Testing the Limits of Language Models: A Conversational Framework for Medical AI Assessment

Authors: Shreya Johri* [1], Jaehwan Jeong* [1,4], Benjamin A. Tran MD [5], Daniel I. Schlessinger MD[6], Shannon Wongvibulsin MD PhD[7], Zhuo Ran Cai MD[3], Roxana Daneshjou MD PhD† [2,3], Pranav Rajpurkar PhD† [1]

Affiliations:

1. Department of Biomedical Informatics, Harvard Medical School, Boston, United States
2. Department of Biomedical Data Science, Stanford University, Stanford, United States
3. Department of Dermatology, Stanford University, Stanford, United States
4. Department of Computer Science, Stanford University, Stanford, United States
5. Medstar Georgetown University Hospital/Washington Hospital Center, Department of Dermatology, Washington, DC, United States
6. Department of Dermatology, Northwestern University, Chicago, IL, United States
7. Division of Dermatology, David Geffen School of Medicine at the University of California, Los Angeles, California, United States

* These authors contributed equally: Shreya Johri, Jaehwan Jeong

† These authors share senior authorship: Roxana Daneshjou, Pranav Rajpurkar

Corresponding author:
Pranav Rajpurkar, PhD
pranav_rajpurkar@hms.harvard.edu

Abstract

Large Language Models (LLMs) show promise for medical diagnosis, but traditional evaluations using static exam questions overlook the complexity of real-world clinical dialogues. We introduce a multi-agent conversational framework where doctor-AI and patient-AI agents interact to diagnose medical conditions, evaluated by a grader-AI agent and medical experts. We assessed the diagnostic accuracy of GPT-4 and GPT-3.5, in conversational versus static settings using 140 cases focusing on skin diseases. Our study revealed a decline in diagnostic accuracy, unmasking key limitations in LLMs' ability to integrate details from conversational interactions to improve diagnostic performance. We introduced Conversational Summarization, a technique that enhanced performance, and expert review identified deficiencies compared to human dermatologists in comprehensive history gathering, appropriate use of terminology, and reliability. Our findings advocate for nuanced, rigorous evaluation of LLMs before clinical integration, and our framework represents a significant advancement toward responsible testing methodologies in medicine.

Introduction

The doctor-patient conversation serves as the linchpin of diagnostic medicine, enabling physicians to uncover key details that guide their clinical decisions. However, the mounting pressure of escalating patient numbers, lack of access to care[1], short consultation times[2,3], and the expedited adoption of telemedicine due to the COVID-19 pandemic[4] have presented formidable challenges to this conventional model of interaction. As these factors risk compromising the quality of history taking and thereby diagnostic accuracy[2], there is an urgent need for innovative solutions that can enhance the efficacy of these crucial conversations.

New advances in generative artificial intelligence, specifically in Large Language Models (LLMs), could present a potential solution to this problem[5–9]. These AI models have the ability to engage in nuanced and complex conversations[10–12], making them ideal candidates for extracting comprehensive patient histories and assisting physicians in generating differential diagnoses. However, a considerable gap remains in assessing these models' readiness for application in real-world clinical scenarios[13–15]. Current evaluation methodologies are primarily centered on their performance on medical exam-type questions[16–18], failing to capture the subtleties inherent in doctor-patient conversations that are critical in actual clinical practice.

In response, we present the **C**onversational **R**easoning **A**ssessment **F**ramework for **T**esting in **Med**icine (CRAFT-MD), a multi-agent conversational framework designed to assess the capabilities of LLMs in medical diagnosis. CRAFT-MD focuses on history-taking and contextual understanding, providing a comprehensive evaluation of LLM's ability to extract and utilize information from complex conversations, similar to the process that physicians undertake in clinical scenarios. We chose to concentrate on skin diseases, some of the most frequent complaints in primary care[19]. The diversity of skin conditions necessitates nuanced and context-dependent reasoning around the onset, progression, associated symptoms, and relevant personal or familial medical histories, thereby providing a rigorous testing ground for AI capabilities.

Within the CRAFT-MD framework, doctor-AI and patient-AI agents emulate the intricate dynamics of a doctor-patient interaction, while the grader-AI agent and medical experts evaluate the conversations. The doctor-AI gathers a history from the patient-AI agent and generates a diagnosis based on this interaction, while the grader-AI agent evaluates the diagnostic accuracy of the doctor-AI agent. Medical experts evaluate the behavior of the three AI agents. We applied CRAFT-MD to assess the clinical reasoning capabilities of two leading LLMs, GPT-4 and GPT-3.5. Our analysis reveals critical limitations in their ability to integrate details from conversations and make accurate diagnostic decisions, unlike human dermatologists. However, LLMs demonstrate potential for extracting comprehensive histories from patient interactions. With focused development, they could augment physician decision-making by identifying clinically relevant information from natural conversations. CRAFT-MD provides a robust methodology for testing LLMs beyond static exams to guide progress. It constitutes an important first step towards realizing the immense promise of conversational AI in medicine.

Results

The CRAFT-MD Evaluation Framework

To evaluate the clinical reasoning skills of LLMs, we introduced CRAFT-MD, a multi-agent conversational framework consisting of a doctor-AI and a patient-AI agent that is engineered to simulate real-world doctor-patient conversations, followed by grader-AI agent and expert evaluation (Figure 1, see Methods).

The doctor-AI agent is an LLM assuming the role of a doctor communicating with a patient. It initiates queries to compile the patient's medical history, current symptoms, medications, and family history. The doctor-AI proceeds to ask pertinent questions until a confident diagnosis can be made.

The patient-AI agent is equipped with a case vignette and embodies the patient role, responding to the doctor-AI's inquiries. It is directed to articulate symptoms in everyday language, disclosing only the information available in the vignette. This restriction curbs the introduction of unrelated details and promotes a natural flow of conversation.

The grader-AI agent determines whether the doctor-AI's concluding diagnosis corresponds with the ground truth diagnosis of the case. It is tasked with correlating disease synonyms and subtypes to account for variations in terminology.

Finally, expert evaluation enables a comprehensive assessment of each of the agent's performance. It determines whether the doctor-AI agent successfully extracts a complete medical history, the patient-AI agent uses layman terminology, and the grader-AI agent's accuracy is reliable.

We used the CRAFT-MD framework to convert 140 multiple choice case vignettes about skin diseases into interactive cases. We compiled 100 case vignettes from an online question bank[20] and developed 40 new case vignettes (see Methods, Data Availability). This strategy enabled the utilization of CRAFT-MD on both existing and newly formulated exam-style questions. The vignettes address a broad spectrum of skin diseases encountered in primary and specialist care environments. By transitioning these static vignettes into interactive cases via our multi-agent conversational framework, we could effectively evaluate LLMs on aspects of clinical reasoning, including relevance of questioning, integration of details, and diagnostic accuracy. Both GPT-4 and GPT-3.5 were evaluated using this framework centered on clinical conversational reasoning abilities.

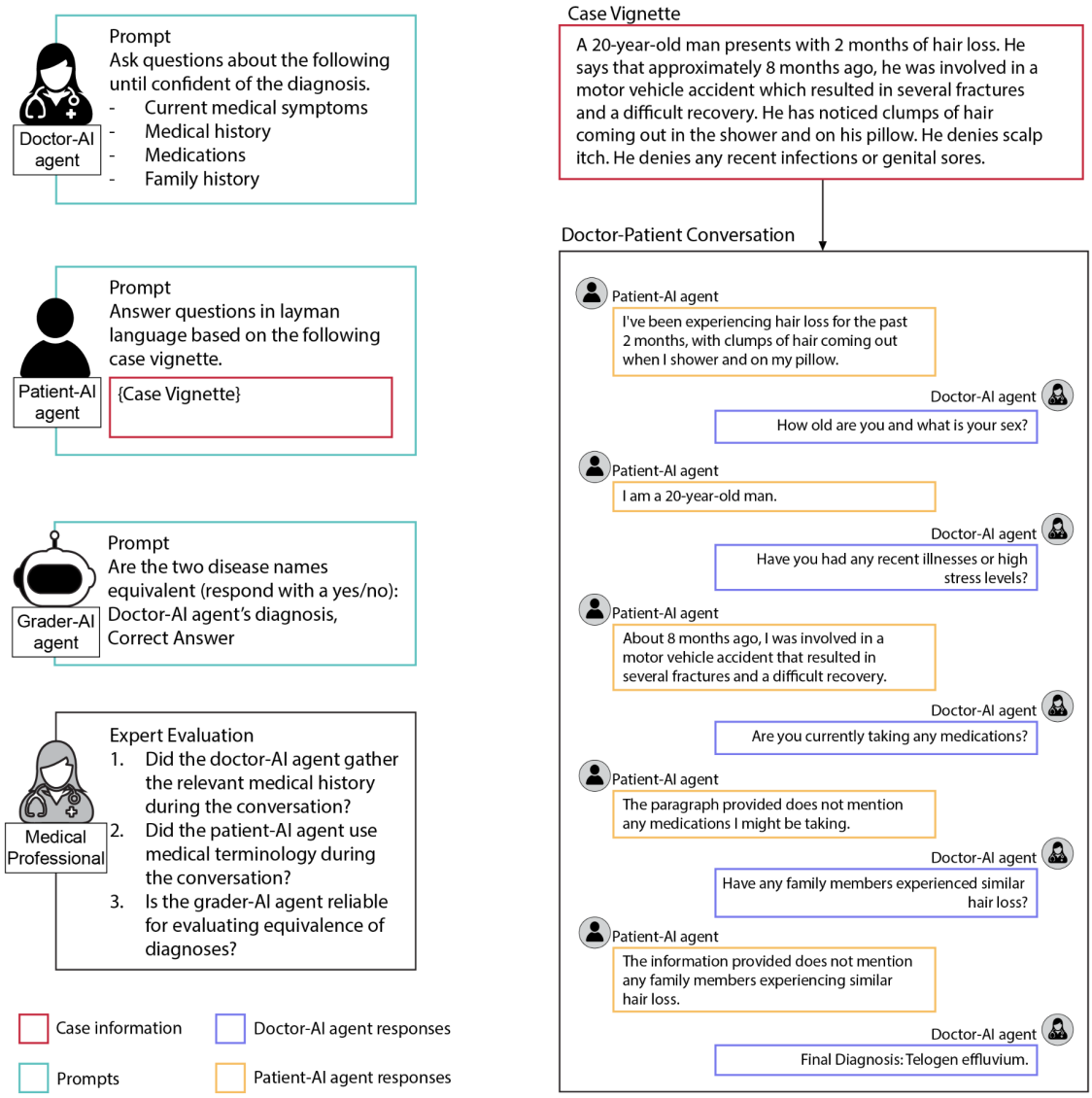


Figure 1: CRAFT-MD, a multi-agent conversational framework for clinical reasoning using LLMs. The doctor-AI agent is prompted to ask questions to extract the relevant medical history and arrive at a diagnosis. The patient-AI agent is provided a case-vignette and prompted to answer the doctor-AI agent's questions. A multi-turn doctor-patient conversation takes place, where the static case vignette is converted into an interactive conversation. A grader-AI agent assesses whether the final diagnosis made by the doctor-AI agent after the conversation matches with the correct answer. A medical professional qualitatively evaluates the conversation for assessing the performance of the doctor-AI, patient-AI and grader-AI agents.

Conversational Interactions Reduce Diagnostic Accuracy

We evaluated whether GPT-4 and GPT-3.5 maintain accuracy when making diagnoses through simulated doctor-patient conversations versus static case vignettes. Using the CRAFT-MD framework, we transformed vignettes into multi-turn conversations between the doctor-AI and patient-AI agents (Figure 2a, 2b; see Methods). If physical exam details were present in the original

vignette, they were provided after the conversation but before the diagnosis. This mirrors real clinical settings in which doctors integrate history gathering with exam findings.

For both GPT-4 and GPT-3.5, diagnostic accuracy dropped when using conversations versus vignettes with 4-choice multiple choice questions (MCQs) (Figure 2c; Supplementary Tables 1-6). The decrease was smaller for GPT-4 (0.919 to 0.854, adjusted p-value < 0.05) than GPT-3.5 (0.833 to 0.724, adjusted p-value < 0.01). To estimate a lower bound for accuracy, we evaluated performance using just physical exam details, which remained high (GPT-4 = 0.747, GPT-3.5 = 0.698; Extended Data Figure 1).

