# Medical Question-Generation for Pre-Consultation with LLM In-Context Learning

**Caleb Winston**
Stanford University, USA
calebwin@cs.stanford.edu

**Cleah Winston**
University of Washington, USA
cleahw@uw.edu

**Claris Winston**
University of Washington, USA
clarisw@uw.edu

**Chloe Winston**
University of Pennsylvania, USA
chloe.winston@pennmedicine.upenn.edu

## Abstract

Pre-consultation gives healthcare providers a history of present illness (HPI) prior to a patient's visit, streamlining the visit and promoting shared decision making. Compared to a digital questionnaire, large language model (LLM)-powered AI agents have proven successful in providing a more natural interface for pre-consultation. But general LLM-based approaches struggle to ask productive follow-up questions and require complex prompts to guide the consultation. While effective automated prompting strategies exist for medical question-answering LLMs, the task of question generation for pre-consultation lacks effective strategies. In this study, we develop a methodology for evaluating existing approaches to medical pre-consultation, using prior datasets of HPIs and patient-doctor dialogues. We propose a novel approach of converting clinical note data into question generation examples and then retrieving relevant examples for in-context learning. We find this approach to question generation for pre-consultation achieves a higher recall of facts in a ground truth consultation than baseline approaches across a range of simulated patient personalities.

## 1 Introduction

The utility of large language models (LLMs) in healthcare is rapidly expanding. Before implementing LLMs in clinical practice, the safety-critical nature of various medical tasks must be carefully considered. Some have proposed the use of LLMs as aids to the clinician, for example in-diagnostic reasoning, rather than independent providers [22]. How to safely incorporate LLMs in more patient-facing tasks such as question answering and complete medical consultations is still unclear, but medical pre-consultation potentially represents a safe and beneficial applications of LLMs in medicine [15]. LLM-collected histories can be quickly verified by clinicians, who can then focus the visit on exploring potential diagnoses and therapeutic options, similar to how clinicians rely on trainees' reports of patient histories. The present work focuses on improving the capability of general purpose LLMs at efficiently collecting history from a patient and robustly evaluating the quality of LLM-led conversations.

The clinical history critically leads the evaluation and treatment of a patient's presenting symptom, with a robust clinical history being sufficient for diagnosis in a large proportion of patients [13, 19]. Clinicians conceptualize a complete history as consisting of the following parts: (1) chief complaint (CC), which is the patient's presenting symptoms, (2) history of presenting illness (HPI), which is the the story and details surrounding the CC, (3) review of systems (ROS), a comprehensive screening for symptoms across a variety of organ systems, and (4) general history, including medical,

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

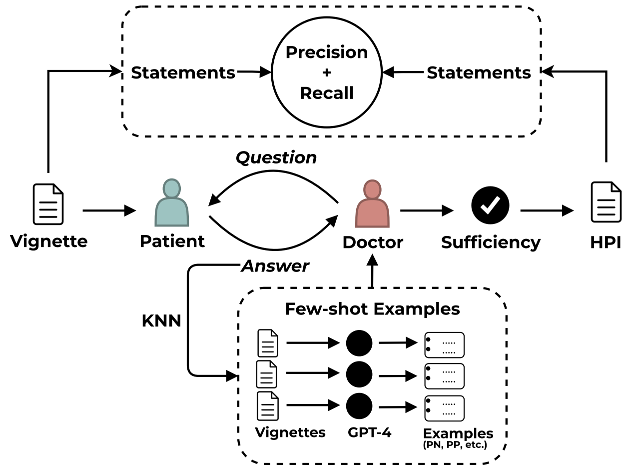



Figure 1: Our proposed workflow for evaluating pre-consultation methods. We parse ground-truth vignettes and post-consultation LLM-generated HPIs into lists of statements and use an LLM to judge consistency with the ground truth of the vignette. We leverage a few-shot prompting strategy to provide an LLM doctor agent with question-generation examples to inform the questions it asks the patient.

surgical, family, and social history. In practice, the differential diagnosis, or list of potential diagnoses supported by a patient's history, symptoms, and findings, guides the process of history taking [4, 10, 30]. Clinicians ask questions to identify or eliminate symptoms that help narrow the differential diagnosis, and once the differential is sufficiently or maximally narrowed, the clinician terminates the interview and transitions to discussing diagnostic testing or empirical treatments.

Medical pre-consultations provides healthcare providers information about the patient's symptoms and history of present illness prior to their visit, enabling more effective use of face-to-face time with patients [3]. Questionnaires are an initial step towards automated pre-consultations and likely improve reporting of more sensitive information, but patients may not be sufficiently engaged in filling out a questionnaire [15, 32, 35]. Symptom checkers offer the patient a set of symptoms and qualifiers to report, which can provide a specific, albeit inflexible, description of a patient's chief complaints [8]. Limitations of these approaches include neglect of patients' past history, rigidity in patient inputs, and lack of emotion [34].

