










An evaluation framework for clinical use of large language models in patient interaction tasks

Received: 8 August 2023

Accepted: 1 October 2024

Published online: 02 January 2025

 Check for updates

Shreya Johri ^{1,10}, Jaehwan Jeong^{1,2,10}, Benjamin A. Tran³, Daniel I. Schlessinger ⁴, Shannon Wongvibulsin⁵, Leandra A. Barnes⁶, Hong-Yu Zhou ¹, Zhuo Ran Cai⁶, Eliezer M. Van Allen ⁷, David Kim ⁸, Roxana Daneshjou ^{6,9,11}  & Pranav Rajpurkar ^{1,11} 

The integration of large language models (LLMs) into clinical diagnostics has the potential to transform doctor–patient interactions. However, the readiness of these models for real-world clinical application remains inadequately tested. This paper introduces the Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD) approach for evaluating clinical LLMs. Unlike traditional methods that rely on structured medical examinations, CRAFT-MD focuses on natural dialogues, using simulated artificial intelligence agents to interact with LLMs in a controlled environment. We applied CRAFT-MD to assess the diagnostic capabilities of GPT-4, GPT-3.5, Mistral and LLaMA-2-7b across 12 medical specialties. Our experiments revealed critical insights into the limitations of current LLMs in terms of clinical conversational reasoning, history-taking and diagnostic accuracy. These limitations also persisted when analyzing multimodal conversational and visual assessment capabilities of GPT-4V. We propose a comprehensive set of recommendations for future evaluations of clinical LLMs based on our empirical findings. These recommendations emphasize realistic doctor–patient conversations, comprehensive history-taking, open-ended questioning and using a combination of automated and expert evaluations. The introduction of CRAFT-MD marks an advancement in testing of clinical LLMs, aiming to ensure that these models augment medical practice effectively and ethically.

Patient history collection is the foundation of medical diagnosis, enabling physicians to identify key information that guides their clinical decisions. However, the mounting pressure of escalating patient numbers, lack of access to care¹, short consultation times^{2,3} and the expedited adoption of telemedicine due to the coronavirus disease 2019 (COVID-19) pandemic⁴ have presented formidable challenges to this conventional model of interaction. As these factors risk compromising the quality of history-taking and, thereby, diagnostic accuracy²,

there is a need for innovative solutions that can enhance the efficacy of these clinical conversations.

New advances in generative artificial intelligence (AI), specifically in large language models (LLMs), present a potential solution to this problem^{5–9}. These AI models have the ability to engage in nuanced conversations, making them ideal candidates for extracting comprehensive patient histories and assisting physicians in generating differential diagnoses^{10–12}. However, a considerable gap remains in

assessing the readiness of these models for application in real-world clinical scenarios^{13–15}. The predominant method for evaluating LLMs in medicine involves medical examination-style questions, with a strong emphasis on multiple-choice formats^{16–18}. Although there are instances where LLMs are tested on free-response and reasoning tasks^{12,19,20} or for medical conversation summarization and care plan generation²¹, these assessments are less common. Importantly, these assessments do not explore the ability of LLMs to engage in interactive patient conversations, which could enhance telehealth and virtual medical visits, help emergency room physicians triage patients and facilitate medical education by teaching medical students best practices for history-taking.

Addressing this evaluative shortfall, we propose a new framework for evaluation of clinical LLMs, called the Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD). As opposed to the conventional reliance on structured medical examinations, CRAFT-MD evaluates a clinical LLM by simulating active collection and integration of information through a doctor–patient conversation, similar to a physician’s interaction with patients. This simulation is achieved through a patient-AI agent that interacts with the clinical LLM. A grader-AI agent then evaluates the conversation for correctness of diagnosis, and medical experts assess the reliability of each AI agent. CRAFT-MD substantially enhances the scalability of evaluations, enabling broader and faster testing to keep pace with the rapid evolution of LLMs. It addresses the challenges of using human testers alone and mitigates potential ethical and safety concerns of early LLM interactions with real patients, reducing the risk of harm from such engagements.

We applied CRAFT-MD to assess the clinical diagnostic capabilities of commercial and open-source LLMs, including GPT-4 (ref. 22), GPT-3.5 (ref. 23), Mistral (ref. 24) and LLaMA-2-7b (ref. 25), as well as multimodal LLMs, such as GPT-4V (refs. 26,27). Our evaluations encompassed medical conditions common in both primary and specialist care settings across 12 medical specialties. The experiments highlight the limitations of current LLMs in incorporating details from conversational interactions for accurate diagnosis and medical image interpretation. Supported by this empirical evidence, we further developed a comprehensive set of recommendations for evaluating the conversational reasoning capabilities of clinical LLMs. CRAFT-MD, therefore, provides a robust framework for evaluating the proficiency of LLMs in medical information processing, critical thinking and decision-making—skills essential in clinical settings—ultimately supporting the development of LLMs tailored to the complexities of healthcare.

Results

The CRAFT-MD framework

CRAFT-MD is a framework designed to evaluate the conversational reasoning abilities of clinical LLMs in simulated doctor–patient interactions. At its core, CRAFT-MD assesses the capacity of a clinical LLM to conduct medical interviews, synthesize information and formulate diagnoses in a realistic clinical context. The framework employs a multi-agent approach comprising four components (Fig. 1): the clinical LLM being evaluated, a patient-AI agent that simulates patient responses, a grader-AI agent that assesses diagnostic accuracy and medical experts who validate the process. This design allows for comprehensive evaluation of any clinical LLM, as the model being tested can be easily switched out.

The clinical LLM interacts with the patient-AI agent, asking questions about current LLM symptoms, medical history, medications and family history to formulate a differential diagnosis. The patient-AI agent responds in layman’s terms, based on a detailed case vignette. The grader-AI agent evaluates the clinical LLM’s diagnosis in free text for accuracy against the correct diagnosis provided in the vignette, accounting for synonyms and disease variants. Finally, medical experts review a subset of the simulated dialogues for qualitative insights into the limitations of the clinical LLM and determine the reliability of each AI agent. The clinical LLM is evaluated on its ability to gather relevant

medical information and symptoms to arrive at the most likely diagnosis. The patient-AI agent is assessed on its ability to avoid medical jargon, similar to real patients, and the grader-AI agent is judged on the precision of its grading (Methods). This bears similarities to the Objective Structured Clinical Examination (OSCE) while also introducing unique advantages, such as scalability and rapidity of evaluations. The simulation of doctor–patient conversations enables clinically meaningful evaluation across various medical specialties, and assessments by medical experts quantify confidence in the results obtained.

The CRAFT-MD framework was evaluated on a total of 2,000 case vignettes (see ‘Data availability’). Of these, 1,800 were sourced from MedQA-United States Medical Licensing Examination (USMLE)²⁸, encompassing medical conditions common in primary and specialist care across 12 medical specialties: Dermatology, Hematology and Oncology, Neurology, Gastroenterology, Pediatrics and Neonatology, Cardiology, Infectious Disease, Obstetrics and Gynecology, Urology and Nephrology, Endocrinology, Rheumatology and Others (Extended Data Fig. 1). One hundred case vignettes were included from an online question bank²⁹ (referred to as Derm-Public), and 100 newly generated private cases (referred to as Derm-Private) were also included to study trends across data sources and focused evaluation on skin diseases. Commercial models, including GPT-4 (6 November 2024 version) and GPT-3.5 (6 November 2024 version), and open-source models, including LLaMA-2-7b, Mistral-v1-7b and Mistral-v2-7b, were evaluated for their clinical conversational reasoning skills. Dataset contamination estimation of the 2,000 case vignettes using Memorization Effects Levenshtein Detector (MELD) analysis⁶ did not reveal overlap with the GPT-4 training dataset (Extended Data Fig. 1), although it is noted that MELD has high precision but unknown recall. For evaluation of the multimodal LLM GPT-4V, case vignettes and their associated images were sourced from the NEJM Image Challenge dataset (see ‘Data availability’).

CRAFT-MD considerably outpaces traditional human-centric evaluation methods in efficiency and scale. It processes 10,000 multi-turn conversations in 48–72 h (API calls being the primary constraint), plus 15–16 h of expert evaluation. In contrast, human-based approaches would require extensive recruitment and an estimated 500 h for patient simulations (~3 min per conversation) and about 650 h for expert evaluations (~4 min per conversation). This demonstrates the capacity of CRAFT-MD to markedly reduce time and resources in large-scale clinical LLM assessments.

Conversational interactions reduce diagnostic accuracy

We evaluated whether LLMs maintain accuracy when making diagnoses through conversations versus static case vignettes in the four-choice multiple choice questions (MCQs) setting. Using the CRAFT-MD framework, we transformed vignettes into multi-turn conversations between the clinical LLM and patient-AI agents (Fig. 2a,b and Methods). For all the evaluated LLMs (GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b), diagnostic accuracy dropped when using conversations versus vignettes (Fig. 2c and Supplementary Tables 1 and 2). Performance drops were 0.193 for GPT-4 (0.820 to 0.627), 0.19 for GPT-3.5 (0.657 to 0.467), 0.211 for Mistral-v2-7b (0.637 to 0.426) and 0.076 for LLaMA-2-7b (0.395 to 0.319), all with adjusted *P* values less than 0.0001. Therefore, despite their impressive capabilities on static inputs, current LLMs are limited in adapting to the dynamic conversations for four-choice MCQs.

We next quantified the impact of follow-up questions by the clinical LLM in multi-turn conversations. For this, we evaluated the performance in single-turn conversations (Fig. 2d and Methods), where the clinical LLM based its diagnosis solely on initially described symptoms without asking follow-up questions to the patient-AI agent. Accuracy in four-choice MCQs for single-turn versus multi-turn conversations decreased by 0.107 for GPT-4 (0.627 to 0.520, adjusted *P* < 0.0001), by 0.032 for GPT-3.5 (0.467 to 0.435, adjusted *P* < 0.0001) and by 0.015 for LLaMA-2-7b (0.319 to 0.304, adjusted *P* < 0.05) and increased by 0.022 for Mistral-v2-7b (0.426 to 0.448, adjusted *P* < 0.001) (Fig. 2c and

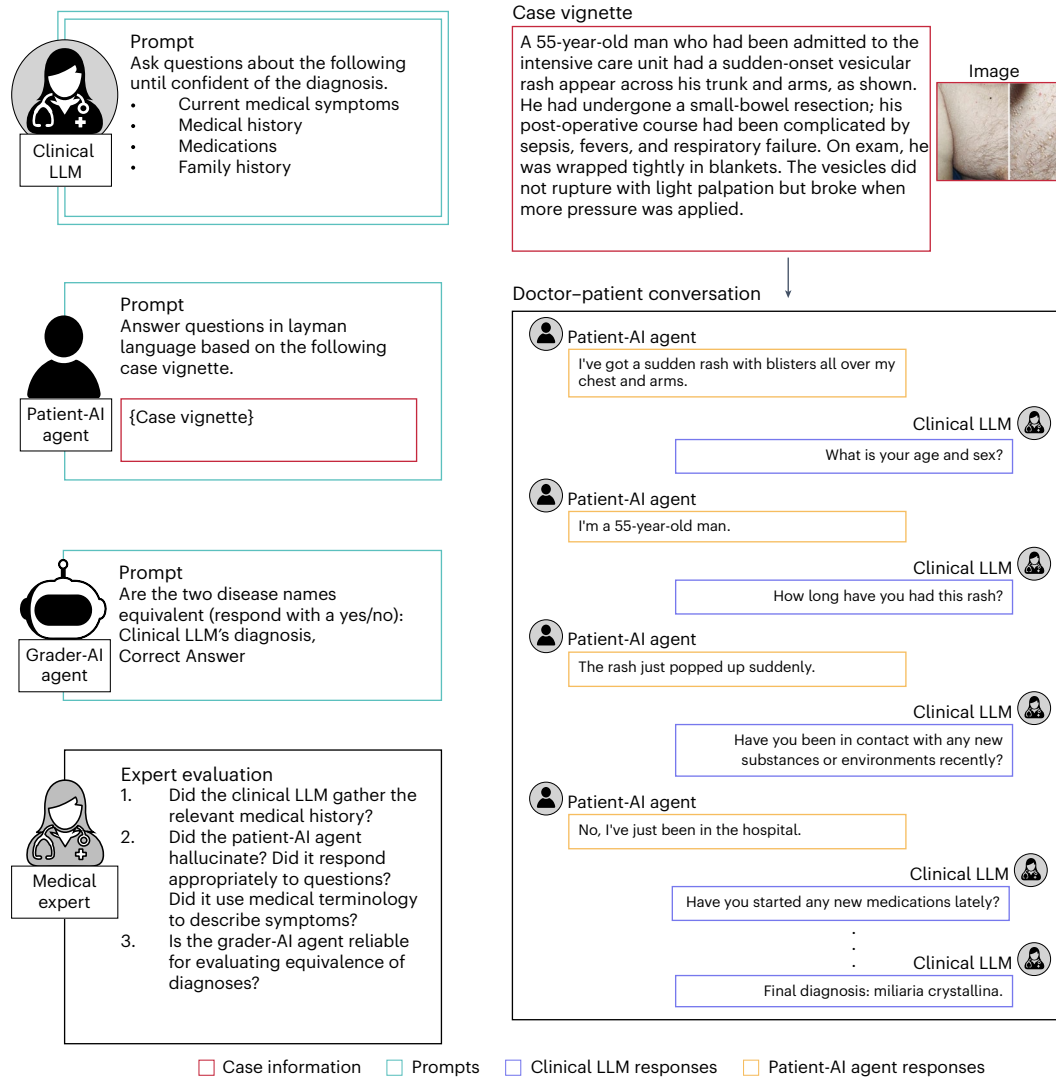


Fig. 1 | CRAFT-MD: a framework for evaluating the conversational abilities of clinical LLMs in medical contexts. The framework simulates doctor-patient interactions to assess the proficiency of a clinical LLM in history-taking, information synthesis and diagnostic accuracy. A patient-AI agent engages the clinical LLM in conversation while a grader-AI agent and medical experts evaluate

the LLM's performance. This multi-agent approach enables comprehensive assessment of the reasoning capabilities of the clinical LLM in a simulated medical environment. Credits: Patient icon reproduced from Adobe Stock. Doctor and Grader-AI icons adapted from Adobe Stock. Image reproduced with permission from ref. 49 Massachusetts Medical Society.

Supplementary Tables 1 and 2). Surprisingly, this decrease in accuracy for GPT-4, GPT-3.5 and LLaMA-2-7b was lower than anticipated, despite the relevance of the follow-up questions to the final diagnosis.

Conversational Summarization improves the limited reasoning of LLMs across multiple dialogues

We hypothesized that the minimal changes in accuracy between single-turn and multi-turn conversations could result from difficulties in synthesizing information across multiple dialogues. This issue could emerge if the training dataset predominantly features vignette-like examples rather than extended dialogues. To test this hypothesis, we developed a technique called Conversational Summarization, which transforms multi-turn conversations into vignette-like summaries, consolidating all details into a single paragraph (hereafter called 'summarized conversation') (Fig. 2e, Extended Data Fig. 2 and Methods). The summarized conversation is different from the vignette itself because only the details revealed by the patient-AI agent are transformed.

We observed an increase in accuracy when the clinical LLM was provided with summarized conversations compared to multi-turn conversations, for all evaluated models in the four-choice MCQ setting

(GPT-4 = 0.627 to 0.669, adjusted $P < 0.0001$; GPT-3.5 = 0.467 to 0.507, adjusted $P < 0.0001$; Mistral-v2-7b = 0.426 to 0.513, adjusted $P < 0.0001$; LLaMA-2-7b = 0.319 to 0.335, adjusted $P < 0.05$) (Fig. 2c and Supplementary Tables 1 and 2). These observations indicate that transforming scattered multi-turn conversations to concise vignette-like formats (that is, summarized conversations) may be useful for more accurate diagnoses.

Trends persist in open-ended diagnoses and across specialties

The four-choice MCQs used in medical licensing examinations 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 evaluated the conversational reasoning of clinical LLMs without answer choices—that is, free-response questions (FRQs) (Fig. 2a,b,d,e and Methods). All the free text responses by the clinical LLM were evaluated using the grader-AI agent.

Removing answer options leads to decrease in accuracy. The accuracy of all models considerably decreased in the FRQ format compared to the four-choice MCQ format (Fig. 2c,f and Supplementary Table 3).

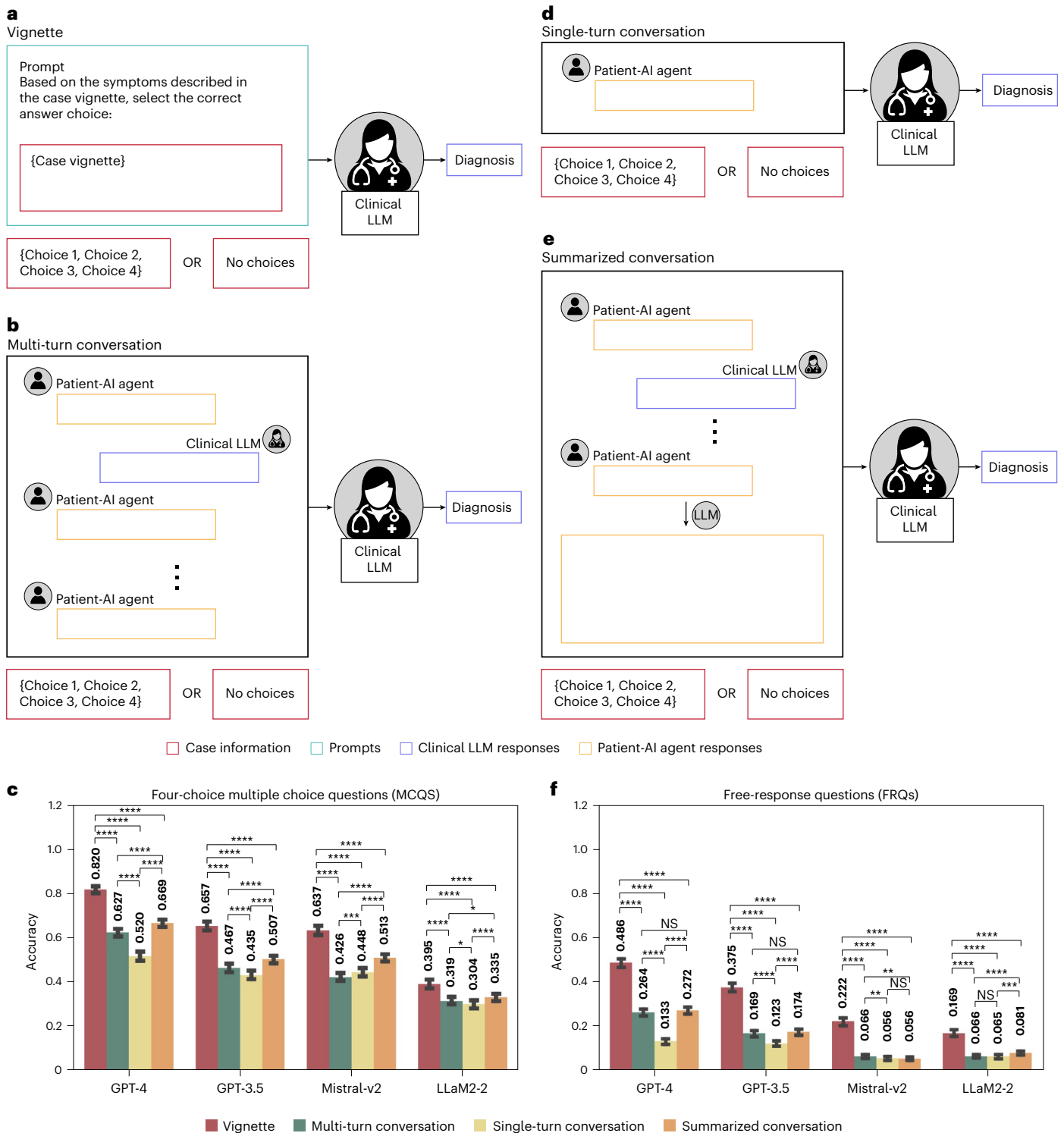


Fig. 2 | Effect of replacing case vignettes with simulated doctor–patient conversations in four-choice MCQs and FRQs. Experimental setup for diagnosis using case vignettes (a), multi-turn conversations (b), single-turn conversations (d) and summarized conversations (e), followed by four-choice MCQ or FRQ (no choices). c, Diagnostic accuracy for four experimental setups—vignette + four-choice MCQs, multi-turn conversation + four-choice MCQs, single-turn conversation + four-choice MCQs and summarized conversation + four-choice MCQs—across four evaluated LLMs (GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b). f, Diagnostic accuracy for four experimental setups—vignette

+ FRQs, multi-turn conversation + FRQs, single-turn conversation + FRQs and summarized conversation + FRQs—across four evaluated LLMs (GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b). Error bars represent 95% confidence intervals on 10,000 samples, and numbers represent the mean accuracy. NS, non-significant; * ≤ 0.05; ** ≤ 0.01; *** ≤ 0.001; **** ≤ 0.0001. All P values were calculated using a two-sided bootstrapping test, followed by Holm–Bonferroni correction (Methods and Supplementary Tables 1–5). Credits: Patient icon reproduced from Adobe Stock. Doctor and Grader-AI icons adapted from Adobe Stock.

For vignettes, the accuracy of GPT-4 decreased by 0.334 (from 0.820 to 0.486), GPT-3.5 by 0.282 (from 0.657 to 0.375), Mistral-v2-7b by 0.415 (from 0.637 to 0.222) and LLaMA-2-7b by 0.226 (from 0.395 to 0.169), all with adjusted *P* values less than 0.0001. Similar decreases were also observed for multi-turn conversations (GPT-4 = 0.627 to 0.264; GPT-3.5 = 0.467 to 0.169; Mistral-v2-7b = 0.426 to 0.066; LLaMA-2-7b = 0.319 to 0.066); single-turn conversations (GPT-4 = 0.520 to 0.133; GPT-3.5 = 0.435 to 0.123, Mistral-v2-7b = 0.448 to 0.056; LLaMA-2-7b = 0.304 to 0.065); and summarized conversations (GPT-4 = 0.669 to 0.272; GPT-3.5 = 0.507 to 0.174; Mistral-v2-7b = 0.513 to 0.056; LLaMA-2-7b = 0.335 to 0.081) (Extended Data Table 1), all with adjusted *P* values less than 0.0001. These findings indicate that removing predefined answer options significantly lowers diagnostic accuracy across all models and conversation types, underscoring the difficulty in handling open-ended clinical diagnostic tasks.

Conversational interactions continue underperforming vignettes.

Replacing vignettes with multi-turn conversations in the FRQ format resulted in a substantial decline in accuracy, similar to the four-choice MCQ format. Accuracy dropped from 0.486 to 0.264 for GPT-4, from 0.375 to 0.169 for GPT-3.5, from 0.222 to 0.066 for Mistral-v2-7b and from 0.169 to 0.066 for LLaMA-2-7b, all with adjusted *P* values less than 0.0001. The difference between multi-turn and single-turn accuracies was significant for GPT-4 (0.264 to 0.133, adjusted *P* < 0.0001), for GPT-3.5 (0.169 to 0.123, adjusted *P* < 0.0001) and for Mistral-v2-7b (0.066 to 0.056; adjusted *P* < 0.01) but not for LLaMA-2-7b (0.066 to 0.065). Notably, although Mistral-v2-7b showed higher single-turn accuracy than multi-turn in the four-choice MCQ setting, this trend did not persist in the FRQ setting. Additionally, the difference in accuracy between summarized and multi-turn conversations without answer choices was significant only for open-source models (Mistral-v2-7b = 0.066 to 0.056, adjusted *P* < 0.01; LLaMA-2-7b = 0.066 to 0.081, adjusted *P* < 0.0001) but for not commercial models (GPT-4 = 0.264 to 0.272; GPT-3.5 = 0.169 to 0.174) (Fig. 2f and Supplementary Tables 4 and 5).

Trends in conversational diagnostic accuracy persist across medical specialties.

For each of the 12 medical specialties in our dataset, we observed similar trends between different conversational formats for both four-choice MCQ and FRQ settings (Extended Data Figs. 3 and 4 and Supplementary Tables 6–9). A notable decrease in accuracy occurs when vignettes are replaced by multi-turn conversations. Additionally, summarized conversations maintain higher accuracy than multi-turn conversations but fall short of the accuracy achieved with vignettes. This consistency underscores the robustness of these observed trends.

A case study in skin diseases

For a detailed analysis with medical experts, we chose to concentrate on skin diseases, which are frequent complaints in primary care³⁰. 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.