Multi-Turn Conversations Do Not Enhance Diagnostic Accuracy as Expected

Given that patient histories often contain subtle details revealed across an extended conversation, we expected multi-turn conversations to enhance diagnostic accuracy over single-turn interactions. Multi-turn conversations simulate the framework's doctor-AI agent engaging in back-and-forth questioning of the patient-AI agent until confident in a diagnosis. In contrast, single-turn interactions contain only the patient-AI agent's initial statement summarizing symptoms. Surprisingly, single-turn conversations did not increase accuracy for either GPT-4 or GPT-3.5 compared to multi-turn conversations (Figure 2c, Supplementary Tables 1-6). That is, the multi-turn structure did not enhance the integration of details from the interaction as expected. This reveals limitations in conversational reasoning capabilities.

Conversational Summarization Improves Accuracy

We developed a technique called Conversational Summarization to condense the multi-turn conversations into vignette-like summaries that consolidated all the details into one paragraph. For Conversational Summarization, we extracted all of the patient-AI agent's conversations from the full multi-turn conversation and summarized them into a coherent vignette. During this process, we used few-shot prompting to encourage the model to remove any artifacts from the conversational format such as references to "paragraphs" or "AI language models" (see Methods).

When the doctor-AI agent was provided with these summarized conversation vignettes instead of the multi-turn conversations, we did not observe a significant difference in accuracy for GPT-4 conversations (multi-turn = 0.854, summarized = 0.856). However, GPT-3.5's diagnostic accuracy improved significantly, increasing from 0.724 to 0.810 (adjusted p-value <0.001) (Figure 2c, Supplementary Tables 1-6, Extended Data Figure 1). This indicates that the prolonged, scattered conversation format was more difficult for GPT-3.5 to comprehend and reason through compared to having all the details presented together. Condensing conversations into vignette-like summaries may be a valuable technique to aid LLMs in integrating details from conversational interactions for improved reasoning.

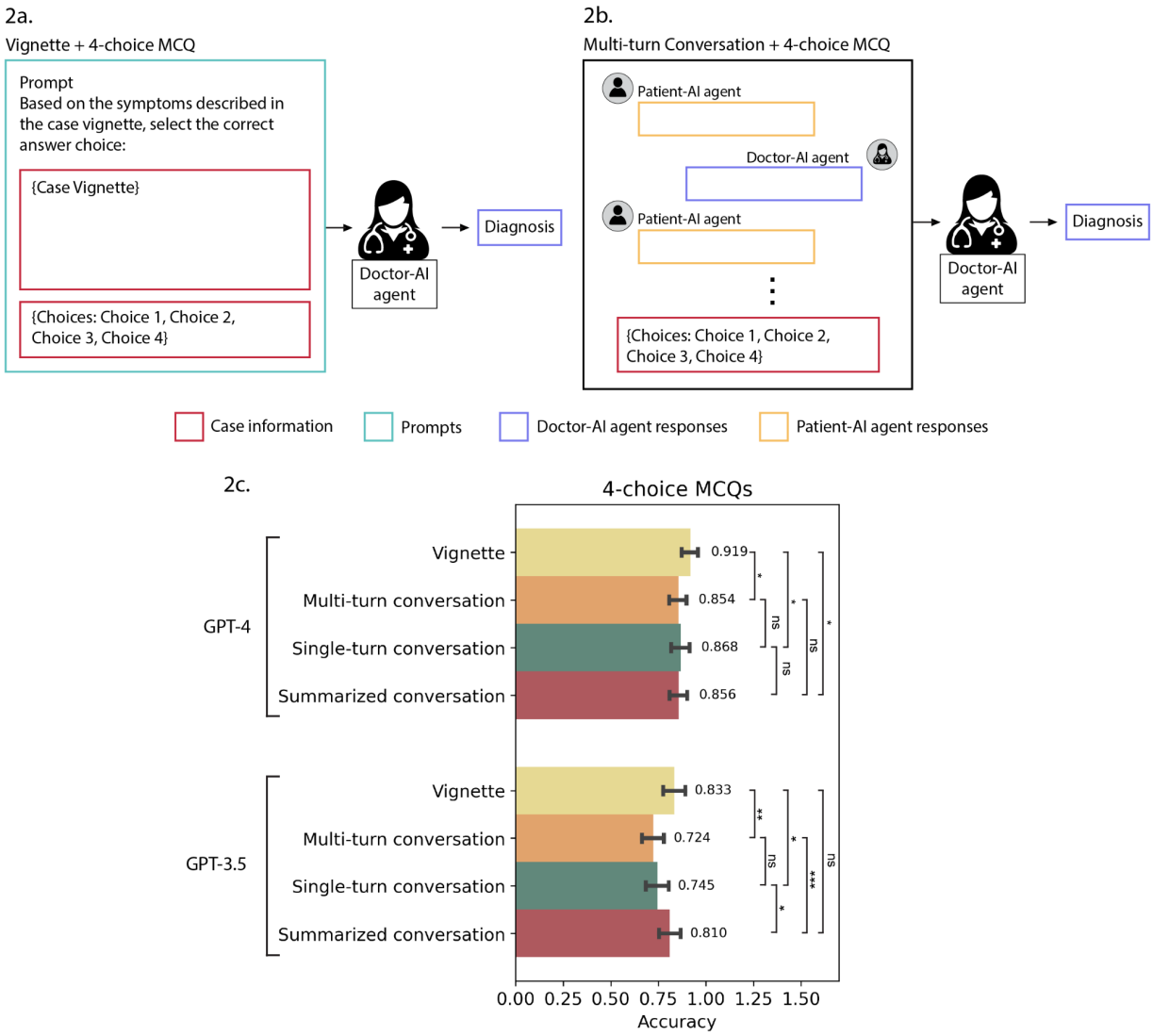


Figure 2: Effect of replacing case vignettes with simulated clinical interactions in Multiple Choice Questions (MCQs). **(a)** Experimental setup for diagnosis using case vignette, and **(b)** simulated doctor-patient conversations, followed by 4-choice MCQ. **(c)** Diagnostic accuracy using GPT-4 and GPT-3.5 for five experimental setups: vignette + 4-choice MCQs, multi-turn conversation + 4-choice MCQs, single-turn conversation + 4-choice MCQs, and summarized conversation + 4-choice MCQs. Error bars represent 95% confidence intervals, and numbers represent the mean accuracy (ns = not significant, * = <0.05, ** = <0.01, *** = <0.001).

Expert Evaluation Reveals Critical Reasoning Gaps

While accuracy metrics provide valuable insights, expert assessment reveals qualitative gaps not captured in scores alone. As part of the CRAFT-MD framework, dermatology experts annotated 120 GPT-4 and GPT-3.5 multi-turn conversations to assess the performance of the doctor-AI, patient-AI and grader-AI agents (see Methods). This highlighted several core deficiencies:

Incomplete Medical Histories: Human review categorized 26.6% and 30.0% of GPT-4 and GPT-3.5 conversations as eliciting incomplete histories through questioning (see Methods), struggling to extract all relevant details even when prompted (Extended Data Figure 2). With GPT-4, incorrect diagnoses remained similar between multi-turn (11.6%) and summarized (15%) conversations regardless of medical history completeness. However, for GPT-3.5, errors dropped from 31.6% multi-turn to 18.3% summarized, with a large decrease observed for complete histories (23.3% to 13.3%) (Extended Data Figure 2).

Non-Adherence to Prompts: An important part of realistic clinical conversations is responding in an accessible way using everyday language rather than complex medical terminology. However, analysis found non-adherence to this prompted behavior. Specifically, 13.3% of GPT-4 conversations and 10% of GPT-3.5 conversations were found to incorporate technical medical language in the patient-AI agent's responses (Extended Data Figure 2). Examples include use of terms such as “pearly papules”, “diffuse shoddy lymphadenopathy”, and “seizure prophylaxis”. While concerning, this did represent an improvement compared to the case vignettes, 100% of which contained specialized medical vocabulary. The models appear to have partially learned to rephrase the vignette details into more layman-friendly responses when explicitly prompted. Furthermore, the patient-AI agent used phrases such as “the paragraph provided does not mention” in its responses (Figure 1), despite being prompted to not break character and reveal the use of a case vignette for describing symptoms.

Reliability of Grader-AI: The grader-AI showed high correlation with dermatologist judgments on FRQ accuracy, as assessed by evaluating with three dermatology residents for different experimental setups. The average correlation was 0.94 for GPT-4 experiments and 0.912 for GPT-3.5 experiments (Supplementary Table 21; see Methods).

Qualitative Observations: The doctor-AI agent occasionally failed to inquire about prompted critical details like patient age, sex, treatments, and medications. In some cases, the doctor-AI agent did not pose relevant follow-up questions beyond the initial prompt that a human doctor likely would have; for instance, patient occupation and travel history can provide crucial clinical context but were often not elicited. The patient-AI agent sometimes refused to provide requested information, offered only partial responses, or took over questioning instead of answering. Because the patient-AI agent was designed to strictly adhere to the case vignette information, the agent often responded with "I don't have that information" to queries beyond the scope of the case vignette. While not prevalent, these examples of role confusion and selective unresponsiveness reveal gaps in comprehension and appropriate conversation dynamics.

Assessing Diagnostic Accuracy Beyond Multiple Choice

The multiple choice questions used in medical licensing exams do not reflect the open-ended diagnosis process in real clinical settings. To evaluate conversational reasoning in a more realistic

scenario as part of the CRAFT-MD framework, we expanded beyond the standard 4-choice MCQs in two ways.

First, we increased the number of MCQ answer choices to 381, encompassing all disease conditions in the dataset. This forced the doctor-AI agent to select from a larger set of choices (many-choice MCQ). Second, we removed answer choices entirely, forcing the doctor-AI agent to generate a diagnosis without predefined options (free-response questions (FRQ)) (Figure 3a, 3b; see Methods).

To evaluate the FRQ accuracy, the grader-AI agent categorized the doctor-AI agent's diagnosis into: single diagnosis, multiple diagnoses, or no diagnosis. For single diagnoses, the grader-AI agent matched the response to the ground truth using fuzzy matching to account for synonymous conditions. It was prompted using examples of disease synonyms like eczema/atopic dermatitis to recognize alternative terminologies (see Methods). A senior dermatology resident categorized each case vignette as having a single diagnosis, one most likely diagnosis, or many possible diagnoses. The diagnostic accuracy measurement allowed for alternative acceptable selections in the latter two categories.

Performance Trends Persist with Open-Ended Diagnosis

Expanding beyond multiple choice questions provided critical insights into how well multi-turn conversation performance trends generalized to FRQ diagnosis. Using the CRAFT-MD framework, we evaluated both increasing answer choices and removing them entirely.

Increasing Multiple Choice Options Significantly Reduces Accuracy: We observed that increasing the multiple choice options from 4 to 381 led to substantial declines in diagnostic accuracy for both models, in the multi-turn conversation setting. For GPT-4, accuracy dropped significantly from 0.854 with 4 choices to 0.572 with 381 choices (adjusted p-value < 0.001, Figure 3c, Supplementary Tables 7-12). This represents a large decrease of 0.282. Similarly, GPT-3.5 accuracy decreased significantly from 0.724 with 4 choices to 0.201 with 381 choices (adjusted p-value < 0.001, Figure 3c, Supplementary Tables 7-12). This constitutes a very substantial reduction of 0.523. The models clearly struggled to select the correct diagnosis when provided with a large set of choices versus the standard 4 choices.

Removing All Choices Has Divergent Effects: Interestingly, removing multiple choice options entirely through FRQs led to divergent results between the two models, in the multi-turn conversation setting. For GPT-4, FRQs further decreased accuracy significantly from 0.572 with many choices to 0.399 with no choices (adjusted p-value < 0.001, Figure 3c, Supplementary Tables 7-12). However, for GPT-3.5, FRQs increased accuracy significantly from 0.201 with many choices to 0.444 with no choices (adjusted p-value < 0.001, Figure 3c, Supplementary Tables 7-12). This difference could potentially be attributed to factors such as GPT-3.5 overfitting to the multiple choice format after extensive pre-training, challenges generating FRQ responses within context length constraints, or inherent

model architecture differences affecting few-shot learning. Further analysis is warranted to definitively determine the causes.