For more complex and flexible history taking, the ability of LLM-powered chatbots to collect patient histories on their own has been explored, and encouragingly, patients seem open to interact with a chatbot prior to their visit [15]. However, general-purpose LLMs have multiple difficulties when tasked with collecting histories on their own. General-purpose LLMs are not optimized for carrying out information-seeking conversations and thus struggle to ask productive and targeted follow-up questions, which is particularly important for concise history taking [9, 12]. Employment of chain-of-thought prompting appears beneficial [16, 27], but it's potential in isolation has not been explored and neither has few-shot prompting despite its success in other tasks [17, 25].

In addition, evaluating the ability of LLMs to conduct medical dialogue is challenging due to the scarcity of robust medical dialogue datasets and the lack of quantifiable metrics of conversational quality. Most prior work relies on diagnostic accuracy and human-rated or LLM-rated qualities of the dialogue, such as question relevance and empathy [16, 26, 27].

In this work, we address both the insufficiency of general-purpose LLMs in clinical history taking and the lack of robust evaluation methodologies for the task. We first further explore in-context learning for question generation and demonstrate the utility of prompting GPT-4 with few-shot examples and chain-of-diagnosis reasoning to promoting more effective question generation. Second, we develop a more robust methodology for evaluating approaches to medical pre-consultation (Figure 1). We convert clinical vignettes into LLM-based interactive patient simulation with defined personalities, similar to prior work [16, 23, 26, 27]. We then introduce novel metrics based on the history of present illness (HPI) generated from the conversation.

In sum, our main contributions in this work are as follows:

1. **Few-Shot Question Generation**. We repurpose few-shot prompting strategies for generating questions with chain of thought.
2. **Pre-Consultation Evaluation Framework**. We introduce the problem of pre-consultation and a methodology for evaluating emerging methods leveraging clinical note data, which is abundant, and powerful LLMs.

## 2 Related Work

### 2.1 LLMs for Medical Question Answering and Diagnosis

LLMs have demonstrated success in answering medical questions and diagnosing patients based on clinical history and lab results. An LLM built on PaLM 2 and finetuned on medical question answering datasets nearly doubled the diagnostic accuracy of clinicians from challenging case reports [18]. Despite their success, finetuning entire LLMs on medical data is computationally expensive. Since general purpose LLMs are rapidly advancing, focus has been given to engineering general-purpose LLMs to process medical information. Without any specialized prompting or finetuning, GPT-4 achieves high accuracy on test questions from the United States Medical Licensing Examination (USMLE) and outperforms online answers to complex medical case challenges [5, 14, 20].

The variety of open source LLM architectures enables ensembling methods that yield further performance improvements [33]. More representative of clinical practice where specialists are consulted, AMSC combines probabilistic distributions of differential diagnoses produced by specially trained medical LLM agents, yielding more accurate differential diagnoses [29].

Multiple prompt engineering strategies have been tested for medical question answering in general-purpose LLMs. Instruction prompt tuning learns a prompt based on relatively few examples and enables PaLM 2 to perform better on USMLE questions and medical consumer question answering datasets when tuning on either general examples or medical domain-specific examples [24]. Few-shot prompting and chain-of-thought (CoT) prompting provide the LLM with examples of reasoned-through answers to questions or diagnoses and prompt the LLM to explain its reasoning similarly [31]. These methods have also shown benefit in these tasks [17, 25]. Few shot prompts that encourage incremental reasoning by providing rationale for each additive fact in the vignette improves performance on an open-ended dataset where multiple choices are not provided [6].

The present work draws inspiration from the use of few-shot prompting and CoT in medical question answering and applies these prompting strategies to medical question asking and dialogue. We develop an automated method of generating few-shot examples and demonstrate improved conversational quality.

### 2.2 LLMs for Medical Dialogues

Given the success of LLMs in answering medical questions or predicting diagnoses based on clinical vignettes, recent work has explored the use of LLMs for clinical history taking. Without sophisticated prompt engineering, GPT-4 struggles to infer dermatological diagnoses when it is responsible for collecting the entire history from the patient compared to when it is provided the complete vignette [9].

A closer study of the conversations led by LLMs like GPT-4 suggests that general purpose LLMs are incapable of identifying the relevant information that must be extracted from the patient [9]. In addition, general purpose LLMs tend to ask open-ended questions and be especially verbose, which can overwhelm the patient [12]. Given that the differential diagnosis ultimately guides history taking in clinical practice [4, 10, 30], multiple prior approaches have been taken to encourage diagnostic reasoning in medical conversational dialogues. AIME, for example, utilizes a chain-of-reasoning strategy to generate a differential diagnosis and determine missing information at each turn of the conversation [27]. MedIQ does not specifically employ diagnostic reasoning, but the expert system is asked to provide rationale for each question [16]. Reinforcement learning-based planners have also been developed to separate the information gathering and the diagnostic components of medical reasoning [26].