Consistent trends across the datasets. Across the three evaluated datasets—MedQA-USMLE (*n* = 117), Derm-Public (*n* = 100) and Derm-Private (*n* = 100)—vignettes consistently had higher accuracy compared to conversational formats (Fig. 3 and Supplementary Tables 10–13). We noted that a subset of case vignettes obtained from public datasets had multiple possible diagnoses when answer options were removed. Medical experts determined that additional details on symptoms, medications or physical examinations were necessary for conclusive diagnosis in these cases. Consequently, we also evaluated clinical LLM diagnostic accuracies in FRQ settings for cases with single possible diagnoses, finding higher accuracies and highlighting the need

for improved design of case vignettes for FRQ evaluations (Extended Data Fig. 4 and Supplementary Tables 14–17). Notably, the dermatologists' diagnostic accuracy on the dermatology case vignettes was consistent across formats, achieving 86% accuracy on four-choice MCQs and 87% on FRQs (see 'Data availability'). They expressed uncertainty about many cases from the MedQA-USMLE and Derm-Public datasets, indicating that an image would be required for diagnostic certainty.

Medical expert evaluations. To evaluate each of the LLM agents (patient-AI and grader-AI) in the CRAFT-MD framework, medical experts assessed a subset of the conversations (*n* = 180) evenly distributed among the four evaluated models (GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b) and the three datasets (MedQA-USMLE, Derm-Public and Derm-Private) (Methods). Two dermatologists conducted the evaluations, with a third of the conversations being dual-annotated to estimate expert agreement. In cases where the two dermatologists disagreed, a third dermatologist resolved the tie (Extended Data Table 2).

We first assessed the reliability of the patient-AI agent and grader-AI agent. When responding to questions posed by the clinical LLM, the patient-AI agent provided accurate answers 99.995% of the time when the question was within the scope of the case vignette. For questions beyond the vignette's scope, the agent either indicated unavailability of information or denied symptoms. Relevant and complete answers were provided 94.25% of the time, with incomplete answers typically occurring when multiple questions were posed within the same dialogue. Additionally, 7.22% of conversations included technical medical language in the agent's responses, compared to 100% of the case vignettes. Furthermore, the grader-AI agent agreed with medical experts at a high rate of 93.35% (see 'Data availability').

We next qualitatively evaluated the clinical LLM for the ability to lead clinical conversations and gather complete medical histories. For assessing the clinical LLM's understanding of when to continue asking questions for clinical information and when sufficient information had been gathered to make a diagnosis, we calculated the percentage of conversations where a medical expert could identify a single most likely diagnosis, regardless of the correctness of the diagnosis. We found substantial variance across the evaluated models: GPT-4 achieved 53.33%, GPT-3.5 achieved 31.11%, Mistral-v2-7b achieved 11.11% and LLaMA-2-7b achieved 35.55% (Fig. 3i and Supplementary Table 18). With regard to gathering complete medical history during conversations, there was again a considerable variance among models: GPT-4 achieved 71.11%, GPT-3.5 achieved 31.11%, Mistral-v2-7b achieved 8.88% and LLaMA-2-7b achieved 51.11% (Fig. 3j and Supplementary Table 18). These results could indicate potential gaps in the medical knowledge of these LLMs that affect their ability to effectively lead clinical conversations.

Multimodal models are limited in image comprehension

Medical diagnosis often relies on visual examination, through either direct observation or imaging techniques. This necessitates robust multimodal LLMs capable of accurate image interpretation alongside natural language conversation³¹. We evaluated GPT-4V (Methods) using the CRAFT-MD framework to assess its combined visual and conversational abilities. Our study compared diagnostic accuracy between vignette and conversational formats, both with and without image inputs (Fig. 4a,b). This approach allowed us to evaluate the ability of the clinical LLM to lead medical conversations when provided with an image of the affected area upfront, contrasting it with scenarios where no image was available, as is the case with traditional LLMs.

To evaluate the medical image interpretation capabilities of GPT-4V, we curated 74 (image and case vignette) pairs from the NEJM Image Challenge dataset³² (Methods). This dataset is particularly suitable for our evaluation because each case vignette's diagnosis heavily depends on the corresponding medical image. We hypothesized that if GPT-4V possesses strong medical image interpretation skills, it would

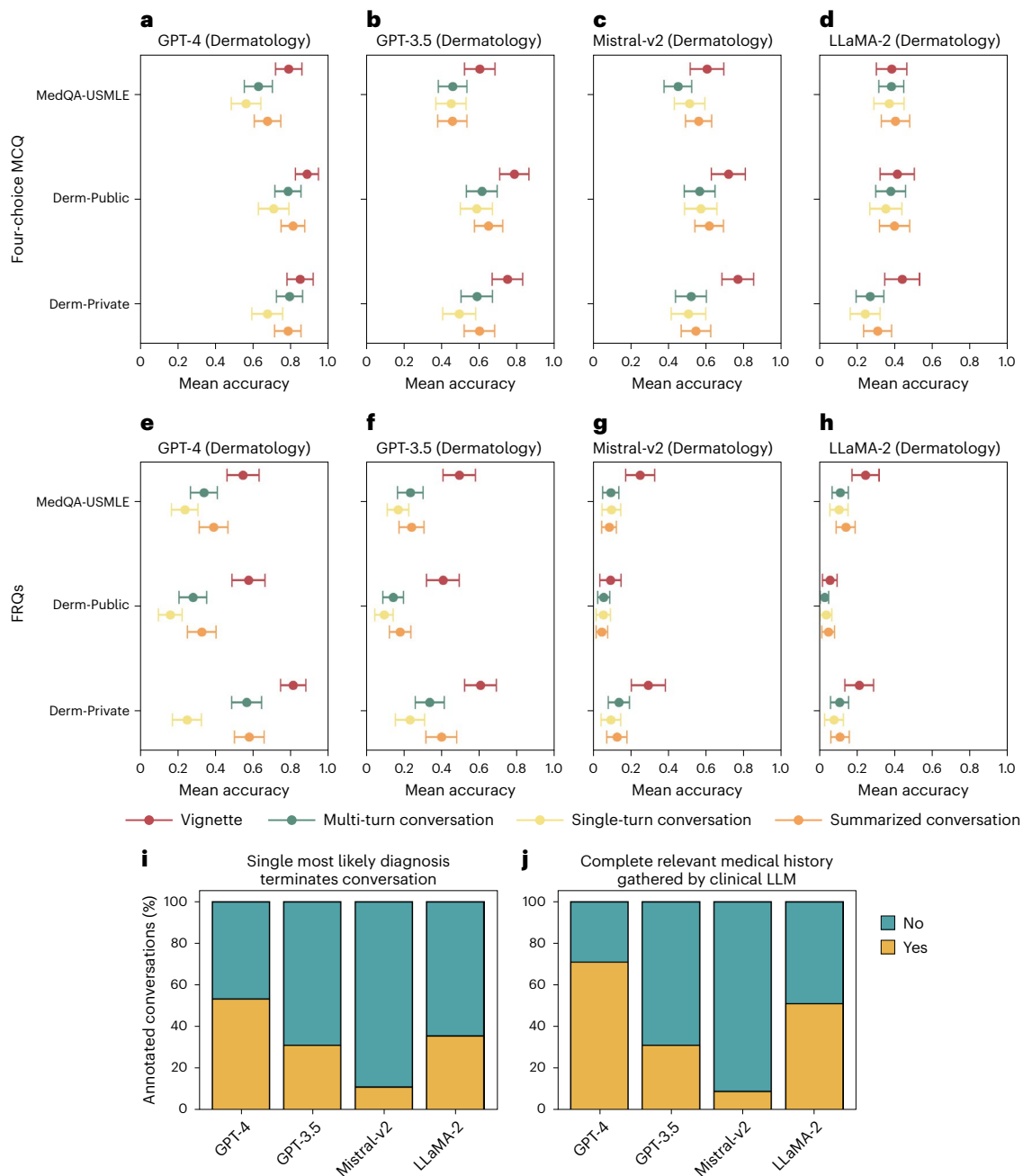


Fig. 3 | Trends in vignette and conversational formats across skin disease datasets. Results for four-choice MCQ (a–d) and FRQ (e–h) persist across the three datasets—MedQA-USMLE, Derm-Public and Derm-Private. Error bars represent 95% confidence intervals on 585 samples for MedQA-USMLE and 500 samples each for Derm-Public and Derm-Private. **i**, Percentage of annotated

conversations where the conversation terminated (that is, clinical LLM stopped asking questions) when single most likely diagnosis was possible. **j**, Percentage of annotated conversations with complete relevant medical history for the four evaluated models (GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b) as assessed by medical experts.

demonstrate significantly higher diagnostic accuracy when presented with both the image and the case vignette, compared to scenarios where only the textual information is provided.

Our findings revealed a small decrease in accuracy across all experimental setups (vignette, multi-turn, single-turn and summarized conversations) when images were removed, in both four-choice MCQ and FRQ settings (Fig. 4c–j and Supplementary Tables 19 and 20). In the four-choice MCQ format, we observed decreases of 0.055 for vignettes, 0.024 for multi-turn conversations, 0.074 for single-turn conversations and 0.044 for summarized conversations. Similarly, in the FRQ format, decreases were 0.021 for vignettes, 0.058 for multi-turn conversations, 0.024 for single-turn conversations

and 0.055 for summarized conversations. Although consistent, these decreases were not statistically significant (Supplementary Table 21).

Need for continuous monitoring of LLMs

The rapid development of LLMs and frequent release of new versions necessitates continuous monitoring of their evolving capabilities. We employed CRAFT-MD to evaluate the proficiency in leading clinical conversations across two versions of the open-source model Mistral (v1 and v2).

Mistral-v1-7b exhibited similar accuracy trends between vignette and conversational formats as Mistral-v2-7b (Fig. 5 and Supplementary Tables 22 and 23). In the four-choice MCQ setting, Mistral-v1-7b

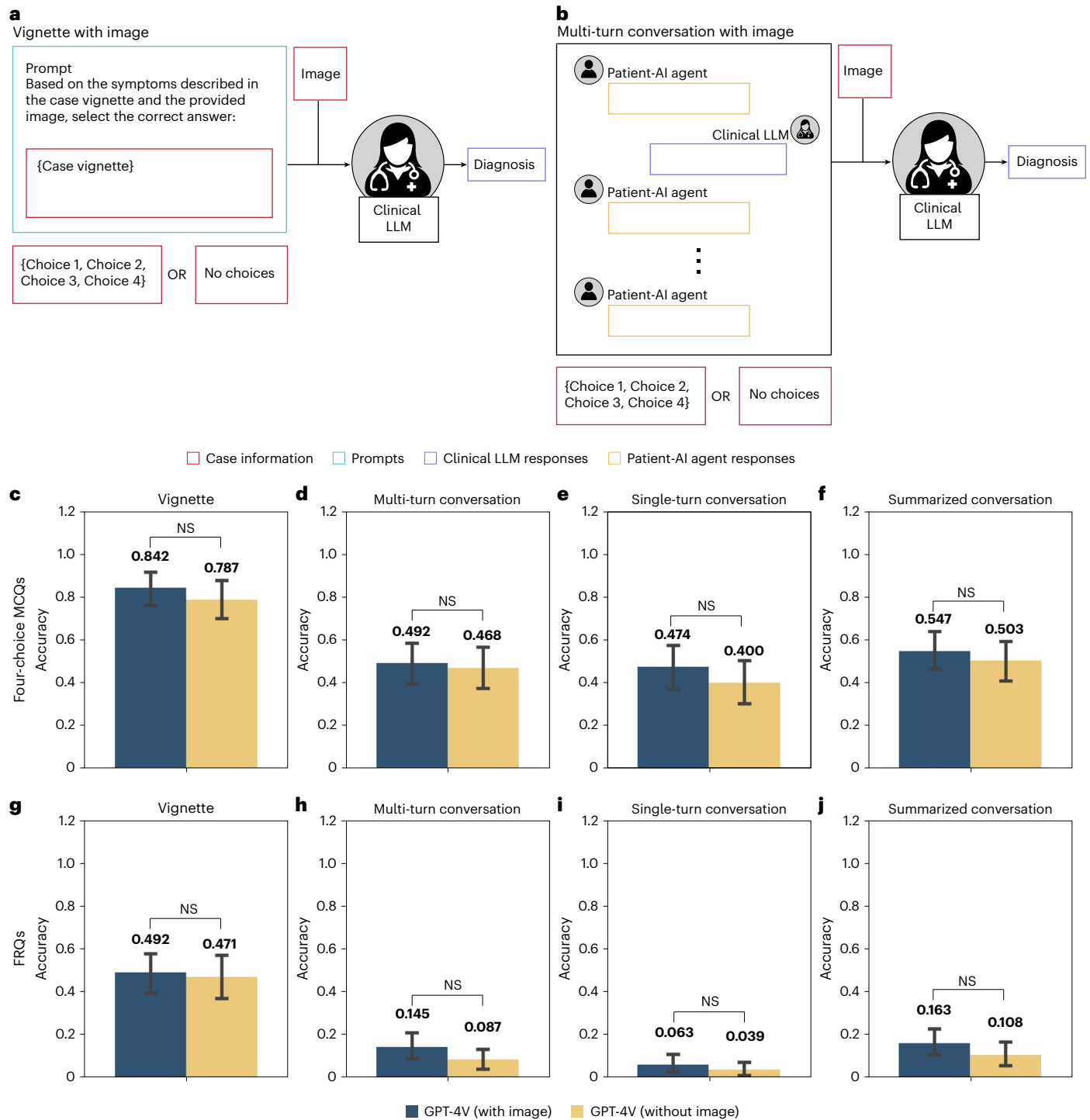


Fig. 4 | Evaluation of coupled image interpretation and conversational capabilities of GPT-4V. a, b. Schematic showing the vignette and multi-turn conversation setup with image input. Bar plot showing mean accuracy for four-choice MCQ setting (c–f) and FRQ setting (g–j), for vignette and conversational formats (multi-turn, single-turn and summarized). Error bars represent 95% confidence intervals over 370 data points, and numbers represent the mean

accuracy; NS, non-significant; * ≤ 0.05 ; ** ≤ 0.01 ; *** ≤ 0.001 ; **** ≤ 0.0001 . All *P* values were calculated using a two-sided bootstrapping test, followed by Holm–Bonferroni correction (Methods and Supplementary Tables 19–21). Credits: Patient icon reproduced from Adobe Stock. Doctor and Grader-AI icons adapted from Adobe Stock.

showed a significant decrease in accuracy from vignette to multi-turn conversations (adjusted $P < 0.0001$), followed by a significant increase from multi-turn to summarized conversations (adjusted $P < 0.0001$). The FRQ setting displayed similar trends. Notably, the accuracy of Mistral-v1-7b did not significantly differ between single-turn and multi-turn conversations (adjusted $P > 0.05$), whereas Mistral-v2-7b

demonstrated significantly higher accuracy in single-turn compared to multi-turn conversations.

Comparing the two versions, mean accuracies increased from Mistral-v1-7b to Mistral-v2-7b across all formats in the four-choice MCQ setting (vignette = 0.196, multi-turn = 0.095, single-turn = 0.124, summarized = 0.152). However, in the FRQ setting, only the vignette

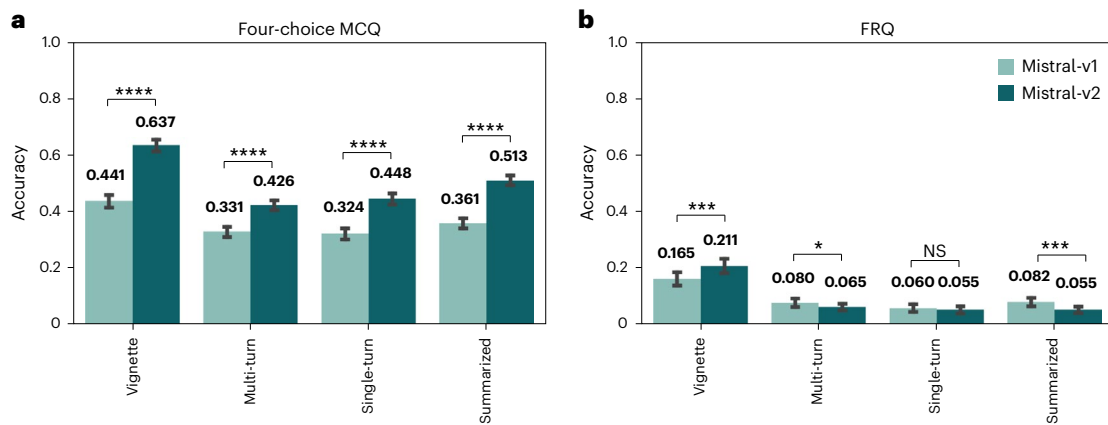


Fig. 5 | Continuous monitoring of LLMs. Evolution of mean diagnostic accuracy between the two versions of Mistral (v1 and v2) for vignettes and conversational settings (multi-turn, single-turn and summarized) for four-choice MCQ (a) and FRQ (b) settings. Error bars represent 95% confidence intervals over 10,000 data points, and numbers represent the mean accuracy; NS, non-significant; * ≤ 0.05 ; ** ≤ 0.01 ; *** ≤ 0.001 ; **** ≤ 0.0001 . All *P* values were calculated using a two-sided bootstrapping test, followed by Holm–Bonferroni correction (Methods and Supplementary Tables 22 and 23).

Table 1 | Proposed recommendations for evaluation of clinical LLMs

Recommendation	Description
Recommendation 1	Evaluate Diagnostic Accuracy Through Realistic Doctor–Patient Conversations: Assess LLMs in dynamic, conversational settings that reflect real-world clinical interactions, moving beyond the limitations of traditional static examinations to capture the complexities of medical dialogues.
Recommendation 2	Employ Open-Ended Questions for Evaluating Diagnostic Reasoning: Move away from multiple-choice questions to open-ended questioning that mimics the complexities of actual medical practice, thereby capturing the diagnostic reasoning of LLMs in real-world scenarios more effectively.
Recommendation 3	Assess Comprehensive History Taking Skills: Critically evaluate LLMs for their ability to conduct thorough medical interviews and gather essential information through conversations, acknowledging the importance of interactive dialogue in understanding patient conditions.
Recommendation 4	Evaluate LLMs on the Synthesis of Information Over Multiple Dialogues: Examine the ability of LLMs to integrate and comprehend information presented over extended interactions, addressing the shortfall in current assessments that focus on immediate responses to queries.
Recommendation 5	Incorporate Multimodal Information Available to Physicians to Enhance LLM Performance: Bridge the gap in the information available to LLMs compared to physicians, including visual assessments, physical examinations or laboratory test results. Work toward better multimodal integration for a balanced approach.
Recommendation 6	Continuous Evaluation of Conversational Abilities for Guiding Development of Clinical LLMs: Monitor evolving capabilities of LLMs across different model versions to guide future training of models.
Recommendation 7	Test and Refine Prompting Strategies to Enhance LLM Performance: Routinely evaluate and refine different prompt structures and styles, including responses to clarifications and follow-ups, to guide LLMs toward more accurate and contextually relevant responses.
Recommendation 8	Implement Patient–LLM Interactions for Ethical and Scalable Testing: Use simulated AI agents for clinical LLM evaluation to enable large-scale, rapid testing in a controlled environment, mitigating ethical and safety concerns and enhancing the efficiency of the evaluation process.
Recommendation 9	Combine Automated and Expert Evaluations for Comprehensive Insights: Merge the efficiency of automated systems with focused expert reviews for in-depth analysis, assessing not just the correctness of the diagnosis but also the process by which the LLM arrived at the diagnosis.
Recommendation 10	Encourage Collection of Public Datasets Covering Diverse Medical Scenarios, Suited for Open-ended Evaluation: Expand the focus to a greater diversity of clinical cases and address concerns regarding LLMs potentially memorizing training dataset cases by demanding transparency from AI developers about training methodologies and data and incorporating entirely new cases in studies.

format showed improvement (increase = 0.048), whereas all conversational formats saw declines (multi-turn = -0.015, single-turn = -0.005, summarized = -0.027) (Fig. 5 and Extended Data Table 3). These findings underscore the importance of comprehensive evaluation across different formats when training LLMs to align improvements with real-world use cases.

Discussion

Clinical LLMs claim proficiency in various medical tasks, yet their validation is still largely based on static, structured assessments, such as multiple-choice questions. Although these assessments showcase certain capabilities, they do not capture the dynamic complexity

of real-world clinical practice. Our evaluation using the CRAFT-MD framework revealed that LLMs perform notably worse in conversational settings compared to examination-based evaluations. This discrepancy highlights the need for more realistic testing approaches before LLMs can be confidently integrated into clinical workflows. We propose several recommendations to align LLM evaluations with the demands of clinical practice, for their potential use as future diagnostic tools (Table 1).

Medical conversations are inherently more complex than static examination questions, requiring iterative information exchange, clarification of symptoms and continuous diagnostic reasoning. Therefore, studies demonstrating high accuracy of commercial or open-source

LLMs on examination-style medical cases^{16–18} may present an overly optimistic outlook. Our findings show a consistent decline in diagnostic accuracy when LLMs are evaluated in conversational contexts, emphasizing the importance of using a doctor–patient interaction framework for testing these models (Recommendation 1).

In these conversational contexts, evaluating the open-ended diagnostic reasoning of LLMs is crucial. Models must be able to ask relevant questions for comprehensive history-taking, reason through scattered information and interpret multimodal data, such as images. Current evaluations^{16,33–37} often focus on immediate, structured unimodal queries—such as multiple-choice questions—and overlook these more complex requirements. In line with previous studies^{20,38,39}, we found that LLMs perform worse when confronted with open-ended questions instead of MCQs, suggesting that they rely heavily on the structure provided by traditional formats. We recommend transitioning to open-ended questioning⁴⁰, which more accurately mirrors the unstructured nature of real clinical reasoning (Recommendation 2). Additionally, our findings showed that LLMs frequently missed critical details during history-taking, considerably impairing their diagnostic abilities. This underscores the need for evaluations that assess the model's capacity to ask the right questions and extract essential information (Recommendation 3).

The diagnostic accuracy of LLMs also dropped significantly when information was spread across multiple dialogues rather than presented as a concise vignette. This could be due to challenges in processing extended textual contexts⁴¹ or the predominance of structured vignettes in training data. Future development should focus on improving context comprehension and information integration for more effective use in clinical conversations (Recommendation 4), potentially through techniques such as chain-of-thought⁴². We also observed limited success in using images for diagnostic purposes, revealing the need for better integration of verbal histories with visual examination findings⁴³ and possibly other diagnostic data, such as electrocardiograms and blood tests (Recommendation 5). Moving forward, continuous evaluation of both conversational and multimodal interpretation skills should be prioritized in the development of LLMs (Recommendation 6). Additionally, refining the structure of prompts that guide model responses could further enhance their performance (Recommendation 7). We advocate for a balanced approach where LLMs complement, rather than replace, the nuanced diagnostic process of physicians⁴⁴.

Beyond diagnostic reasoning, ensuring scalability and reliability in evaluations is paramount. One key challenge in conversational evaluations involving human participants⁴⁵—whether real patients or individuals posing as one—is that these evaluations are resource intensive. The CRAFT-MD framework addresses this limitation by using LLMs as primary evaluators, reserving human involvement for confidence estimations. It uses AI agents^{46,47} to simulate patient interactions, allowing for large-scale, rapid testing without risking real patient exposure to unverified LLMs. These AI agents simulate realistic interactions, where patients disclose information only when prompted, mimicking OSCE-style assessments. However, our study revealed that these agents were sometimes unreliable when answering questions beyond the scope of the case vignette, potentially underestimating the accuracy of LLMs. To resolve this, future work should focus on developing more sophisticated AI agents that can interpret non-verbal cues, such as facial expressions, tone and body language (Recommendation 8). Additionally, periodically involving human evaluators to assess the reliability of the LLMs remains essential for their real-world deployment (Recommendation 9). The flexible design of CRAFT-MD allows for the integration of improved patient-AI models as they become available, ensuring continuous advancement of the evaluation process.

Finally, the evaluation framework itself relies on diverse, publicly available datasets. Although our study spanned multiple medical specialties, it did not assess the impact of race and ethnicity on

diagnosis due to limited diversity in the datasets. Additionally, many case vignettes lacked sufficient details for precise diagnoses without answer options. We performed MELD analysis and generated a private case vignette dataset to address concerns about training dataset memorization⁶. However, we were unable to conduct a more comprehensive analysis because training datasets for many open-source and commercial LLMs were unavailable⁴⁸. We recommend developing case vignettes that enable open-ended analysis and evaluate potential biases in LLMs to better assess their diagnostic performance across diverse populations. Full transparency, including public access to both model weights and training datasets, should be encouraged (Recommendation 10). These recommendations lay the groundwork for a more nuanced and comprehensive approach to evaluating LLMs, aligning our assessment methods with the complexities and subtleties of real-world medical practice.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-024-03328-5>.