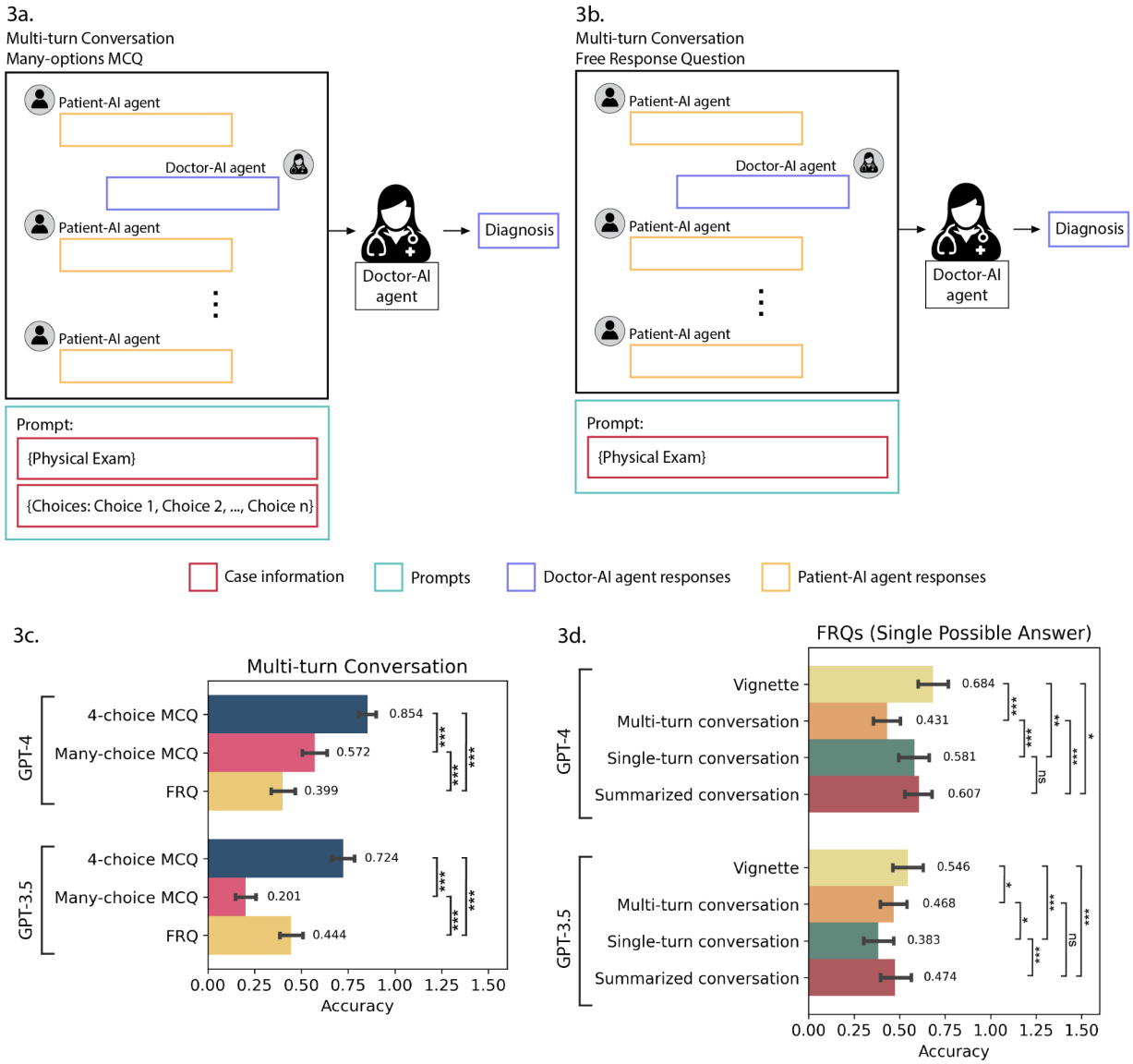


Figure 3: Effect of varying the number of answer choices in MCQs. **(a)** Experimental setup for diagnosis using simulated doctor-patient conversation followed by many-choice MCQ, and **(b)** FRQ. **(c)** Diagnostic accuracy for the experimental setups, for both GPT-4 and GPT-3.5, with multiplicity in answers accounted for in many-choice MCQs and FRQs. **(d)** Diagnostic accuracy using GPT-4 and GPT-3.5 for five experimental setups for only cases which have single possible answer: vignette + FRQs, summarized conversation + FRQs, multi-turn conversation + FRQs, single-turn conversation + FRQs, and physical exam + FRQs (ns = not significant, * = <0.05, ** = <0.01, *** = <0.001)

Conversational Interactions Continue Underperforming Vignettes: Importantly, conversational interactions continued to significantly underperform as compared to case vignettes, when cases with single possible answer were compared (Figure 3d, Supplementary Tables 13-18, Extended Data Figure

3). Multi-turn conversations decreased GPT-4 accuracy from 0.684 to 0.431 versus vignettes (adjusted p-value < 0.001, Figure 3d). GPT-3.5 accuracy changed from 0.546 to 0.468 multi-turn versus vignettes (adjusted p-value < 0.05, Figure 3d). The impact of conversation format depended on the model. Single turn conversations increased accuracy for GPT-4 from 0.431 to 0.581 (adjusted p-value < 0.001, Figure 3d) but reduced accuracy for GPT-3.5 from 0.468 to 0.383 (adjusted p-value < 0.05, Figure 3d). Summarized conversations increased GPT-4 accuracy from 0.431 to 0.607 (adjusted p-value < 0.001, Figure 3d) but did not affect GPT-3.5. This suggests that conversation structure can both improve and hinder accuracy depending on the model. Furthermore, for the dermatologist-annotated conversations, we found that requesting the top 3 differential diagnoses decreased the incorrect diagnoses from 60% to 33.3% for GPT-4, with a large decrease for conversations with complete histories (from 45% to 26.66%). The decrease in incorrect diagnoses was smaller for GPT-3.5 (top 1 = 56.66%, top 3 = 46.66%) (Extended Data Figure 4).

Evaluating Patient Self-Diagnosis by Removing Physical Exam Details

So far in the CRAFT-MD framework, the doctor-AI agent was provided with physical exam details from the case vignettes after conversing with the patient-AI agent, mirroring a healthcare visit in which some in-person evaluation occurs. However, with increasing use of AI chatbots for informal self-diagnosis prior to doctor visits[21], we investigated the impact of removing physical exam information from the framework across all formats (Figure 4a, 4b; see Methods).

For 4-choice MCQs, accuracy declined significantly from 0.854 to 0.774 for GPT-4 when eliminating physical exam details (adjusted p-value <0.001, Figure 4c, Supplementary Tables 19-20), and from 0.724 to 0.642 for GPT-3.5 (adjusted p-value <0.01, Figure 4c, Supplementary Tables 19-20). Similarly, for many-choice MCQs, GPT-4 accuracy dropped substantially from 0.572 with physical exam details to 0.476 without (adjusted p-value < 0.001, Figure 4c, Supplementary Tables 19-20). GPT-3.5 accuracy was 0.201 with physical exam information and 0.203 without this format. The reduction was also significant for FRQ, with accuracy decreasing from 0.399 to 0.324 for GPT-4 (adjusted p-value < 0.001, Figure 4c, Supplementary Tables 19-20) and 0.444 to 0.318 for GPT-3.5 (adjusted p-value < 0.001, Figure 4c, Supplementary Tables 19-20) when physical exam details were removed. Overall, eliminating physical exam details from the conversations significantly reduced diagnostic accuracy across all experimental formats for both models, which can be explained by the presence of classic exam style descriptors in the physical exams. This finding is also in line with the high accuracy achieved using the physical exam alone (Extended Data Figure 1, Extended Data Figure 3) in all of our experimental settings (4-choice MCQ, many-choice MCQ and FRQs). Overall, this highlights the value of in-person clinical evaluation or visual exam in telemedicine settings for optimal diagnosis even with AI assistance.

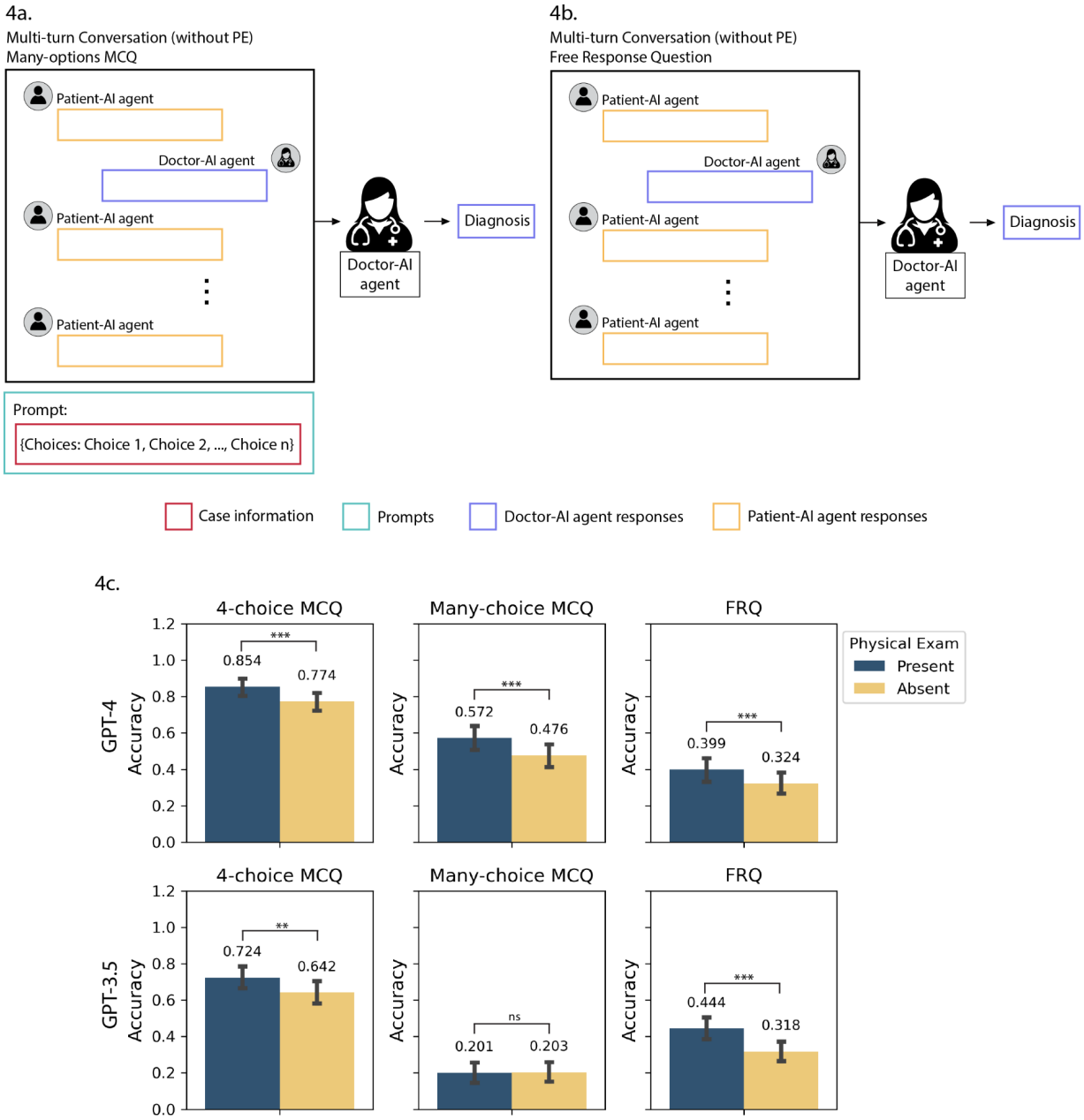


Figure 4: Conversation without physical exam (PE), in an informal pre-doctor's visit consultation setup. **(a)** Experimental setup for diagnosis using simulated doctor-patient conversation followed by many-choice MCQ, and **(b)** FRQ. **(c)** Diagnostic accuracy across GPT-3.5 and GPT-4 for conversation without physical exam (PE) for three experimental setups: 4-choice MCQs, many-choice MCQs, and FRQs (ns = not significant, * = <0.05, ** = <0.01, *** = <0.001).