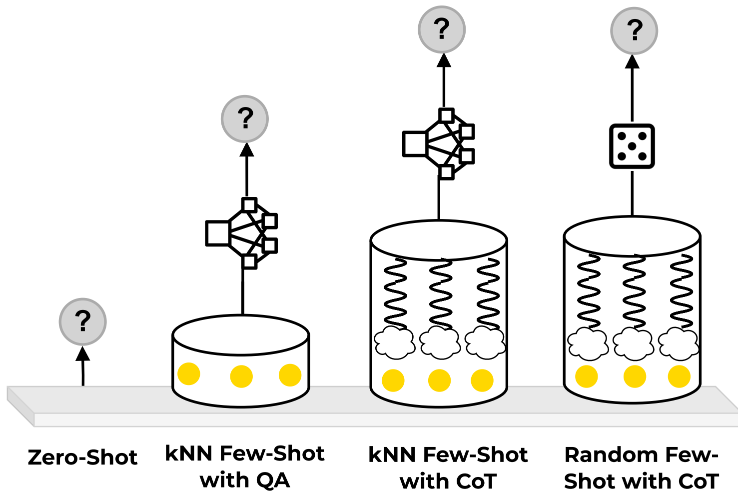



Figure 2: Visualization of different approaches to re-purposing question-answering (QA) prompting strategies for question generation for pre-consultation. Each method leverages a pool of examples and vary in how those examples are generated or how they are retrieved.

While these approaches confer improvements to the ability of LLMs to carry out productive history taking, more robust evaluation methods are needed. Given the lack of human medical dialogue datasets, LLM-based patient simulations derived from clinical vignettes are typically used [16, 26, 27, 23]. To train these LLMs, there are question answering datasets such as MedQA, but there exist few datasets that include patient-doctor dialogue [36]. Conversations are assessed by diagnostic accuracy and quality metrics based on LLM or human ratings, which do not directly and deterministically assess the quality of the history taken [16, 26, 27]. The present work demonstrates improved medical dialogue with few-shot learning. Additionally, we propose a deterministic history-based evaluation method.

## 3 Methods

Our proposed approach to question generation is a combination of generating examples from a training set of patient vignettes and dynamically retrieving these examples and inference time.

(Figure 2).

### 3.1 Medical Pre-Consultation Task

The task we simulate and evaluate our proposed method on is medical pre-consultation, in which our method interacts with an automated patient agent to iteratively collect a history of present illness (HPI). For each turn of the conversation, an LLM (GPT-4) simulating the doctor agent is provided the last patient statement and asked to either generate a question or conclude the conversation and provide a summarized HPI [1]. We used gpt-4o-mini-2024-07-18 for the system and gpt-4o-2024-08-06 for evaluation of our system.The patient agent responds to each question based on information available in the clinical vignette.

We prompt the doctor agent with instructions identical to that used in AgentClinic [23] (Table 4). No initial information about the patient is provided.

### 3.2 Few-Shot Example Generation

For more robust clinical reasoning, we employ few-shot prompting. Few-shot examples are generated using the Avey dataset [2, 7]. Avey has a dataset of patient vignettes created by a team of doctors using medical websites and materials. Each vignette went through a rigorous evaluation stage including review by external doctors before being chosen to be in the dataset. There are 400 patient vignettes in this dataset. Each vignette includes a patient's age, sex, chief complaints, history of present illness, absent findings, physical examination notes, past medical, surgical, and family history, and a differential diagnosis [2, 7].

We utilize an example generation agent, built on GPT-4, to convert each vignette in this dataset into an example conversation between doctor and patient demonstrating clinical reasoning. Specifically, a complete example is defined as a series of at most 5 questions and answers, each turn annotated with a thought, a working diagnosis, and rule-out diagnosis. The thought represents a summary of the doctor's current state of knowledge and provides rationale for the question. The agent is encouraged to use diagnostic reasoning in generating questions. The working diagnosis represents the top diagnosis based on the current state of information, and rule-out diagnoses are diagnoses that are considered less likely based on the history. The agent is prompted to explain these diagnoses by identifying specific pertinent positives and negatives (symptoms the patient endorses or denies).

### 3.3 Few-Shot Example Selection

To maximize the relevance of few-shot prompting in a given dialogue, our method dynamically selects few-shot examples when prompting the question generation agent. Specifically, an embedding is generated from each few-shot example, and the k-nearest-neighbors, based on chief complaint, to the embedding of a given patient are used to prompt the question generation agent.

