arXiv:2402.00234v1 [cs.HC] 31 Jan 2024

Are Generative AI systems Capable of Supporting Information Needs of Patients?

SHREYA RAJAGOPAL, University of Michigan, USA
SUBHASHIS HAZARIKA, SRI International, USA
SOOYUNG KIM, SRI International, USA
YAN-MING CHIOU, SRI International, USA
JAE HO SOHN, University of California, San Francisco, USA
HARI SUBRAMONYAM, Stanford University, USA
SHIWALI MOHAN, SRI International, USA

Patients managing a complex illness such as cancer face a complex information challenge where they not only must learn about their illness but also how to manage it. Close interaction with healthcare experts (radiologists, oncologists) can improve patient learning and thereby, their disease outcome. However, this approach is resource intensive and takes expert time away from other critical tasks. Given the recent advancements in Generative AI models aimed at improving the healthcare system, our work investigates whether and how generative visual question answering systems can responsibly support patient information needs in the context of radiology imaging data. 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 visual language models 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.

CCS Concepts: • **Human-centered computing** → **User studies**; • **Computing methodologies** → *Knowledge representation and reasoning*; Natural language processing.

Additional Key Words and Phrases: Generative AI, Healthcare, Medicine, LLM, Visual Question Answering

ACM Reference Format:
Shreya Rajagopal, Subhashis Hazarika, Sooyung Kim, Yan-Ming Chiou, Jae Ho Sohn, Hari Subramonyam, and Shiwali Mohan. 2024. Are Generative AI systems Capable of Supporting Information Needs of Patients?. 1, 1 (February 2024), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Patients who are better educated about their disease have better outcomes, especially in serious illnesses such as cancer [1, 7, 11, 18]. However, they and their caregivers 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

---

Authors' addresses: Shreya Rajagopal, shreyara@umich.edu, University of Michigan, 500 S State St, Ann Arbor, Michigan, USA, 48109; Subhashis Hazarika, SRI International, 3333 Coyote Hill Rd, Palo Alto, California, USA, 94304, subhashis.hazarika@sri.com; Sooyung Kim, SRI International, 3333 Coyote Hill Rd, Palo Alto, California, USA, 94304, soo.kim@sri.com; Yan-Ming Chiou, SRI International, 3333 Coyote Hill Rd, Palo Alto, California, USA, 94304, yan-ming.chiou@sri.com; Jae Ho Sohn, University of California, San Francisco, USA, JaeHo.Sohn@ucsf.edu; Hari Subramonyam, Stanford University, USA, harihars@stanford.edu; Shiwali Mohan, SRI International, 3333 Coyote Hill Rd, Palo Alto, California, USA, 94304, shiwali.mohan@sri.com.

to anxiety and stress from 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. A crucial resource is meaningful and empathetic interactions with healthcare providers supporting the patient's journey through the healthcare system [14]. 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 [12] and forums [2].

Recently, conversational generative AI systems built upon large language models and their multi-modal counterparts have shown great promise as tools that enable people to access information through natural conversations [13]. Generative models for medicine such as MedFlamingo [29], GPT-4v [49], LLava-Med [21] have shown impressive performance in visual question answering tasks comprising of medical scans including Computed Tomograhpy Scan (CT scan) and accompanying medical reports [46]. Purportedly, these models have medical common sense at par with doctors as evidenced by their high scores in standard medical knowledge evaluation paradigms like the United States Medical Licensing Examination (USMLE) [19]. However, unlike usecases involving factual knowledge for experts, integrating generative models in patient-facing tools requires that the models also consider teaching and explanation of 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. Our goal is assess whether generative AI systems can support patients in their informational needs as they learn more about their illness and reason about next steps.

In this work, we study how caregivers interact with radiologists to gain a better understanding of their scans and reports. This use case is relevant for multi-modal generative AI systems that can combine the visual channel of scan images and the linguistic channel of reports. 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 multi-modal generative AI systems - MedFlamingo [29] and ChatGPT-4V [49] - against the radiologist's responses in the realistic interview task across four high-level themes of conversation identified. 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 in healthcare settings. This 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. The 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, adopting a much-needed lens missing from the medical generative AI research landscape. We introduce a novel qualitative answer evaluation paradigm that measures both factual correctness and relevance of answers generated by AI systems. Our research lays the groundwork for a rigorous, quantitative approach to measuring the suitability of a generative AI system for addressing patients' informational needs. Our key contributions include: (1)

a set of conversational themes central to patient-facing LLM systems and (2) an evaluation of quality of responses from two state-of-the-art medical generative models.