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* **96**, 1300–1307 (2011).
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. <https://doi.org/10.48550/arXiv.2303.12712> (2023).
6. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. <https://doi.org/10.48550/arXiv.2303.13375> (2023).
7. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (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 Digit. Med.* **6**, 135 (2023).
15. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and adoption of large language models in medicine. *JAMA* **330**, 866–869 (2023).

16. Ali, R. et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery* **93**, 1090–1098 (2023).
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 Digit. Health* **2**, e0000198 (2023).
19. Han, T. et al. MedAlpaca—an open-source collection of medical conversational AI models and training data. <https://doi.org/10.48550/arXiv.2304.08247> (2023).
20. Strong, E. et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Intern. Med.* **183**, 1028–1030 (2023).
21. Nair, V., Schumacher, E., Tso, G. & Kannan, A. DERA: enhancing large language model completions with dialog-enabled resolving agents. <https://doi.org/10.48550/arXiv.2303.17071> (2023).
22. OpenAI et al. GPT-4 technical report. <https://doi.org/10.48550/arXiv.2303.08774> (2023).
23. Brown, T. B. et al. Language models are few-shot learners. <https://doi.org/10.48550/arXiv.2005.14165> (2020).
24. Jiang, A. Q. et al. Mistral 7B. <https://doi.org/10.48550/arXiv.2310.06825> (2023).
25. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. <https://doi.org/10.48550/arXiv.2307.09288> (2023).
26. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf (2023).
27. Yang, Z. et al. The dawn of LLMs: preliminary explorations with GPT-4V(ision). <https://doi.org/10.48550/arXiv.2309.17421> (2023).
28. Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
29. Martinez, V., Schleicher, S., Falci, M. & Rau, J. PA & NP medical guidance. *Clinical Advisor*. <https://www.clinicaladvisor.com/> (2019).
30. Lowell, B. A., Froelich, C. W., Federman, D. G. & Kirsner, R. S. Dermatology in primary care: prevalence and patient disposition. *J. Am. Acad. Dermatol.* **45**, 250–255 (2001).
31. Buckley, T., Diao, J. A., Rodman, A. & Manrai, A. K. Accuracy of a vision-language model on challenging medical cases. <https://doi.org/10.48550/arXiv.2311.05591> (2023).
32. Image Challenge. *N. Engl. J. Med.* <https://www.nejm.org/image-challenge> (2024).
33. 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).
34. 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* **37**, 3694–3695 (2023).
35. 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).
36. 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).
37. Lai, U. H. et al. Evaluating the performance of ChatGPT-4 on the United Kingdom medical licensing assessment. *Front. Med.* **10**, 1240915 (2023).
38. Liu, K.-C. et al. Performance of ChatGPT on Chinese Master's Degree Entrance Examination in Clinical Medicine. *PLoS One* **19**, e0301702 (2024).
39. Ueda, D. et al. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. *BMC Digit. Health* **2**, 4 (2023).
40. Tu, T. et al. Towards conversational diagnostic AI. <https://doi.org/10.48550/arXiv.2401.05654> (2024).
41. Liu, N. F. et al. Lost in the middle: how language models use long contexts. *Trans. Assoc. Comput. Linguist.* **12**, 157–173 (2024).
42. Introducing OpenAI o1-preview. <https://openai.com/index/introducing-openai-o1-preview/> (2024).
43. Rajpurkar, P. & Lungren, M. P. The current and future state of AI interpretation of medical images. *N. Engl. J. Med.* **388**, 1981–1990 (2023).
44. 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).
45. Chiang, C.-H. & Lee, H.-Y. Can large language models be an alternative to human evaluations? In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. <https://doi.org/10.18653/v1/2023.acl-long.870> (Association for Computational Linguistics, 2023).
46. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623**, 493–498 (2023).
47. 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).
48. OpenAI. GPT-4 Technical Report. <https://cdn.openai.com/papers/gpt-4.pdf> (2023).
49. Goldberger, T. & Armoni-Weiss, G. Miliaria crystallina. *N. Engl. J. Med.* **388**, e68 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2025

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ²Department of Computer Science, Stanford University, Stanford, CA, USA. ³Department of Dermatology, Medstar Georgetown University Hospital/Washington Hospital Center, Washington, DC, USA. ⁴Department of Dermatology, Northwestern University, Chicago, IL, USA. ⁵Division of Dermatology, David Geffen School of Medicine at the University of California, Los Angeles, Los Angeles, CA, USA. ⁶Department of Dermatology, Stanford University, Stanford, CA, USA. ⁷Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. ⁸Department of Emergency Medicine, Stanford University, Stanford, CA, USA. ⁹Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ¹⁰These authors contributed equally: Shreya Johri, Jaehwan Jeong. ¹¹These authors jointly supervised this work: Roxana Daneshjou, Pranav Rajpurkar. ✉ e-mail: roxanad@stanford.edu; pranav_rajpurkar@hms.harvard.edu

Methods

Dataset

Evaluation of text-only LLMs. We evaluated text-only LLMs using 2,000 case vignette-based questions, each with four answer options (see ‘Data availability’). Of these, 1,800 were sourced from the MedQA-USMLE²⁸ dataset, covering 12 medical specialties: Dermatology, Hematology and Oncology, Neurology, Gastroenterology, Pediatrics and Neonatology, Cardiology, Infectious Disease, Obstetrics and Gynecology, Urology and Nephrology, Endocrinology, Rheumatology and Others (Extended Data Fig. 1). Specialties were categorized using GPT-4 (version 6 November 2023). Each vignette included key patient details, such as age, sex, symptoms, medical history, medications and, in some cases, physical examination and laboratory results.

To streamline the dataset, we filtered the original MedQA-USMLE dataset—initially containing 14,369 questions—down to 1,804 by identifying case vignettes that focused on diagnosis. Four questions were reserved for prompt optimization, leaving 1,800 questions for final evaluation.

We also focused on collecting additional case vignettes for skin diseases from both public and private data sources. This included 100 questions from an online question bank²⁹ (Derm-Public) and 100 newly created questions (Derm-Private) developed by three dermatologists: D1 (10 questions), D2 (10 questions) and D3 (80 questions). The dermatologists created questions similar to Derm-Public but covering different conditions. They ensured that each vignette had one most likely diagnosis, and D3 reviewed vignettes from MedQA-USMLE ($n = 117$) and Derm-Public ($n = 100$) to determine whether each had a single diagnosis or multiple possible diagnoses.

Evaluation of multimodal LLMs. For multimodal LLMs (GPT-4V), we used the NEJM Image Challenge dataset³², consisting of case vignettes paired with images. We manually downloaded 100 image–vignette pairs, but 26 were excluded due to GPT-4V’s content filter, leaving 74 for final evaluation. These cases relied heavily on the provided images for diagnosis, unlike the text-based evaluations.

MELD analysis

To analyze dataset contamination, we applied MELD to the 2,000 case vignettes. MELD evaluates the similarity between a model’s generated response and the actual answer by calculating the inverse of the length-normalized Levenshtein distance. A similarity score of 0.95 or higher suggests that the test question was likely part of the model’s training set⁶.

Although MELD is precise in detecting matches, its recall is unknown. Therefore, a detected match indicates likely memorization of the test data, but the absence of a match does not guarantee exclusion from the training set, as some instances of memorization may go undetected.

The Levenshtein distance between the original case vignette and the model-generated completion was calculated using the ‘Levenshtein’ Python package. The following prompt was used to get model-generated completions:

You are given the first half of a medical case vignette. Generate the second half of the case vignette. You do not have to give the diagnosis. Generate only {word_count} words.

****Case Vignette**:** {first_half_case_vignette}

Accessing LLMs

The results presented in this paper were generated using two commercial models and two open-source models. For the commercial models (GPT-4 and GPT-3.5), we accessed the 6 November 2023 versions through the Azure API. For the open-source models

(LLaMA-2-7b, Mistral-v1-7b and Mistral-v2-7b), we employed the HuggingFace implementations: ‘meta-llama/Llama-2-7b-chat-hf’, ‘mistralai/Mistral-7B-Instruct-v0.2’ and ‘mistralai/Mistral-7B-Instruct-v0.1’, respectively.

Multi-agent AI conversational framework

We introduced a multi-agent AI conversational framework, including a clinical LLM assessed through interactions with patient-AI and grader-AI agents, along with medical experts. The patient-AI agents simulated conversations for each of the 2,000 case vignettes. The prompts for each AI agent and clinical LLM were iteratively developed in collaboration with clinicians. To optimize performance, prompts were tested on a 0.2% development set ($n = 4$), refining them for the four models evaluated (GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b) and kept consistent for the full dataset.

Each conversation was repeated five times per vignette to capture response variability, generating 10,000 simulated conversations. The grader-AI agent assessed diagnostic accuracy, and medical experts reviewed AI agent performance. Conversations were conducted in two formats—multiple-choice (MCQ) and free-response (FRQ)—and were terminated when the clinical LLM provided a ‘Final Diagnosis’ or failed to ask a follow-up question (indicated by the absence of a question mark in the dialogue).

Clinical LLM

The clinical LLM was designed to ask the patient relevant questions about current symptoms, medical history, medications and, when necessary, family history. It continued asking follow-up questions until it reached what it perceived to be a confidence diagnosis. The following prompt was used:

*You are an AI doctor specializing in <insert medical specialty>. Arrive at a diagnosis of a patient’s medical condition. Ask only one question at a time, and it should not be more than 1 line. Continue asking questions until you’re 100% confident of the diagnosis. Do not ask the same question multiple times. Ask different questions to cover more information. 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 questions short and brief to not confuse the patient. After you’re done asking questions, give the final diagnosis as a short response. Do not explain, only give the diagnosis name. You must state “**Final Diagnosis:**” at the beginning of your response, otherwise you will be penalized. You must give only 1 diagnosis otherwise you will be penalized.*

Patient AI agent

The patient-AI agent was given a case vignette and tasked with answering the clinical LLM’s follow-up questions. It was explicitly instructed to provide information incrementally, responding only to the questions asked rather than revealing the entire case at once. To ensure accurate responses, the patient-AI agent was also 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. You have to describe your symptoms from the given case vignette based on the questions asked. Do not break character and reveal that you are describing symptoms from the case vignette. Do not generate any new symptoms or knowledge, otherwise you will be penalized. Do not reveal more information than what the question asks. Keep your answer short, to only 1 sentence. Simplify terminology used in the given paragraph to layman language.

****Case Vignette**:** <insert case_vignette>

Grader-AI agent

We used the grader-AI agent to quantify the diagnostic accuracy for FRQ experiments. The grader-AI agent was instantiated using GPT-4 to maintain consistency of evaluation across all experiments and models. This was done in two-steps.

Step 1: diagnosis name extraction. The grader-AI agent first categorized the clinical LLM's final diagnosis according to the following three categories: (1) single diagnosis, (2) multiple diagnoses and (3) no diagnosis. We estimated the error rate for this step through manual verification to be less than 0.5% (one mistake in ~200 conversations). The following prompt was used:

*Identify and return the dermatology diagnosis name from the given **Paragraph**. If there are more than one diagnoses present, return **Multiple**. If there are no diagnoses present, then return **None**. If there is a main diagnosis with a concurrent minor diagnosis, return the name of the main diagnosis. Do not explain.*

Paragraph: <insert clinical LLM response>

For the clinical LLM's responses that 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 an accuracy of 0.

Step 2: compare extracted diagnosis to correct answer. For comparing the clinical LLM's responses to the correct answers across experiments, few-shot prompting was used to accommodate alternative medical terminologies. If the vignette answer was a subtype of the clinical LLM's diagnosis, the response was marked as correct. However, if the clinical LLM's diagnosis was a more specific subtype than the vignette answer, it was marked as incorrect. This is because it is possible that the clinical LLM could have made an unsupported leap to a more specific diagnosis, which may not be justified by the information available in the vignette. This could lead to potential underestimation of the model's performance, which is also a limitation of vignette-based evaluations. The following prompt was used:

*Identify if the two query medical diagnoses are equivalent or synonymous names of the disease. Respond with a yes/no. Do not explain. Also, if **Diagnosis 1** is a subtype of **Diagnosis 2** respond with yes, but if **Diagnosis 2** is a subtype of **Diagnosis 1** respond with no.*

Example 1: **Diagnosis 1**: eczema, **Diagnosis 2**: eczema. They are the same, so respond Yes.

Example 2: **Diagnosis 1**: eczema, **Diagnosis 2**: onychomycosis. They are different, so respond No.

Example 3: **Diagnosis 1**: toe nail fungus, **Diagnosis 2**: onychomycosis. They are synonymous, so return Yes.

Example 4: **Diagnosis 1**: wart, **Diagnosis 2**: verruca vulgaris. They are synonymous, so return Yes.

Example 5: **Diagnosis 1**: lymphoma, **Diagnosis 2**: hodgkin's lymphoma. Diagnosis 2 is subtype of Diagnosis 1, so return No.

Example 6: **Diagnosis 1**: hodgkin's lymphoma, **Diagnosis 2**: lymphoma. Diagnosis 1 is subtype of Diagnosis 2, so return Yes.

Example 7: **Diagnosis 1**: melanoma, **Diagnosis 2**: None. They are different, so respond No.

Example 8: **Diagnosis 1**: melanoma, **Diagnosis 2**: Multiple. They are different, so respond No.

Query Diagnosis 1: <insert correct answer>

Query Diagnosis 2: <insert clinical LLM's extracted 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 examination.

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

Single-turn conversations. The patient-AI agent's initial symptom summary (that is, first dialogue in a multi-turn conversation) was used as a single-turn conversation. The clinical LLM had to make the diagnosis without asking any follow-up questions in this case.

Summarized conversations. These were generated using the Conversational Summarization technique. All the patient-AI agent's dialogues were extracted from the multi-turn conversations. GPT-3.5 was used with few-shot prompting to generate the summarized conversations. The following prompt was used:

*Convert the following **Query Vignette** into 3rd person. Do not add any new information otherwise you will be penalized. A demonstrative **Example** is provided after the query vignette.*

Query Vignette: <insert patient-AI agent dialogues>

For example:

Original Vignette - 'I have painful sores on my penis and swelling in my left groin that began 10 days ago. I am 22 years old. No, I haven't had symptoms like this before. My female partner was diagnosed with chlamydia last year, but I haven't been checked for it. No, I don't have any other medical conditions and I'm not taking any medications. There's no mention of a family history of skin conditions or autoimmune diseases in my case.'

Converted Vignette - 'A patient presents to the clinic with several concerns. The patient is 22 years old and has not had symptoms like this before. The patient's female partner was diagnosed with chlamydia last year, but the patient has not been checked for it. The patient does not have any other medical conditions and is not taking any medications. There's no family history of skin conditions or autoimmune diseases.'

Varying number of answer choices

Four-choice MCQs. For four-choice MCQs, the clinical LLM was provided with the answer choices after case vignette or conversation (multi-turn, single-turn and summarized). In case of multi-turn conversations, the response containing the final diagnosis was removed before providing the answer choices.

For vignette and summarized conversation followed by four-choice MCQ, the following prompt was used:

You are an AI doctor specializing in <specialty>. You are given the patient's symptoms and a list of possible answer choices. Only one of the choices is correct. Select the correct choice, and give the answer as a short response. Do not explain.

****Symptoms****: <case_vignette>

****Choices****: <choices>

For single-turn and multi-turn conversation followed by four-choice MCQ, the following prompt was used:

Choose the correct option based on the patient's above symptoms and a list of possible options. Only one of the choices is correct. Give the answer as a short response. Do not explain.

****Choices****: <insert answer choices>

The clinical LLM refused to select a 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.

FRQs. The clinical LLM was presented with a case vignette or conversation (multi-turn, single-turn and summarized) or a physical examination and was asked to give a diagnosis. In case of multi-turn conversations, the final response of the clinical LLM containing the diagnosis was removed.

For vignettes and summarized conversation (FRQ setting), 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 single-turn and multi-turn conversations (FRQ setting), the following prompt was used:

Based on the patient's above symptoms, give the diagnosis as a short response. Do not explain.

Expert evaluation

Accuracy of medical experts on case vignettes. To have a human baseline for benchmarking the performance of various LLMs, the accuracy of two board-certified dermatologists (D2 and D5) was assessed. Different board-certified dermatologists graded the four-choice MCQ and FRQ experiments respectively to prevent biased grading of FRQs due to familiarity with answer choices, and they were intentionally chosen to be different from the dermatologists who created the Derm-Private case vignettes.

Assessment of clinical LLMs. We performed expert evaluations of 180 multi-turn conversations with four dermatologists (D1, D3, D4 and D5). These conversations were equally distributed among the four evaluated models (GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b) and based on dermatology case vignettes that had a single most likely diagnosis (15 each from the three datasets: MedQA-USMLE, Derm-Public and Derm-Private). Dermatologists D1, D3 and D5 evaluated these conversations, with a third of them being dual annotated to also estimate expert agreement. In cases when the two dermatologists disagreed, a third dermatologist (D4) broke the tie.

The following questions were asked for clinical LLM evaluation:

- Did the clinical LLM stop asking questions when only a single most likely diagnosis was possible?
- Did the clinical LLM elicit the relevant medical history from the vignette (excluding the physical exam, lab results)?

Assessment of patient-AI agent. To assess the reliability of the patient-AI agent, medical experts (dermatologists D1, D3 and D5) were asked the following questions. D4 broke the tie when there was a disagreement.

- Is the patient-AI agent using medical terminology? Please respond with a yes/no. Medical terminology includes primary and secondary morphological descriptive terms (for example, macule, papule, pustule, plaque, erosions, lichenification) while examples of spot, bump, blister, ulcer and pus bump are examples of expected non-medical patient terminology. Additional medical terminology includes non-skin exam findings such as shoddy lymphadenopathy and terminology referencing anatomic locations such as glabella rather than forehead and subcutaneous nodules on my shins rather than bumps on the shins/legs.
- Was the patient-AI agent's answer to clinical LLM's questions based on information provided in the case vignette? Please note the hallucination in the comments, if not.
- Did the patient-AI agent provide complete information related to the question asked? Please note the missing information in the comments, if not.

Assessment of grader-AI agent. The correlation between accuracies of the clinical LLM as annotated by grader-AI and dermatologists was compared. The vignette + FRQ experiment was annotated by a dermatologist (D4) for the dermatology case vignettes (Derm-Public, Derm-Private and Derm-USMLE; $n = 317$) to assess the correlation. The following question was asked:

Is the clinical LLM's diagnosis equivalent to the correct vignette answer?

If the clinical LLM's diagnosis is a subtype of the correct answer, then it is incorrect. If the correct answer is a subtype of the clinical LLM's diagnosis, then it is correct. Below are some examples -

Example 1: Clinical-LLM agent's diagnosis = eczema, correct vignette answer = onychomycosis. They are different, so incorrect.

Example 2: Clinical-LLM agent's diagnosis = toenail fungus, correct vignette answer = onychomycosis. They are synonyms, so correct.

Example 3: Clinical-LLM agent's diagnosis = wart, correct vignette answer = verruca vulgaris. They are synonyms, so correct.

Example 4: Clinical-LLM agent's diagnosis = lymphoma, correct vignette answer = hodgkin's lymphoma. Correct answer is a subtype of clinical LLM's diagnosis, so correct.

Example 5: Clinical-LLM agent's diagnosis = hodgkin's lymphoma, correct vignette answer = lymphoma. Clinical-LLM agent's diagnosis is a subtype of correct vignette answer, so incorrect.

Statistical tests

Bootstrap testing. *P* values were computed using the bootstrap method (see 'Code availability') to determine the statistical significance of the differences between two paired samples. The bootstrap

procedure was repeated 10,000 times to generate a distribution of differences. The observed test statistic was calculated as the mean of the differences between the two samples. For each iteration, we computed the differences between the paired samples and generated bootstrap samples by randomly sampling with replacement from these differences. To calculate the *P* value for a two-tailed test, we counted the number of bootstrap sample statistics that were as extreme or more extreme than the observed test statistic. The *P* value was then computed using the formula $(\text{extreme_count} + 1) / (\text{num_bootstrap_samples} + 1)$, adjusting to include the observed statistic. The random seed was set to ensure reproducibility. To control the family-wise error rate, the final reported *P* values were adjusted using the Holm–Bonferroni correction method. In cases where the *P* value was less than 0.0001, it was reported as $P < 0.0001$.

McNemar test. The McNemar test was used to evaluate the statistical significance of differences in binary paired data. The ‘statsmodels’ package was used to perform this test.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

MedQA-USMLE case vignettes were downloaded from <https://github.com/jind11/MedQA>. Derm-Public case vignettes were downloaded from <https://www.clinicaladvisor.com/>. The images and corresponding vignettes for the NEJM Image Challenge were downloaded from <https://www.nejm.org/image-challenge>. The private dataset generated as a part of our study can be found at <https://github.com/rajpurkarlab/craft-md>. All case vignettes used in the study are also available in the following repository: <https://github.com/rajpurkarlab/craft-md>.

Code availability

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

Acknowledgements

S.J. is supported by the Quad Fellowship. This research project has benefitted from the Microsoft Accelerate Foundation Models Research grant program awarded to P.R., and we especially thank K. Takeda for facilitating access to resources. We are also appreciative of Harvard Medical School’s Dean’s Innovation Award (awarded to P.R.) for funding this study.

Author contributions

P.R. and R.D. conceived the study. S.J. planned and performed all experiments. S.J. and J.J. performed data analysis. B.A.T., D.I.S. and Z.R.C. created new case vignettes for the private dataset. B.A.T., D.I.S., S.W. and L.A.B. performed assessment of the clinical LLMs and patient-AI agent. S.W. performed assessment of the grader-AI agent. L.A.B. and Z.R.C. performed experiments to assess human expert accuracy. S.J., R.D. and P.R. contributed to the interpretation of findings and drafted the paper. H.-Y.Z., D.K. and E.M.V.A. provided

critical feedback on the technical and clinical content. All authors provided critical feedback and substantially contributed to the revision of the paper. All authors read and approved the final paper.

Competing interests

R.D. reports receiving personal fees from DWA, personal fees from Pfizer, personal fees from L’Oreal, personal fees from VisualDx and stock options from MDA Algorithms and Revea outside the submitted work and has a patent for Truelmage pending. D.I.S. is the co-founder of FixMySkin Healing Balms, a shareholder in Appiell Inc. and K-Health, a consultant for Appiell Inc. and LuminDx and an investigator for AbbVie and Sanofi. E.M.V.A. serves as an advisor to Enara Bio, Manifold Bio, Monte Rosa, Novartis Institute for Biomedical Research and Serinus Bio. E.M.V.A. provides research support to Novartis, Bristol Myers Squibb, Sanofi and NextPoint. E.M.V.A. holds equity in Tango Therapeutics, Genome Medical, Genomic Life, Enara Bio, Manifold Bio, Microsoft, Monte Rosa, Riva Therapeutics, Serinus Bio and Syapse. E.M.V.A. has filed for institutional patents on chromatin mutations and immunotherapy response and methods for clinical interpretation and provides intermittent legal consulting on patents to Foaley & Hoag. E.M.V.A. also serves on the editorial board of *Science Advances*. The other authors declare no competing interests.

Ethics Declaration

The CRAFT-MD framework is designed to enable faster evaluation of LLMs for leading clinical conversations and to uncover limitations to guide future model development. These LLMs could enhance clinical workflows by engaging in preliminary conversations with patients, collecting and summarizing relevant medical information and presenting these data to doctors before patient visits, potentially improving the effectiveness of doctor–patient interactions. These LLMs could be more effective than the pre-visit questionnaires, given their ability to lead dynamic conversations. However, this will require not only developing more capable LLMs but also making them more fault tolerant and cognizant of appropriate empathetic behavior.