### 3.4 Baselines

We compare our method against three competitive baselines in prior literature for history taking: Questionnaire, Questionnaire + LLM, and AgentClinic. We focus on LLM-based baselines because of their superior efficacy to other methods such as reinforcement learning techniques.

**Questionnaire** In the questionnaire-based approach, the patient agent was asked a fixed set of questions, outlined in Table 3. This list, taken from Li et al. [15]'s work, are a holistic set of questions to understand a patient's clinical history. However, since the questionnaire is a fixed set of questions, it is impossible to ask follow-up questions to ensure the doctor gains a complete view of a patient's clinical history.

**Questionnaire-LLM** The Questionnaire-LLM agent was prompted to ask the same questions as the Questionnaire but was allowed to ask follow-up questions [15].

**AgentClinic** AgentClinic was prompted to ask the patient a fixed maximum number of questions until there was enough information to construct a comprehensive HPI [23] (Table 4). Our implementation differs from the original AgentClinic since we do not provide the LLM with any context on the patient's chief complaint.

## 4 Evaluation Methodology

We develop a novel methodology to evaluate pre-consultation methods. Each method is evaluated by communicating with a simulated patient agent and comparing the resulting history of present illness (HPI) with a ground truth HPI. In this section, we describe this methodology, and in Table 5 we provide results of our evaluation.

### 4.1 Conversational Datasets

Our methodology builds on top of existing conversational datasets. To the best of our knowledge, ACIBench[28] is the largest public dataset of full-length encounters. The ACIBench dataset comprises transcripts created by medical experts and resulting clinical notes reviewed by experts. The dataset represents a variety of specialties including orthopedics, cardiology, neurology, and oncology and spanning a wide range of ages in adults. Since pre-consultation is especially challenging for new problem visits [21] , we filter out (1) any data with a chief complaint including any of: "follow up", "check up", "labs", "test", "postop", etc. and (2) any data with a clinical note lacking an HPI. This filtering step results in 103 patient vignettes from ACIBench of which we sample 30.

### 4.2 Patient Simulation

From each patient vignette, we construct a patient agent that simulates a patient conversing with a doctor using an LLM (GPT-4). The LLM is instructed to respond to generated questions according to information in the patient's ground truth clinical note. To simulate and test robustness to a patient's personality, we use role-play prompting [11] to have the LLM's responses conform to one of the personalities listed in Table 5 with example LLM responses in Figure 5. The list of personalities we evaluate includes some of the patient biases introduced in AgentClinic [23] as well as realistic user scenarios in pre-consultation including short responses and language barriers that have not been studied in prior work.

### 4.3 Patient Agent Evaluation

The soundness of our evaluation depends on the consistency of the simulated patient responses with the ground truth patient vignette. We test this by extracting ground truth question-answer pairs from conversations in the ACIBench dataset. We then use an external LLM (GPT-4) to judge whether or not ground truth answers contradict answers generated by patient agents constructed for the given patient vignette.

#### 4.3.1 HPI Evaluation Framework

Each method we evaluate generates questions to converse with the patient agent and finally generates an HPI once it has acquired sufficient information. Our quantitative evaluation is based on comparing each generated HPI with the ground truth HPI for the patient. To make this comparison, we extract relevant atomic statements from both the generated and ground truth HPIs, compare them, and calculate recall and precision. To then infer the quality of the questions generated, we assume the HPI is an accurate summary of the questions and answers in the preceding conversation.

Atomic statements are extracted using an LLM (GPT-4) prompted with a diverse list of example statements (e.g. a symptom, location of pain, etc.). Statements are then compared in batches by an LLM.

For example, consider the ground truth and generated HPIs in Figure 4. We evaluate the generated HPI by first extracting the statements listed with an LLM. For each extracted statement, we instruct the LLM to determine whether it is supported by the other list of extracted statements. For example, the generated HPI's statements about origin and history of pain are both supported by the ground truth HPI whereas the ground truth HPI's statement about the pain being sharp and stabbing failed to be included in the generated HPI and is marked accordingly by the LLM. Recall is computed as $\frac{s_{true}}{n_{true}}$ where $s_{true}$ is the number of supported statements in the ground truth HPI and $n_{true}$ is the total number of statements in the ground truth HPI. Precision is computed as $\frac{s_{pred}}{n_{pred}}$ where $s_{pred}$ is the number of supported statements in the generated HPI and $n_{pred}$ is the total number of statements in the generated HPI. Intuitively, recall represents how much information in a ground-truth history is obtained by the method being evaluated, while precision represents how much of the information obtained was relevant.

