like or if they adequately address the humans' information needs. To address this shortcoming in AI & ML evaluation paradigm, we will develop a quantitative measurement scale based on our the qualitative evaluation paradigm proposed here. Finally, we are invested in developing intelligent assistive systems for the healthcare. Our future work will explore how the shortcoming highlighted in this paper can be alleviated with modern and classical AI methods to develop useful conversational agents.

Despite its limitations, our makes several important contributions towards sustainable and productive deployment of generative AI systems. First, the performance of generative AI systems is typically measured on domain-agnostic, task-agnostic question-answering datasets. While these methods are useful in comparing algorithms against each other, they provide little insight about how useful the

answers are for humans. Our research situates question answering in a specific context of people seeking information to enhance understanding in healthcare settings. This enables us measure the degree to which GenAI systems are useful. Second, GenAI systems are typically evaluated on datasets curated by and for clinicians. These datasets contain questions and answers that are most interesting to clinicians while performance measurement includes criterion such as diagnostic accuracy. Such an approach introduces an implicit bias in the design and analysis of these systems - they are designed to replace clinicians by mimicking the output of their problem-solving process, following the *AI-is-automation* philosophy. However, to design a good information support tool, paying attention to who the seeker and the consumer of that information is critical. Patients and caregivers have vastly different needs than clinicians who are experts of terminology, diagnostics, and treatment. Our research takes the *AI-is-augmentation* view and centers patients and caregivers, adopting a much needed lens missing from the medical generative AI research landscape. Finally, this paper lays the groundwork for a principled, within-context evaluation approach for generative AI systems. We study the space of information needs of a patient/caregiver to understand what types of questions are asked by the patient/caregiver. We introduce a novel qualitative answer evaluation paradigm that measures both correctness and relevance of answers generated by AI systems. Our research can be considered a precursor to a rigorous, quantitative approach for measuring suitability of a generative AI system for addressing humans' informational needs.

References

[1] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, et al., "So What if ChatGPT Wrote it?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy, International Journal of Information Management 71 (2023) 102642.

[2] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, P. Szolovits, What disease does this patient have? A Large-scale Open Domain Question Answering Dataset from Medical Exams, Applied Sciences 11 (2021) 6421.

[3] M. Cascella, J. Montomoli, V. Bellini, E. Bignami, Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios, Journal of medical systems 47 (2023) 33.

[4] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen, et al., The future landscape of large language models in medicine, Communications medicine 3 (2023) 141.

[5] R. J. Adams, Improving Health Outcomes with Better Patient Understanding and Education, Risk management and healthcare policy (2010) 61–72.

[6] N. D. Berkman, S. L. Sheridan, K. E. Donahue, D. J. Halpern, A. Viera, K. Crotty, A. Holland, M. Brasure, K. N. Lohr, E. Harden, E. Tant, I. Wallace, M. Viswanathan, Health Literacy Interventions and Outcomes: An Updated Systematic Review, Evidence report/technology assessment (2011) 1—941.

[7] D. Howell, T. Harth, J. Brown, C. Bennett, S. Boyko, Self-management Education Interventions for Patients with Cancer: A Systematic Review, Supportive Care in Cancer 25 (2017) 1323–1355.

[8] K. Christiansen, L. Buswell, T. Fadelu, A Systematic Review of Patient Education Strategies for Oncology Patients in Low-and Middle-Income Countries, The Oncologist 28 (2023) 2–11.

[9] T. Gilligan, N. Coyle, R. M. Frankel, D. L. Berry, K. Bohlke, R. M. Epstein, E. Finlay, V. A. Jackson, C. S. Lathan, C. L. Loprinzi, et al., Patient-clinician communication: American Society of Clinical Oncology Consensus Guideline, Obstetrical & Gynecological Survey 73 (2018) 96–97.

[10] L. De Groot, I. Harris, G. Regehr, A. Tekian, P.-A. Ingledew, Quality of Online Resources for Pancreatic Cancer Patients, Journal of Cancer Education 34 (2019) 223–228.

[11] M. Alarifi, T. Patrick, A. Jabour, M. Wu, J. Luo, Understanding Patient Needs and Gaps in Radiology Reports Through Online Discussion Forum Analysis, Insights into Imaging 12 (2021) 1–9.

[12] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, L. Wang, The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), arXiv preprint arXiv:2309.17421 9 (2023).

[13] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling Language Modeling with Pathways, Journal of Machine Learning Research 24 (2023) 1–113.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language Models are Unsupervised Multitask Learners, OpenAI blog 1 (2019) 9.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models are Few-shot Learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[16] OpenAI, :, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, et al., Gpt-4 Technical Report, 2023. `arXiv:2303.08774`.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and Efficient Foundation Language Models, arXiv preprint arXiv:2302.13971 (2023).

[18] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open Foundation and Fine-tuned Chat Models, arXiv preprint arXiv:2307.09288 (2023).

[19] Anthropic, Claude 3.5 sonnet, 2024. URL: https://www.anthropic.com/news/claude-3-5-sonnet.

[20] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A Survey of Large Language Models, arXiv preprint arXiv:2303.18223 (2023).

[21] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, X. Hu, Harnessing the Power of LLMs in Practice: A Survey on Chatgpt and Beyond, arXiv preprint arXiv:2304.13712 (2023).

[22] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large Language Models Encode Clinical Knowledge, Nature 620 (2023) 172–180.