Discussion

In this study, we introduced the CRAFT-MD framework, a novel multi-agent conversational framework for evaluating the clinical reasoning capabilities of Large Language Models (LLMs) like GPT-4 and GPT-3.5 within a medical conversational context. By transforming conventional multiple-choice vignettes into interactive and real-world doctor-patient dialogues, the framework offers a marked departure from traditional standardized evaluations. Our analysis using leading LLMs revealed several important insights:

Diagnostic Accuracy and Conversational Challenges: Converting static exam-style case vignettes into conversational interactions significantly reduced diagnostic accuracy for both models. Recent studies which show that LLMs like GPT-4 and GPT-3.5 can achieve high accuracy on medical cases[16–18] may present an overly optimistic outlook, overlooking the nuanced challenges associated with dynamic, medical conversations as opposed to static, clearly defined questions. We found that the models struggled to arrive at the right diagnosis when required to make the nuanced back-and-forth of gathering a history versus having all information consolidated in a single vignette. This observation is further deepened by expert analysis, which exposed critical gaps not captured in scores alone, including incomplete history gathering, non-adherence to layman language prompts, and unreliable responding. This qualitatively reveals deficiencies in comprehension, reasoning, and appropriate conversation dynamics compared to human experts. Together, these findings argue for a more cautious and nuanced understanding of the capabilities of LLMs in clinical settings. Rigorous analysis of actual model conversations, along with a recognition of the complex challenges of conversational reasoning, are essential to guide development and responsible deployment.

Multi-turn Conversations and Lack of Improvement: Contrary to expectations, multi-turn conversations did not enhance accuracy over single-turn interactions. The prolonged back-and-forth conversation format did not help the models better comprehend and utilize the details shared. This reveals limitations in simulated conversation reasoning capabilities compared to human experts who continually integrate new details. One explanation for this lack of improvement could be the models' struggle to effectively process and understand information presented over longer textual contexts[22]. In the context of multi-turn conversations, the scattered presentation of relevant details over longer conversational lengths could create a challenge in integrating information into a coherent understanding. Additionally, the presence of extraneous information and conversational noise could easily divert the models' attention from key symptoms and patient history. Further development of context comprehension and information integration is needed before deployment in real clinical conversations.

Conversational Summarization and Insights into Model Behavior: We introduced conversational summarization, a technique to condense the information of a conversation into a succinct summary. This method allowed for more focused interpretation and reasoning, evident in the performance of GPT-3.5, where conversational summarization enabled a more efficient processing of multi-turn dialogues, minimizing distractions and improving diagnostic accuracy. The observed increase in

accuracy using conversational summarization supports the notion that by condensing relevant information into a succinct and focused summary, the distraction from irrelevant details is minimized, aiding in the interpretation and reasoning processes. This insight into the behavior of the models also highlights the importance of mitigating length and distraction effects to aid conversational reasoning.

From Multiple Choice Options towards Open-ended Reasoning: Increasing multiple choice options from 4 to 381 substantially decreased accuracy, while removing choices[23–25] entirely had divergent effects depending on the model. This demonstrates the perils of evaluating models solely through constrained multiple choice questions that fail to reflect open-ended clinical decision-making. Traditional criteria have used these models' accuracy on medical licensing exams as a performance measure, often focused on standardized test questions with defined answer choices[16,26–30]. These benchmarks may offer an overly optimistic assessment of the abilities of LLMs to function effectively in complex, real-world medical scenarios where there are no multiple-choice answers. The variation in results according to task format in our study underscores the importance of varied and extensive testing methodologies like the CRAFT-MD framework for a comprehensive evaluation of LLMs. It highlights the need to move beyond static exams to free response conversational reasoning, reflecting a more authentic clinical decision-making process.

Potential Deployment Scenarios and the Importance of Physical Exams: Our findings point to potential deployment scenarios for LLMs in clinical settings. These models could serve as tools for collecting patient history prior to a doctor's visit, streamlining the process and saving time and human resources. Furthermore, they could also be used as helpful tools for medical education, enabling medical students to acquire critical skills of gathering patient history. These models also have the potential to assist in low-resource settings with unavailability of specialist doctors, by helping generate differential diagnoses and ultimately improve overall patient care by reducing the frequency of misdiagnoses. However, the observed reduction in diagnostic accuracy during conversational interactions highlights the need for improvements in the models' conversational reasoning capabilities before such deployment can occur. Additionally, without physical exam details, accuracy declined considerably across all experimental formats. This underscores the continued importance of in-person clinical evaluation even with AI assistance. Visual and physical assessment remains invaluable for optimal diagnosis. Future work on conversational agents should explore multimodal integration of verbal histories and visual exam findings[31]. The combined insights emphasize the potential and limitations of LLMs in clinical settings, advocating for a balanced approach that recognizes the value of human expertise and contextual information[32].

Limitations: Our study has certain limitations. First, the focus on skin disease, while offering an initial domain for rigorous evaluation given the contextual nuances involved, can be expanded to use a greater diversity of cases across other clinical concerns like hypertension, diabetes, respiratory infections, and mental health disorders. Incorporating a larger and more comprehensive set of cases could provide greater power to detect deficiencies and a more representative assessment of how conversational reasoning performance translates across different clinical scenarios. Second, the use of

simulated doctor-patient conversations, while an innovative technique, does not fully capture the dynamic intricacies and unpredictability of actual human dialog. Testing performance with real patients in clinics and actual providers in medical settings will be an important future direction to gain a more authentic evaluation of the models' capabilities. A significant concern arises from the potential for large pre-trained models like GPT-4 to memorize training dataset cases[6], coupled with the lack of transparency into GPT-4's full training data corpus[33]. While we developed 40 entirely new cases to mitigate this issue in our study, the memorization of specific cases or conversational patterns in the training data could still influence model responses in unpredictable ways, potentially skewing the results. This challenge highlights the urgent need for transparency from AI developers regarding the precise training methodologies and data utilized in model development. In addition, more rigorous control over training and testing datasets on the part of researchers could help mitigate such biases in evaluation results.

In conclusion, our study introduces CRAFT-MD, a novel framework that represents a significant departure from traditional methods of evaluating large language models (LLMs) in clinical contexts. Through the use of multiple AI agents[34], our approach uncovers new dimensions of understanding, revealing both the potential and the limitations of LLMs like GPT-4 and GPT-3.5. While the findings highlight areas for improvement, they also pave the way for a more nuanced and iterative evaluation process that resonates with the complexities of real-world clinical settings. This work not only emphasizes the need for responsible, comprehensive evaluation but also hints at a broader shift in perspective, guiding future research towards a more integrative and human-centric approach.

Methods

Dataset

The dataset consisted of 140 questions, each structured as a case vignette followed by four answer choices. Among these, we obtained 100 questions from a reputable online question bank[20] (https://dermnetnz.org/quizzes). Three dermatology residents (D1 = 10 questions, D2 = 10 questions, D3 = 20 questions) constructed an additional 40 questions. The dermatologists were shown the 100 questions from the online question bank and instructed to create similar questions with corresponding 4 answer choices, but covering different dermatology diseases to increase the diversity of our dataset. Together, these 140 questions encompassed a wide range of dermatological conditions. A dermatology resident (D3) rigorously evaluated all 140 case vignettes (without the associated 4 answer options) and categorized them as having a single diagnosis, one most likely diagnosis, or many possible diagnoses.

Accessing GPT-4 and GPT-3.5

We generated all the results in this paper using the March 14, 2023 version of GPT-4 ("gpt-4-0314") and the March 1, 2023 version of GPT-3.5 ("gpt-3.5-turbo-0301"). We accessed these models through the OpenAI API.

Multi-agent AI conversational framework

We introduced a multi-agent AI conversational framework comprising doctor-AI, patient-AI agent, and grader-AI agents along with medical experts. The doctor-AI and patient-AI agent were used to simulate conversations from each of the 140 case vignettes in our dataset. Conversations were repeated 10 times for each case vignette to model the variance in doctor-AI agent's and patient-AI agent's responses, and quantify the frequency with which the doctor-AI agent is able to arrive at a correct final diagnosis. The grader-AI agent assessed the diagnostic accuracy of doctor-AI agent's responses, and medical experts assessed each of the AI agents' performance. Conversations were set up in three ways - 4-choice MCQ, many-choice MCQ and FRQ. The conversation terminated when the doctor-AI agent's response contained the phrase 'Final Diagnosis'. Alternatively, the conversation was terminated if the doctor-AI agent's response did not contain a follow-up question (i.e., if there was no '?' present in the doctor-AI agent's responses, the conversation was terminated).

Doctor-AI agent

The doctor-AI agent was prompted to ask the patient-AI agent questions related to current symptoms, medical history of illness and medications, and relevant family history, if necessary. It keeps asking questions until it is confident of the diagnosis. The following prompt was used.

You are an AI doctor specializing in dermatology. Arrive at a differential diagnosis of a patient's medical condition. Ask simple 1 line questions, one question at a time. Only stop asking questions when you are 100% confident of the diagnosis, otherwise continue asking questions. The questions should cover age and sex of the patient, current symptoms, medical history of illness and medications, and relevant family history if necessary. Keep your responses very minimal and brief to not confuse the patient. When you arrive at the differential diagnosis, you must state 'Final Diagnosis:' in the beginning of your response, otherwise you will be penalized.

Patient-AI agent

The patient-AI agent was provided with a case vignette and tasked with answering follow-up questions posed by the doctor-AI agent. It was explicitly prompted to not reveal the entire contents of the paragraph at once and only answer the questions asked. Additionally, the patient-AI agent is incentivized to avoid creating new symptoms by imposing a negative penalty for doing so.

You are a patient. You do not have any medical knowledge. Based upon questions asked, you have to describe your symptoms from the following paragraph: <case_vignette>. Do not break character and reveal that you are describing symptoms from a paragraph. Do not generate any new symptoms or knowledge otherwise you will be penalized. Do not reveal more knowledge than what the question asks. Keep your answer to only 1 sentence. Simplify terminology used in the given paragraph to layman language.

Grader-AI agent

We used a grader-AI agent to quantify the diagnostic accuracy for many-choice MCQ and FRQ experiments. In all experiments, GPT-4 was used for the grader-AI agent. For the conversation + FRQ experiments, the grader-AI agent first categorized the doctor-AI agent's final diagnosis according to the following three categories: (i) single diagnosis, (ii) multiple diagnoses, and (iii) no diagnosis. We estimated the error rate for this step through manual verification to be <0.5% (1 mistake in ~200 conversations). We additionally manually categorized incomplete conversations (multi-turn, single-turn, summarized) into the 'no diagnosis' category. The following prompt was used:

Identify and return the dermatology diagnosis name from the given paragraph. If there are more than one diagnosis present, return 'Multiple'. If there are no diagnoses present, then return 'None'. Do not explain.
Paragraph : <insert doctor-AI agent's diagnosis containing response>.

For the doctor-AI agent's responses which contained a single diagnosis, the grader-AI agent matched the diagnosis to the correct answer, accounting for alternative medical terminologies for the conditions. The conversations with 'no diagnosis' and 'multiple diagnosis' responses were assigned accuracy of 0.

For comparing between doctor-AI agent's response and correct answer for all experiments (vignette + 4-choice MCQ, vignette + many-choice MCQ, vignette + FRQ, multi-turn conversation + 4-choice MCQ, multi-turn conversation + many-choice MCQ, multi-turn conversation + FRQ, single-turn conversation + 4-choice MCQ, single-turn conversation + FRQ, summarized conversation + 4-choice MCQ, summarized conversation + FRQ, physical exam (PE) + 4-choice MCQ, physical exam (PE) + FRQ, multi-turn conversation (without PE) + 4-choice MCQ, multi-turn conversation (without PE) + many-choice MCQ, multi-turn conversation (without PE) + FRQ), the grader-AI agent was specifically prompted to account for alternative medical terminologies. The following prompt was used:

> *Are the two dermatology conditions the same or have synonymous names of diseases? Respond with a yes/no. Do not explain.*
> *Example: Choice 1: eczema Choice 2: eczema They are the same, so return yes.*
> *Example: Choice 1: wart Choice 2: wart They are the same, so return yes.*
> *Example: Choice 1: eczema Choice 2: onychomycosis They are different, so return no.*
> *Example: Choice 1: wart Choice 2: alopecia areata They are different, so return no.*
> *Example: Choice 1: eczema Choice 2: atopic dermatitis They are synonymous, so return yes.*
> *Example: Choice 1: benign nevus Choice 2: mole They are synonymous, so return yes.*
> *Example: Choice 1: toe nail fungus Choice 2: onychomycosis They are synonymous, so return yes.*
> *Example: Choice 1: wart Choice 2: verruca vulgaris They are synonymous, so return yes.*
>
> *Choice 1: <insert extracted disease name>*
> *Choice 2: <insert correct answer>*

Experimental Setups

Varying format of presented medical information

Case Vignette

The case vignette was structured as a paragraph, and contained all or a subset of the following information: age and sex of the patient, current symptoms, medical history of illness and medications, relevant family history, and physical exam.