2 Background

Recently, large language models (LLMs) like PaLM [10], GPTs [8, 31, 35], LLaMA [42, 43] have significantly advanced the state-of-the-art in various natural language processing (NLP) tasks [48, 53]. This has inspired the development of many medical LLMs [16, 26, 36, 37, 41, 50] to assist the healthcare professionals and improve patient care [4, 34, 39]. Based on publicly available LLMs, specialized medical versions of the models like ChatDoctor [50], MedAlpacha [16], PMC-LLaMA [45], BenTsao [44], and Clinical Camel [41] have come up in the past year. Models like Med-PaLM [36] and Med-PaLM-2 [37] built on top of PaLM have achieved close to human expert scores in the United States Medical Licensing Examination (USMLE) [19]. Apart for only text-based models several multimodal models [5, 23, 27, 49] 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 utilize MedFlamingo [29] and GPT-4V [49] as our generative AI models to evaluate human expert responses in patient-radiologist interactions.

MedFlamingo [29] leverages the Flamingo architecture [3] as its foundation. It is pre-trained on paired medical image-text data from medical publications and textbooks. This multimodal few-shot learner has been demonstrated to have remarkable performance in medical visual question answering (VQA) tasks, surpassing existing models by up to 20% in clinician evaluation score as reported in the original work. [29] Unlike previous efforts in multimodal medical foundation models, such as ChexZero[40] and BiomedCLIP[51], Med-Flamingo is designed specifically for in-context learning in the medical domain and can learn tasks from few examples during prompting. More recently, GPT-4 model with vision capabilities (GPT-4V) [49] 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[28]. The research, which referenced various sources, thoroughly examined GPT-4V across a broad spectrum of standard radiology text-based tasks, including MS-CXR-T[6] and CheXbert[38]. The findings revealed that GPT-4 either surpasses or matches the performance of specialized fine-tuned radiology models. A detailed analysis in a Microsoft explorations paper[49] on GPT-4V specifically investigated 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.

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 [17, 47, 53] 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 [19] and MedMCQA [32] 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 [47]. It is imperative to prepare more domain and task specific evaluation protocols to adjudicate the properties of GenAI models. Singhal et al. [36] created a benchmark comprising the most commonly searched health queries to evaluate LLM responses. TruthfulQA [25] and HaluEval [22] 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 and evaluate strengths and weakness of generative AI system responses against human experts in a real-world patient-doctor interaction.

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 [15, 20].

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 plenty of 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 effective multi-modal 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

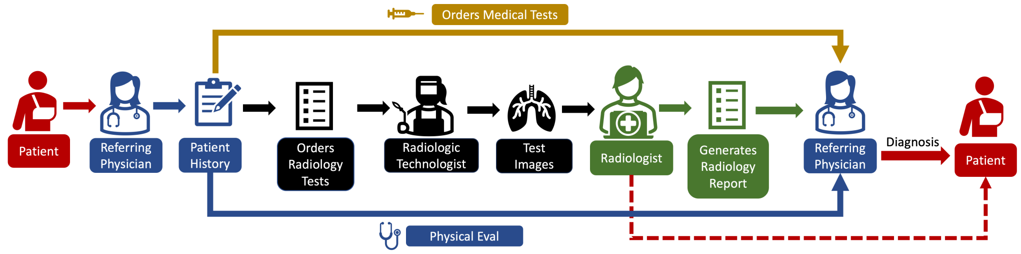


Fig. 1. Patient journey through the medical system highlight various interactions they have with different healthcare experts.

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

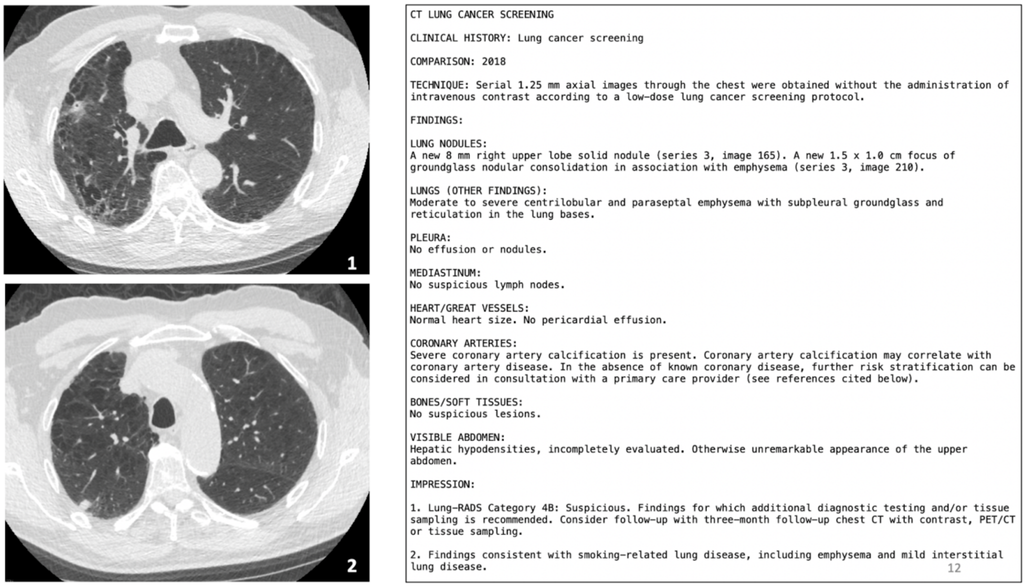
We designed a formative study to closely resemble the medical context patients and caregivers find themselves in as they navigate the healthcare system (Section 3) - the timepoint when they have access to their scans and reports and are preparing to meet with the PCP. We formulated the study as an interview 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

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:

(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.