[23] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, et al., Towards Expert-level Medical Question Answering with Large Language Models, arXiv preprint arXiv:2305.09617 (2023).

[24] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, K. K. Bressem, MedAlpaca: An Open-Source Collection of Medical Conversational AI Models and Training Data, arXiv preprint arXiv:2304.08247 (2023).

[25] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, B. Wang, Clinical Camel: An Open-Source Expert-level Medical Language Model with Dialogue-based Knowledge Encoding, arXiv preprint arXiv:2305.12031 (2023).

[26] L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, Z. You, Chatdoctor: A Medical Chat Model Fine-tuned on llama Model using Medical Domain Knowledge, arXiv preprint arXiv:2303.14070 (2023).

[27] F. Liu, T. Zhu, X. Wu, B. Yang, C. You, C. Wang, L. Lu, Z. Liu, Y. Zheng, X. Sun, et al., A Medical Multimodal Large Language Model for Future Pandemics, NPJ Digital Medicine 6 (2023) 226.

[28] A. Arora, A. Arora, The Promise of Large Language Models in Healthcare, The Lancet 401 (2023) 641.

[29] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large Language Models in Medicine, Nature medicine 29 (2023) 1930–1940.

[30] S. B. Patel, K. Lam, ChatGPT: The Future of Discharge Summaries?, The Lancet Digital Health 5 (2023) e107–e108.

[31] C. Wu, X. Zhang, Y. Zhang, Y. Wang, W. Xie, Pmc-llama: Further Finetuning llama on Medical Papers, arXiv preprint arXiv:2304.14454 (2023).

[32] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, T. Liu, Huatuo: Tuning llama Model with Chinese Medical Knowledge, arXiv preprint arXiv:2304.06975 (2023).

[33] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual Instruction Tuning, arXiv preprint arXiv:2304.08485 (2023).

[34] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, et al., Openflamingo: An Open-source Framework for Training Large Autoregressive Vision-language Models, arXiv preprint arXiv:2308.01390 (2023).

[35] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping Language-image Pre-training with Frozen

Image Encoders and Large Language Models, arXiv preprint arXiv:2301.12597 (2023).

[36] Q. Liu, S. Hyland, S. Bannur, K. Bouzid, D. C. Castro, M. T. Wetscherek, R. Tinn, H. Sharma, F. Pérez-García, A. Schwaighofer, et al., Exploring the Boundaries of GPT-4 in Radiology, arXiv preprint arXiv:2310.14573 (2023).

[37] S. Bannur, S. Hyland, Q. Liu, F. Pérez-García, M. Ilse, D. C. de Castro, B. Boecking, H. Sharma, K. Bouzid, A. Schwaighofer, et al., MSCXR-T: Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing, 2023.

[38] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, M. P. Lungren, CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling using BERT, arXiv preprint arXiv:2004.09167 (2020).

[39] R. Xie, S. Palayew, A. Toma, G. Bader, B. Wang, WangLab at MEDIQA-M3G 2024: Multimodal medical answer generation using large language models, in: T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, D. Bitterman (Eds.), Proceedings of the 6th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 624–634. URL: https://aclanthology.org/2024.clinicalnlp-1.60. doi:10.18653/v1/2024.clinicalnlp-1.60.

[40] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, E. Cambria, A Survey of Large Language Models for Healthcare: From Data, Technology, and Applications to Accountability and Ethics, arXiv preprint arXiv:2310.05694 (2023).

[41] Q. Xie, E. J. Schenck, H. S. Yang, Y. Chen, Y. Peng, F. Wang, Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond, medRxiv (2023).

[42] A. Pal, L. K. Umapathi, M. Sankarasubbu, Medmcqa: A Large-scale Multi-subject Multi-choice Dataset for Medical Domain Question Answering, in: Conference on Health, Inference, and Learning, PMLR, 2022, pp. 248–260.

[43] S. Lin, J. Hilton, O. Evans, Truthfulqa: Measuring How Models Mimic Human Falsehoods, arXiv preprint arXiv:2109.07958 (2021).

[44] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, J.-R. Wen, Halueval: A Large-scale Hallucination Evaluation Benchmark for Large Language Models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 6449–6464.

[45] G. M. Glazer, J. A. Ruiz-Wibbelsmann, The Invisible Radiologist, Radiology 258 (2011) 18–22.

[46] J. L. Kemp, M. C. Mahoney, V. P. Mathews, M. Wintermark, J. Yee, S. D. Brown, Patient-Centered Radiology: Where are We, Where do We Want to Be, and How do We Get There?, Radiology 285 (2017) 601–608.

[47] L. S. Nowell, J. M. Norris, D. E. White, N. J. Moules, Thematic Analysis: Striving to Meet the Trustworthiness Criteria, International journal of qualitative methods 16 (2017) 1609406917733847.

[48] OpenAI, Hello gpt-4o, 2024. URL: https://openai.com/index/hello-gpt-4o/.

[49] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al., Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, arXiv preprint arXiv:2406.01574 (2024).

[50] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.

[51] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder for english, in: Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations, 2018, pp. 169–174.

[52] H. Grice, Logic and conversation, Syntax and semantics 3 (1975).

[53] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., Sparks of artificial general intelligence: Early experiments with gpt-4, arXiv preprint arXiv:2303.12712 (2023).

[54] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[55] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation

with BERT, in: International Conference on Learning Representations, 2020.