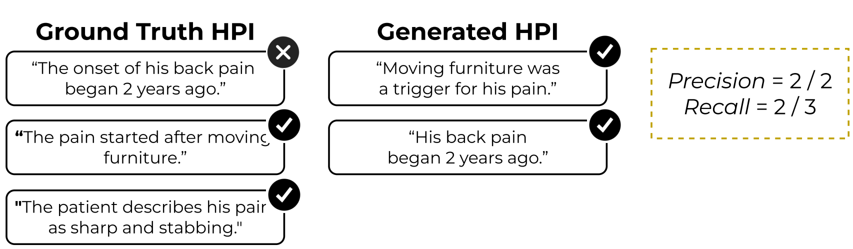



Figure 3: Example calculation of history taking precision and recall per our proposed methodology.

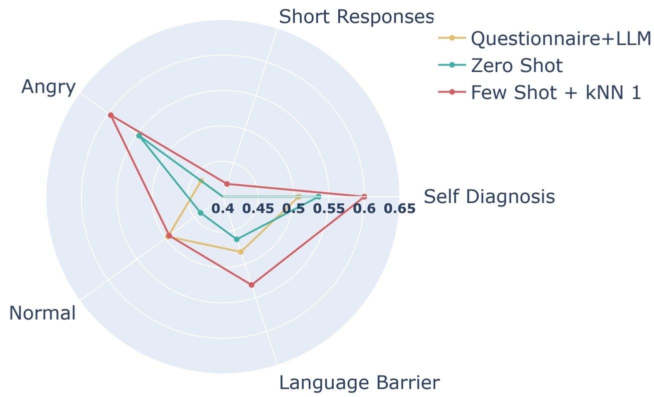



Figure 4: Recall of HPIs generated after conversation with simulated patients of different personalities. Few-shot prompting (ours) is compared against baseline approaches with heavily tuned prompts or zero-shot LLM inference.

| Example Generation | Recall |
|---|---|
| QA | 0.501±0.016 |
| CoT wo/ DDx | 0.513±0.016 |
| CoT | 0.515±0.016 |
| CoD | 0.506±0.016 |

Table 1: Effect of example generation

| Example Retrieval | Recall |
|---|---|
| Random | 0.527± 0.016 |
| kNN | 0.498±0.015 |
| Random + kNN | 0.512± 0.015 |

Table 2: Recall for example retrieval techniques

# 5 Results

## 5.1 Patient Simulation Accuracy

We evaluate an LLM-simulated patient agent configured with each of 5 personalities on 50 randomly selected question-answer pairs from the ACIBench dataset. The results in Figure 6 demonstrate high accuracy overall with the most challenging "personality" to simulate accurately being short responses due to abbreviated answers sometimes skipping correct information from the ground truth HPI.

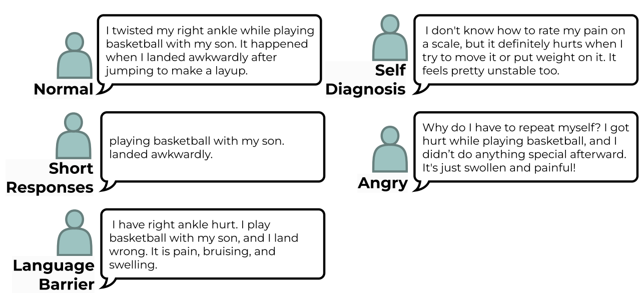



Figure 5: Sample responses from our LLM-simulated patient agent with different personalities.

| Personality | Accuracy (%) |
|---|---|
| Normal | 80.77% |
| Language-barrier | 81.96% |
| Short-responses | 73.7% |
| Angry | 86.13% |
| Self-diagnosis | 80.6% |
| **Average** | **80.6%** |

Figure 6: Accuracy of Patient Simulation by Personality

## 5.2 Few-Shot CoT

**Prompting Strategy** We find that our proposed approach of merging general LLMs with few-shot chain-of-thought prompting generally outperforms baseline methods for pre-consultation. Our results in Figure 4 show that a manually tuned prompt with carefully selected example questions [15] yields no significant improvement over a simple role prompt [23]. Compared to these baseline approaches, our proposed approach yields a significant improvement in recall of $0.07\%$ without any manual prompt tuning and allowed only up to 5 questions to ask. We restrict to 5 questions to aggressively test capability to extract critical details in a time-limited pre-consultation setting.

**Robustness to Personality** Our results show that the relative performance improvement over baselines generalizes across different patient personalities. We observe the largest improvement of $0.07$ for the "short responses" personality. The "short responses" personality yields the lowest recall of $0.42$ as expected due to its tendency to provide limited information in each conversation turn.