Physical Exam

In the case vignette, relevant information pertaining to the physical exam, such as "physical examination," "laboratory tests," or any explicit mention of examination results, was manually extracted. 'None' was stored in case no physical exam was present.

Multi-turn conversations

The multi-agent AI conversational framework was used to generate a multi-turn conversation between the doctor-AI agent and patient-AI. The conversation terminated when the doctor-AI agent's response contained the phrase 'Final Diagnosis'. Alternatively, the conversation was terminated if the doctor-AI agent's response did not contain a follow-up question.

Single-turn conversations

The patient-AI agent's initial symptom summary (i.e., first dialogue in a multi-turn conversation) was used as a single-turn conversation.

Summarized conversations

These were generated using the technique Conversational Summarization. All the patient-AI agent's dialogues were extracted from the GPT-4 and GPT-3.5 multi-turn conversations, and artifacts such as references to "paragraphs" or "AI language model" were removed. GPT-3.5 was used in this process, and few-shot prompting was used to improve the model output. Different prompts were used for GPT-4 and GPT-3.5 conversations due to the differing nature of artifacts in the two models' outputs.

The following prompt was used for converting all GPT-4 multi-turn conversations into summaries:

> *Convert the following vignette into 3rd person. It contains information from a patient describing their medical symptoms. Do not use the words 'AI language model' or references to the 'paragraph' mentioned in the vignette. Do not create new information. - <insert patient-AI agent's dialogues>*
>
> *For example:*
>
> *Original Vignette - 'I have a hard, yellowish-white horn-like growth on my head that started as a small, hard bump a few months ago and has grown bigger, and hurts when accidentally hit. I am a 60-year-old man. I'm sorry, but I do not have that information as I am an AI language model. I am an AI language model, and based on the given paragraph, there is no information available regarding previous growths or history of skin cancer or other cancers. As an AI language model, it is not mentioned in the given paragraph whether I have tried any treatments or remedies for the horn-like growth. The growth has a very firm texture, and it is located superficially on the skin, just beside the midline on the superior aspect of the skull. As described in the given paragraph, the patient denies any pain at rest, but experiences pain when the lesion is accidentally struck; there is no mention of itch or bleeding. As an AI language model, there is no information available in the given paragraph regarding any comorbidities or previous skin disease diagnoses. As mentioned in the given paragraph, the patient noticed a small, hard lesion a few months ago, which has grown progressively larger from that time to now. However, there is no mention of any changes in color or shape over time.'*

Converted Vignette - 'A 60-year-old man reports a hard, yellowish-white horn-like growth on his head that started as a small, hard bump a few months ago. He says that it has grown bigger and hurts when accidently hit. He says that there are no previous growths or history of skin cancer or other cancers. He does not know if he has tried any treatments or remedies for the horn-like growth. He says that the The growth has a very firm texture, and it is located superficially on the skin, just beside the midline on the superior aspect of the skull. He denies any pain at rest, but experiences pain when the lesion is accidentally struck. There is no mention of itch or bleeding. He does not have information regarding any comorbidities or previous skin disease diagnoses. There is no mention of any changes in color or shape over time.'

All the summarized conversations went through manual evaluation to remove remaining references to artifacts such as 'paragraph' and 'AI language model'. For the case used as an example in the prompt above (public_case09), a different example from public_case02 was used in the prompt to generate the summarized conversation.

The following prompt was used for converting all GPT-3.5 multi-turn conversations into summaries:

Convert the vignette into 3rd person. It contains information from a patient describing their medical symptoms. Do not use the words 'AI language model' or references to the 'paragraph' mentioned in the vignette. Do not create new information. - < insert patient-AI agent's dialogues>

For example:

Original Vignette - 'I have noticed that the skin around my hands, forearms, and face has been getting thicker over the past several months. I am 47 years old. I am a woman. I didn't mention any joint pain or stiffness in my described symptoms. I have no family history of autoimmune disease. I have not mentioned having any red or purple patches on my skin. I haven't mentioned any rapid weight gain in my face, arms, or legs. Yes, I am experiencing thickening of my skin around my hands, forearms, and face. I am not taking any medications.'

Converted Vignette - 'A 47-year-old woman presents to the clinic with several concerns. Over the past several months, she has experienced thickening of her skin around her hands, forearms, and face. She does not have any joint pain, stiffness, or family history of autoimmune diseases. She also does not have any red or purple patches on her skin. She denies any rapid weight gain on her face, arms or legs. She denies taking any medications.'

All the summarized conversations went through manual evaluation to remove remaining references to artifacts such as 'paragraph' and 'AI language model'. For the case used as an example in the

prompt above (public_case75), a different example from public_case02 was used in the prompt to generate the summarized conversation.

Varying number of answer choices

4-choice and many-choice Multiple Choice Questions

For both 4-choice and many-choice MCQs, the doctor-AI agent was provided with the answer choices after case vignette, or conversation (multi-turn, single-turn, summarized), or physical exam. In case of multi-turn conversations, the response containing the final diagnosis was removed.

For vignette, single-turn conversation or summarized conversation, followed by 4-choice or many-choice MCQ, the following prompt was used:

> *You are an AI doctor specializing in dermatology. You are given the patient's symptoms and a list of possible diagnosis choices. Only one of the choices is correct. Select the correct choice, and give the answer as a short response. Do not explain.*
> *Symptoms: <insert symptoms>*
> *Choices: <insert comma separated list of answer choices>*

For physical exam followed by 4-choice or many-choice MCQ, the following prompt was used:

> *You are an AI doctor specializing in dermatology. You are given the patient's age, sex, physical examination result, and a list of possible diagnosis choices. Only one of the choices is choices. Select the correct choice, and give the answer as a short response. Do not explain.*
> *Age: <insert age>*
> *Sex: <insert sex>*
> *Physical Examination: <insert physical examination>*
> *Choices: <insert comma separated list of answer choices>*

For multi-turn conversation followed by 4-choice or many-choice MCQ, the following prompt was used:

> *Based on the patient's symptoms described above and a list of possible diagnosis choices, select the correct choice and give the answer as a short response. Do not explain*
> *Choices: <insert comma separated list of answer choices>*

The doctor-AI agent refused to select diagnosis from one of the choices in cases where the multi-turn conversations did not provide sufficient information. In such cases, the final diagnosis was marked as incorrect.

Free-Response Questions

The doctor-AI agent was presented with a case vignette, or conversation (multi-turn, single-turn, summarized), or physical examination, and asked to give a diagnosis. Except for multi-turn conversations, the final response of the doctor-AI agent containing the diagnosis was removed.

For vignette + FRQ, single-turn conversation + FRQ, and summarized conversation + FRQ, the following prompt was used:

> *You are an AI doctor specializing in dermatology. You are given the patient's symptoms. Give the name of the correct diagnosis as a short answer. Do not explain.*
> *Symptoms: <insert symptoms>*

For physical exam + FRQ, the following prompt was used:

> *You are an AI doctor specializing in dermatology. You are given the patient's age, sex and physical examination result. Give the name of the correct diagnosis as a short answer. Do not explain.*
> *Age: <insert age>*
> *Sex: <insert sex>*
> *Physical Examination: <insert physical examination>*

For multi-turn conversation + FRQ, the doctor-AI agent's initial prompt contained instructions for giving a diagnosis, therefore no further prompting was required.

Human Evaluation

To assess the performance of doctor-AI, patient-AI and grader-AI agents, expert evaluation was done by dermatology residents (D1, D3, D4).

Doctor-AI agent's performance: 120 multi-turn conversations, 60 generated by GPT-4 and 60 by GPT-3.5, were also evaluated for the presence of complete medical history. A senior dermatology resident (D3) annotated the conversations for the presence or absence of important information present in the case vignette required for arriving at the correct diagnosis.

Patient-AI agent's performance: A senior dermatology resident (D3) assessed each of the 120 multi-turn conversations for the presence or absence of medical terminology. Other qualitative observations were also recorded for each conversation.

Grader-AI agent's performance: The correlation between accuracies of the doctor-AI agent as annotated by grader-AI and dermatology residents was compared. To assess this, three experiments (vignette + FRQ, multi-turn conversation + FRQ, multi-turn conversation without physical exam +

FRQ) were simultaneously annotated by D1, D3 and D4. Only public cases (n=100) from the dataset were used.

Statistical Tests

P-values were computed using the bootstrap method. To compare the experiment arms, samples were drawn with replacement from each arm, and the difference in means was estimated. This process was repeated 10,000 times to generate a distribution of differences. The p-value was then calculated based on the number of bootstrap samples that had a difference in mean greater than or equal to the observed statistic (original difference in means of the two experiment arms), considering both tails. To control the family wise error rate, final reported p-values were adjusted using the Holm-Bonferroni correction method (see Code Availability).

Correlations

Spearman correlation was used for quantifying the concordance between dermatologists' and grader-AI agent's annotations. 'spearmanr' from scipy.stats was used to calculate the correlation value, and the associated p-value (see Code Availability).

Data Availability

Data used in the study is available on the following repository:
https://github.com/rajpurkarlab/craft-md

Code Availability

All code for reproducing our analysis is available on the following repository:
https://github.com/rajpurkarlab/craft-md

Author information

These authors contributed equally: Shreya Johri, Jaehwan Jeong.
These authors share senior authorship: Roxana Daneshjou, Pranav Rajpurkar.

Authors and Affiliations

Department of Biomedical Informatics, Harvard Medical School, Boston, United States
Shreya Johri, Jaehwan Jeong, Pranav Rajpurkar

Department of Biomedical Data Science, Stanford University, Stanford, United States
Roxana Daneshjou

Department of Dermatology, Stanford University, Stanford, United States
Zhuo Ran Cai, Roxana Daneshjou

Department of Computer Science, Stanford University, Stanford, United States
Jaehwan Jeong

Medstar Georgetown University Hospital/Washington Hospital Center, Department of Dermatology, Washington, DC, United States
Benjamin A. Tran

Department of Dermatology, Northwestern University, Chicago, IL, United States
Daniel I. Schlessinger

Division of Dermatology, David Geffen School of Medicine at the University of California, Los Angeles, California, United States
Shannon Wongvibulsin

Contributions

P.R. and R.D. conceived the study. S.J. and J.J. planned and executed the experiments and data analysis. B.A.T., D.I.S., and Z.R.C. created new case vignettes for the dataset. B.A.T., D.IS., and S.W. interpreted dermatologic results. S.J., J.J., R.D., and P.R. contributed to the interpretation of findings. S.J. and P.R. drafted the manuscript. All authors provided critical feedback and substantially contributed to the revision of the manuscript. All authors read and approved the manuscript.

Corresponding author

Correspondence to Pranav Rajpurkar, PhD (pranav_rajpurkar@hms.harvard.edu)

Ethics declarations

Competing interests

D.I.S. is the co-founder of FixMySkin Healing Balms, a shareholder in Appiell Inc., and a consultant with LuminDx. R.D. reported receiving personal fees from DWA, personal fees from Pfizer, personal fees from L'Oreal, personal fees from VisualDx, stock options from MDAlgorithms and Revea outside the submitted work, and a patent for TrueImage pending.

References

1. Lasser, K. E., Himmelstein, D. U. & Woolhandler, S. Access to Care, Health Status, and Health

Disparities in the United States and Canada: Results of a Cross-National Population-Based Survey. *Am. J. Public Health* (2011) doi:10.2105/AJPH.2004.059402.

2. Irving, G. *et al.* International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ Open* **7**, e017902 (2017).
3. Wong, J. L. C., Vincent, R. C. & Al-Sharqi, A. Dermatology consultations: how long do they take? *Future Hosp J* **4**, 23–26 (2017).
4. Shaver, J. The State of Telehealth Before and After the COVID-19 Pandemic. *Prim. Care* **49**, 517–530 (2022).
5. Bubeck, S. *et al.* Sparks of Artificial General Intelligence: Early experiments with GPT-4. (2023).
6. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. (2023).
7. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* 1–9 (2023).
8. Sarraju, A. *et al.* Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA* **329**, 842–844 (2023).
9. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
10. Lee, P., Bubeck, S. & Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).
11. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
12. Ayers, J. W. *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
13. Au Yeung, J. *et al.* AI chatbots not yet ready for clinical use. *Front Digit Health* **5**, 1161098 (2023).

14. Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine* **6**, 1–10 (2023).

15. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and Adoption of Large Language Models in Medicine. *JAMA* (2023) doi:10.1001/jama.2023.14217.