CT LUNG CANCER SCREENING

CLINICAL HISTORY: Lung cancer screening

COMPARISON: 2018

TECHNIQUE: Serial 1.25 mm axial images through the chest were obtained without the administration of intravenous contrast according to a low-dose lung cancer screening protocol.

FINDINGS:

LUNG NODULES:
A new 8 mm right upper lobe solid nodule (series 3, image 165). A new 1.5 x 1.0 cm focus of groundglass nodular consolidation in association with emphysema (series 3, image 210).

LUNGS (OTHER FINDINGS):
Moderate to severe centrilobular and paraseptal emphysema with subpleural groundglass and reticulation in the lung bases.

PLEURA:
No effusion or nodules.

MEDIASTINUM:
No suspicious lymph nodes.

HEART/GREAT VESSELS:
Normal heart size. No pericardial effusion.

CORONARY ARTERIES:
Severe coronary artery calcification is present. Coronary artery calcification may correlate with coronary artery disease. In the absence of known coronary disease, further risk stratification can be considered in consultation with a primary care provider (see references cited below).

BONES/SOFT TISSUES:
No suspicious lesions.

VISIBLE ABDOMEN:
Hepatic hypodensities, incompletely evaluated. Otherwise unremarkable appearance of the upper abdomen.

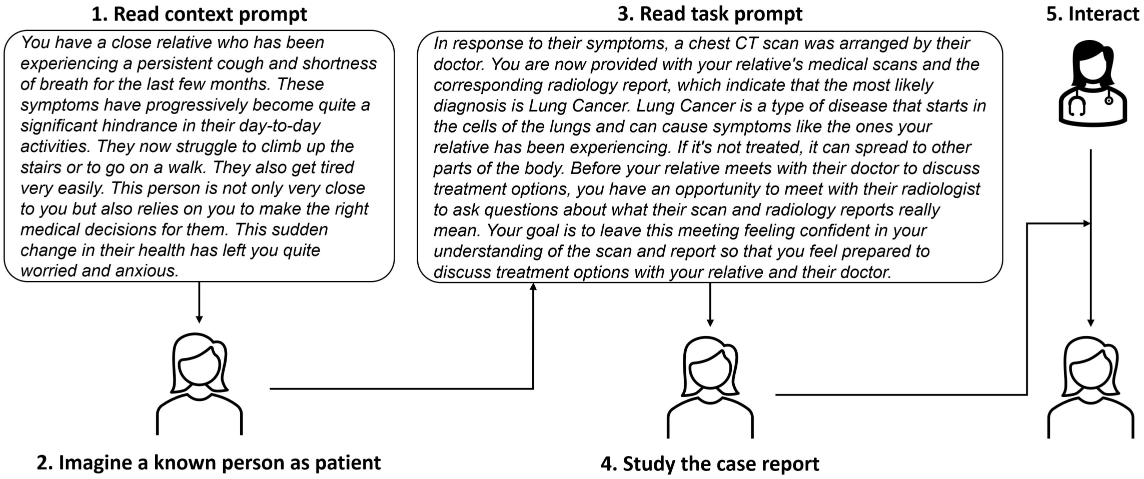
IMPRESSION:

1. Lung-RADS Category 4B: Suspicious. Findings for which additional diagnostic testing and/or tissue sampling is recommended. Consider follow-up with three-month follow-up chest CT with contrast, PET/CT or tissue sampling.

2. Findings consistent with smoking-related lung disease, including emphysema and mild interstitial lung disease.

12

Fig. 2. 2 axial slices of the lung CT and the radiology report presented to study participants



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.

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.