Additional information

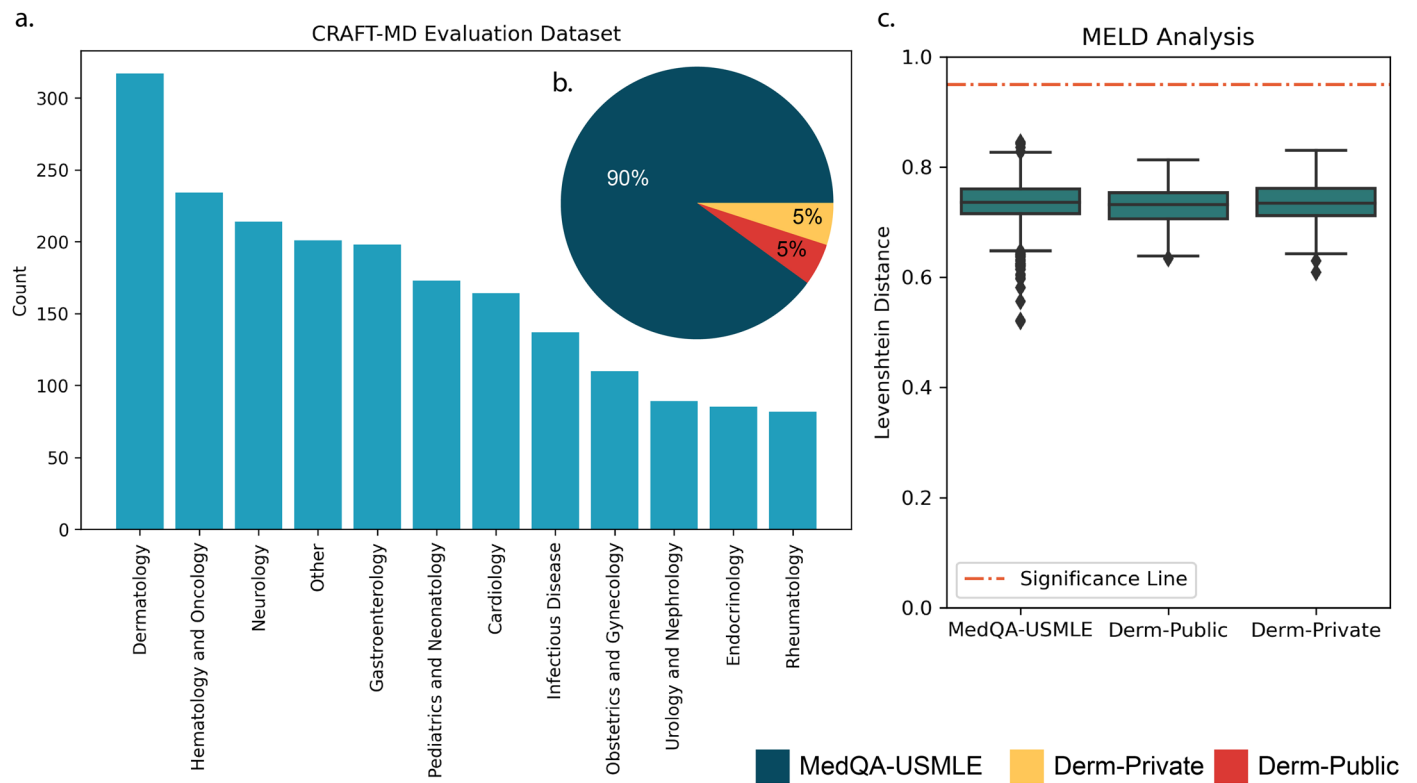
Extended data is available for this paper at <https://doi.org/10.1038/s41591-024-03328-5>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-024-03328-5>.

Correspondence and requests for materials should be addressed to Roxana Daneshjou or Pranav Rajpurkar.

Peer review information *Nature Medicine* thanks Milica Gasic, Pearse A. Keane and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Distribution of case vignettes across medical specialties and source datasets. (a) CRAFT-MD evaluation dataset, showing the distribution of case vignettes across 12 medical specialties - Dermatology, Hematology and Oncology, Neurology, Gastroenterology, Pediatrics and Neonatology, Cardiology, Infectious Disease, Obstetrics and Gynecology,

Urology and Nephrology, Endocrinology, Rheumatology and Others. **(b)** Inset pie chart showing the proportion of case vignettes based on source of curation (MedQA-USMLE, Derm-Public and Derm-Private). **(c)** MELD analysis showing Levenshtein Distance between original and GPT-4 completed case vignettes.

(i) Case Vignette

A 42-year-old woman presents to the clinic for a recurrent rash that has remitted and relapsed over the last 2 years. The patient states that she has tried multiple home remedies when she has flare-ups, to no avail. The patient is wary of medical care and has not seen a doctor in at least 15 years. On examination, she has multiple disc-shaped, erythematous lesions on her neck, progressing into her hairline. The patient notes no other symptoms. Lab work is performed and is positive for antinuclear antibodies.

(ii) Case Vignette

A 38-year-old woman with a history of systemic lupus erythematosus, obesity, and hyperlipidemia presents to her primary care physician for evaluation of new bilateral blisters on her hands. She says that she first noticed these blisters shortly after returning from a weekend trip to the beach two weeks prior. She denies any fevers, joint pains, or other skin rash during this period. The patient works in the adult entertainment industry. On examination, there are multiple flaccid blisters of the bilateral dorsal hands with hemorrhagic crusts, mild scarring, and hyperpigmentation. In addition, increased hair growth is noted on the bilateral malar cheeks.

(iii) Case Vignette

A 5-year-old boy presents to the pediatrician office with complaints of multiple spots that began on the abdomen and spread to the arms and legs. The spots are not itchy or painful, unless he scratches them. Sometimes the skin surrounding the lesion will become very red and painful, but then the lesion and redness fade. The parents report they have tried topical corticosteroids but the lesions do not go away. Mom endorses a history of atopic dermatitis in her sister and herself. Physical exam reveals multiple dome shaped skin colored to light pink papules with **central umbilication and yellow core** across the chest, abdomen, and back.

Extended Data Fig. 2 | Examples of summarized conversations. (i) an ideal summarized conversation containing all relevant medical history, without hallucinations or use of medical terminology; (ii) a summarized conversation

Summarized Conversation

A 42-year-old woman has been experiencing a recurring rash for the last 2 years. The rash consists of round, red patches on her neck and is spreading into her hairline. There is no mention of scaling or flaking, itching, or pain associated with the rash. The woman has not noticed any specific triggers for the rash and has not seen a doctor in at least 15 years. She has tried multiple home remedies when experiencing flare-ups, but they have not been effective. There is no mention of any allergies or sensitivities in her case.

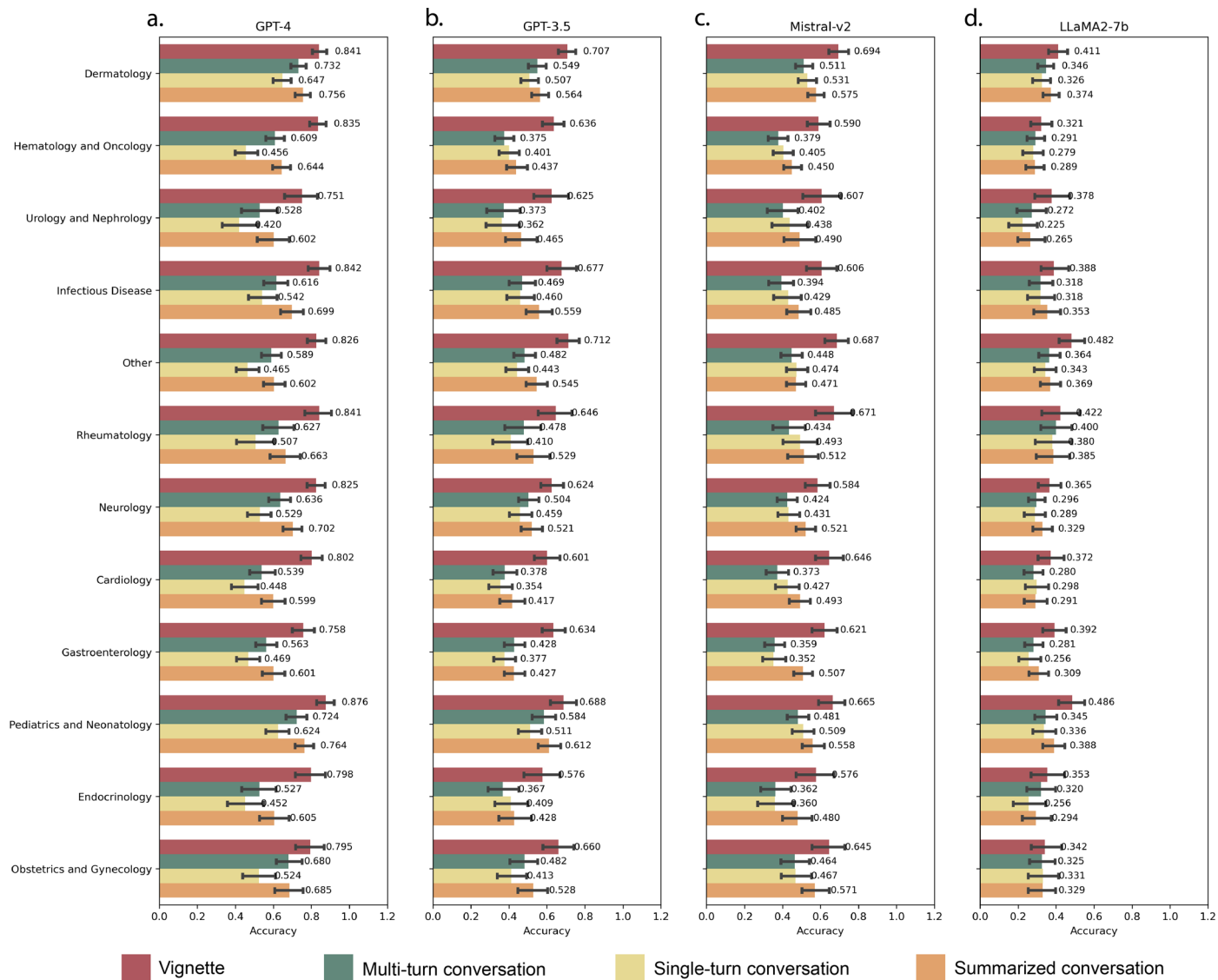
Summarized Conversation

A 38-year-old woman presents to her primary care physician with new blisters on both of her hands and increased hair growth on her cheeks. She first noticed the blisters shortly after a weekend trip to the beach. She is unable to provide information on her family history regarding skin diseases or hormonal disorders. **Mild scarring and hyperpigmentation** is present around the blisters on her hands, and she has not experienced fevers, joint pains, or other skin rash. She denies any itchiness, pain, or changes in urine color and reports that the blisters have not healed, prompting her visit to the physician for evaluation.

Summarized Conversation

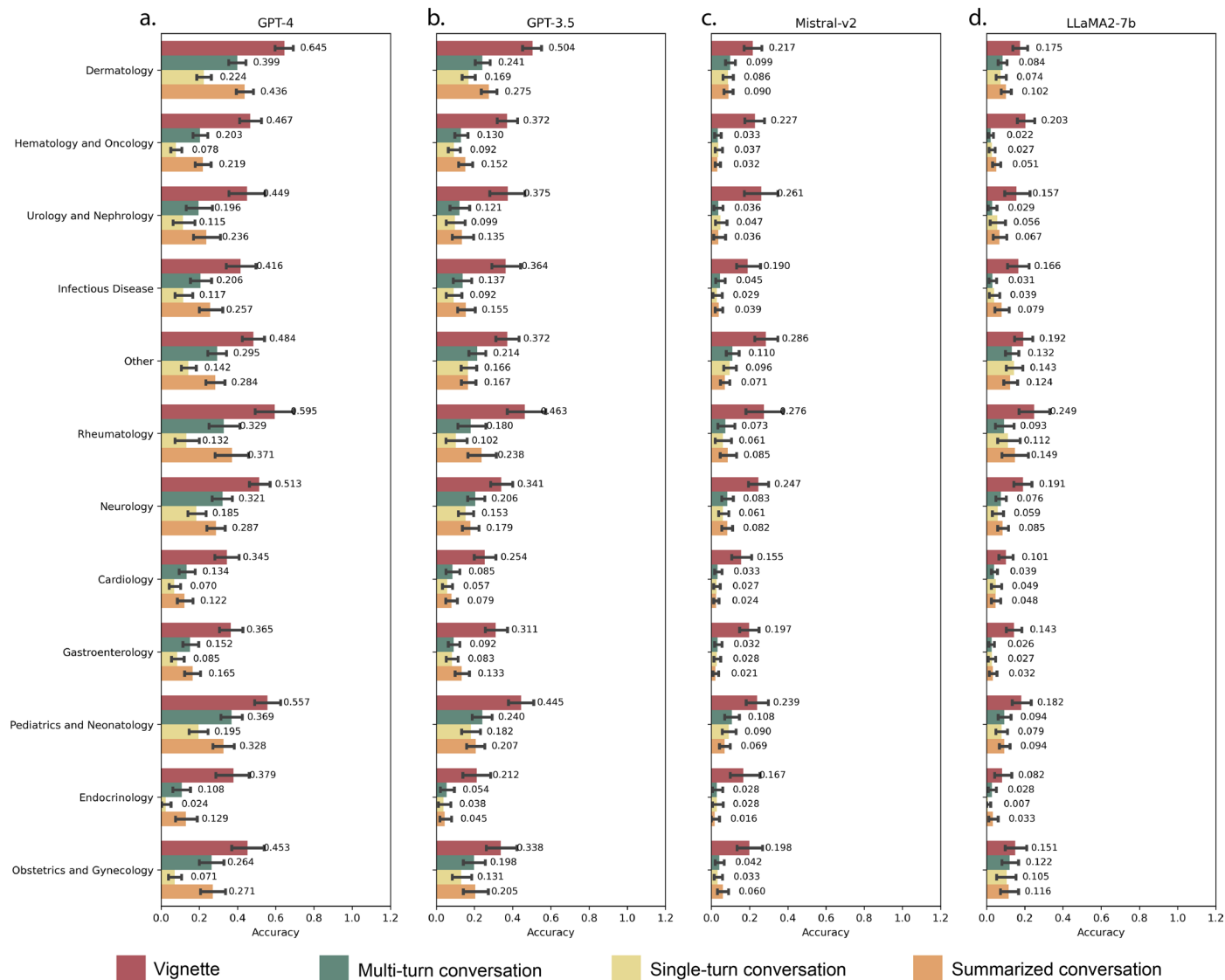
A 5-year-old child presents to the clinic with multiple non-itchy spots on the tummy that have spread to the arms and legs. The spots have sometimes become red and painful, then improved. The child's mother and aunt have a history of atopic dermatitis, suggesting a family history of skin allergies. The spots have persisted despite treatment with cream, and the surrounding skin can become red and painful at times. The symptoms come and go over time. The specific diagnosis of psoriasis has not been mentioned in the case, and the patient is not aware of any such medical history.

with medical terminology use; red highlight indicates use of medical terminology (iii) a summarized conversation with incomplete medical history; red highlight demarcates missing information that is crucial for the diagnosis.



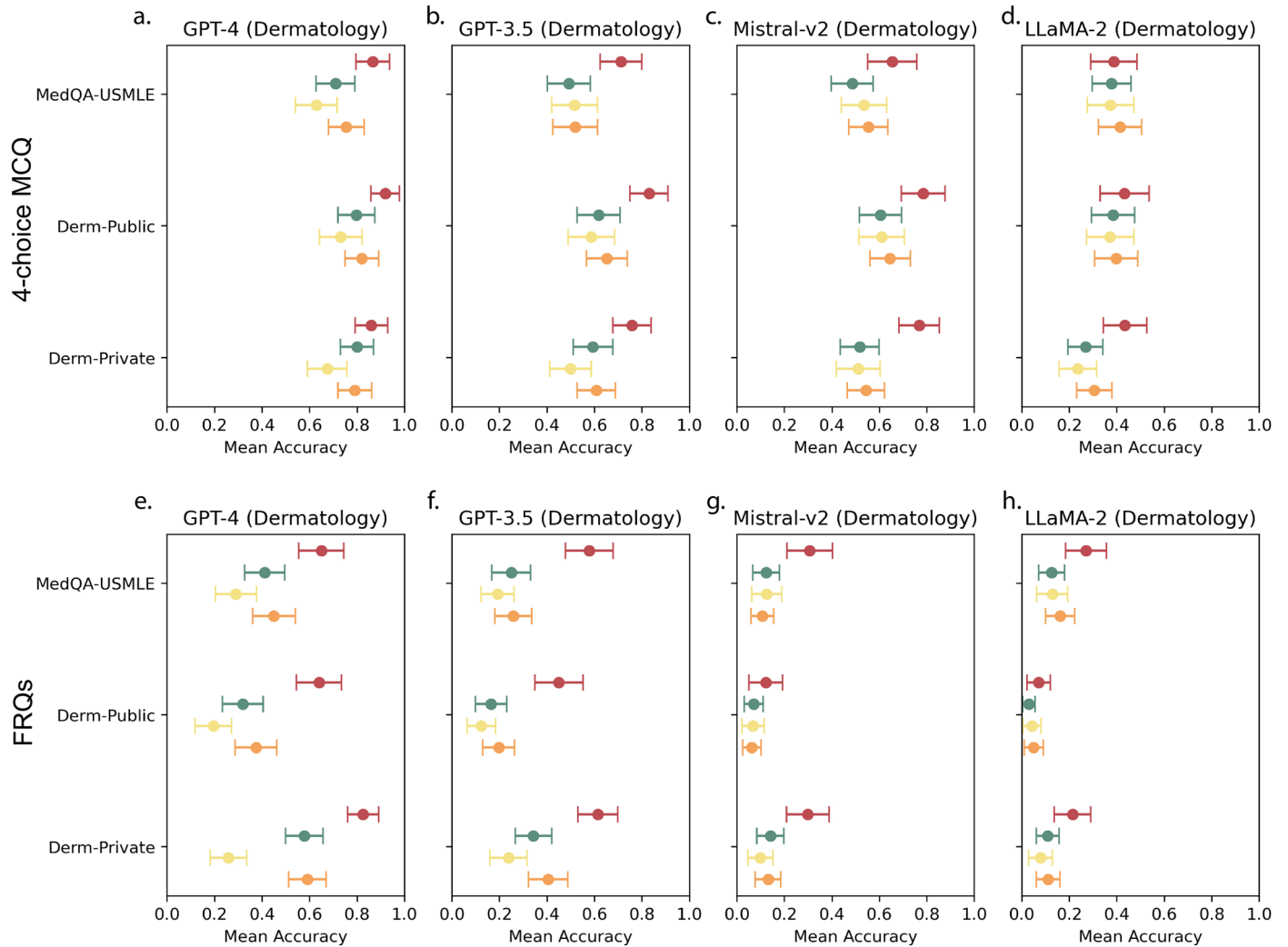
Extended Data Fig. 3 | Distribution of clinical LLM's accuracy in 4-choice MCQ across the medical specialties. Distribution of clinical LLM's accuracy in 4-choice MCQ across the 12 medical specialties for (a) GPT-4, (b) GPT-3.5, (c) Mistral-v2-7b, and (d) LLaMA2-7b. Trends for the 4 experimental settings (vignette, multi-turn conversation, single-turn conversation and summarized conversation) are consistent to the combined accuracy for all 12 specialties

- Dermatology, Hematology and Oncology, Neurology, Gastroenterology, Pediatrics and Neonatology, Cardiology, Infectious Disease, Obstetrics and Gynecology, Urology and Nephrology, Endocrinology, Rheumatology, and Others. Error bars represent 95% confidence intervals, and numbers represent the mean accuracy.



Extended Data Fig. 4 | Distribution of clinical LLM's accuracy in FRQ across the medical specialties. Distribution of clinical LLM's accuracy in FRQs across the 12 medical specialties for (a) GPT-4, (b) GPT-3.5, (c) Mistral-v2-7b, and (d) LLaMA-2-7b. Trends for the 4 experimental settings (vignette, multi-turn conversation, single-turn conversation and summarized conversation)

are consistent to the combined accuracy for all 12 specialties - Dermatology, Hematology and Oncology, Neurology, Gastroenterology, Pediatrics and Neonatology, Cardiology, Infectious Disease, Obstetrics and Gynecology, Urology and Nephrology, Endocrinology, Rheumatology, and Others. Error bars represent 95% confidence intervals, and numbers represent the mean accuracy.



Extended Data Fig. 5 | Trends in vignette and conversational formats in dermatology datasets, for cases with single most likely diagnosis. Trends in vignette and conversational formats persist across skin disease datasets

(MedQA-USMLE, Derm-Public and Derm-Private) for cases with single most likely diagnosis. Results are shown for both (a,b,c,d) 4-choice MCQ and (e,f,g,h) FRQ settings. Error bars represent 95% confidence intervals.

Extended Data Table 1 | Decrease in accuracy between four-choice MCQ and FRQ settings for vignette and conversational formats

	GPT-4	GPT-3.5	Mistral-v2-7b	LLaMA-2-7b
Vignette	0.334	0.284	0.415	0.226
Multi-turn conversation	0.363	0.298	0.36	0.254
Single-turn conversation	0.387	0.313	0.392	0.24
Summarized conversation	0.399	0.335	0.457	0.255

Magnitude of decrease in accuracy between four-choice MCQ and FRQ settings for vignette and conversational formats (multi-turn, single-turn and summarized) across GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b.

Extended Data Table 2 | Comparison between Mistral-v1-7b and Mistral-v2-7b accuracies

	GPT-4	GPT-3.5	Mistral-v2-7b	LLaMA-2-7b
Q1	13/15	12/15	12/15	12/15
Q2	12/15	13/15	11/15	12/15
Q3	12/15	14/15	15/15	15/15

Mean accuracy and adjusted *P* value for difference in mean accuracies for Mistral-v1-7b and Mistral-v2-7b.

Extended Data Table 3 | Inter-rater agreement for medical expert annotations

	Experiment	Mean accuracy (Mistral-v1-7b)	Mean accuracy (Mistral-v2-7b)	Difference in accuracy (v2-v1)	Adjusted p-value
4-choice MCQ	Vignette	0.441	0.637	0.196	<0.0001
	Multi-turn conversation	0.331	0.426	0.095	<0.0001
	Single-turn conversation	0.324	0.448	0.124	<0.0001
	Summarized conversation	0.361	0.513	0.152	<0.0001
FRQ	Vignette	0.165	0.211	0.048	0.0006
	Multi-turn conversation	0.08	0.065	-0.015	0.0407
	Single-turn conversation	0.06	0.055	-0.005	0.4298
	Summarized conversation	0.082	0.055	-0.027	0.0004

Inter-rater agreement for medical expert annotations to assess clinical LLM and patient-AI agent. Each cell in the table represents the number of evaluations with inter-rater agreement/total number of evaluations for the different models (GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b) and questions (Q1: Did the clinical LLM stop asking questions when only a single most likely diagnosis was possible? Q2: Did the clinical LLM elicit the relevant medical history from the vignette? Q3: Did the patient-AI agent use medical terminology in its responses?) (Methods)

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection | No software was used for data collection.

Data analysis | Python version 3.9.19 was used for performing all data analysis. python-Levenshtein package (version 0.25.1) was used for performing MELD analysis described in the methods. Seaborn (version 0.13.2) and matplotlib (version 3.8.4) packages were used for data visualization. Scipy (version 1.13.1) and statsmodel (version 0.14.2) packages was used for performing statistical tests for p-value calculation. Transformers package (version 4.39.2) was used to access models (Mistral-v1, Mistral-v2, LLaMA-2-7b) from huggingface.