16. Ali, R. *et al.* Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery* (2023) doi:10.1227/neu.0000000000002551.

17. Fijačko, N., Gosak, L., Štiglic, G., Picard, C. T. & John Douma, M. Can ChatGPT pass the life support exams without entering the American heart association course? *Resuscitation* **185**, 109732 (2023).

18. Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* **2**, e0000198 (2023).

19. Dermatology in primary care: Prevalence and patient disposition. *J. Am. Acad. Dermatol.* **45**, 250–255 (2001).

20. Quizzes. https://dermnetnz.org/quizzes.

21. Shahsavar, Y. & Choudhury, A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Hum Factors* **10**, e47564 (2023).

22. Liu, N. F. *et al.* Lost in the Middle: How Language Models Use Long Contexts. (2023).

23. Liu, X. *et al.* Performance of ChatGPT on Clinical Medicine Entrance Examination for Chinese Postgraduate in Chinese. *medRxiv* 2023.04.12.23288452 (2023) doi:10.1101/2023.04.12.23288452.

24. Ueda, D. *et al.* Evaluating GPT-4-based ChatGPT's Clinical Potential on the NEJM Quiz. *medRxiv* 2023.05.04.23289493 (2023) doi:10.1101/2023.05.04.23289493.

25. Strong, E. *et al.* Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. *JAMA Intern. Med.* (2023) doi:10.1001/jamainternmed.2023.2909.

26. Takagi, S., Watari, T., Erabi, A. & Sakaguchi, K. Performance of GPT-3.5 and GPT-4 on the

Japanese Medical Licensing Examination: Comparison Study. *JMIR Med Educ* **9**, e48002 (2023).

27. Lin, J. C., Younessi, D. N., Kurapati, S. S., Tang, O. Y. & Scott, I. U. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. *Eye* (2023) doi:10.1038/s41433-023-02564-2.
28. Giannos, P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurol Open* **5**, e000451 (2023).
29. Moshirfar, M., Altaf, A. W., Stoakes, I. M., Tuttle, J. J. & Hoopes, P. C. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. *Cureus* **15**, e40822 (2023).
30. Angel, M., Patel, A., Alachkar, A. & Baldi, P. Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Pharmacy: A Comparative Study on the NAPLEX Exam. *bioRxiv* 2023.06.07.544055 (2023) doi:10.1101/2023.06.07.544055.
31. Rajpurkar, P. & Lungren, M. P. The Current and Future State of AI Interpretation of Medical Images. *N. Engl. J. Med.* **388**, 1981–1990 (2023).
32. Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. *Combining human expertise with artificial intelligence: Experimental evidence from radiology*. http://www.nber.org/papers/w31422.pdf (2023) doi:10.3386/w31422.
33. OpenAI. GPT-4 Technical Report. (2023).
34. de Zarzà, I., de Curtò, J., Roig, G., Manzoni, P. & Calafate, C. T. Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs. *Electronics* **12**, 2722 (2023).

Extended Data

Contents

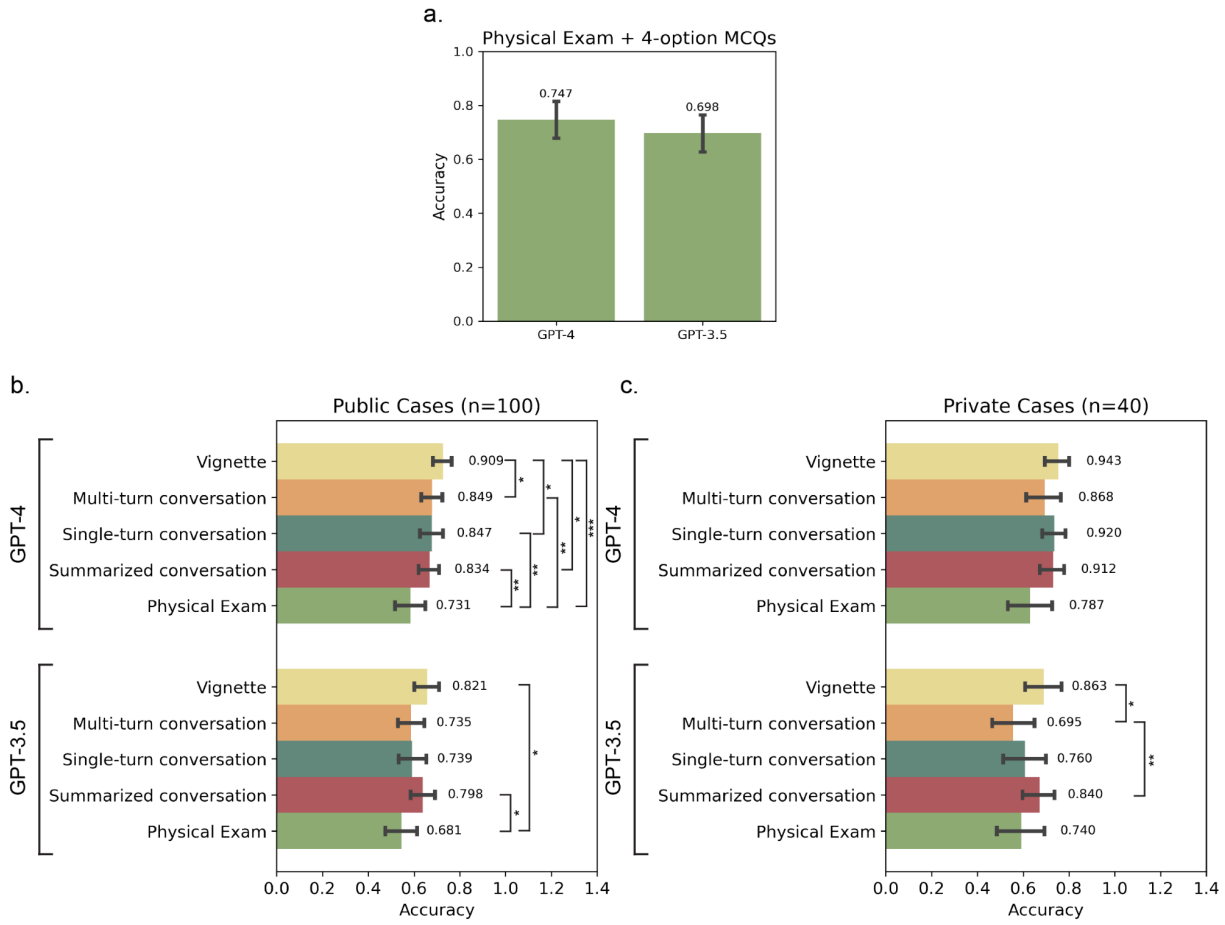
Extended Data Figure 1: (a) Diagnostic accuracy for physical exam followed by 4-choice MCQs. **(b, c)** Diagnostic accuracy using GPT-4 and GPT-3.5, separated by public and private cases in the dataset, for five experimental setups: vignette + 4-choice MCQs, multi-turn conversation + 4-choice MCQs, single-turn conversation + 4-choice MCQs, summarized conversation + 4-choice MCQs, and physical exam + 4-choice MCQs. Error bars represent 95% confidence intervals, and numbers represent the mean accuracy. Statistically significant results have been annotated with brackets (* = <0.05, ** = <0.01, *** = <0.001).

Extended Data Figure 2: (a) Fraction of annotated GPT-4 and GPT-3.5 conversations with use of medical terminology by the patient-AI agent. Distribution of annotated GPT-4 **(b, c)** and GPT-3.5 **(d, e)** conversations for completeness of medical history extracted by the doctor-AI agent (evaluated by a senior dermatology resident) and the correctness of diagnosis (evaluated by grader-AI) in 4-choice MCQ setup.

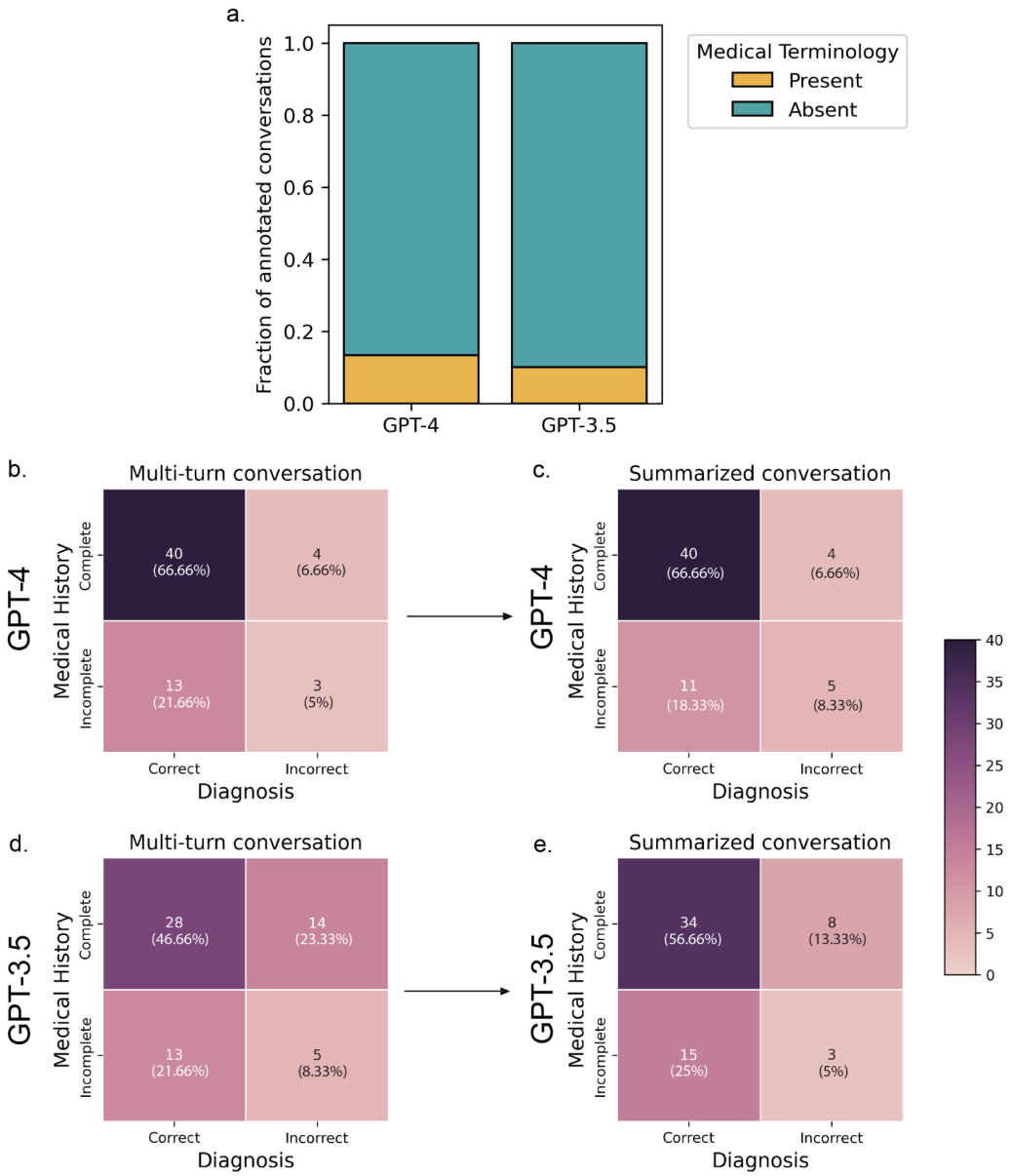
Extended Data Figure 3: (a, b) Diagnostic accuracy using GPT-4 and GPT-3.5, separated by public and private cases in the dataset, when multi-turn conversation is followed by 4-choice, many-choice and no-choice MCQs. **(c, d)** Diagnostic accuracy using GPT-4 and GPT-3.5, separated by public and private cases in the dataset, for five experimental setups: vignette + FRQs, multi-turn conversation + FRQs, single-turn conversation + FRQs, summarized conversation + FRQs, and physical exam + FRQs. Error bars represent 95% confidence intervals, and numbers represent the mean accuracy. Statistically significant results have been annotated with brackets (* = <0.05, ** = <0.01, *** = <0.001). **(e)** Diagnostic accuracy for combined dataset using physical exam followed by FRQs.