Fig. 3. Protocol showing the context and task prompt as well as participant actions and interactions

(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 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 [30]. 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 analysing 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 3 describes the proportion of interview content each theme occupied 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 from the graph:

- 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

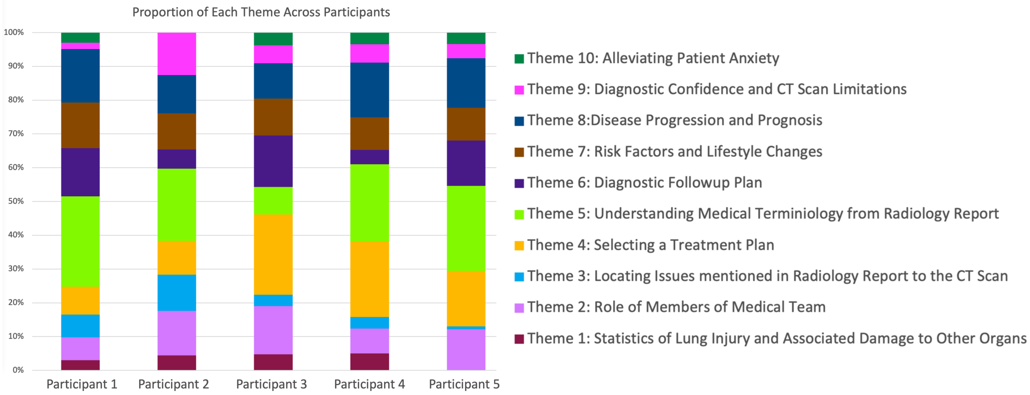


Fig. 4. Stacked bar chart describing the proportion of each theme in each participant's interaction with the radiologist, with the portion of the conversation occupied by each theme on the y-axis, and participants along the x-axis.

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.

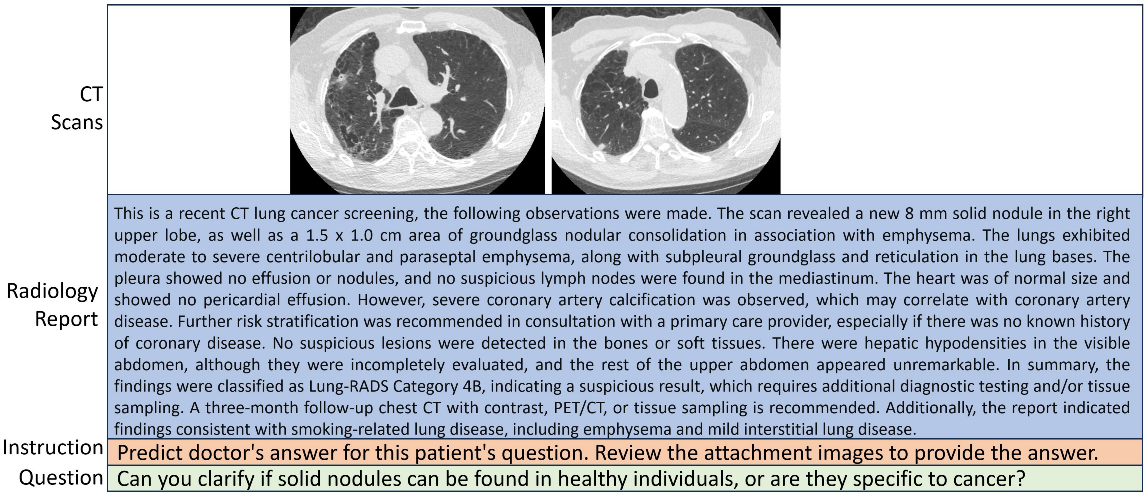
- *Alleviating Anxiety* and *Statistics of Lung Injury and Associated Damage to Other Organs* themes constitute the smallest portions of the interactions. This is interesting because participants seem minimally interested in statistical facts about the likelihood of different aspects of the disease, and also in emotionally soothing statements by the radiologist that only serve the purpose of alleviating anxiety.
- *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

To effectively use Generative AI systems to support patients' understanding of their medical reports, they must be able to respond reliably, factually, and meaningfully to diverse questions asked by patients. Here, we evaluate if the current generative AI models can meet this desiderata. Our evaluation focuses on two state-of-art foundational models for the medical visual questions answering task: MedFlamingo [29] and ChatGPT-4V [49], elaborated in Section 2.

5.1 Method

Our evaluation paradigm is based on the materials in our study (Section 4) and our inductive thematic analysis of the patient-provider conversation. From the 10 themes identified in Section 4.2, we selected 4 themes to use in our evaluation of generative models above: theme 4, selecting a treatment plan; theme 3: locating issues mentioned in the radiology report to the CT scan; theme 5, understanding medical terminology from radiology reports, and theme 6, diagnostic follow-up plan. We chose these 4 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



This is a recent CT lung cancer screening, the following observations were made. The scan revealed a new 8 mm solid nodule in the right upper lobe, as well as a 1.5 x 1.0 cm area of groundglass nodular consolidation in association with emphysema. The lungs exhibited moderate to severe centrilobular and paraseptal emphysema, along with subpleural groundglass and reticulation in the lung bases. The pleura showed no effusion or nodules, and no suspicious lymph nodes were found in the mediastinum. The heart was of normal size and showed no pericardial effusion. However, severe coronary artery calcification was observed, which may correlate with coronary artery disease. Further risk stratification was recommended in consultation with a primary care provider, especially if there was no known history of coronary disease. No suspicious lesions were detected in the bones or soft tissues. There were hepatic hypodensities in the visible abdomen, although they were incompletely evaluated, and the rest of the upper abdomen appeared unremarkable. In summary, the findings were classified as Lung-RADS Category 4B, indicating a suspicious result, which requires additional diagnostic testing and/or tissue sampling. A three-month follow-up chest CT with contrast, PET/CT, or tissue sampling is recommended. Additionally, the report indicated findings consistent with smoking-related lung disease, including emphysema and mild interstitial lung disease.