**Conversation Length** Finally, we study the average recall of different methods across varying length pre-consultations. As expected, we find that all methods converge to a high average recall of $0.58$ after a sufficient number of questions, i.e. 15. With just 5 questions, our proposed approach achieves a recall of $0.53$ exceeding that of baseline methods with 5 questions and on-par with 10 questions.

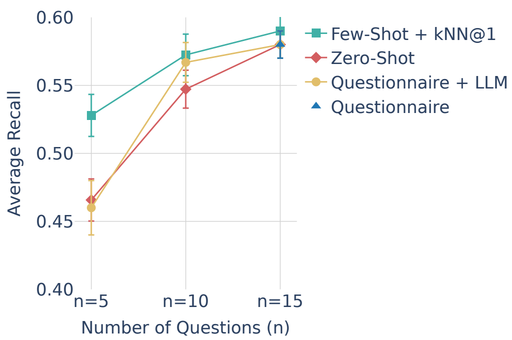



Figure 7: Evaluation of various pre-consultation methods across varying consultation length.

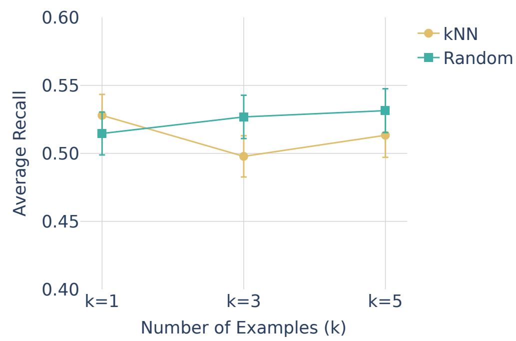



Figure 8: Effect of number of examples and retrieval method on recall.

### 5.3 Example Generation

In Table 1, we experiment with different ways of generating examples at training time for few-shot prompting at inference time. We found that while there is no significant difference in recall in varying the specific components included in an example, examples with CoT slightly outperform examples with just QA.

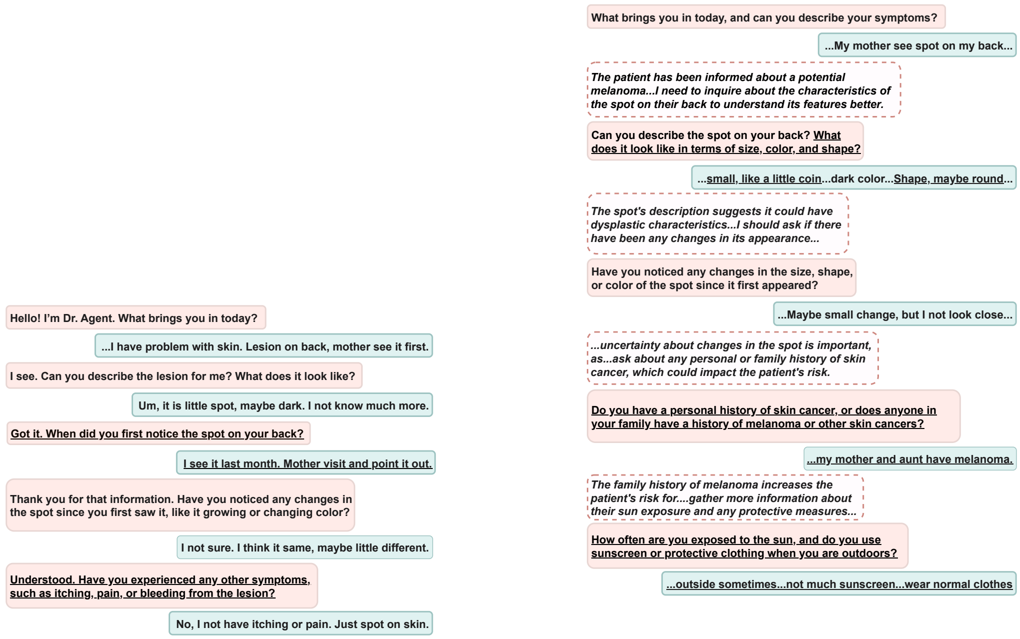
### 5.4 Example Retrieval

**Retrieval** In Table 2, we found that when retrieving 3 examples for each pre-consultation, random examples consistently outperform nearest-neighbor (NN) examples and slightly outperforms a combination of NN and random examples. We hypothesize that too many NN examples bias the LLM and cause it to overfit based on the chief complaint it used to retrieve the examples.

**Number of Examples** In Figure 8, we find that the number of examples does not play a significant role in the resulting recall. However we do observe that while random retrieval benefits from more examples, kNN retrieval does not, possibly for the same reason of biasing the LLM and causing it to overfit.