Extended Data Figure 4: Distribution of annotated GPT-4 **(a, b, c)** and GPT-3.5 **(d, e, f)** conversations for completeness of medical history extracted by the doctor-AI agent (evaluated by a senior dermatology resident) and the correctness of diagnosis (evaluated by grader-AI) in **(a, d)** multi-turn conversation + no-choice MCQs. **(b, e)** summarized conversation + no-choice MCQs, and **(c, f)** multi-turn conversation + no-choice MCQs, when the doctor-AI agent is prompted to give the top-3 possible diagnoses instead of top-1.

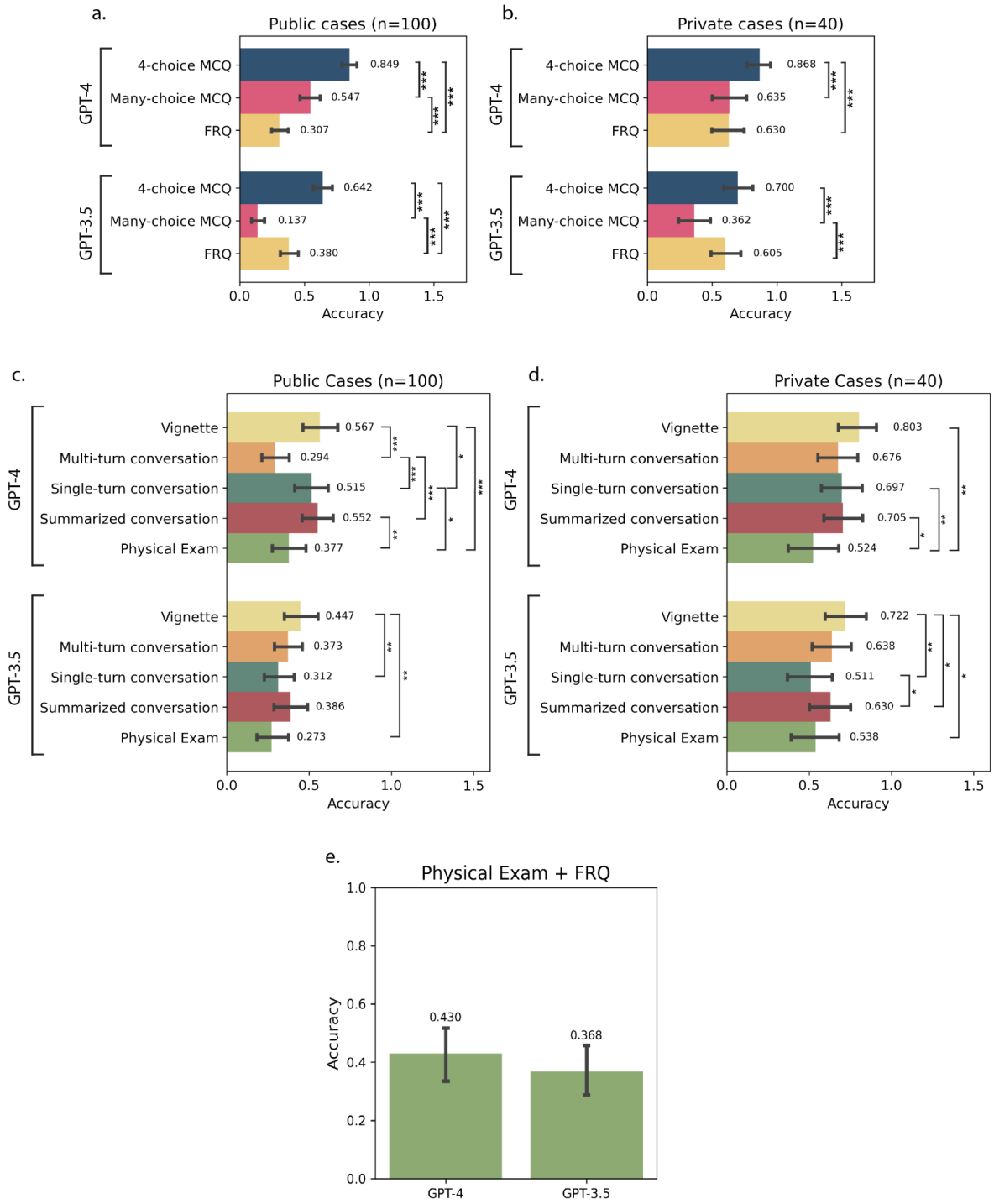
Extended Data Figures



Extended Data Figure 1: (a) Diagnostic accuracy for physical exam followed by 4-choice MCQs. **(b, c)** Diagnostic accuracy using GPT-4 and GPT-3.5, separated by public and private cases in the dataset, for five experimental setups: vignette + 4-choice MCQs, multi-turn conversation + 4-choice MCQs, single-turn conversation + 4-choice MCQs, summarized conversation + 4-choice MCQs, and physical exam + 4-choice MCQs. Error bars represent 95% confidence intervals, and numbers represent the mean accuracy. Statistically significant results have been annotated with brackets (* = <0.05, ** = <0.01, *** = <0.001).

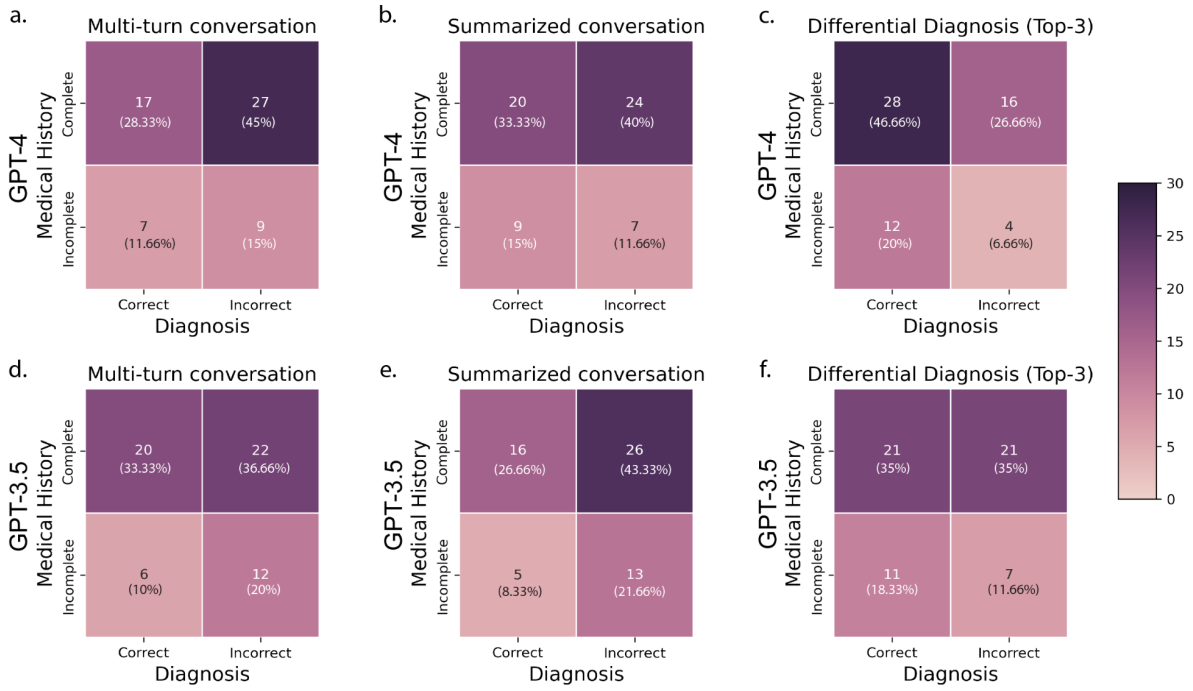


Extended Data Figure 2: (a) Fraction of annotated GPT-4 and GPT-3.5 conversations with use of medical terminology by the patient-AI agent. Distribution of annotated GPT-4 (b, c) and GPT-3.5 (d, e) conversations for completeness of medical history extracted by the doctor-AI agent (evaluated by a senior dermatology resident) and the correctness of diagnosis (evaluated by grader-AI) in 4-choice MCQ setup.



Extended Data Figure 3: (a, b) Diagnostic accuracy using GPT-4 and GPT-3.5, separated by public and private cases in the dataset, when multi-turn conversation is followed by 4-choice, many-choice and no-choice MCQs. (c, d) Diagnostic accuracy using GPT-4 and GPT-3.5, separated by public and private cases in the dataset, for five experimental setups: vignette + FRQs, multi-turn conversation + FRQs, single-turn conversation + FRQs, summarized conversation + FRQs, and physical exam + FRQs. Error bars represent 95% confidence intervals,

and numbers represent the mean accuracy. Statistically significant results have been annotated with brackets (* = <0.05, ** = <0.01, *** = <0.001). (e) Diagnostic accuracy for combined dataset using physical exam followed by FRQs.



Extended Data Figure 4: Distribution of annotated GPT-4 **(a, b, c)** and GPT-3.5 **(d, e, f)** conversations for completeness of medical history extracted by the doctor-AI agent (evaluated by a senior dermatology resident) and the correctness of diagnosis (evaluated by grader-AI) in **(a, d)** multi-turn conversation + no-choice MCQs. **(b, e)** summarized conversation + no-choice MCQs, and **(c, f)** multi-turn conversation + no-choice MCQs, when the doctor-AI agent is prompted to give the top-3 possible diagnoses instead of top-1.

Supplementary Information

Contents

Supplementary Table 1 : Statistical significance for all cases (n=140) between different pairs of GPT-4 experiments for 4-choice MCQ.

Supplementary Table 2 : Statistical significance for all public cases (n=100) between different pairs of GPT-4 experiments for 4-choice MCQ.

Supplementary Table 3 : Statistical significance for all private cases (n=40) between different pairs of GPT-4 experiments for 4-choice MCQ.

Supplementary Table 4 : Statistical significance for all cases (n=140) between different pairs of GPT-3.5 experiments for 4-choice MCQ.

Supplementary Table 5 : Statistical significance for public cases (n=100) between different pairs of GPT-3.5 experiments for 4-choice MCQ.

Supplementary Table 6 : Statistical significance for private cases (n=40) between different pairs of GPT-3.5 experiments for 4-choice MCQ.

Supplementary Table 7 : Statistical significance for all cases (n=140) between different pairs of GPT-4 experiments for multi-turn conversations.

Supplementary Table 8 : Statistical significance for public cases (n=100) between different pairs of GPT-4 experiments for multi-turn conversations.

Supplementary Table 9 : Statistical significance for private cases (n=40) between different pairs of GPT-4 experiments for multi-turn conversations.

Supplementary Table 10 : Statistical significance for all cases (n=140) between different pairs of GPT-3.5 experiments for multi-turn conversations.

Supplementary Table 11 : Statistical significance for public cases (n=100) between different pairs of GPT-3.5 experiments for multi-turn conversations.

Supplementary Table 12 : Statistical significance for private cases (n=40) between different pairs of GPT-3.5 experiments for multi-turn conversations.

Supplementary Table 13 : Statistical significance for all cases (n=140) between different pairs of GPT-4 experiments for FRQs.

Supplementary Table 14 : Statistical significance for public cases (n=100) between different pairs of GPT-4 experiments for FRQs.

Supplementary Table 15 : Statistical significance for private cases (n=40) between different pairs of GPT-4 experiments for FRQs.

Supplementary Table 16 : Statistical significance for all cases (n=140) between different pairs of GPT-3.5 experiments for FRQs.

Supplementary Table 17 : Statistical significance for public cases (n=100) between different pairs of GPT-3.5 experiments for FRQs.

Supplementary Table 18 : Statistical significance for private cases (n=40) between different pairs of GPT-3.5 experiments for FRQs.

Supplementary Table 19 : Statistical significance for all cases (n=140) between different pairs of GPT-4 experiments for multi-turn conversations with and without physical exam (PE).

Supplementary Table 20 : Statistical significance for all cases (n=140) between different pairs of GPT-3.5 experiments for multi-turn conversations with and without physical exam (PE).

Supplementary Table 21 : Correlation between grader-AI and dermatologists' accuracy annotation of the doctor-AI agent (PE = physical exam, FRQ = Free Response Question).