Predict doctor's answer for this patient's question. Review the attachment images to provide the answer.

Can you clarify if solid nodules can be found in healthy individuals, or are they specific to cancer?

Fig. 5. An example of a medical visual question answering problem in our evaluation

to questions from the selected 4 themes. From each of these themes, we selected 5 questions that were asked by the participants representative of the diversity in each theme, maximizing variety both in terms of the content of the question and the participant asking the question. These questions and the corresponding answers given by the radiologist formed the evaluation set for the generative AI models.

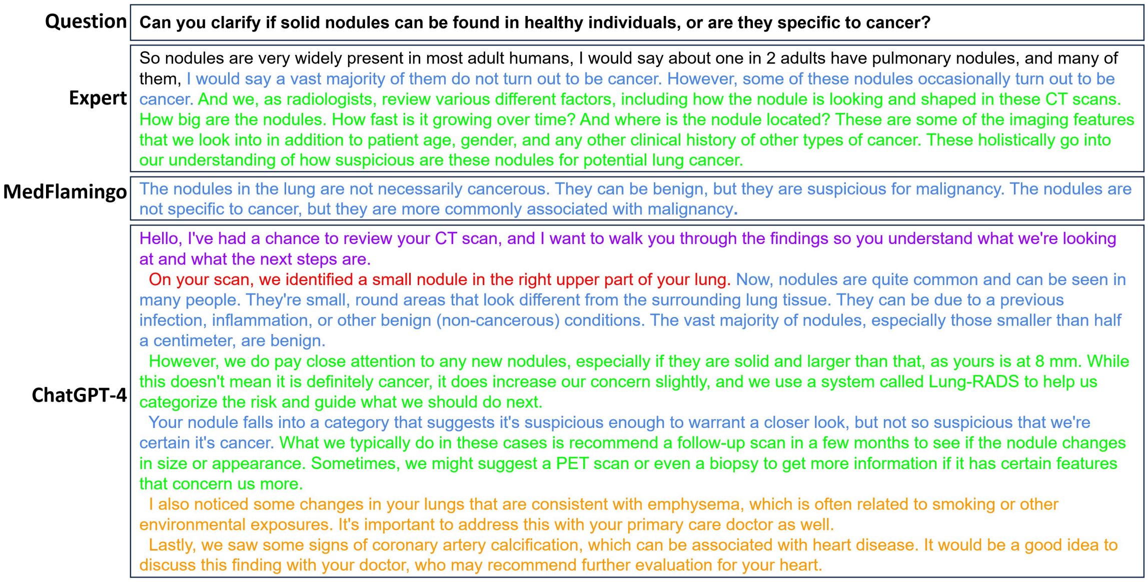
We structured the evaluation as a visual questions-answering problem shown in Figure 5. A problem consisted of (1) two images from radiology scans, (2) corresponding radiology report, and (3) a question. A major concern with the models we evaluated is the character limit on the length of the prompt - if the prompt becomes too lengthy, it is truncated internally in the model. To fit the medical report in the given character limit, we manually summarized it and validated it for accuracy and correctness. Another challenge is how patients phrase their questions in a natural conversation. Often, the questions are not grammatical, have disfluencies, and may be convoluted. Generative AI models we evaluated are not designed to handle such natural conversational patters. Therefore, we paraphrased the question such that they are grammatical and clearly stated. Our final evaluation set consisted of 20 question-answer problems. Each model was prompted with a problem and the answer it generated was recorded.

It is noteworthy that generative AI models are malleable and their behavior can be modulated through in-context learning and few-shot prompting such that it follows a desired style, format, or reasoning approach. Our evaluation protocol doesn't leverage this capability, but evaluates the nascent behavior of these models. However, it has the advantage of being simple and straightforward leading to a clearer analysis approach. In future, we will study various in-context and few-shot learning prompts and their impact on the model responses in a principled fashion.

5.2 Analysis

We analyzed the answers generated by MedFlamingo and ChatGPT-4V with respect to those provided by the radiologists during our interview study on two dimensions, qualitatively. Figure 6 shows an example question from our evaluation set and corresponding answer provided by the expert as well as those generated by MedFlamingo and ChaptGPT-4V.