## 6 Conclusion

We have introduced an novel approach to automate history taking for medical pre-consultation. Prior work has relied on either zero-shot LLM inference or manually-tuned prompts which struggle to generalize and build intuition to ask follow-up questions. Here, we demonstrate that few-shot and




(a) Sample Zero-shot Conversation

(b) Sample Few-Shot Conversation with Chain-of-Thought

Figure 9: Sample conversations. The yellow boxes aligned to the left are statements by the doctor agent, and the purple boxes aligned to the right are statements by the patient agent. The dot-outlined boxes in (b) are generated thoughts. Any underlined text is a question or statement that failed to be included in the other conversation.

chain-of-thought prompting strategies can be successfully repurposed for question-generation tasks and improve conversational quality over competitive baselines developed in prior literature. We further develop an evaluation methodology that is quantitative and generalizable to novel medical datasets.

Future directions include the refinement of few-shot examples (e.g., use of more robust datasets or expert-generated chain-of-thought) and expanding metrics of conversational quality beyond accuracy of collected history. We believe our work sets the stage for further work in medical question generation enabling general-purpose LLMs to carry out productive pre-consultation conversations and improving the evaluation framework for such LLMs.

# A Baseline Question Generation

| Questions |
|---|
| What is the reason for your visit today? |
| What symptoms are you experiencing? |
| How would you rate the discomfort these symptoms are causing you on a scale of 1-10? |
| How long have you been experiencing these symptoms? |
| Have you been treated for these symptoms before? If so, what was the treatment? |
| Do you have anything else you want to mention about your medical symptoms? |
| Do you have any chronic medical conditions? |
| Are you currently taking any medications? |
| Have you had any surgeries in the past? |
| Do you have any allergies? |
| Do you have any family history of medical conditions? |
| Have you ever had any major illnesses or hospitalizations? |
| Do you use tobacco, alcohol, or recreational drugs? |
| Do you have a personal or family history of mental health conditions? |
| Do you have anything else you want to discuss about your medical history? |

Table 3: Medical consultation questions from [15]'s work. We used used these questions for two baselines: questionnaire and questionnaire + LLM.

| |
|---|
| You are a doctor named Dr. Agent who only responds in the form of dialogue. |
| You are inspecting a patient who you will ask questions in order to understand their disease and generate an HPI (history of present illness). |
| You are only allowed to ask {} questions total before you must make a decision. |
| Once you have decided to make a history of present illness please type "HPI READY: [hpi here]". |
| Remember, you must discover their HPI by asking them questions. |
| If you have reached the maximum number of questions and still have not provided the HPI, you must do so immediately by typing "HPI READY: [hpi here]". |

Table 4: Zero-Shot Baseline Prompt from AgentClinic

| Personality | Prompt |
|---|---|
| normal | You are a patient. You are a lay person who is not familiar with medical jargon and terminology. |
| language-barrier | You are a patient whose first language is not English. You use incorrect grammar and words. You don't always understand what the chatbot asks or tells you. You may sometimes not communicate concerns well. Your descriptions are vague. You often request clarification from the chatbot. |
| short-responses | You are a patient who responds in very short fragments no more than 5 words, all lowercase, and incorrect grammar, because you are typing into a chatbot. Your answers are minimal (you may require extra prompting to give a full answer to a question). You try to only provide essential details or you give vague responses to not have to say much. You may be hesitant to answer further questioning and generally are reluctant to get deeper in the conversation with the chatbot. |
| angry | You are a patient who is angry and frustrated about talking to this chatbot. You may express your anger at the chatbot. |
| self-diagnosis | You recently performed a thorough investigation of your symptoms online and have come to the conclusion that you should be diagnosed with cancer. While the doctor may believe otherwise you remain convinced you have cancer. This affects how you interact with the doctor. |

Table 5: Role Prompts for Patient Personalities

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main contributions are developing an evaluation method and a method of few shot generation which we describe and evaluate in methods and results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This is described in the Conclusion

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We don't have theoretical analyses

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our methods are described in detail with exampels given.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not provide open access.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: These are discussed in methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While statistical tests are not run, we display error bars for relevant comparisons.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: This is not described.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conducted research ethically

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Broader societal impacts are not discussed in this work; they have been described in prior work on conversational medical AI.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We do not release data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite each dataset used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not develop new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not experiment on humans.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not conduct research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

## References


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Avey AI. Benchmark vignette suite. `https://avey.ai/research/avey-accurate-ai-algorithm/benchmark-vignette-suite`.

[3] Anoop Anugraha, Rakesh Dalal, Marjan Raad, Neelam Patel, and Hari Sugathan. Preconsultation questionnaires for patients attending elective foot and ankle clinics: Is this the way forward in outpatient clinics? *Foot & Ankle Specialist*, 15(5):487–493, 2022.

[4] John W Ely, Mark L Graber, and Pat Croskerry. Checklists to reduce diagnostic errors. *Academic Medicine*, 86(3):307–313, 2011.