All code used for data analysis is also available on the github repo, with instructions to reproduce our results - <https://github.com/rajpurkarlab/craft-md>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

MedQA-USMLE case vignettes can be downloaded from - <https://github.com/jind11/MedQA>. Derm-Public case vignettes were downloaded from - <https://www.clinicaladvisor.com/>. The images and corresponding vignettes for the NEJM Image Challenge can be downloaded from their website. (<https://www.nejm.org/image-challenge>). The private dataset generated as a part of our study can be found on <https://github.com/rajpurkarlab/craft-md>. All case vignettes used in the study are also available on the following repository: <https://github.com/rajpurkarlab/craft-md>.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	<input type="text" value="Not Applicable."/>
Population characteristics	<input type="text" value="Not Applicable."/>
Recruitment	<input type="text" value="Not Applicable."/>
Ethics oversight	<input type="text" value="Not Applicable."/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<p>For text-only evaluation of LLMs, all case vignette like questions were extracted from MedQA-USMLE dataset for this. This led to a total of 2000 case vignettes structured as 4-choice MCQs for the final evaluation, with four case vignettes used for prompt optimization.</p> <p>For multimodal evaluation of LLMs, 100 case vignettes, images pairs were downloaded from the NEJM Image Challenge website. Of these, only 26 images were restricted by the GPT-4V content filter, therefore only 74 cases were used for the final evaluation.</p> <p>These sample sizes are sufficient for detecting statistically significant differences using Mann-Whitney U Test.</p>
Data exclusions	<input type="text" value="No data/case vignettes were excluded."/>
Replication	All statistics calculations in the study had controlled seeds to ensure reproducibility (number of bootstrap samples = 10000). All attempts at replication of statistical results were successful. Large Language Models (LLMs) cannot be seeded, therefore exact prompts are provided in the Methods section of the manuscript and raw outputs for all experiments with LLMs are provided the github repository - https://github.com/rajpurkarlab/craft-md . Each LLM experiment was performed 5 times.
Randomization	Randomization was not relevant to this study, since there were no group comparisons.
Blinding	There was no group allocation - all case vignettes were evaluated across the selected LLMs. Therefore, there was no blinding in this study. Medical experts were not blinded to the LLM they were evaluating to allow them to make conclusions about overall trends across model outputs.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

- | n/a | Involvement |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |

- | n/a | Involvement |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

An evaluation framework for clinical use of large language models in patient interaction tasks

In the format provided by the authors and unedited

Supplementary Information

Contents

Supplementary Table 1: Mean accuracy and 95% confidence intervals for 4-choice MCQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 2: Adjusted p-values for 4-choice MCQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 3: Adjusted p-values between 4-choice MCQ and FRQ settings for each experimental setup (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 4: Mean accuracy and 95% confidence intervals for FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 5: Adjusted p-values for FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 6: Medical specialty wise mean accuracy and 95% confidence intervals for 4-choice MCQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 7: Medical specialty wise adjusted p-values for 4-choice MCQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 8: Medical specialty wise mean accuracy and 95% confidence intervals for FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 9: Medical specialty wise adjusted p-values for FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 10: Mean accuracy and 95% confidence intervals for Dermatology, calculated by dataset source for 4-choice MCQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 11: Adjusted p-values for Dermatology, calculated by dataset source for the 4-choice MCQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 12: Mean accuracy and 95% confidence intervals for Dermatology, calculated by dataset source for FRQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 13: Adjusted p-values for Dermatology, calculated by dataset source for the FRQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 14: Mean accuracy and 95% confidence intervals for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for 4-choice MCQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 15: Adjusted p-values for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for the 4-choice MCQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 16: Mean accuracy and 95% confidence intervals for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for FRQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 17: Adjusted p-values for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for the FRQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 18: P-values between pairs of evaluated models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b) for clinical LLM assessment by medical experts. All p-values were calculated using a McNemar test (see Methods).

Supplementary Table 19: Mean accuracy and 95% confidence intervals for 4-choice MCQ and FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) for GPT-4V and GPT-4V-without-image.

Supplementary Table 20: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to GPT-4V and GPT-4V-without-image. All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 21: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated models (GPT-4V and GPT-4V-without-image) corresponding to each of the experimental setups (vignette,

multi-turn conversation, single-turn conversation and summarized conversation). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 22: Mean accuracy and 95% confidence intervals for 4-choice MCQ and FRQ setting, across experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) for Mistral-v1-7b and Mistral-v2-7b.

Supplementary Table 23: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to Mistral-v1-7b. All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Tables

Model	Experiment	Mean Accuracy	95% C.I.
GPT-4	Vignette	0.82	(0.804, 0.837)
	Multi-turn conversation	0.627	(0.609, 0.645)
	Single-turn conversation	0.52	(0.5, 0.539)
	Summarized conversation	0.669	(0.652, 0.686)
GPT-3.5	Vignette	0.657	(0.637, 0.676)
	Multi-turn conversation	0.467	(0.448, 0.485)
	Single-turn conversation	0.435	(0.416, 0.454)
	Summarized conversation	0.507	(0.489, 0.526)
Mistral-v2-7b	Vignette	0.637	(0.616, 0.658)
	Multi-turn conversation	0.426	(0.409, 0.443)
	Single-turn conversation	0.448	(0.429, 0.468)
	Summarized conversation	0.513	(0.496, 0.529)
LLaMA-2-7b	Vignette	0.395	(0.376, 0.415)
	Multi-turn conversation	0.319	(0.303, 0.335)
	Single-turn conversation	0.304	(0.286, 0.323)
	Summarized conversation	0.335	(0.318, 0.352)

Supplementary Table 1: Mean accuracy and 95% confidence intervals for 4-choice MCQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0001	0.0001
	Multi-turn conversation	Summarized conversation	0.0001	0.0001
	Single-turn conversation	Summarized conversation	0.0001	0.0001
GPT-3.5	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0001	0.0001
	Multi-turn conversation	Summarized conversation	0.0001	0.0001
	Single-turn conversation	Summarized conversation	0.0001	0.0001
Mistral-v2-7b	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0009	0.001
	Multi-turn conversation	Summarized conversation	0.0001	0.0001
	Single-turn conversation	Summarized conversation	0.0001	0.0001
LLaMA-2-7b	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0314	0.0314
	Multi-turn conversation	Summarized conversation	0.011	0.0115
	Single-turn conversation	Summarized conversation	0.0001	0.0001

Supplementary Table 2: Adjusted p-values for 4-choice MCQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	p-value	Adjusted p-value
GPT-4	Vignette	0.0001	0.0001
	Multi-turn conversation	0.0001	0.0001
	Single-turn conversation	0.0001	0.0001
	Summarized conversation	0.0001	0.0001
GPT-3.5	Vignette	0.0001	0.0001
	Multi-turn conversation	0.0001	0.0001
	Single-turn conversation	0.0001	0.0001
	Summarized conversation	0.0001	0.0001
Mistral-v2-7b	Vignette	0.0001	0.0001
	Multi-turn conversation	0.0001	0.0001
	Single-turn conversation	0.0001	0.0001
	Summarized conversation	0.0001	0.0001
LLaMA-2-7b	Vignette	0.0001	0.0001
	Multi-turn conversation	0.0001	0.0001
	Single-turn conversation	0.0001	0.0001
	Summarized conversation	0.0001	0.0001

Supplementary Table 3: Adjusted p-values between 4-choice MCQ and FRQ settings for each experimental setup (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Mean Accuracy	95% C.I.
GPT-4	Vignette	0.486	(0.804, 0.837)
	Multi-turn conversation	0.264	(0.609, 0.645)
	Single-turn conversation	0.133	(0.5, 0.539)
	Summarized conversation	0.272	(0.652, 0.686)
GPT-3.5	Vignette	0.375	(0.637, 0.676)
	Multi-turn conversation	0.169	(0.448, 0.485)
	Single-turn conversation	0.123	(0.416, 0.454)
	Summarized conversation	0.174	(0.489, 0.526)
Mistral-v2-7b	Vignette	0.222	(0.616, 0.658)
	Multi-turn conversation	0.066	(0.409, 0.443)
	Single-turn conversation	0.056	(0.429, 0.468)
	Summarized conversation	0.056	(0.496, 0.529)
LLaMA-2-7b	Vignette	0.169	(0.376, 0.415)
	Multi-turn conversation	0.066	(0.303, 0.335)
	Single-turn conversation	0.065	(0.286, 0.323)
	Summarized conversation	0.081	(0.318, 0.352)

Supplementary Table 4: Mean accuracy and 95% confidence intervals for FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0001	0.0001
	Multi-turn conversation	Summarized conversation	0.0936	0.107
	Single-turn conversation	Summarized conversation	0.0001	0.0001
GPT-3.5	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0001	0.0001
	Multi-turn conversation	Summarized conversation	0.2743	0.2992
	Single-turn conversation	Summarized conversation	0.0001	0.0001
Mistral-v2-7b	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0015	0.0018
	Multi-turn conversation	Summarized conversation	0.0013	0.0016
	Single-turn conversation	Summarized conversation	0.8302	0.8302
LLaMA-2-7b	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.7258	0.7574
	Multi-turn conversation	Summarized conversation	0.0001	0.0001
	Single-turn conversation	Summarized conversation	0.0002	0.0003

Supplementary Table 5: Adjusted p-values for FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment Name	Specialty	Mean Accuracy	95% C.I.
GPT-4	Vignette	Cardiology	0.802	(0.744, 0.86)
		Dermatology	0.841	(0.802, 0.88)
		Endocrinology	0.798	(0.713, 0.882)
		Gastroenterology	0.758	(0.699, 0.816)
		Hematology and Oncology	0.835	(0.79, 0.881)
		Infectious Disease	0.842	(0.783, 0.902)
		Neurology	0.825	(0.777, 0.873)
		Obstetrics and Gynecology	0.795	(0.72, 0.869)
		Other	0.826	(0.775, 0.877)
		Pediatrics and Neonatology	0.876	(0.829, 0.923)
		Rheumatology	0.841	(0.766, 0.917)
		Urology and Nephrology	0.751	(0.662, 0.839)
	Multi-turn conversation	Cardiology	0.539	(0.473, 0.605)
		Dermatology	0.732	(0.691, 0.774)
		Endocrinology	0.527	(0.439, 0.616)
		Gastroenterology	0.563	(0.506, 0.619)
		Hematology and Oncology	0.609	(0.559, 0.658)
		Infectious Disease	0.616	(0.547, 0.685)
		Neurology	0.636	(0.58, 0.691)
		Obstetrics and Gynecology	0.68	(0.608, 0.752)
		Other	0.589	(0.534, 0.644)
		Pediatrics and Neonatology	0.724	(0.669, 0.778)
		Rheumatology	0.627	(0.543, 0.711)
		Urology and Nephrology	0.528	(0.435, 0.621)
	Single-turn conversation	Cardiology	0.448	(0.377, 0.518)
		Dermatology	0.647	(0.601, 0.694)
		Endocrinology	0.452	(0.357, 0.547)
		Gastroenterology	0.469	(0.406, 0.532)
		Hematology and Oncology	0.456	(0.398, 0.513)
		Infectious Disease	0.542	(0.465, 0.618)
		Neurology	0.529	(0.468, 0.59)
		Obstetrics and Gynecology	0.524	(0.44, 0.607)
		Other	0.465	(0.402, 0.528)

		Pediatrics and Neonatology	0.624	(0.561, 0.686)
		Rheumatology	0.507	(0.409, 0.605)
		Urology and Nephrology	0.42	(0.325, 0.515)
	Summarized conversation	Cardiology	0.599	(0.536, 0.663)
		Dermatology	0.756	(0.716, 0.795)
		Endocrinology	0.605	(0.521, 0.688)
		Gastroenterology	0.601	(0.542, 0.659)
		Hematology and Oncology	0.644	(0.595, 0.694)
		Infectious Disease	0.699	(0.636, 0.761)
		Neurology	0.702	(0.651, 0.753)
		Obstetrics and Gynecology	0.685	(0.611, 0.758)
		Other	0.602	(0.548, 0.657)
		Pediatrics and Neonatology	0.764	(0.713, 0.816)
		Rheumatology	0.663	(0.579, 0.748)
		Urology and Nephrology	0.602	(0.517, 0.688)
GPT-3.5	Vignette	Cardiology	0.601	(0.533, 0.67)
		Dermatology	0.707	(0.661, 0.754)
		Endocrinology	0.576	(0.476, 0.677)
		Gastroenterology	0.634	(0.575, 0.694)
		Hematology and Oncology	0.636	(0.579, 0.693)
		Infectious Disease	0.677	(0.603, 0.751)
		Neurology	0.624	(0.563, 0.686)
		Obstetrics and Gynecology	0.66	(0.575, 0.745)
		Other	0.712	(0.654, 0.771)
		Pediatrics and Neonatology	0.688	(0.622, 0.754)
		Rheumatology	0.646	(0.55, 0.743)
		Urology and Nephrology	0.625	(0.533, 0.716)
	Multi-turn conversation	Cardiology	0.378	(0.315, 0.441)
		Dermatology	0.549	(0.502, 0.595)
		Endocrinology	0.367	(0.28, 0.454)
Gastroenterology		0.428	(0.371, 0.485)	
Hematology and Oncology		0.375	(0.323, 0.427)	
Infectious Disease		0.469	(0.399, 0.538)	
Neurology		0.504	(0.446, 0.561)	
Obstetrics and Gynecology	0.482	(0.404, 0.56)		

		Other	0.482	(0.424, 0.539)
		Pediatrics and Neonatology	0.584	(0.522, 0.645)
		Rheumatology	0.478	(0.386, 0.57)
		Urology and Nephrology	0.373	(0.283, 0.463)
	Single-turn conversation	Cardiology	0.354	(0.288, 0.419)
		Dermatology	0.507	(0.459, 0.556)
		Endocrinology	0.409	(0.319, 0.499)
		Gastroenterology	0.377	(0.318, 0.435)
		Hematology and Oncology	0.401	(0.346, 0.456)
		Infectious Disease	0.46	(0.385, 0.535)
		Neurology	0.459	(0.399, 0.518)
		Obstetrics and Gynecology	0.413	(0.331, 0.494)
		Other	0.443	(0.382, 0.503)
		Pediatrics and Neonatology	0.511	(0.446, 0.576)
		Rheumatology	0.41	(0.315, 0.504)
		Urology and Nephrology	0.362	(0.275, 0.449)
	Summarized conversation	Cardiology	0.417	(0.352, 0.483)
		Dermatology	0.564	(0.518, 0.609)
		Endocrinology	0.428	(0.34, 0.517)
		Gastroenterology	0.427	(0.37, 0.484)
		Hematology and Oncology	0.437	(0.383, 0.49)
		Infectious Disease	0.559	(0.491, 0.628)
		Neurology	0.521	(0.463, 0.578)
		Obstetrics and Gynecology	0.528	(0.448, 0.607)
		Other	0.545	(0.488, 0.603)
		Pediatrics and Neonatology	0.612	(0.551, 0.673)
		Rheumatology	0.529	(0.44, 0.618)
		Urology and Nephrology	0.465	(0.38, 0.551)
Mistral-v2-7b	Vignette	Cardiology	0.646	(0.572, 0.72)
		Dermatology	0.694	(0.643, 0.745)
		Endocrinology	0.576	(0.469, 0.684)
		Gastroenterology	0.621	(0.553, 0.689)
		Hematology and Oncology	0.59	(0.526, 0.653)
		Infectious Disease	0.606	(0.523, 0.689)
		Neurology	0.584	(0.518, 0.651)

		Obstetrics and Gynecology	0.645	(0.555, 0.736)
		Other	0.687	(0.622, 0.751)
		Pediatrics and Neonatology	0.665	(0.594, 0.736)
		Rheumatology	0.671	(0.567, 0.775)
		Urology and Nephrology	0.607	(0.503, 0.71)
	Multi-turn conversation	Cardiology	0.373	(0.316, 0.43)
		Dermatology	0.511	(0.466, 0.556)
		Endocrinology	0.362	(0.281, 0.444)
		Gastroenterology	0.359	(0.306, 0.411)
		Hematology and Oncology	0.379	(0.33, 0.428)
		Infectious Disease	0.394	(0.327, 0.461)
		Neurology	0.424	(0.37, 0.478)
		Obstetrics and Gynecology	0.464	(0.387, 0.541)
		Other	0.448	(0.391, 0.504)
		Pediatrics and Neonatology	0.481	(0.421, 0.541)
		Rheumatology	0.434	(0.347, 0.521)
		Urology and Nephrology	0.402	(0.319, 0.485)
		Single-turn conversation	Cardiology	0.427
	Dermatology		0.531	(0.482, 0.58)
	Endocrinology		0.36	(0.266, 0.454)
	Gastroenterology		0.352	(0.292, 0.411)
	Hematology and Oncology		0.405	(0.35, 0.46)
	Infectious Disease		0.429	(0.353, 0.505)
	Neurology		0.431	(0.373, 0.489)
	Obstetrics and Gynecology		0.467	(0.385, 0.55)
	Other		0.474	(0.411, 0.537)
	Pediatrics and Neonatology		0.509	(0.444, 0.574)
	Rheumatology		0.493	(0.396, 0.59)
	Urology and Nephrology		0.438	(0.344, 0.532)
	Summarized conversation		Cardiology	0.493
		Dermatology	0.575	(0.533, 0.618)
		Endocrinology	0.48	(0.397, 0.563)
		Gastroenterology	0.507	(0.454, 0.56)
Hematology and Oncology		0.45	(0.402, 0.497)	
Infectious Disease		0.485	(0.421, 0.549)	

		Neurology	0.521	(0.471, 0.572)
		Obstetrics and Gynecology	0.571	(0.498, 0.644)
		Other	0.471	(0.418, 0.523)
		Pediatrics and Neonatology	0.558	(0.5, 0.617)
		Rheumatology	0.512	(0.43, 0.594)
		Urology and Nephrology	0.49	(0.409, 0.571)
LLaMA-2-7 b	Vignette	Cardiology	0.372	(0.304, 0.44)
		Dermatology	0.411	(0.362, 0.461)
		Endocrinology	0.353	(0.26, 0.446)
		Gastroenterology	0.392	(0.331, 0.453)
		Hematology and Oncology	0.321	(0.266, 0.376)
		Infectious Disease	0.388	(0.312, 0.465)
		Neurology	0.365	(0.307, 0.424)
		Obstetrics and Gynecology	0.342	(0.26, 0.424)
		Other	0.482	(0.417, 0.546)
		Pediatrics and Neonatology	0.486	(0.418, 0.553)
		Rheumatology	0.422	(0.321, 0.523)
		Urology and Nephrology	0.378	(0.286, 0.469)
		Multi-turn conversation	Cardiology	0.28
	Dermatology		0.346	(0.305, 0.388)
	Endocrinology		0.32	(0.242, 0.398)
	Gastroenterology		0.281	(0.231, 0.331)
	Hematology and Oncology		0.291	(0.245, 0.338)
	Infectious Disease		0.318	(0.257, 0.38)
	Neurology		0.296	(0.249, 0.343)
	Obstetrics and Gynecology		0.325	(0.256, 0.395)
	Other		0.364	(0.309, 0.42)
	Pediatrics and Neonatology		0.345	(0.286, 0.403)
	Rheumatology		0.4	(0.319, 0.481)
	Urology and Nephrology		0.272	(0.194, 0.35)
	Single-turn conversation	Cardiology	0.298	(0.234, 0.362)
		Dermatology	0.326	(0.279, 0.373)
		Endocrinology	0.256	(0.168, 0.345)
		Gastroenterology	0.256	(0.199, 0.312)
		Hematology and Oncology	0.279	(0.226, 0.333)

		Infectious Disease	0.318	(0.247, 0.39)
		Neurology	0.289	(0.233, 0.344)
		Obstetrics and Gynecology	0.331	(0.248, 0.414)
		Other	0.343	(0.283, 0.404)
		Pediatrics and Neonatology	0.336	(0.274, 0.399)
		Rheumatology	0.38	(0.281, 0.48)
		Urology and Nephrology	0.225	(0.147, 0.302)
	Summarized conversation	Cardiology	0.291	(0.231, 0.352)
		Dermatology	0.374	(0.33, 0.418)
		Endocrinology	0.294	(0.215, 0.373)
		Gastroenterology	0.309	(0.257, 0.361)
		Hematology and Oncology	0.289	(0.24, 0.337)
		Infectious Disease	0.353	(0.284, 0.423)
		Neurology	0.329	(0.277, 0.381)
		Obstetrics and Gynecology	0.329	(0.254, 0.405)
		Other	0.369	(0.314, 0.425)
		Pediatrics and Neonatology	0.388	(0.328, 0.449)
		Rheumatology	0.385	(0.298, 0.472)
		Urology and Nephrology	0.265	(0.189, 0.341)

Supplementary Table 6: Medical specialty wise mean accuracy and 95% confidence intervals for 4-choice MCQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Medical specialty	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Summarized conversation	0.0001	0.0003
	Cardiology	Multi-turn conversation	Single-turn conversation	0.0003	0.0007
	Cardiology	Multi-turn conversation	Summarized conversation	0.0001	0.0003
	Cardiology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Summarized conversation	0.0001	0.0003
	Dermatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Dermatology	Multi-turn conversation	Summarized conversation	0.0351	0.0513
	Dermatology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Endocrinology	Vignette	Multi-turn conversation	0.0001	0.0003
	Endocrinology	Vignette	Single-turn conversation	0.0001	0.0003
	Endocrinology	Vignette	Summarized conversation	0.0002	0.0005
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.0272	0.0417
	Endocrinology	Multi-turn conversation	Summarized conversation	0.0069	0.0117
	Endocrinology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0003
	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0003

Gastroenterology	Vignette	Summarized conversation	0.0001	0.0003
Gastroenterology	Multi-turn conversation	Single-turn conversation	0.0007	0.0015
Gastroenterology	Multi-turn conversation	Summarized conversation	0.0182	0.0286
Gastroenterology	Single-turn conversation	Summarized conversation	0.0001	0.0003
Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0003
Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0003
Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0003
Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.0094	0.0156
Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0001	0.0003
Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0003
Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0003
Infectious Disease	Vignette	Summarized conversation	0.0001	0.0003
Infectious Disease	Multi-turn conversation	Single-turn conversation	0.0287	0.0433
Infectious Disease	Multi-turn conversation	Summarized conversation	0.0002	0.0005
Infectious Disease	Single-turn conversation	Summarized conversation	0.0001	0.0003
Neurology	Vignette	Multi-turn conversation	0.0001	0.0003
Neurology	Vignette	Single-turn conversation	0.0001	0.0003
Neurology	Vignette	Summarized conversation	0.0001	0.0003
Neurology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
Neurology	Multi-turn conversation	Summarized conversation	0.0002	0.0005
Neurology	Single-turn	Summarized	0.0001	0.0003

		conversation	conversation		
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0011	0.0022
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0003
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0023	0.0043
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.0002	0.0005
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.8411	0.8714
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Other	Vignette	Multi-turn conversation	0.0001	0.0003
	Other	Vignette	Single-turn conversation	0.0001	0.0003
	Other	Vignette	Summarized conversation	0.0001	0.0003
	Other	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Other	Multi-turn conversation	Summarized conversation	0.3659	0.4157
	Other	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0003
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0057	0.0099
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0003
	Rheumatology	Vignette	Summarized conversation	0.0001	0.0003

	Rheumatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Rheumatology	Multi-turn conversation	Summarized conversation	0.1444	0.1857
	Rheumatology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Summarized conversation	0.0006	0.0013
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.0044	0.0077
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0174	0.0275
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.0001	0.0003
GPT-3.5	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Summarized conversation	0.0001	0.0003
	Cardiology	Multi-turn conversation	Single-turn conversation	0.2938	0.344
	Cardiology	Multi-turn conversation	Summarized conversation	0.0713	0.0978
	Cardiology	Single-turn conversation	Summarized conversation	0.0124	0.0202
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Summarized conversation	0.0001	0.0003
	Dermatology	Multi-turn conversation	Single-turn conversation	0.0158	0.0253
	Dermatology	Multi-turn conversation	Summarized conversation	0.3195	0.3695
	Dermatology	Single-turn conversation	Summarized conversation	0.0051	0.0089
	Endocrinology	Vignette	Multi-turn	0.0004	0.0009

			conversation		
Endocrinology	Vignette	Single-turn conversation	0.0085	0.0142	
Endocrinology	Vignette	Summarized conversation	0.0033	0.006	
Endocrinology	Multi-turn conversation	Single-turn conversation	0.2877	0.3382	
Endocrinology	Multi-turn conversation	Summarized conversation	0.0443	0.0632	
Endocrinology	Single-turn conversation	Summarized conversation	0.6448	0.7008	
Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0003	
Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0003	
Gastroenterology	Vignette	Summarized conversation	0.0001	0.0003	
Gastroenterology	Multi-turn conversation	Single-turn conversation	0.0183	0.0286	
Gastroenterology	Multi-turn conversation	Summarized conversation	0.9662	0.9696	
Gastroenterology	Single-turn conversation	Summarized conversation	0.0475	0.0669	
Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0003	
Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0003	
Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0003	
Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.2599	0.3093	
Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.0036	0.0065	
Hematology and Oncology	Single-turn conversation	Summarized conversation	0.119	0.1537	
Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0003	
Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0003	
Infectious Disease	Vignette	Summarized conversation	0.0008	0.0017	
Infectious Disease	Multi-turn conversation	Single-turn conversation	0.7508	0.7921	