Summary Statistics In our evaluation set, the radiologist's responses had $M = 136.95$ ($SD = 75.53$) words. MedFlamingo's responses tended to be significantly more concise with $M = 52.05$ ($SD = 22.28$) words. ChatGPT4 was the most verbose of all responses with $M = 301.42$ ($SD = 78.48$) words on average.



Question	**Can you clarify if solid nodules can be found in healthy individuals, or are they specific to cancer?**
Expert	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.
MedFlamingo	The nodules in the lung are not necessarily cancerous. They can be benign, but they are suspicious for malignancy. The nodules are not specific to cancer, but they are more commonly associated with malignancy.
ChatGPT-4	Hello, I've had a chance to review your CT scan, and I want to walk you through the findings so you understand what we're looking at and what the next steps are. On your scan, we identified a small nodule in the right upper part of your lung. Now, nodules are quite common and can be seen in many people. They're small, round areas that look different from the surrounding lung tissue. They can be due to a previous infection, inflammation, or other benign (non-cancerous) conditions. The vast majority of nodules, especially those smaller than half a centimeter, are benign. However, we do pay close attention to any new nodules, especially if they are solid and larger than that, as yours is at 8 mm. While this doesn't mean it is definitely cancer, it does increase our concern slightly, and we use a system called Lung-RADS to help us categorize the risk and guide what we should do next. Your nodule falls into a category that suggests it's suspicious enough to warrant a closer look, but not so suspicious that we're certain it's cancer. What we typically do in these cases is recommend a follow-up scan in a few months to see if the nodule changes in size or appearance. Sometimes, we might suggest a PET scan or even a biopsy to get more information if it has certain features that concern us more. I also noticed some changes in your lungs that are consistent with emphysema, which is often related to smoking or other environmental exposures. It's important to address this with your primary care doctor as well. Lastly, we saw some signs of coronary artery calcification, which can be associated with heart disease. It would be a good idea to discuss this finding with your doctor, who may recommend further evaluation for your heart.

Fig. 6. Question and corresponding answers provided by the expert, MedFlamingo, and ChatGPT-4

Evaluation Criterion We qualitatively evaluated the answers generated by Generative AI models on two dimensions: *correctness* and *relevance*. Correctness evaluates if the answers generated by a model was judged to be right given the patient's by an expert evaluator (a radiologist and a co-author on this paper). We adopt 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. Relevance evaluates how well the question is answered. For this, we characterized the information in sentences 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 evaluators of the relevant dimension.

Finding 1 On the correctness dimension, we found that out of 20 questions, MedFlamingo was judged to be incorrect on 14 (70% error rate) while ChatGPT-4V on 7 (35% error rate). 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. In one case, ChatGPT identified itself as a radiologist. In at least one instance where both MedFlamingo and ChatGPT were judged to be correct, the expert preferred the answer generated by MedFlamingo because it was concise. In a couple of cases where ChatGPT4's answer was judged to be correct, the expert noted that it *"it rambles on and on about risk stratification and coronary calcification when the patient simply asked where the nodule is located"*. In another case, the highly detailed answer provided by ChatGPT-4V prompted the expert to say *"but given the detailed answer it gave, I felt that it would be necessary for it to have mentioned what staging system (such as TNM staging) that it used"*. Both these cases indicate that ChatGPT-4V is unable to distinguish between relevant and irrelevant information and indiscriminately produces both.

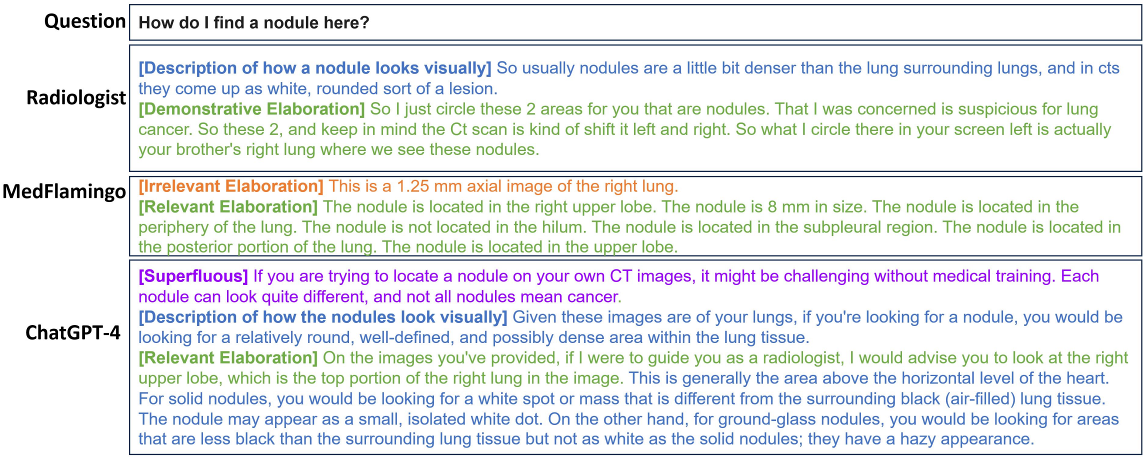
Finding 2 Our analysis of the relevance dimension revealed several insights about how doctors address their patient's questions. In a large number of cases (11) the doctor's response consisted of sentences that directly answer the patient question followed by a relevant elaboration. An example is shown 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.