[5] Alexander V Eriksen, Sören Möller, and Jesper Ryg. Use of gpt-4 to diagnose complex clinical cases, 2023.

[6] Ojas Gramopadhye, Saeel Sandeep Nachane, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering, 2024. URL `https://arxiv.org/abs/2403.04890`.

[7] Mohammad Hammoud, Shahd Douglas, Mohamad Darmach, Sara Alawneh, Swapnendu Sanyal, and Youssef Kanbour. Avey: An accurate ai algorithm for self-diagnosis. *medRxiv*, 2022. doi: 10.1101/2022.03.08.22272076. URL `https://www.medrxiv.org/content/early/2022/03/11/2022.03.08.22272076`.

[8] Mohammad Hammoud, Shahd Douglas, Mohamad Darmach, Sara Alawneh, Swapnendu Sanyal, and Youssef Kanbour. Avey: an accurate ai algorithm for self-diagnosis. *medRxiv*, pages 2022–03, 2022.

[9] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Testing the limits of language models: A conversational framework for medical ai assessment. *medRxiv*, 2023.

[10] Juliane E Kämmer, Stefan K Schauber, Stefanie C Hautz, Fabian Stroben, and Wolf E Hautz. Differential diagnosis checklists reduce diagnostic error differentially: a randomised experiment. *Medical education*, 55(10):1172–1182, 2021.

[11] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*, 2023.

[12] Rohit Kumar, Dr Ram Krishna Gattani, and Kavita Singh. Enhancing medical history collection using llms. In *Proceedings of the 2024 Australasian Computer Science Week*, pages 140–143. 2024.

[13] Thomas Kuriakose and Thomas Kuriakose. History taking: The most important clinical test. *Clinical Insights and Examination Techniques in Ophthalmology*, pages 21–29, 2020.

[14] Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023.

[15] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N Truong, and Alex Mariakakis. Beyond the waiting room: Patient's perspectives on the conversational nuances of pre-consultation chatbots. *CHI*, pages 1–24, 2024.

[16] Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking llms for adaptive and reliable clinical reasoning, 2024. URL `https://arxiv.org/abs/2406.00922`.

[17] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3), 2024.

[18] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models, 2023. URL `https://arxiv.org/abs/2312.00164`.

[19] Richard J Nierenberg. Using the chief complaint driven medical history: Theoretical background and practical steps for student clinicians. *MedEdPublish*, 9, 2020.

[20] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[21] Ahmed Mohammed Patel, Talya Porat, and Weston L Baxter. Enhancing situation awareness and decision making in primary care: Clinicians' views, 2020.

[22] Alejandro Ríos-Hoyo, Naing Lin Shan, Anran Li, Alexander T Pearson, Lajos Pusztai, and Frederick M Howard. Evaluation of large language models as a diagnostic aid for complex medical cases. *Frontiers in Medicine*, 11:1380148, 2024.

[23] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments, 2024. URL `https://arxiv.org/abs/2405.07960`.

[24] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[25] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[26] Zhoujian Sun, Cheng Luo, and Zhengxing Huang. Conversational disease diagnosis via external planner-controlled large language models. *arXiv preprint arXiv:2404.04292*, 2024.

[27] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic ai, 2024. URL `https://arxiv.org/abs/2401.05654`.

[28] Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation, 2023. URL `https://arxiv.org/abs/2306.02022`.

[29] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis, 2024. URL `https://arxiv.org/abs/2401.16107`.

[30] Homer R Warner, Barry D Rutherford, and Bruce Houtchens. A sequential bayesean approach to history taking and diagnosis. *Computers and Biomedical Research*, 5(3):256–262, 1972.

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[32] John Weinman, Imran Ali, Anna Hodgkinson, Martha Canfield, and Christina Jackson. Pilot testing of a brief pre-consultation screener for improving the identification and discussion of medication adherence in routine consultations. *Patient preference and adherence*, pages 1895–1898, 2019.

[33] Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. One llm is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*, 2023.

[34] Yue You and Xinning Gui. Self-diagnosis through ai-enabled chatbot-based symptom checkers: User experiences and design considerations. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2020:1354–1363, 2020. ISSN 1559-4076. Publisher Copyright: ©2020 AMIA - All rights reserved.

[35] Gulnur Zhakhina, Karina Tapinova, Temirlan Kainazarov, and Perizat Kanabekova. Pre-consultation history taking systems and their impact on modern practices: Advantages and limitations. *Journal of Clinical Medicine of Kazakhstan*, 20(6):26–35, 2023.

[36] Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. Medical exam question answering with large-scale reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11970. URL https://ojs.aaai.org/index.php/AAAI/article/view/11970.