	Infectious Disease	Multi-turn conversation	Summarized conversation	0.0002	0.0005
	Infectious Disease	Single-turn conversation	Summarized conversation	0.0005	0.0011
	Neurology	Vignette	Multi-turn conversation	0.0002	0.0005
	Neurology	Vignette	Single-turn conversation	0.0001	0.0003
	Neurology	Vignette	Summarized conversation	0.0001	0.0003
	Neurology	Multi-turn conversation	Single-turn conversation	0.0471	0.0668
	Neurology	Multi-turn conversation	Summarized conversation	0.3892	0.4396
	Neurology	Single-turn conversation	Summarized conversation	0.0118	0.0194
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0001	0.0003
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0003
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0003	0.0007
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.0284	0.043
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.1076	0.1409
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0022	0.0042
	Other	Vignette	Multi-turn conversation	0.0001	0.0003
	Other	Vignette	Single-turn conversation	0.0001	0.0003
	Other	Vignette	Summarized conversation	0.0001	0.0003
	Other	Multi-turn conversation	Single-turn conversation	0.0476	0.0669
	Other	Multi-turn conversation	Summarized conversation	0.0027	0.005
	Other	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0006	0.0013
	Pediatrics and Neonatology	Vignette	Single-turn	0.0001	0.0003

			conversation		
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0128	0.0207
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0022	0.0042
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.1542	0.1956
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0007	0.0015
	Rheumatology	Vignette	Multi-turn conversation	0.002	0.0038
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0003
	Rheumatology	Vignette	Summarized conversation	0.0122	0.02
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.0582	0.0814
	Rheumatology	Multi-turn conversation	Summarized conversation	0.1087	0.1417
	Rheumatology	Single-turn conversation	Summarized conversation	0.0004	0.0009
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Summarized conversation	0.0002	0.0005
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.6949	0.7486
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0004	0.0009
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.001	0.0021
Mistral-v2-7b	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Summarized conversation	0.0002	0.0005
	Cardiology	Multi-turn conversation	Single-turn conversation	0.013	0.0209
	Cardiology	Multi-turn conversation	Summarized conversation	0.0001	0.0003

Cardiology	Single-turn conversation	Summarized conversation	0.0268	0.0413
Dermatology	Vignette	Multi-turn conversation	0.0001	0.0003
Dermatology	Vignette	Single-turn conversation	0.0001	0.0003
Dermatology	Vignette	Summarized conversation	0.0001	0.0003
Dermatology	Multi-turn conversation	Single-turn conversation	0.2429	0.2903
Dermatology	Multi-turn conversation	Summarized conversation	0.0002	0.0005
Dermatology	Single-turn conversation	Summarized conversation	0.0231	0.0358
Endocrinology	Vignette	Multi-turn conversation	0.0011	0.0022
Endocrinology	Vignette	Single-turn conversation	0.0028	0.0051
Endocrinology	Vignette	Summarized conversation	0.0904	0.12
Endocrinology	Multi-turn conversation	Single-turn conversation	0.9471	0.9571
Endocrinology	Multi-turn conversation	Summarized conversation	0.0058	0.01
Endocrinology	Single-turn conversation	Summarized conversation	0.0062	0.0106
Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0003
Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0003
Gastroenterology	Vignette	Summarized conversation	0.0018	0.0035
Gastroenterology	Multi-turn conversation	Single-turn conversation	0.735	0.7782
Gastroenterology	Multi-turn conversation	Summarized conversation	0.0001	0.0003
Gastroenterology	Single-turn conversation	Summarized conversation	0.0001	0.0003
Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0003
Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0003
Hematology and Oncology	Vignette	Summarized	0.0002	0.0005

			conversation		
Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.1834	0.2287	
Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.0014	0.0028	
Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0674	0.0929	
Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0003	
Infectious Disease	Vignette	Single-turn conversation	0.0004	0.0009	
Infectious Disease	Vignette	Summarized conversation	0.0087	0.0145	
Infectious Disease	Multi-turn conversation	Single-turn conversation	0.1983	0.2441	
Infectious Disease	Multi-turn conversation	Summarized conversation	0.0013	0.0026	
Infectious Disease	Single-turn conversation	Summarized conversation	0.0858	0.1144	
Neurology	Vignette	Multi-turn conversation	0.0001	0.0003	
Neurology	Vignette	Single-turn conversation	0.0001	0.0003	
Neurology	Vignette	Summarized conversation	0.0372	0.0541	
Neurology	Multi-turn conversation	Single-turn conversation	0.6992	0.7486	
Neurology	Multi-turn conversation	Summarized conversation	0.0001	0.0003	
Neurology	Single-turn conversation	Summarized conversation	0.0002	0.0005	
Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0002	0.0005	
Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0006	0.0013	
Obstetrics and Gynecology	Vignette	Summarized conversation	0.0797	0.1078	
Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.8467	0.8714	
Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.0007	0.0015	
Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0043	0.0076	

Other	Vignette	Multi-turn conversation	0.0001	0.0003
Other	Vignette	Single-turn conversation	0.0001	0.0003
Other	Vignette	Summarized conversation	0.0001	0.0003
Other	Multi-turn conversation	Single-turn conversation	0.1483	0.1898
Other	Multi-turn conversation	Summarized conversation	0.2698	0.3185
Other	Single-turn conversation	Summarized conversation	0.9117	0.9245
Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0003
Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0004	0.0009
Pediatrics and Neonatology	Vignette	Summarized conversation	0.0026	0.0048
Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.1787	0.2238
Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0019	0.0037
Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0652	0.0907
Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0003
Rheumatology	Vignette	Single-turn conversation	0.0025	0.0047
Rheumatology	Vignette	Summarized conversation	0.0005	0.0011
Rheumatology	Multi-turn conversation	Single-turn conversation	0.106	0.1394
Rheumatology	Multi-turn conversation	Summarized conversation	0.0213	0.0332
Rheumatology	Single-turn conversation	Summarized conversation	0.6076	0.673
Urology and Nephrology	Vignette	Multi-turn conversation	0.0002	0.0005
Urology and Nephrology	Vignette	Single-turn conversation	0.0035	0.0063
Urology and Nephrology	Vignette	Summarized conversation	0.033	0.0487
Urology and Nephrology	Multi-turn	Single-turn	0.3229	0.372

		conversation	conversation		
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0166	0.0264
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.2055	0.2508
LLaMA-2-7b	Cardiology	Vignette	Multi-turn conversation	0.0018	0.0035
	Cardiology	Vignette	Single-turn conversation	0.0307	0.0461
	Cardiology	Vignette	Summarized conversation	0.0037	0.0066
	Cardiology	Multi-turn conversation	Single-turn conversation	0.4995	0.5598
	Cardiology	Multi-turn conversation	Summarized conversation	0.6265	0.6913
	Cardiology	Single-turn conversation	Summarized conversation	0.8348	0.8711
	Dermatology	Vignette	Multi-turn conversation	0.0042	0.0075
	Dermatology	Vignette	Single-turn conversation	0.0007	0.0015
	Dermatology	Vignette	Summarized conversation	0.0406	0.0582
	Dermatology	Multi-turn conversation	Single-turn conversation	0.1844	0.2289
	Dermatology	Multi-turn conversation	Summarized conversation	0.1012	0.1337
	Dermatology	Single-turn conversation	Summarized conversation	0.0075	0.0126
	Endocrinology	Vignette	Multi-turn conversation	0.4788	0.5387
	Endocrinology	Vignette	Single-turn conversation	0.0387	0.0557
	Endocrinology	Vignette	Summarized conversation	0.188	0.2324
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.0745	0.1012
	Endocrinology	Multi-turn conversation	Summarized conversation	0.3524	0.4027
	Endocrinology	Single-turn conversation	Summarized conversation	0.3062	0.357
	Gastroenterology	Vignette	Multi-turn conversation	0.0005	0.0011

Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0003
Gastroenterology	Vignette	Summarized conversation	0.0008	0.0017
Gastroenterology	Multi-turn conversation	Single-turn conversation	0.2663	0.3156
Gastroenterology	Multi-turn conversation	Summarized conversation	0.1747	0.2197
Gastroenterology	Single-turn conversation	Summarized conversation	0.0311	0.0464
Hematology and Oncology	Vignette	Multi-turn conversation	0.2238	0.2708
Hematology and Oncology	Vignette	Single-turn conversation	0.0811	0.1091
Hematology and Oncology	Vignette	Summarized conversation	0.1145	0.1485
Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.5763	0.6433
Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.899	0.9149
Hematology and Oncology	Single-turn conversation	Summarized conversation	0.6075	0.673
Infectious Disease	Vignette	Multi-turn conversation	0.0319	0.0474
Infectious Disease	Vignette	Single-turn conversation	0.0664	0.0919
Infectious Disease	Vignette	Summarized conversation	0.222	0.2698
Infectious Disease	Multi-turn conversation	Single-turn conversation	0.9875	0.9875
Infectious Disease	Multi-turn conversation	Summarized conversation	0.1598	0.2019
Infectious Disease	Single-turn conversation	Summarized conversation	0.2261	0.2725
Neurology	Vignette	Multi-turn conversation	0.002	0.0038
Neurology	Vignette	Single-turn conversation	0.0006	0.0013
Neurology	Vignette	Summarized conversation	0.0828	0.1109
Neurology	Multi-turn conversation	Single-turn conversation	0.7036	0.7505
Neurology	Multi-turn	Summarized	0.0723	0.0987

		conversation	conversation		
	Neurology	Single-turn conversation	Summarized conversation	0.038	0.055
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.6343	0.6954
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.7654	0.8045
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.6576	0.712
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.8472	0.8714
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.8743	0.8961
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.9586	0.9653
	Other	Vignette	Multi-turn conversation	0.0001	0.0003
	Other	Vignette	Single-turn conversation	0.0001	0.0003
	Other	Vignette	Summarized conversation	0.0001	0.0003
	Other	Multi-turn conversation	Single-turn conversation	0.3146	0.3653
	Other	Multi-turn conversation	Summarized conversation	0.8103	0.8486
	Other	Single-turn conversation	Summarized conversation	0.1992	0.2441
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0003
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.7241	0.7695
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0279	0.0425
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0334	0.0491
	Rheumatology	Vignette	Multi-turn conversation	0.6379	0.6959
	Rheumatology	Vignette	Single-turn conversation	0.3666	0.4157

	Rheumatology	Vignette	Summarized conversation	0.3378	0.3876
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.635	0.6954
	Rheumatology	Multi-turn conversation	Summarized conversation	0.6992	0.7486
	Rheumatology	Single-turn conversation	Summarized conversation	0.8946	0.9136
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0061	0.0105
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Summarized conversation	0.0015	0.003
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.1506	0.1919
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.846	0.8714
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.232	0.2784

Supplementary Table 7: Medical specialty wise adjusted p-values for 4-choice MCQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment Name	Specialty	Mean Accuracy	95% C.I.
GPT-4	Vignette	Cardiology	0.345	(0.281, 0.409)
		Dermatology	0.645	(0.597, 0.693)
		Endocrinology	0.379	(0.286, 0.472)
		Gastroenterology	0.365	(0.304, 0.425)
		Hematology and Oncology	0.467	(0.409, 0.525)
		Infectious Disease	0.416	(0.339, 0.493)
		Neurology	0.513	(0.453, 0.574)
		Obstetrics and Gynecology	0.453	(0.366, 0.54)
		Other	0.484	(0.422, 0.545)
		Pediatrics and Neonatology	0.557	(0.49, 0.624)
		Rheumatology	0.595	(0.492, 0.698)
		Urology and Nephrology	0.449	(0.354, 0.545)
	Multi-turn conversation	Cardiology	0.134	(0.093, 0.176)
		Dermatology	0.399	(0.354, 0.443)
		Endocrinology	0.108	(0.063, 0.154)
		Gastroenterology	0.152	(0.11, 0.193)
		Hematology and Oncology	0.203	(0.163, 0.243)
		Infectious Disease	0.206	(0.15, 0.261)
		Neurology	0.321	(0.267, 0.376)
		Obstetrics and Gynecology	0.264	(0.194, 0.333)
		Other	0.295	(0.242, 0.347)
		Pediatrics and Neonatology	0.369	(0.31, 0.428)
		Rheumatology	0.329	(0.244, 0.414)
		Urology and Nephrology	0.196	(0.126, 0.265)
	Single-turn conversation	Cardiology	0.07	(0.039, 0.1)
		Dermatology	0.224	(0.184, 0.264)
		Endocrinology	0.024	(-0.003, 0.05)
		Gastroenterology	0.085	(0.051, 0.118)
		Hematology and Oncology	0.078	(0.05, 0.105)
		Infectious Disease	0.117	(0.071, 0.163)
		Neurology	0.185	(0.138, 0.232)
		Obstetrics and Gynecology	0.071	(0.034, 0.107)
		Other	0.142	(0.101, 0.184)

		Pediatrics and Neonatology	0.195	(0.145, 0.246)
		Rheumatology	0.132	(0.068, 0.195)
		Urology and Nephrology	0.115	(0.056, 0.173)
	Summarized conversation	Cardiology	0.122	(0.082, 0.161)
		Dermatology	0.436	(0.391, 0.482)
		Endocrinology	0.129	(0.073, 0.186)
		Gastroenterology	0.165	(0.121, 0.208)
		Hematology and Oncology	0.219	(0.177, 0.261)
		Infectious Disease	0.257	(0.196, 0.318)
		Neurology	0.287	(0.239, 0.336)
		Obstetrics and Gynecology	0.271	(0.2, 0.342)
		Other	0.284	(0.236, 0.333)
		Pediatrics and Neonatology	0.328	(0.272, 0.383)
		Rheumatology	0.371	(0.282, 0.459)
		Urology and Nephrology	0.236	(0.162, 0.31)
GPT-3.5	Vignette	Cardiology	0.254	(0.196, 0.311)
		Dermatology	0.504	(0.454, 0.554)
		Endocrinology	0.212	(0.135, 0.288)
		Gastroenterology	0.311	(0.252, 0.371)
		Hematology and Oncology	0.372	(0.317, 0.426)
		Infectious Disease	0.364	(0.29, 0.437)
		Neurology	0.341	(0.284, 0.398)
		Obstetrics and Gynecology	0.338	(0.256, 0.421)
		Other	0.372	(0.312, 0.432)
		Pediatrics and Neonatology	0.445	(0.378, 0.512)
		Rheumatology	0.463	(0.357, 0.57)
		Urology and Nephrology	0.375	(0.284, 0.467)
	Multi-turn conversation	Cardiology	0.085	(0.051, 0.119)
		Dermatology	0.241	(0.202, 0.28)
		Endocrinology	0.054	(0.016, 0.093)
Gastroenterology		0.092	(0.061, 0.123)	
Hematology and Oncology		0.13	(0.096, 0.164)	
Infectious Disease		0.137	(0.09, 0.185)	
Neurology		0.206	(0.16, 0.251)	
Obstetrics and Gynecology	0.198	(0.14, 0.256)		

		Other	0.214	(0.168, 0.26)
		Pediatrics and Neonatology	0.24	(0.186, 0.295)
		Rheumatology	0.18	(0.111, 0.25)
		Urology and Nephrology	0.121	(0.07, 0.172)
	Single-turn conversation	Cardiology	0.057	(0.029, 0.085)
		Dermatology	0.169	(0.133, 0.205)
		Endocrinology	0.038	(0.005, 0.071)
		Gastroenterology	0.083	(0.05, 0.116)
		Hematology and Oncology	0.092	(0.061, 0.124)
		Infectious Disease	0.092	(0.05, 0.134)
		Neurology	0.153	(0.111, 0.195)
		Obstetrics and Gynecology	0.131	(0.077, 0.184)
		Other	0.166	(0.123, 0.21)
		Pediatrics and Neonatology	0.182	(0.131, 0.232)
		Rheumatology	0.102	(0.044, 0.161)
		Urology and Nephrology	0.099	(0.046, 0.151)
	Summarized conversation	Cardiology	0.079	(0.047, 0.111)
		Dermatology	0.275	(0.234, 0.315)
		Endocrinology	0.045	(0.013, 0.077)
		Gastroenterology	0.133	(0.096, 0.171)
		Hematology and Oncology	0.152	(0.116, 0.189)
		Infectious Disease	0.155	(0.107, 0.203)
		Neurology	0.179	(0.135, 0.222)
		Obstetrics and Gynecology	0.205	(0.144, 0.267)
		Other	0.167	(0.128, 0.206)
		Pediatrics and Neonatology	0.207	(0.158, 0.256)
		Rheumatology	0.238	(0.16, 0.315)
		Urology and Nephrology	0.135	(0.077, 0.193)
Mistral-v2-7b	Vignette	Cardiology	0.155	(0.103, 0.207)
		Dermatology	0.217	(0.173, 0.261)
		Endocrinology	0.167	(0.091, 0.243)
		Gastroenterology	0.197	(0.143, 0.251)
		Hematology and Oncology	0.227	(0.176, 0.278)
		Infectious Disease	0.19	(0.128, 0.252)
		Neurology	0.247	(0.192, 0.302)

		Obstetrics and Gynecology	0.198	(0.128, 0.269)
		Other	0.286	(0.226, 0.345)
		Pediatrics and Neonatology	0.239	(0.178, 0.301)
		Rheumatology	0.276	(0.179, 0.372)
		Urology and Nephrology	0.261	(0.171, 0.35)
	Multi-turn conversation	Cardiology	0.033	(0.015, 0.051)
		Dermatology	0.099	(0.073, 0.125)
		Endocrinology	0.028	(-0.002, 0.059)
		Gastroenterology	0.032	(0.014, 0.05)
		Hematology and Oncology	0.033	(0.015, 0.051)
		Infectious Disease	0.045	(0.02, 0.07)
		Neurology	0.083	(0.053, 0.113)
		Obstetrics and Gynecology	0.042	(0.018, 0.066)
		Other	0.11	(0.076, 0.145)
		Pediatrics and Neonatology	0.108	(0.068, 0.147)
		Rheumatology	0.073	(0.027, 0.12)
		Urology and Nephrology	0.036	(0.013, 0.059)
	Single-turn conversation	Cardiology	0.027	(0.009, 0.045)
		Dermatology	0.086	(0.06, 0.113)
		Endocrinology	0.028	(-0.002, 0.058)
		Gastroenterology	0.028	(0.01, 0.046)
		Hematology and Oncology	0.037	(0.017, 0.057)
		Infectious Disease	0.029	(0.004, 0.055)
		Neurology	0.061	(0.034, 0.087)
		Obstetrics and Gynecology	0.033	(0.01, 0.055)
		Other	0.096	(0.062, 0.129)
		Pediatrics and Neonatology	0.09	(0.052, 0.129)
Rheumatology		0.061	(0.015, 0.107)	
Urology and Nephrology		0.047	(0.015, 0.08)	
Summarized conversation	Cardiology	0.024	(0.01, 0.039)	
	Dermatology	0.09	(0.066, 0.114)	
	Endocrinology	0.016	(-0.005, 0.038)	
	Gastroenterology	0.021	(0.006, 0.036)	
	Hematology and Oncology	0.032	(0.017, 0.046)	
	Infectious Disease	0.039	(0.018, 0.061)	

		Neurology	0.082	(0.054, 0.111)
		Obstetrics and Gynecology	0.06	(0.028, 0.092)
		Other	0.071	(0.046, 0.096)
		Pediatrics and Neonatology	0.069	(0.04, 0.098)
		Rheumatology	0.085	(0.041, 0.13)
		Urology and Nephrology	0.036	(0.004, 0.067)
LLaMA-2-7 b	Vignette	Cardiology	0.101	(0.063, 0.139)
		Dermatology	0.175	(0.136, 0.213)
		Endocrinology	0.082	(0.034, 0.131)
		Gastroenterology	0.143	(0.102, 0.185)
		Hematology and Oncology	0.203	(0.156, 0.251)
		Infectious Disease	0.166	(0.11, 0.223)
		Neurology	0.191	(0.144, 0.237)
		Obstetrics and Gynecology	0.151	(0.091, 0.211)
		Other	0.192	(0.146, 0.239)
		Pediatrics and Neonatology	0.182	(0.13, 0.233)
		Rheumatology	0.249	(0.163, 0.334)
		Urology and Nephrology	0.157	(0.091, 0.224)
	Multi-turn conversation	Cardiology	0.039	(0.022, 0.056)
		Dermatology	0.084	(0.061, 0.107)
		Endocrinology	0.028	(0.007, 0.049)
		Gastroenterology	0.026	(0.01, 0.042)
		Hematology and Oncology	0.022	(0.009, 0.035)
		Infectious Disease	0.031	(0.01, 0.051)
		Neurology	0.076	(0.048, 0.104)
		Obstetrics and Gynecology	0.122	(0.077, 0.166)
		Other	0.132	(0.097, 0.168)
		Pediatrics and Neonatology	0.094	(0.059, 0.128)
		Rheumatology	0.093	(0.045, 0.141)
		Urology and Nephrology	0.029	(0.008, 0.051)
	Single-turn conversation	Cardiology	0.049	(0.021, 0.077)
		Dermatology	0.074	(0.049, 0.1)
		Endocrinology	0.007	(-0.007, 0.021)
		Gastroenterology	0.027	(0.007, 0.048)
		Hematology and Oncology	0.027	(0.011, 0.044)

		Infectious Disease	0.039	(0.011, 0.068)
		Neurology	0.059	(0.031, 0.087)
		Obstetrics and Gynecology	0.105	(0.053, 0.158)
		Other	0.143	(0.1, 0.186)
		Pediatrics and Neonatology	0.079	(0.046, 0.111)
		Rheumatology	0.112	(0.05, 0.174)
		Urology and Nephrology	0.056	(0.015, 0.098)
	Summarized conversation	Cardiology	0.048	(0.023, 0.072)
		Dermatology	0.102	(0.075, 0.128)
		Endocrinology	0.033	(0.006, 0.059)
		Gastroenterology	0.032	(0.011, 0.053)
		Hematology and Oncology	0.051	(0.03, 0.072)
		Infectious Disease	0.079	(0.043, 0.115)
		Neurology	0.085	(0.056, 0.114)
		Obstetrics and Gynecology	0.116	(0.07, 0.163)
		Other	0.124	(0.09, 0.158)
		Pediatrics and Neonatology	0.094	(0.064, 0.124)
		Rheumatology	0.149	(0.08, 0.217)
		Urology and Nephrology	0.067	(0.027, 0.108)

Supplementary Table 8: Medical specialty wise mean accuracy and 95% confidence intervals for FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Medical specialty	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Summarized conversation	0.0001	0.0002
	Cardiology	Multi-turn conversation	Single-turn conversation	0.0003	0.0006
	Cardiology	Multi-turn conversation	Summarized conversation	0.4286	0.486
	Cardiology	Single-turn conversation	Summarized conversation	0.0057	0.0089
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Summarized conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Summarized conversation	0.0013	0.0023
	Dermatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Endocrinology	Vignette	Multi-turn conversation	0.0001	0.0002
	Endocrinology	Vignette	Single-turn conversation	0.0001	0.0002
	Endocrinology	Vignette	Summarized conversation	0.0001	0.0002
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Endocrinology	Multi-turn conversation	Summarized conversation	0.1678	0.2129
	Endocrinology	Single-turn conversation	Summarized conversation	0.0003	0.0006
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0002
	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0002

Gastroenterology	Vignette	Summarized conversation	0.0001	0.0002
Gastroenterology	Multi-turn conversation	Single-turn conversation	0.0002	0.0004
Gastroenterology	Multi-turn conversation	Summarized conversation	0.2019	0.2517
Gastroenterology	Single-turn conversation	Summarized conversation	0.0001	0.0002
Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0002
Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0002
Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0002
Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.2584	0.314
Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0001	0.0002
Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0002
Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0002
Infectious Disease	Vignette	Summarized conversation	0.0001	0.0002
Infectious Disease	Multi-turn conversation	Single-turn conversation	0.0014	0.0025
Infectious Disease	Multi-turn conversation	Summarized conversation	0.0042	0.0067
Infectious Disease	Single-turn conversation	Summarized conversation	0.0001	0.0002
Neurology	Vignette	Multi-turn conversation	0.0001	0.0002
Neurology	Vignette	Single-turn conversation	0.0001	0.0002
Neurology	Vignette	Summarized conversation	0.0001	0.0002
Neurology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
Neurology	Multi-turn conversation	Summarized conversation	0.0553	0.0788
Neurology	Single-turn	Summarized	0.0001	0.0002

		conversation	conversation		
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0001	0.0002
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.7675	0.7923
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Other	Vignette	Multi-turn conversation	0.0001	0.0002
	Other	Vignette	Single-turn conversation	0.0001	0.0002
	Other	Vignette	Summarized conversation	0.0001	0.0002
	Other	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Other	Multi-turn conversation	Summarized conversation	0.5739	0.6167
	Other	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0002
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0402	0.0585
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Summarized conversation	0.0001	0.0002

	Rheumatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Rheumatology	Multi-turn conversation	Summarized conversation	0.0472	0.068
	Rheumatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.0007	0.0013
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0226	0.0334
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.0001	0.0002
GPT-3.5	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Summarized conversation	0.0001	0.0002
	Cardiology	Multi-turn conversation	Single-turn conversation	0.0211	0.0313
	Cardiology	Multi-turn conversation	Summarized conversation	0.5866	0.628
	Cardiology	Single-turn conversation	Summarized conversation	0.0838	0.1155
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Summarized conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Summarized conversation	0.0038	0.0061
	Dermatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Endocrinology	Vignette	Multi-turn	0.0003	0.0006

			conversation		
Endocrinology	Vignette	Single-turn conversation	0.0001	0.0002	
Endocrinology	Vignette	Summarized conversation	0.0001	0.0002	
Endocrinology	Multi-turn conversation	Single-turn conversation	0.2519	0.3087	
Endocrinology	Multi-turn conversation	Summarized conversation	0.5935	0.6331	
Endocrinology	Single-turn conversation	Summarized conversation	0.689	0.7216	
Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0002	
Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0002	
Gastroenterology	Vignette	Summarized conversation	0.0001	0.0002	
Gastroenterology	Multi-turn conversation	Single-turn conversation	0.5041	0.5532	
Gastroenterology	Multi-turn conversation	Summarized conversation	0.0024	0.0041	
Gastroenterology	Single-turn conversation	Summarized conversation	0.0034	0.0056	
Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0002	
Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0002	
Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0002	
Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.003	0.005	
Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.056	0.0794	
Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0003	0.0006	
Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0002	
Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0002	
Infectious Disease	Vignette	Summarized conversation	0.0001	0.0002	
Infectious Disease	Multi-turn conversation	Single-turn conversation	0.0245	0.036	

	Infectious Disease	Multi-turn conversation	Summarized conversation	0.3615	0.4249
	Infectious Disease	Single-turn conversation	Summarized conversation	0.0043	0.0068
	Neurology	Vignette	Multi-turn conversation	0.0001	0.0002
	Neurology	Vignette	Single-turn conversation	0.0001	0.0002
	Neurology	Vignette	Summarized conversation	0.0001	0.0002
	Neurology	Multi-turn conversation	Single-turn conversation	0.002	0.0035
	Neurology	Multi-turn conversation	Summarized conversation	0.0857	0.1175
	Neurology	Single-turn conversation	Summarized conversation	0.1727	0.2181
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0004	0.0008
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0001	0.0002
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.0002	0.0004
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.7258	0.7546
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0035	0.0057
	Other	Vignette	Multi-turn conversation	0.0001	0.0002
	Other	Vignette	Single-turn conversation	0.0001	0.0002
	Other	Vignette	Summarized conversation	0.0001	0.0002
	Other	Multi-turn conversation	Single-turn conversation	0.0023	0.0039
	Other	Multi-turn conversation	Summarized conversation	0.0063	0.0097
	Other	Single-turn conversation	Summarized conversation	0.9552	0.9653
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Single-turn	0.0001	0.0002

			conversation		
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0002
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0013	0.0023
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.1245	0.1668
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.2319	0.2866
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Summarized conversation	0.0001	0.0002
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.0063	0.0097
	Rheumatology	Multi-turn conversation	Summarized conversation	0.0031	0.0051
	Rheumatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.1299	0.1708
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.5503	0.5958
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.0462	0.0669
Mistral-v2-7b	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Summarized conversation	0.0001	0.0002
	Cardiology	Multi-turn conversation	Single-turn conversation	0.4526	0.5013
	Cardiology	Multi-turn conversation	Summarized conversation	0.2343	0.2884

Cardiology	Single-turn conversation	Summarized conversation	0.7807	0.803
Dermatology	Vignette	Multi-turn conversation	0.0001	0.0002
Dermatology	Vignette	Single-turn conversation	0.0001	0.0002
Dermatology	Vignette	Summarized conversation	0.0001	0.0002
Dermatology	Multi-turn conversation	Single-turn conversation	0.1451	0.1891
Dermatology	Multi-turn conversation	Summarized conversation	0.2174	0.2699
Dermatology	Single-turn conversation	Summarized conversation	0.7464	0.7732
Endocrinology	Vignette	Multi-turn conversation	0.0006	0.0011
Endocrinology	Vignette	Single-turn conversation	0.0006	0.0011
Endocrinology	Vignette	Summarized conversation	0.0003	0.0006
Endocrinology	Multi-turn conversation	Single-turn conversation	1.0	1.0
Endocrinology	Multi-turn conversation	Summarized conversation	0.1031	0.1401
Endocrinology	Single-turn conversation	Summarized conversation	0.3447	0.4085
Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0002
Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0002
Gastroenterology	Vignette	Summarized conversation	0.0001	0.0002
Gastroenterology	Multi-turn conversation	Single-turn conversation	0.4457	0.4975
Gastroenterology	Multi-turn conversation	Summarized conversation	0.1289	0.1708
Gastroenterology	Single-turn conversation	Summarized conversation	0.4384	0.4932
Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0002
Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0002
Hematology and Oncology	Vignette	Summarized	0.0001	0.0002

			conversation		
Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.6697	0.7039	
Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.8405	0.8614	
Hematology and Oncology	Single-turn conversation	Summarized conversation	0.613	0.6491	
Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0002	
Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0002	
Infectious Disease	Vignette	Summarized conversation	0.0001	0.0002	
Infectious Disease	Multi-turn conversation	Single-turn conversation	0.1297	0.1708	
Infectious Disease	Multi-turn conversation	Summarized conversation	0.4056	0.4669	
Infectious Disease	Single-turn conversation	Summarized conversation	0.3286	0.3911	
Neurology	Vignette	Multi-turn conversation	0.0001	0.0002	
Neurology	Vignette	Single-turn conversation	0.0001	0.0002	
Neurology	Vignette	Summarized conversation	0.0001	0.0002	
Neurology	Multi-turn conversation	Single-turn conversation	0.0041	0.0066	
Neurology	Multi-turn conversation	Summarized conversation	0.9391	0.9523	
Neurology	Single-turn conversation	Summarized conversation	0.0745	0.1042	
Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0001	0.0002	
Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0002	
Obstetrics and Gynecology	Vignette	Summarized conversation	0.0002	0.0004	
Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.4919	0.5428	
Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.1661	0.2118	
Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.1326	0.1736	

Other	Vignette	Multi-turn conversation	0.0001	0.0002
Other	Vignette	Single-turn conversation	0.0001	0.0002
Other	Vignette	Summarized conversation	0.0001	0.0002
Other	Multi-turn conversation	Single-turn conversation	0.1878	0.2352
Other	Multi-turn conversation	Summarized conversation	0.0089	0.0136
Other	Single-turn conversation	Summarized conversation	0.1592	0.2062
Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0002
Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0002
Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0002
Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0802	0.1116
Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0057	0.0089
Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.1087	0.147
Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0002
Rheumatology	Vignette	Single-turn conversation	0.0001	0.0002
Rheumatology	Vignette	Summarized conversation	0.0001	0.0002
Rheumatology	Multi-turn conversation	Single-turn conversation	0.5254	0.5732
Rheumatology	Multi-turn conversation	Summarized conversation	0.4354	0.4917
Rheumatology	Single-turn conversation	Summarized conversation	0.1656	0.2118
Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0002
Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0002
Urology and Nephrology	Vignette	Summarized conversation	0.0001	0.0002
Urology and Nephrology	Multi-turn	Single-turn	0.4232	0.4829

		conversation	conversation		
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.9699	0.9767
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.6103	0.6486
LLaMA-2-7b	Cardiology	Vignette	Multi-turn conversation	0.0009	0.0017
	Cardiology	Vignette	Single-turn conversation	0.0057	0.0089
	Cardiology	Vignette	Summarized conversation	0.0065	0.01
	Cardiology	Multi-turn conversation	Single-turn conversation	0.3935	0.4551
	Cardiology	Multi-turn conversation	Summarized conversation	0.4069	0.4669
	Cardiology	Single-turn conversation	Summarized conversation	0.8954	0.9145
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Summarized conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Single-turn conversation	0.3827	0.4462
	Dermatology	Multi-turn conversation	Summarized conversation	0.0672	0.0944
	Dermatology	Single-turn conversation	Summarized conversation	0.0193	0.0288
	Endocrinology	Vignette	Multi-turn conversation	0.0125	0.0188
	Endocrinology	Vignette	Single-turn conversation	0.0025	0.0042
	Endocrinology	Vignette	Summarized conversation	0.0052	0.0082
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.082	0.1135
	Endocrinology	Multi-turn conversation	Summarized conversation	0.7129	0.7439
	Endocrinology	Single-turn conversation	Summarized conversation	0.0988	0.1349
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0002

Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0002
Gastroenterology	Vignette	Summarized conversation	0.0001	0.0002
Gastroenterology	Multi-turn conversation	Single-turn conversation	0.9041	0.9201
Gastroenterology	Multi-turn conversation	Summarized conversation	0.3672	0.4299
Gastroenterology	Single-turn conversation	Summarized conversation	0.4409	0.4941
Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0002
Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0002
Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0002
Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.4521	0.5013
Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.0024	0.0041
Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0012	0.0022
Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0002
Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0002
Infectious Disease	Vignette	Summarized conversation	0.0003	0.0006
Infectious Disease	Multi-turn conversation	Single-turn conversation	0.4242	0.4829
Infectious Disease	Multi-turn conversation	Summarized conversation	0.0004	0.0008
Infectious Disease	Single-turn conversation	Summarized conversation	0.0026	0.0044
Neurology	Vignette	Multi-turn conversation	0.0001	0.0002
Neurology	Vignette	Single-turn conversation	0.0001	0.0002
Neurology	Vignette	Summarized conversation	0.0001	0.0002
Neurology	Multi-turn conversation	Single-turn conversation	0.1278	0.1704
Neurology	Multi-turn	Summarized	0.3604	0.4249

		conversation	conversation		
	Neurology	Single-turn conversation	Summarized conversation	0.0495	0.0709
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.3211	0.3837
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.1807	0.2273
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.1662	0.2118
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.2929	0.353
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.6524	0.6882
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.5052	0.5532
	Other	Vignette	Multi-turn conversation	0.0037	0.006
	Other	Vignette	Single-turn conversation	0.0325	0.0475
	Other	Vignette	Summarized conversation	0.0012	0.0022
	Other	Multi-turn conversation	Single-turn conversation	0.5344	0.5808
	Other	Multi-turn conversation	Summarized conversation	0.5592	0.6032
	Other	Single-turn conversation	Summarized conversation	0.2583	0.314
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0007	0.0013
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0002	0.0004
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.1597	0.2062
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.985	0.9884
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.2708	0.3277
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Single-turn conversation	0.0013	0.0023

	Rheumatology	Vignette	Summarized conversation	0.0093	0.0141
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.3063	0.3676
	Rheumatology	Multi-turn conversation	Summarized conversation	0.0022	0.0038
	Rheumatology	Single-turn conversation	Summarized conversation	0.1176	0.1583
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Single-turn conversation	0.0022	0.0038
	Urology and Nephrology	Vignette	Summarized conversation	0.002	0.0035
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.0573	0.0809
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0109	0.0164
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.3912	0.4543

Supplementary Table 9: Medical specialty wise adjusted p-values for FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Dataset	Mean Accuracy	95% C.I.
GPT-4	Vignette	MedQA-USMLE	0.791	(0.721, 0.862)
	Vignette	Derm-Private	0.852	(0.782, 0.922)
	Vignette	Derm-Public	0.888	(0.826, 0.95)
	Multi-turn conversation	MedQA-USMLE	0.631	(0.557, 0.705)
	Multi-turn conversation	Derm-Private	0.796	(0.727, 0.865)
	Multi-turn conversation	Derm-Public	0.788	(0.719, 0.857)
	Single-turn conversation	MedQA-USMLE	0.566	(0.487, 0.645)
	Single-turn conversation	Derm-Private	0.678	(0.596, 0.76)
	Single-turn conversation	Derm-Public	0.712	(0.631, 0.793)
	Summarized conversation	MedQA-USMLE	0.679	(0.609, 0.749)
	Summarized conversation	Derm-Private	0.787	(0.717, 0.857)
	Summarized conversation	Derm-Public	0.814	(0.751, 0.877)
GPT-3.5	Vignette	MedQA-USMLE	0.603	(0.522, 0.685)
	Vignette	Derm-Private	0.75	(0.669, 0.831)
	Vignette	Derm-Public	0.786	(0.709, 0.863)
	Multi-turn conversation	MedQA-USMLE	0.46	(0.383, 0.536)
	Multi-turn conversation	Derm-Private	0.588	(0.505, 0.671)
	Multi-turn conversation	Derm-Public	0.614	(0.532, 0.696)
	Single-turn conversation	MedQA-USMLE	0.451	(0.371, 0.531)
	Single-turn conversation	Derm-Private	0.494	(0.407, 0.581)
	Single-turn conversation	Derm-Public	0.586	(0.501, 0.671)
	Summarized conversation	MedQA-USMLE	0.458	(0.381, 0.536)
	Summarized conversation	Derm-Private	0.602	(0.521, 0.683)
	Summarized conversation	Derm-Public	0.65	(0.574, 0.725)
Mistral-v2-7b	Vignette	MedQA-USMLE	0.607	(0.517, 0.697)
	Vignette	Derm-Private	0.77	(0.686, 0.854)
	Vignette	Derm-Public	0.72	(0.63, 0.81)
	Multi-turn conversation	MedQA-USMLE	0.453	(0.379, 0.527)
	Multi-turn conversation	Derm-Private	0.522	(0.441, 0.603)
	Multi-turn conversation	Derm-Public	0.568	(0.487, 0.649)
	Single-turn conversation	MedQA-USMLE	0.515	(0.433, 0.596)
	Single-turn conversation	Derm-Private	0.508	(0.416, 0.6)
	Single-turn conversation	Derm-Public	0.574	(0.489, 0.659)

	Summarized conversation	MedQA-USMLE	0.562	(0.493, 0.631)
	Summarized conversation	Derm-Private	0.548	(0.47, 0.626)
	Summarized conversation	Derm-Public	0.618	(0.542, 0.694)
LLaMA-2-7b	Vignette	MedQA-USMLE	0.385	(0.304, 0.465)
	Vignette	Derm-Private	0.44	(0.349, 0.531)
	Vignette	Derm-Public	0.414	(0.323, 0.505)
	Multi-turn conversation	MedQA-USMLE	0.383	(0.316, 0.45)
	Multi-turn conversation	Derm-Private	0.27	(0.197, 0.343)
	Multi-turn conversation	Derm-Public	0.38	(0.301, 0.459)
	Single-turn conversation	MedQA-USMLE	0.371	(0.291, 0.451)
	Single-turn conversation	Derm-Private	0.244	(0.165, 0.323)
	Single-turn conversation	Derm-Public	0.354	(0.268, 0.44)
	Summarized conversation	MedQA-USMLE	0.405	(0.33, 0.48)
	Summarized conversation	Derm-Private	0.31	(0.236, 0.384)
	Summarized conversation	Derm-Public	0.4	(0.319, 0.481)

Supplementary Table 10: Mean accuracy and 95% confidence intervals for Dermatology, calculated by dataset source for 4-choice MCQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Dataset	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0003	0.001
		Multi-turn conversation	Single-turn conversation	0.0234	0.0443
		Multi-turn conversation	Summarized conversation	0.0181	0.0372
		Single-turn conversation	Summarized conversation	0.0003	0.001
	Derm-Public	Vignette	Multi-turn conversation	0.0025	0.0062
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0132	0.028
		Multi-turn conversation	Single-turn conversation	0.0219	0.0426
		Multi-turn conversation	Summarized conversation	0.1866	0.2742
		Single-turn conversation	Summarized conversation	0.0015	0.004
	Derm-Private	Vignette	Multi-turn conversation	0.0065	0.0142
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0033	0.0079
		Multi-turn conversation	Single-turn conversation	0.0002	0.0007
		Multi-turn conversation	Summarized conversation	0.5506	0.6293
		Single-turn conversation	Summarized conversation	0.0011	0.0032
GPT-3.5	MedQA-USMLE	Vignette	Multi-turn conversation	0.0013	0.0036
		Vignette	Single-turn conversation	0.0004	0.0013
		Vignette	Summarized conversation	0.0002	0.0007
		Multi-turn conversation	Single-turn conversation	0.7558	0.8003
		Multi-turn conversation	Summarized conversation	0.9599	0.9599
		Single-turn conversation	Summarized conversation	0.8409	0.8649
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0002	0.0007
		Multi-turn conversation	Single-turn conversation	0.3121	0.4086
		Multi-turn conversation	Summarized conversation	0.1919	0.2763
		Single-turn conversation	Summarized conversation	0.067	0.1149
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0001	0.0004

		Multi-turn conversation	Single-turn conversation	0.0047	0.0106
		Multi-turn conversation	Summarized conversation	0.5018	0.5923
		Single-turn conversation	Summarized conversation	0.0023	0.0059
Mistral-v2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.001	0.003
		Vignette	Single-turn conversation	0.0736	0.1232
		Vignette	Summarized conversation	0.307	0.4086
		Multi-turn conversation	Single-turn conversation	0.0584	0.1051
		Multi-turn conversation	Summarized conversation	0.0001	0.0004
		Single-turn conversation	Summarized conversation	0.1977	0.2791
	Derm-Public	Vignette	Multi-turn conversation	0.0002	0.0007
		Vignette	Single-turn conversation	0.004	0.0093
		Vignette	Summarized conversation	0.0189	0.0378
		Multi-turn conversation	Single-turn conversation	0.8129	0.8482
		Multi-turn conversation	Summarized conversation	0.0602	0.1057
		Single-turn conversation	Summarized conversation	0.1336	0.2186
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0001	0.0004
		Multi-turn conversation	Single-turn conversation	0.6331	0.7122
		Multi-turn conversation	Summarized conversation	0.3858	0.4873
		Single-turn conversation	Summarized conversation	0.1858	0.2742
LLaMA-2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.9505	0.9599
		Vignette	Single-turn conversation	0.7234	0.7774
		Vignette	Summarized conversation	0.4754	0.5705
		Multi-turn conversation	Single-turn conversation	0.6705	0.7427
		Multi-turn conversation	Summarized conversation	0.4286	0.523
		Single-turn conversation	Summarized conversation	0.2417	0.3347
	Derm-Public	Vignette	Multi-turn conversation	0.4243	0.523
		Vignette	Single-turn conversation	0.1789	0.2741
		Vignette	Summarized conversation	0.6821	0.7441
		Multi-turn conversation	Single-turn conversation	0.3237	0.4162
		Multi-turn conversation	Summarized conversation	0.5145	0.5975
		Single-turn conversation	Summarized conversation	0.155	0.248
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004

		Vignette	Summarized conversation	0.0001	0.0004
		Multi-turn conversation	Single-turn conversation	0.2841	0.3859
		Multi-turn conversation	Summarized conversation	0.1611	0.2522
		Single-turn conversation	Summarized conversation	0.0381	0.0703

Supplementary Table 11: Adjusted p-values for Dermatology, calculated by dataset source for the 4-choice MCQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Dataset	Mean Accuracy	95% C.I.
GPT-4	Vignette	MedQA-USMLE	0.552	(0.467, 0.637)
	Vignette	Derm-Private	0.818	(0.751, 0.885)
	Vignette	Derm-Public	0.582	(0.495, 0.669)
	Multi-turn conversation	MedQA-USMLE	0.345	(0.274, 0.416)
	Multi-turn conversation	Derm-Private	0.572	(0.493, 0.651)
	Multi-turn conversation	Derm-Public	0.288	(0.215, 0.361)
	Single-turn conversation	MedQA-USMLE	0.244	(0.174, 0.314)
	Single-turn conversation	Derm-Private	0.256	(0.18, 0.332)
	Single-turn conversation	Derm-Public	0.168	(0.105, 0.231)
	Summarized conversation	MedQA-USMLE	0.397	(0.321, 0.472)
	Summarized conversation	Derm-Private	0.586	(0.507, 0.664)
	Summarized conversation	Derm-Public	0.334	(0.258, 0.41)
GPT-3.5	Vignette	MedQA-USMLE	0.496	(0.41, 0.582)
	Vignette	Derm-Private	0.608	(0.524, 0.692)
	Vignette	Derm-Public	0.41	(0.324, 0.496)
	Multi-turn conversation	MedQA-USMLE	0.238	(0.17, 0.306)
	Multi-turn conversation	Derm-Private	0.34	(0.263, 0.417)
	Multi-turn conversation	Derm-Public	0.146	(0.091, 0.201)
	Single-turn conversation	MedQA-USMLE	0.173	(0.116, 0.23)
	Single-turn conversation	Derm-Private	0.236	(0.159, 0.313)
	Single-turn conversation	Derm-Public	0.098	(0.05, 0.146)
	Summarized conversation	MedQA-USMLE	0.244	(0.178, 0.31)
	Summarized conversation	Derm-Private	0.401	(0.32, 0.483)
	Summarized conversation	Derm-Public	0.183	(0.127, 0.24)
Mistral-v2-7b	Vignette	MedQA-USMLE	0.253	(0.176, 0.33)
	Vignette	Derm-Private	0.296	(0.207, 0.385)
	Vignette	Derm-Public	0.096	(0.039, 0.153)
	Multi-turn conversation	MedQA-USMLE	0.097	(0.054, 0.141)
	Multi-turn conversation	Derm-Private	0.14	(0.084, 0.196)
	Multi-turn conversation	Derm-Public	0.06	(0.028, 0.092)
	Single-turn conversation	MedQA-USMLE	0.101	(0.052, 0.15)
	Single-turn conversation	Derm-Private	0.098	(0.046, 0.15)
	Single-turn conversation	Derm-Public	0.058	(0.02, 0.096)

	Summarized conversation	MedQA-USMLE	0.089	(0.05, 0.127)
	Summarized conversation	Derm-Private	0.13	(0.077, 0.183)
	Summarized conversation	Derm-Public	0.05	(0.019, 0.081)
LLaMA-2-7b	Vignette	MedQA-USMLE	0.244	(0.173, 0.316)
	Vignette	Derm-Private	0.212	(0.136, 0.288)
	Vignette	Derm-Public	0.056	(0.017, 0.095)
	Multi-turn conversation	MedQA-USMLE	0.111	(0.069, 0.154)
	Multi-turn conversation	Derm-Private	0.108	(0.06, 0.156)
	Multi-turn conversation	Derm-Public	0.028	(0.006, 0.05)
	Single-turn conversation	MedQA-USMLE	0.104	(0.056, 0.153)
	Single-turn conversation	Derm-Private	0.078	(0.028, 0.128)
	Single-turn conversation	Derm-Public	0.036	(0.005, 0.067)
	Summarized conversation	MedQA-USMLE	0.14	(0.09, 0.19)
	Summarized conversation	Derm-Private	0.11	(0.061, 0.159)
	Summarized conversation	Derm-Public	0.048	(0.014, 0.082)

Supplementary Table 12: Mean accuracy and 95% confidence intervals for Dermatology, calculated by dataset source for FRQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Dataset	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0002	0.0004
		Multi-turn conversation	Summarized conversation	0.0122	0.0201
		Single-turn conversation	Summarized conversation	0.0001	0.0002
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0001	0.0002
		Multi-turn conversation	Summarized conversation	0.0382	0.0598
		Single-turn conversation	Summarized conversation	0.0001	0.0002
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0001	0.0002
		Multi-turn conversation	Summarized conversation	0.4567	0.5391
		Single-turn conversation	Summarized conversation	0.0001	0.0002
GPT-3.5	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0018	0.0032
		Multi-turn conversation	Summarized conversation	0.7346	0.7665
		Single-turn conversation	Summarized conversation	0.004	0.007
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0143	0.0229
		Multi-turn conversation	Summarized conversation	0.0615	0.0942
		Single-turn conversation	Summarized conversation	0.0005	0.001
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002

		Multi-turn conversation	Single-turn conversation	0.0013	0.0025
		Multi-turn conversation	Summarized conversation	0.0042	0.0072
		Single-turn conversation	Summarized conversation	0.0001	0.0002
Mistral-v2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.8394	0.8634
		Multi-turn conversation	Summarized conversation	0.4993	0.5706
		Single-turn conversation	Summarized conversation	0.4866	0.5651
	Derm-Public	Vignette	Multi-turn conversation	0.2204	0.2811
		Vignette	Single-turn conversation	0.2225	0.2811
		Vignette	Summarized conversation	0.1207	0.1704
		Multi-turn conversation	Single-turn conversation	0.869	0.8812
		Multi-turn conversation	Summarized conversation	0.3195	0.3899
		Single-turn conversation	Summarized conversation	0.6318	0.6789
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0123	0.0201
		Multi-turn conversation	Summarized conversation	0.5561	0.6256
		Single-turn conversation	Summarized conversation	0.1073	0.1577
LLaMA-2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0003	0.0006
		Multi-turn conversation	Single-turn conversation	0.7042	0.7456
		Multi-turn conversation	Summarized conversation	0.0998	0.1497
		Single-turn conversation	Summarized conversation	0.1191	0.1704
	Derm-Public	Vignette	Multi-turn conversation	0.1541	0.2061
		Vignette	Single-turn conversation	0.2914	0.3617
		Vignette	Summarized conversation	0.6131	0.6688
		Multi-turn conversation	Single-turn conversation	0.5778	0.64
		Multi-turn conversation	Summarized conversation	0.1383	0.1915
		Single-turn conversation	Summarized conversation	0.4087	0.4904
	Derm-Private	Vignette	Multi-turn conversation	0.0014	0.0026
		Vignette	Single-turn conversation	0.0002	0.0004