MedFlamingo always provided a direct response to the question with no elaborations. Its answers tended to be very concise and expressed in clinical language. Additionally, in 6 responses it used first-person pronouns to recommend a specific action. For example, in response to the questions *"Will there be a follow-up in 3 months, or will you contact me again if there is an acute need for treatment?" "I would recommend a follow-up CT scan in 3 months. If there is any acute treatment need, we will contact you."*. This is in stark contrast with the radiologist's response who doesn't express their medical opinion but describes what potential next steps might be.

ChatGPT-4V generated 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, elaborations about emphysema and coronary artery calicification are not pertinent to the question asked, even though they were mentioned in the report. Additionally, it generated some superfluous information shown in purple that prefaces its actual response but doesn't add any useful information to the response. We found that it was unable to distinguish relevant elaborations from irrelevant information that it generated based on terms used in the report.

Finding 3 Both AI systems struggled with responding to questions in theme 3: *locating issues mentioned in the radiology report to the CT scan.* 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 MedFlamingo and ChatGPT-4V responses. It is to be expected. Medical visual question answering systems are not designed to highlight which piece of visual evidence contributed to answer generation. Additionally, MedFlamingo introduced clinical language and attempted to generate a clinically comprehensive answer about the nodule's location which is unlikely to support the caregiver's understanding. ChatGPT-4V on the other hand generated generic information about how to find various types of nodules in the scans. Both missed the fact that there are in fact two nodules in the scan as is clear from the radiologist's response.

Finding 4 We also found that both AI systems were sensitive to the way the question was framed and demonstrated an over-reliance on the textual report as opposed to applying reasoning to interpret the images. For example, when asked the question "How do I find *a* nodule here?", in addition to providing a brief description of how to spot a nodule in a scan, both models went on to describe only one of the two nodules present. The nodule they described was the one that was associated with the phrase "solid nodule" in the report, and the one they missed was the one described as "ground glass nodular consolidation", suggesting that they simply looked to the terminology in the report instead of studying the image. When asked "Can we discuss the specific *areas* where the suspicious *nodules* might be located?", both models mention both the nodules. However, they struggled to describe the location of the "ground glass nodular consolidation" as its location was not explicitly stated in the report."



Question	**How do I find a nodule here?**
Radiologist	**[Description of how a nodule looks visually]** 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. **[Demonstrative Elaboration]** 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.
MedFlamingo	**[Irrelevant Elaboration]** This is a 1.25 mm axial image of the right lung. **[Relevant Elaboration]** The nodule is located in the right upper lobe. The nodule is 8 mm in size. The nodule is located in the periphery of the lung. The nodule is not located in the hilum. The nodule is located in the subpleural region. The nodule is located in the posterior portion of the lung. The nodule is located in the upper lobe.
ChatGPT-4	**[Superfluous]** If you are trying to locate a nodule on your own CT images, it might be challenging without medical training. Each nodule can look quite different, and not all nodules mean cancer. **[Description of how the nodules look visually]** Given these images are of your lungs, if you're looking for a nodule, you would be looking for a relatively round, well-defined, and possibly dense area within the lung tissue. **[Relevant Elaboration]** On the images you've provided, if I were to guide you as a radiologist, I would advise you to look at the right upper lobe, which is the top portion of the right lung in the image. This is generally the area above the horizontal level of the heart. For solid nodules, you would be looking for a white spot or mass that is different from the surrounding black (air-filled) lung tissue. The nodule may appear as a small, isolated white dot. On the other hand, for ground-glass nodules, you would be looking for areas that are less black than the surrounding lung tissue but not as white as the solid nodules; they have a hazy appearance.

Fig. 7. Theme 3 question and corresponding answers provided by the expert, MedFlamingo, and ChatGPT-4V

In general, we found that MedFlamingo tended to generate concise answers in clinical language which may be hard for the patients and caregivers to understand. On the other hand, ChatGPT-4V tended to generate long answers with generic descriptions of terms used in the report, generic lists of risks, treatment options, etc. without any focus on the specifics of the case under discussion and with no ability to filter relevant details from irrelevant ones.

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: ChatGPT-4V and MedFlamingo on two dimensions relevant to our usecase, *correctness* and *relevance*.

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 (35% for ChatGPT-4V, 70% in MedFlamingo) 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 information systems and frequently produced irrelevant and superfluous elaborations that didn't specifically answer the question asked but rather elaborated terms used in the medical report. 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 [9], 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. First, we will expand the study described here to cover additional medical contexts to come up with an exhaustive set of informational needs patients have. We will use this set to guide the development of an evaluation benchmark that reflects the structure of this real-world task. Second, prominent metrics such like BLEU [33], ROGUE [24], BERT Score [52] while commonly used to evaluate LLM responses, fall short of measuring if the responses are useful in addressing human 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 accountable and transparent systems.

Research Ethics and Social Impact

Research materials used for conducting the study in this paper are based upon sensitive medical data obtained from real patients accessing the public university hospital for care. The data was modified and redacted as per the hospital's guidelines before it was shared with the research team. This step is critical for protecting patients' privacy.

The study described in this paper is sensitive in nature. It asks the participants to imagine themselves in a role of helping a close relative with medical decisions pertaining to a serious illness. This simulation can trigger panic and anxiety from past trauma if the participant or their loved ones have gone through similar experiences. Our recruitment emails highlighted this characteristics of the study as a trigger warning. Our consent forms also highlighted this concern, encouraging potential participants to reflect on their experiences and only consent if they felt comfortable. Finally, our consent form informed the participants that they can terminate the study at any point without incurring any adverse consequences. The IRB at the research company approved this study with an expedited review process and classified it as minimal risk.

The authors on this paper belong to diverse intellectual backgrounds; from AI & ML, to cognitive science, HCI, and education, to medicine. Our joint interest in building intelligent systems can lead to an optimism bias about the usefulness of AI & ML in healthcare. To overcome that bias, in this paper, we apply a critical lens to evaluate the use of this technology in the healthcare context. Our analysis is centered on human needs and evaluates the state of art on its suitability in addressing those needs.

Our work is best understood as a proposal for an evaluation approach for generative AI systems. An adverse impact this paper can have on the research community is accepting this approach as exhaustive, complete, and the only correct way. The research on evaluating generative AI system is nascent and rapidly evolving. Diverging perspectives on how to measure system performance are useful for the advancement of these systems.

References

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

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

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: A Visual Language Model for Few-shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[4] A. Arora and A. Arora. The Promise of Large Language Models in Healthcare. *The Lancet*, 401(10377):641, 2023.

[5] 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.

[6] 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.

[7] 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, and M. Viswanathan. Health Literacy Interventions and Outcomes: An Updated Systematic Review. *Evidence report/technology assessment*, (199):1—941, March 2011.

[8] 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*, 33:1877–1901, 2020.

[9] 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.

[10] 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(240):1–113, 2023.

[11] K. Christiansen, L. Buswell, and T. Fadelu. A Systematic Review of Patient Education Strategies for Oncology Patients in Low-and Middle-Income Countries. *The Oncologist*, 28(1):2–11, 2023.

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

[13] 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*, 71:102642, 2023.

[14] 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(2):96–97, 2018.

[15] G. M. Glazer and J. A. Ruiz-Wibbelsmann. The Invisible Radiologist. *Radiology*, 258(1):18–22, 2011.

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

[17] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and 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.

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

[19] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14):6421, 2021.

[20] J. L. Kemp, M. C. Mahoney, V. P. Mathews, M. Wintermark, J. Yee, and S. D. Brown. Patient-Centered Radiology: Where are We, Where do We Want to Be, and How do We Get There? *Radiology*, 285(2):601–608, 2017.

[21] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a Large Language-and-vision Assistant for Biomedicine in One Day. *arXiv preprint arXiv:2306.00890*, 2023.

[22] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and 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*, pages 6449–6464, 2023.

[23] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023.

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

[25] S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring How Models Mimic Human Falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[26] 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(1):226, 2023.

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[28] 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.

[29] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar. Medflamingo: A Multimodal Medical Few-shot Learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[30] L. S. Nowell, J. M. Norris, D. E. White, and N. J. Moules. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International journal of qualitative methods*, 16(1):1609406917733847, 2017.

[31] 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.

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

[33] K. Papineni, S. Roukos, T. Ward, and 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*, pages 311–318. Association for Computational Linguistics, 2002.

[34] S. B. Patel and K. Lam. ChatGPT: The Future of Discharge Summaries? *The Lancet Digital Health*, 5(3):e107–e108, 2023.

[35] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019.

[36] 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(7972):172–180, 2023.

[37] 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.

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

[39] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large Language Models in Medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[40] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level Detection of Pathologies from Unannotated Chest X-ray Images via Self-supervised Learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

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

[42] 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.

[43] 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.

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

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

[46] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. Towards Generalist Foundation Model for Radiology. *arXiv preprint arXiv:2308.02463*, 2023.

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

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

[49] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*, 9(1), 2023.

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

[51] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, et al. Large-scale Domain-specific Pretraining for Biomedical Vision-language Processing. *arXiv preprint arXiv:2303.00915*, 2023.

[52] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020.

[53] 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.