Supplementary Tables

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	2.00E-04	0.0018
Physical Exam	Single-turn conversation	2.00E-04	0.0018
Physical Exam	Summarized conversation	2.00E-04	0.0018
Physical Exam	Vignette	1.00E-04	0.001
Multi-turn conversation	Single-turn conversation	0.4303	1
Multi-turn conversation	Summarized conversation	0.8839	1
Multi-turn conversation	Vignette	0.0027	0.0135
Single-turn conversation	Summarized conversation	0.3676	1
Single-turn conversation	Vignette	0.0116	0.0464
Summarized conversation	Vignette	0.0018	0.0108

Supplementary Table 1 : Statistical significance for all cases (n=140) between different pairs of GPT-4 experiments for 4-choice MCQ.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	4.00E-04	0.0036
Physical Exam	Single-turn conversation	7.00E-04	0.0056
Physical Exam	Summarized conversation	7.00E-04	0.0056
Physical Exam	Vignette	1.00E-04	0.001
Multi-turn conversation	Single-turn conversation	0.9051	0.9051
Multi-turn conversation	Summarized conversation	0.2237	0.671
Multi-turn conversation	Vignette	0.0079	0.0325
Single-turn conversation	Summarized conversation	0.4016	0.8031
Single-turn conversation	Vignette	0.0065	0.0325
Summarized conversation	Vignette	0.0019	0.0114

Supplementary Table 2 : Statistical significance for all public cases (n=100) between different pairs of GPT-4 experiments for 4-choice MCQ.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	6.07E-02	0.4249
Physical Exam	Single-turn conversation	5.50E-03	0.055

Physical Exam	Summarized conversation	6.30E-03	0.0567
Physical Exam	Vignette	1.61E-02	0.1288
Multi-turn conversation	Single-turn conversation	0.1595	0.7619
Multi-turn conversation	Summarized conversation	0.1524	0.7619
Multi-turn conversation	Vignette	0.1099	0.6593
Single-turn conversation	Summarized conversation	0.712	1
Single-turn conversation	Vignette	0.5994	1
Summarized conversation	Vignette	0.4213	1

Supplementary Table 3 : Statistical significance for all private cases (n=40) between different pairs of GPT-4 experiments for 4-choice MCQ.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	0.4108	0.8471
Physical Exam	Single-turn conversation	8.03E-02	0.3212
Physical Exam	Summarized conversation	0.0004	0.0032
Physical Exam	Vignette	7.00E-04	0.0049
Multi-turn conversation	Single-turn conversation	0.3824	0.8471
Multi-turn conversation	Summarized conversation	1.00E-04	0.001
Multi-turn conversation	Vignette	2.00E-04	0.0018
Single-turn conversation	Summarized conversation	0.0025	0.015
Single-turn conversation	Vignette	0.0029	0.015
Summarized conversation	Vignette	0.2824	0.8471

Supplementary Table 4 : Statistical significance for all cases (n=140) between different pairs of GPT-3.5 experiments for 4-choice MCQ.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	0.1736	0.5207
Physical Exam	Single-turn conversation	8.33E-02	0.3332
Physical Exam	Summarized conversation	0.0028	0.0252
Physical Exam	Vignette	1.40E-03	0.014
Multi-turn conversation	Single-turn conversation	0.8929	0.8929
Multi-turn conversation	Summarized conversation	9.40E-03	0.0658
Multi-turn conversation	Vignette	7.50E-03	0.06
Single-turn conversation	Summarized conversation	0.0196	0.1176

Single-turn conversation	Vignette	0.0208	0.1176
Summarized conversation	Vignette	0.384	0.7679

Supplementary Table 5 : Statistical significance for public cases (n=100) between different pairs of GPT-3.5 experiments for 4-choice MCQ.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	0.321	0.9629
Physical Exam	Single-turn conversation	6.52E-01	1
Physical Exam	Summarized conversation	0.0279	0.2232
Physical Exam	Vignette	3.08E-02	0.2232
Multi-turn conversation	Single-turn conversation	0.0843	0.3372
Multi-turn conversation	Summarized conversation	7.00E-04	0.007
Multi-turn conversation	Vignette	2.30E-03	0.0207
Single-turn conversation	Summarized conversation	0.0494	0.247
Single-turn conversation	Vignette	0.0363	0.2232
Summarized conversation	Vignette	0.514	1

Supplementary Table 6 : Statistical significance for private cases (n=40) between different pairs of GPT-3.5 experiments for 4-choice MCQ.

Experiment 1	Experiment 2	p-value	Adjusted p-value
No-choice MCQ	4-choice MCQ	1.00E-04	0.0003
No-choice MCQ	Many-choice MCQ	1.00E-04	0.0003
4-choice MCQ	Many-choice MCQ	1.00E-04	0.0003

Supplementary Table 7 : Statistical significance for all cases (n=140) between different pairs of GPT-4 experiments for multi-turn conversations.

Experiment 1	Experiment 2	p-value	Adjusted p-value
No-choice MCQ	4-choice MCQ	1.00E-04	0.0003
No-choice MCQ	Many-choice MCQ	1.00E-04	0.0003
4-choice MCQ	Many-choice MCQ	1.00E-04	0.0003

Supplementary Table 8 : Statistical significance for public cases (n=100) between different pairs of GPT-4 experiments for multi-turn conversations.

Experiment 1	Experiment 2	p-value	Adjusted p-value
No-choice MCQ	4-choice MCQ	1.00E-04	0.0003
No-choice MCQ	Many-choice MCQ	0.9192	0.9192
4-choice MCQ	Many-choice MCQ	0.0002	0.0004

Supplementary Table 9 : Statistical significance for private cases (n=40) between different pairs of GPT-4 experiments for multi-turn conversations.

Experiment 1	Experiment 2	p-value	Adjusted p-value
No-choice MCQ	4-choice MCQ	1.00E-04	0.0003
No-choice MCQ	Many-choice MCQ	1.00E-04	0.0003
4-choice MCQ	Many-choice MCQ	1.00E-04	0.0003

Supplementary Table 10 : Statistical significance for all cases (n=140) between different pairs of GPT-3.5 experiments for multi-turn conversations.

Experiment 1	Experiment 2	p-value	Adjusted p-value
No-choice MCQ	4-choice MCQ	1.00E-04	0.0003
No-choice MCQ	Many-choice MCQ	1.00E-04	0.0003
4-choice MCQ	Many-choice MCQ	1.00E-04	0.0003

Supplementary Table 11 : Statistical significance for public cases (n=100) between different pairs of GPT-3.5 experiments for multi-turn conversations.

Experiment 1	Experiment 2	p-value	Adjusted p-value
No-choice MCQ	4-choice MCQ	0.1495	0.1495
No-choice MCQ	Many-choice MCQ	1.00E-04	0.0003
4-choice MCQ	Many-choice MCQ	1.00E-04	0.0003

Supplementary Table 12 : Statistical significance for private cases (n=40) between different pairs of GPT-3.5 experiments for multi-turn conversations.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	0.988	0.988
Physical Exam	Single-turn conversation	1.00E-04	0.001
Physical Exam	Summarized conversation	1.00E-04	0.001
Physical Exam	Vignette	1.00E-04	0.001

Multi-turn conversation	Single-turn conversation	1.00E-04	0.001
Multi-turn conversation	Summarized conversation	1.00E-04	0.001
Multi-turn conversation	Vignette	1.00E-04	0.001
Single-turn conversation	Summarized conversation	0.2369	0.4738
Single-turn conversation	Vignette	0.0003	0.0012
Summarized conversation	Vignette	0.004	0.012

Supplementary Table 13 : Statistical significance for all cases (n=140) between different pairs of GPT-4 experiments for FRQs.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	0.201	0.402
Physical Exam	Single-turn conversation	2.20E-03	0.011
Physical Exam	Summarized conversation	2.00E-04	0.0012
Physical Exam	Vignette	1.00E-04	0.001
Multi-turn conversation	Single-turn conversation	1.00E-04	0.001
Multi-turn conversation	Summarized conversation	1.00E-04	0.001
Multi-turn conversation	Vignette	1.00E-04	0.001
Single-turn conversation	Summarized conversation	0.2194	0.402
Single-turn conversation	Vignette	0.0081	0.0324
Summarized conversation	Vignette	0.0561	0.1683

Supplementary Table 14 : Statistical significance for public cases (n=100) between different pairs of GPT-4 experiments for FRQs.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	0.0153	0.0918
Physical Exam	Single-turn conversation	1.40E-03	0.0126
Physical Exam	Summarized conversation	3.30E-03	0.0264
Physical Exam	Vignette	3.00E-04	0.003
Multi-turn conversation	Single-turn conversation	5.04E-01	1
Multi-turn conversation	Summarized conversation	2.46E-01	0.7367
Multi-turn conversation	Vignette	8.20E-03	0.0574
Single-turn conversation	Summarized conversation	0.7834	1
Single-turn conversation	Vignette	0.0247	0.1225

Summarized conversation	Vignette	0.0245	0.1225

Supplementary Table 15 : Statistical significance for private cases (n=40) between different pairs of GPT-4 experiments for FRQs.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	0.0061	0.0183
Physical Exam	Single-turn conversation	0.5138	1
Physical Exam	Summarized conversation	0.0007	0.0042
Physical Exam	Vignette	1.00E-04	0.001
Multi-turn conversation	Single-turn conversation	0.0036	0.018
Multi-turn conversation	Summarized conversation	0.8	1
Multi-turn conversation	Vignette	0.0045	0.018
Single-turn conversation	Summarized conversation	1.00E-04	0.001
Single-turn conversation	Vignette	1.00E-04	0.001
Summarized conversation	Vignette	1.00E-04	0.001

Supplementary Table 16 : Statistical significance for all cases (n=140) between different pairs of GPT-3.5 experiments for FRQs.

Experiment 1	Experiment 2	p-value	Adjusted p-value
Physical Exam	Multi-turn conversation	0.0346	0.173
Physical Exam	Single-turn conversation	0.1681	0.3362
Physical Exam	Summarized conversation	0.009	0.0588
Physical Exam	Vignette	7.00E-04	0.0063
Multi-turn conversation	Single-turn conversation	0.0941	0.2823
Multi-turn conversation	Summarized conversation	0.6193	0.6193
Multi-turn conversation	Vignette	0.0357	0.173
Single-turn conversation	Summarized conversation	6.60E-03	0.0528
Single-turn conversation	Vignette	4.00E-04	0.004
Summarized conversation	Vignette	8.40E-03	0.0588

Supplementary Table 17 : Statistical significance for public cases (n=100) between different pairs of GPT-3.5 experiments for FRQs.

Experiment 1	Experiment 2	p-value	Adjusted p-value

Physical Exam	Multi-turn conversation	0.0462	0.1848
Physical Exam	Single-turn conversation	0.5104	1
Physical Exam	Summarized conversation	0.0146	0.073
Physical Exam	Vignette	1.70E-03	0.0119
Multi-turn conversation	Single-turn conversation	0.0098	0.0588
Multi-turn conversation	Summarized conversation	0.8332	1
Multi-turn conversation	Vignette	0.0722	0.2166
Single-turn conversation	Summarized conversation	1.40E-03	0.0117
Single-turn conversation	Vignette	5.00E-04	0.005
Summarized conversation	Vignette	1.30E-03	0.0117

Supplementary Table 18 : Statistical significance for private cases (n=40) between different pairs of GPT-3.5 experiments for FRQs.

Experiment 1	Experiment 2	p-value	Adjusted p-value
4-choice MCQ	4-choice MCQ (without PE)	0.0002	0.0003
many-choice MCQ	many-choice MCQ (without PE)	1.00E-04	0.0003
FRQ	FRQ (without PE)	1.00E-04	0.0003

Supplementary Table 19 : Statistical significance for all cases (n=140) between different pairs of GPT-4 experiments for multi-turn conversations with and without physical exam (PE).

Experiment 1	Experiment 2	p-value	Adjusted p-value
4-choice MCQ	4-choice MCQ (without PE)	0.0006	0.0012
many-choice MCQ	many-choice MCQ (without PE)	0.3473	0.3473
FRQ	FRQ (without PE)	1.00E-04	0.0003

Supplementary Table 20 : Statistical significance for all cases (n=140) between different pairs of GPT-3.5 experiments for multi-turn conversations with and without physical exam (PE).

Experiment	Spearman Correlation	p-value
Vignette + FRQ (GPT-4)	0.954	2.08×10^{-53}
Vignette + FRQ (GPT-3.5)	0.936	2.54×10^{-46}
Multi-turn conversation + FRQ (GPT-4)	0.942	1.63×10^{-48}
Multi-turn conversation (without PE) + FRQ (GPT-4)	0.924	5.74×10^{-43}
Multi-turn conversation (without PE) + FRQ (GPT-3.5)	0.889	4.32×10^{-35}

Supplementary Table 21 : Correlation between grader-AI and dermatologists' accuracy annotation of the doctor-AI agent (PE = physical exam, FRQ = Free Response Question).