		Vignette	Summarized conversation	0.0014	0.0026
		Multi-turn conversation	Single-turn conversation	0.1825	0.2389
		Multi-turn conversation	Summarized conversation	0.9422	0.9422
		Single-turn conversation	Summarized conversation	0.1546	0.2061

Supplementary Table 13: Adjusted p-values for Dermatology, calculated by dataset source for the FRQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Dataset	Mean Accuracy	95% C.I.
GPT-4	Vignette	MedQA-USMLE	0.867	(0.796, 0.937)
	Vignette	Derm-Private	0.861	(0.792, 0.929)
	Vignette	Derm-Public	0.919	(0.859, 0.979)
	Multi-turn conversation	MedQA-USMLE	0.71	(0.628, 0.791)
	Multi-turn conversation	Derm-Private	0.8	(0.73, 0.87)
	Multi-turn conversation	Derm-Public	0.797	(0.72, 0.875)
	Single-turn conversation	MedQA-USMLE	0.629	(0.54, 0.717)
	Single-turn conversation	Derm-Private	0.675	(0.592, 0.758)
	Single-turn conversation	Derm-Public	0.732	(0.642, 0.821)
	Summarized conversation	MedQA-USMLE	0.755	(0.68, 0.83)
	Summarized conversation	Derm-Private	0.791	(0.721, 0.861)
	Summarized conversation	Derm-Public	0.82	(0.749, 0.891)
GPT-3.5	Vignette	MedQA-USMLE	0.712	(0.624, 0.799)
	Vignette	Derm-Private	0.758	(0.677, 0.838)
	Vignette	Derm-Public	0.83	(0.751, 0.91)
	Multi-turn conversation	MedQA-USMLE	0.493	(0.402, 0.584)
	Multi-turn conversation	Derm-Private	0.594	(0.51, 0.677)
	Multi-turn conversation	Derm-Public	0.618	(0.527, 0.708)
	Single-turn conversation	MedQA-USMLE	0.517	(0.42, 0.613)
	Single-turn conversation	Derm-Private	0.499	(0.411, 0.587)
	Single-turn conversation	Derm-Public	0.587	(0.49, 0.685)
	Summarized conversation	MedQA-USMLE	0.519	(0.425, 0.613)
	Summarized conversation	Derm-Private	0.608	(0.528, 0.689)
	Summarized conversation	Derm-Public	0.653	(0.566, 0.739)
Mistral-v2-7b	Vignette	MedQA-USMLE	0.655	(0.551, 0.759)
	Vignette	Derm-Private	0.768	(0.683, 0.852)
	Vignette	Derm-Public	0.785	(0.692, 0.877)
	Multi-turn conversation	MedQA-USMLE	0.486	(0.397, 0.575)
	Multi-turn conversation	Derm-Private	0.517	(0.436, 0.598)
	Multi-turn conversation	Derm-Public	0.605	(0.517, 0.693)
	Single-turn conversation	MedQA-USMLE	0.536	(0.44, 0.631)
	Single-turn conversation	Derm-Private	0.511	(0.419, 0.603)
	Single-turn conversation	Derm-Public	0.61	(0.515, 0.705)

	Summarized conversation	MedQA-USMLE	0.555	(0.472, 0.637)
	Summarized conversation	Derm-Private	0.543	(0.465, 0.622)
	Summarized conversation	Derm-Public	0.646	(0.56, 0.731)
LLaMA-2-7b	Vignette	MedQA-USMLE	0.388	(0.291, 0.485)
	Vignette	Derm-Private	0.434	(0.343, 0.526)
	Vignette	Derm-Public	0.433	(0.33, 0.536)
	Multi-turn conversation	MedQA-USMLE	0.379	(0.297, 0.46)
	Multi-turn conversation	Derm-Private	0.269	(0.195, 0.342)
	Multi-turn conversation	Derm-Public	0.385	(0.294, 0.475)
	Single-turn conversation	MedQA-USMLE	0.374	(0.276, 0.471)
	Single-turn conversation	Derm-Private	0.236	(0.158, 0.315)
	Single-turn conversation	Derm-Public	0.372	(0.272, 0.473)
	Summarized conversation	MedQA-USMLE	0.414	(0.324, 0.505)
	Summarized conversation	Derm-Private	0.305	(0.231, 0.379)
	Summarized conversation	Derm-Public	0.397	(0.307, 0.488)

Supplementary Table 14: Mean accuracy and 95% confidence intervals for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for 4-choice MCQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Dataset	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0003	0.001
		Multi-turn conversation	Single-turn conversation	0.0234	0.0443
		Multi-turn conversation	Summarized conversation	0.0181	0.0372
		Single-turn conversation	Summarized conversation	0.0003	0.001
	Derm-Public	Vignette	Multi-turn conversation	0.0025	0.0062
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0132	0.028
		Multi-turn conversation	Single-turn conversation	0.0219	0.0426
		Multi-turn conversation	Summarized conversation	0.1866	0.2742
		Single-turn conversation	Summarized conversation	0.0015	0.004
	Derm-Private	Vignette	Multi-turn conversation	0.0065	0.0142
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0033	0.0079
		Multi-turn conversation	Single-turn conversation	0.0002	0.0007
		Multi-turn conversation	Summarized conversation	0.5506	0.6293
		Single-turn conversation	Summarized conversation	0.0011	0.0032
GPT-3.5	MedQA-USMLE	Vignette	Multi-turn conversation	0.0013	0.0036
		Vignette	Single-turn conversation	0.0004	0.0013
		Vignette	Summarized conversation	0.0002	0.0007
		Multi-turn conversation	Single-turn conversation	0.7558	0.8003
		Multi-turn conversation	Summarized conversation	0.9599	0.9599
		Single-turn conversation	Summarized conversation	0.8409	0.8649
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0002	0.0007
		Multi-turn conversation	Single-turn conversation	0.3121	0.4086
		Multi-turn conversation	Summarized conversation	0.1919	0.2763
		Single-turn conversation	Summarized conversation	0.067	0.1149
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0001	0.0004

		Multi-turn conversation	Single-turn conversation	0.0047	0.0106
		Multi-turn conversation	Summarized conversation	0.5018	0.5923
		Single-turn conversation	Summarized conversation	0.0023	0.0059
Mistral-v2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.001	0.003
		Vignette	Single-turn conversation	0.0736	0.1232
		Vignette	Summarized conversation	0.307	0.4086
		Multi-turn conversation	Single-turn conversation	0.0584	0.1051
		Multi-turn conversation	Summarized conversation	0.0001	0.0004
		Single-turn conversation	Summarized conversation	0.1977	0.2791
	Derm-Public	Vignette	Multi-turn conversation	0.0002	0.0007
		Vignette	Single-turn conversation	0.004	0.0093
		Vignette	Summarized conversation	0.0189	0.0378
		Multi-turn conversation	Single-turn conversation	0.8129	0.8482
		Multi-turn conversation	Summarized conversation	0.0602	0.1057
		Single-turn conversation	Summarized conversation	0.1336	0.2186
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0001	0.0004
		Multi-turn conversation	Single-turn conversation	0.6331	0.7122
		Multi-turn conversation	Summarized conversation	0.3858	0.4873
		Single-turn conversation	Summarized conversation	0.1858	0.2742
LLaMA-2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.9505	0.9599
		Vignette	Single-turn conversation	0.7234	0.7774
		Vignette	Summarized conversation	0.4754	0.5705
		Multi-turn conversation	Single-turn conversation	0.6705	0.7427
		Multi-turn conversation	Summarized conversation	0.4286	0.523
		Single-turn conversation	Summarized conversation	0.2417	0.3347
	Derm-Public	Vignette	Multi-turn conversation	0.4243	0.523
		Vignette	Single-turn conversation	0.1789	0.2741
		Vignette	Summarized conversation	0.6821	0.7441
		Multi-turn conversation	Single-turn conversation	0.3237	0.4162
		Multi-turn conversation	Summarized conversation	0.5145	0.5975
		Single-turn conversation	Summarized conversation	0.155	0.248
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004

		Vignette	Summarized conversation	0.0001	0.0004
		Multi-turn conversation	Single-turn conversation	0.2841	0.3859
		Multi-turn conversation	Summarized conversation	0.1611	0.2522
		Single-turn conversation	Summarized conversation	0.0381	0.0703

Supplementary Table 15: Adjusted p-values for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for the 4-choice MCQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Dataset	Mean Accuracy	95% C.I.
GPT-4	Vignette	MedQA-USMLE	0.65	(0.555, 0.745)
	Vignette	Derm-Private	0.826	(0.761, 0.892)
	Vignette	Derm-Public	0.641	(0.546, 0.735)
	Multi-turn conversation	MedQA-USMLE	0.412	(0.327, 0.496)
	Multi-turn conversation	Derm-Private	0.578	(0.499, 0.657)
	Multi-turn conversation	Derm-Public	0.319	(0.233, 0.405)
	Single-turn conversation	MedQA-USMLE	0.29	(0.204, 0.377)
	Single-turn conversation	Derm-Private	0.259	(0.182, 0.335)
	Single-turn conversation	Derm-Public	0.195	(0.119, 0.271)
	Summarized conversation	MedQA-USMLE	0.45	(0.36, 0.54)
	Summarized conversation	Derm-Private	0.591	(0.513, 0.67)
	Summarized conversation	Derm-Public	0.375	(0.287, 0.462)
GPT-3.5	Vignette	MedQA-USMLE	0.579	(0.478, 0.68)
	Vignette	Derm-Private	0.614	(0.53, 0.698)
	Vignette	Derm-Public	0.451	(0.349, 0.552)
	Multi-turn conversation	MedQA-USMLE	0.25	(0.168, 0.332)
	Multi-turn conversation	Derm-Private	0.343	(0.266, 0.421)
	Multi-turn conversation	Derm-Public	0.165	(0.099, 0.23)
	Single-turn conversation	MedQA-USMLE	0.193	(0.124, 0.262)
	Single-turn conversation	Derm-Private	0.238	(0.161, 0.316)
	Single-turn conversation	Derm-Public	0.124	(0.064, 0.184)
	Summarized conversation	MedQA-USMLE	0.259	(0.181, 0.336)
	Summarized conversation	Derm-Private	0.406	(0.324, 0.488)
	Summarized conversation	Derm-Public	0.197	(0.13, 0.263)
Mistral-v2-7b	Vignette	MedQA-USMLE	0.307	(0.211, 0.403)
	Vignette	Derm-Private	0.299	(0.209, 0.389)
	Vignette	Derm-Public	0.122	(0.051, 0.192)
	Multi-turn conversation	MedQA-USMLE	0.124	(0.067, 0.181)
	Multi-turn conversation	Derm-Private	0.141	(0.085, 0.198)
	Multi-turn conversation	Derm-Public	0.071	(0.032, 0.11)
	Single-turn conversation	MedQA-USMLE	0.126	(0.063, 0.19)
	Single-turn conversation	Derm-Private	0.099	(0.047, 0.151)
	Single-turn conversation	Derm-Public	0.068	(0.021, 0.115)

	Summarized conversation	MedQA-USMLE	0.107	(0.059, 0.155)
	Summarized conversation	Derm-Private	0.131	(0.078, 0.185)
	Summarized conversation	Derm-Public	0.063	(0.025, 0.102)
LLaMA-2-7b	Vignette	MedQA-USMLE	0.271	(0.186, 0.357)
	Vignette	Derm-Private	0.214	(0.138, 0.291)
	Vignette	Derm-Public	0.071	(0.022, 0.12)
	Multi-turn conversation	MedQA-USMLE	0.126	(0.072, 0.18)
	Multi-turn conversation	Derm-Private	0.109	(0.061, 0.157)
	Multi-turn conversation	Derm-Public	0.03	(0.004, 0.056)
	Single-turn conversation	MedQA-USMLE	0.129	(0.064, 0.193)
	Single-turn conversation	Derm-Private	0.079	(0.028, 0.129)
	Single-turn conversation	Derm-Public	0.043	(0.005, 0.081)
	Summarized conversation	MedQA-USMLE	0.162	(0.101, 0.223)
	Summarized conversation	Derm-Private	0.111	(0.061, 0.161)
	Summarized conversation	Derm-Public	0.051	(0.01, 0.091)

Supplementary Table 16: Mean accuracy and 95% confidence intervals for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for FRQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Dataset	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.0004	0.0009
		Multi-turn conversation	Summarized conversation	0.1376	0.1981
		Single-turn conversation	Summarized conversation	0.0003	0.0007
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.0005	0.0011
		Multi-turn conversation	Summarized conversation	0.0405	0.0663
		Single-turn conversation	Summarized conversation	0.0002	0.0005
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.0001	0.0003
		Multi-turn conversation	Summarized conversation	0.4547	0.528
		Single-turn conversation	Summarized conversation	0.0001	0.0003
GPT-3.5	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.0222	0.0372
		Multi-turn conversation	Summarized conversation	0.7198	0.7735
		Single-turn conversation	Summarized conversation	0.0171	0.0293
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.048	0.0768
		Multi-turn conversation	Summarized conversation	0.1802	0.2317
		Single-turn conversation	Summarized conversation	0.0038	0.0068
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003

		Multi-turn conversation	Single-turn conversation	0.0007	0.0014
		Multi-turn conversation	Summarized conversation	0.0038	0.0068
		Single-turn conversation	Summarized conversation	0.0001	0.0003
Mistral-v2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.9125	0.9254
		Multi-turn conversation	Summarized conversation	0.3342	0.401
		Single-turn conversation	Summarized conversation	0.4193	0.4949
	Derm-Public	Vignette	Multi-turn conversation	0.1719	0.2292
		Vignette	Single-turn conversation	0.1785	0.2317
		Vignette	Summarized conversation	0.1227	0.1803
		Multi-turn conversation	Single-turn conversation	0.8608	0.8982
		Multi-turn conversation	Summarized conversation	0.5251	0.5907
		Single-turn conversation	Summarized conversation	0.8178	0.8659
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0002	0.0005
		Multi-turn conversation	Single-turn conversation	0.0117	0.0205
		Multi-turn conversation	Summarized conversation	0.5379	0.5958
		Single-turn conversation	Summarized conversation	0.1004	0.1506
LLaMA-2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.0003	0.0007
		Vignette	Single-turn conversation	0.0004	0.0009
		Vignette	Summarized conversation	0.0011	0.0021
		Multi-turn conversation	Single-turn conversation	0.9431	0.9431
		Multi-turn conversation	Summarized conversation	0.0912	0.1397
		Single-turn conversation	Summarized conversation	0.2334	0.2948
	Derm-Public	Vignette	Multi-turn conversation	0.0866	0.1355
		Vignette	Single-turn conversation	0.272	0.3377
		Vignette	Summarized conversation	0.3328	0.401
		Multi-turn conversation	Single-turn conversation	0.4691	0.5361
		Multi-turn conversation	Summarized conversation	0.1521	0.2106
		Single-turn conversation	Summarized conversation	0.6356	0.6934
	Derm-Private	Vignette	Multi-turn conversation	0.0015	0.0028
		Vignette	Single-turn conversation	0.0004	0.0009

		Vignette	Summarized conversation	0.0011	0.0021
		Multi-turn conversation	Single-turn conversation	0.1652	0.2244
		Multi-turn conversation	Summarized conversation	0.9063	0.9254
		Single-turn conversation	Summarized conversation	0.1459	0.206

Supplementary Table 17: Adjusted p-values for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for the FRQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Question	Model 1	Model 2	P-value (McNemar's test)
Did the clinical LLM stop asking questions when only a single most likely diagnosis was possible?	GPT-4	GPT-3.5	0.0129
	GPT-4	Mistral-v2-7b	<0.0001
	GPT-4	LLaMA-2-7b	0.0573
	GPT-3.5	Mistral-v2-7b	0.0039
	GPT-3.5	LLaMA-2-7b	0.7744
	Mistral-v2-7b	LLaMA-2-7b	0.0009
Did the clinical LLM elicit the relevant medical history?	GPT-4	GPT-3.5	0.0001
	GPT-4	Mistral-v2-7b	<0.0001
	GPT-4	LLaMA-2-7b	0.0635
	GPT-3.5	Mistral-v2-7b	0.0212
	GPT-3.5	LLaMA-2-7b	0.0635
	Mistral-v2-7b	LLaMA-2-7b	<0.0001

Supplementary Table 18: P-values between pairs of evaluated models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b) for clinical LLM assessment by medical experts. All p-values were calculated using McNemar's test (see Methods).

	Model	Experiment	Mean Accuracy	95% C.I.
4-choice MCQ	GPT-4V	Vignette	0.842	(0.763, 0.921)
		Multi-turn conversation	0.492	(0.397, 0.587)
		Single-turn conversation	0.474	(0.369, 0.578)
		Summarized conversation	0.547	(0.456, 0.639)
	GPT-4V-without-image	Vignette	0.787	(0.695, 0.878)
		Multi-turn conversation	0.468	(0.37, 0.567)
		Single-turn conversation	0.4	(0.297, 0.503)
		Summarized conversation	0.503	(0.405, 0.601)
FRQ	GPT-4V	Vignette	0.492	(0.398, 0.586)
		Multi-turn conversation	0.145	(0.084, 0.205)
		Single-turn conversation	0.063	(0.021, 0.106)
		Summarized conversation	0.163	(0.1, 0.226)
	GPT-4V-without-image	Vignette	0.471	(0.367, 0.575)
		Multi-turn conversation	0.087	(0.037, 0.137)
		Single-turn conversation	0.039	(0.01, 0.069)
		Summarized conversation	0.108	(0.052, 0.164)

Supplementary Table 19: Mean accuracy and 95% confidence intervals for 4-choice MCQ and FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) for GPT-4V and GPT-4V-without-image.

	Model	Experiment 1	Experiment 2	p-values	Adjusted p-value
4-choice MCQ	GPT-4V	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.6586	0.6586
		Multi-turn conversation	Summarized conversation	0.1105	0.1357
		Single-turn conversation	Summarized conversation	0.1131	0.1357
	GPT-4V-without-image	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.1097	0.1357
		Multi-turn conversation	Summarized conversation	0.2719	0.2966
		Single-turn conversation	Summarized conversation	0.019	0.0326
FRQ	GPT-4V	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0002	0.0003
		Multi-turn conversation	Summarized conversation	0.3699	0.3699
		Single-turn conversation	Summarized conversation	0.0007	0.0011
	GPT-4V-without-image	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn	0.0304	0.0365

			conversation		
		Multi-turn conversation	Summarized conversation	0.3625	0.3699
		Single-turn conversation	Summarized conversation	0.0052	0.0069

Supplementary Table 20: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to GPT-4V and GPT-4V-without-image.

	Experiment	Model 1	Model 2	p-value	Adjusted p-value
4-choice MCQ	Vignette	GPT-4V	GPT-4V-wit hout-image	0.0631	0.2064
	Multi-turn conversation	GPT-4V	GPT-4V-wit hout-image	0.5356	0.5356
	Single-turn conversation	GPT-4V	GPT-4V-wit hout-image	0.1032	0.2064
	Summarized conversation	GPT-4V	GPT-4V-wit hout-image	0.1885	0.2513
FRQ	Vignette	GPT-4V	GPT-4V-wit hout-image	0.5309	0.5309
	Multi-turn conversation	GPT-4V	GPT-4V-wit hout-image	0.0148	0.0506
	Single-turn conversation	GPT-4V	GPT-4V-wit hout-image	0.1312	0.1749
	Summarized conversation	GPT-4V	GPT-4V-wit hout-image	0.0253	0.0506

Supplementary Table 21: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated models (GPT-4V and GPT-4V-without-image) corresponding to each of the experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

	Model	Experiment	Accuracy	95% C.I.
4-choice MCQ	Mistral-v1-7b	Vignette	0.441	(0.419, 0.462)
		Multi-turn conversation	0.331	(0.314, 0.349)
		Single-turn conversation	0.324	(0.305, 0.344)
		Summarized conversation	0.361	(0.343, 0.379)
	Mistral-v2-7b	Vignette	0.637	(0.616, 0.658)
		Multi-turn conversation	0.426	(0.409, 0.443)
		Single-turn conversation	0.448	(0.429, 0.468)
		Summarized conversation	0.513	(0.496, 0.529)
FRQ	Mistral-v1-7b	Vignette	0.165	(0.142, 0.189)
		Multi-turn conversation	0.08	(0.065, 0.095)
		Single-turn conversation	0.06	(0.046, 0.074)
		Summarized conversation	0.082	(0.068, 0.097)
	Mistral-v2-7b	Vignette	0.211	(0.186, 0.237)
		Multi-turn conversation	0.065	(0.052, 0.077)
		Single-turn conversation	0.055	(0.043, 0.068)
		Summarized conversation	0.055	(0.044, 0.066)

Supplementary Table 22: Mean accuracy and 95% confidence intervals for 4-choice MCQ and FRQ setting, across experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) for Mistral-v1-7b and Mistral-v2-7b.

	Model	Experiment 1	Experiment 2	p-values	Adjusted p-value
4-choice MCQ	Mistral-v1-7b	Vignette	Multi-turn conversation	0.0001	0.0001
		Vignette	Single-turn conversation	0.0001	0.0001
		Vignette	Summarized conversation	0.0001	0.0001
		Multi-turn conversation	Single-turn conversation	0.2975	0.2975
		Multi-turn conversation	Summarized conversation	0.0001	0.0001
		Single-turn conversation	Summarized conversation	0.0001	0.0001
FRQ	Mistral-v1-7b	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0003	0.0004
		Multi-turn conversation	Summarized conversation	0.5212	0.5686
		Single-turn conversation	Summarized conversation	0.0001	0.0002

Supplementary Table 23: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to Mistral-v1-7b. All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection | No software was used for data collection.

Data analysis | Python version 3.9.19 was used for performing all data analysis. python-Levenshtein package (version 0.25.1) was used for performing MELD analysis described in the methods. Seaborn (version 0.13.2) and matplotlib (version 3.8.4) packages were used for data visualization. Scipy (version 1.13.1) and statsmodel (version 0.14.2) packages was used for performing statistical tests for p-value calculation. Transformers package (version 4.39.2) was used to access models (Mistral-v1, Mistral-v2, LLaMA-2-7b) from huggingface.

All code used for data analysis is also available on the github repo, with instructions to reproduce our results - <https://github.com/rajpurkarlab/craft-md>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

MedQA-USMLE case vignettes can be downloaded from - <https://github.com/jind11/MedQA>. Derm-Public case vignettes were downloaded from - <https://www.clinicaladvisor.com/>. The images and corresponding vignettes for the NEJM Image Challenge can be downloaded from their website. (<https://www.nejm.org/image-challenge>). The private dataset generated as a part of our study can be found on <https://github.com/rajpurkarlab/craft-md>. All case vignettes used in the study are also available on the following repository: <https://github.com/rajpurkarlab/craft-md>.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	Not Applicable.
Population characteristics	Not Applicable.
Recruitment	Not Applicable.
Ethics oversight	Not Applicable.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<p>For text-only evaluation of LLMs, all case vignette like questions were extracted from MedQA-USMLE dataset for this. This led to a total of 2000 case vignettes structured as 4-choice MCQs for the final evaluation, with four case vignettes used for prompt optimization.</p> <p>For multimodal evaluation of LLMs, 100 case vignettes, images pairs were downloaded from the NEJM Image Challenge website. Of these, only 26 images were restricted by the GPT-4V content filter, therefore only 74 cases were used for the final evaluation.</p> <p>These sample sizes are sufficient for detecting statistically significant differences using Mann-Whitney U Test.</p>
Data exclusions	No data/case vignettes were excluded.
Replication	All statistics calculations in the study had controlled seeds to ensure reproducibility (number of bootstrap samples = 10000). All attempts at replication of statistical results were successful. Large Language Models (LLMs) cannot be seeded, therefore exact prompts are provided in the Methods section of the manuscript and raw outputs for all experiments with LLMs are provided the github repository - https://github.com/rajpurkarlab/craft-md . Each LLM experiment was performed 5 times.
Randomization	Randomization was not relevant to this study, since there were no group comparisons.
Blinding	There was no group allocation - all case vignettes were evaluated across the selected LLMs. Therefore, there was no blinding in this study. Medical experts were not blinded to the LLM they were evaluating to allow them to make conclusions about overall trends across model outputs.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

- | n/a | Involvement |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |

- | n/a | Involvement |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |