Towards Better Health Conversations: The Benefits of Context-seeking

RORY SAYRES*†, YUEXING HAO*, ABBI WARD, AMY WANG, BEVERLY FREEMAN, SERENA ZHAN, DIEGO ARDILA, JIMMY LI, I-CHING LEE, ANNA IURCHENKO, SIYI KOU, KARTIKEYA BADOLA, JIMMY HU, BHAWESH KUMAR, KEITH JOHNSON, SUPRIYA VIJAY, JUSTIN KROGUE, AVINATAN HASSIDIM, YOSSI MATIAS, DALE R. WEBSTER, SUNNY VIRMANI, YUN LIU, QUANG DUONG, and MIKE SCHAEKERMANN†, Google Research, Mountain View, USA

Navigating health questions can be daunting in the modern information landscape. Large language models (LLMs) may provide tailored, accessible information, but also risk being inaccurate, biased or misleading. We present insights from 4 mixed-methods studies (total N=163), examining how people interact with LLMs for their own health questions. Qualitative studies revealed the importance of context-seeking in conversational AIs to elicit specific details a person may not volunteer or know to share. Context-seeking by LLMs was valued by participants, even if it meant deferring an answer for several turns. Incorporating these insights, we developed a "Wayfinding AI" to proactively solicit context. In a randomized, blinded study, participants rated the Wayfinding AI as more helpful, relevant, and tailored to their concerns compared to a baseline AI. These results demonstrate the strong impact of proactive context-seeking on conversational dynamics, and suggest design patterns for conversational AI to help navigate health topics.

CCS Concepts: • **Human-centered computing → User studies**; **User interface management systems**; *Empirical studies in HCI*.

Additional Key Words and Phrases: Health Information Seeking, Large Language Model

1 Introduction

Modern information technology offers many resources for laypeople to understand their health concerns; but also poses potential pitfalls. People commonly use Internet search for their health and medical questions, with a recent survey reporting that 58.5% of US-based adults used the Internet for this purpose [70]. But questions persist about how people may interpret low-quality or less-relevant results and how it may influence population health behavior [4]. Meanwhile, large language models (LLMs) have been demonstrated to encode medical knowledge, and can perform well on a range of medical tasks [25, 62, 63]. A growing ecosystem of LLM-based tools has the potential to help provide more tailored or accessible information, and anecdotal reports have illustrated cases where LLMs surfaced crucial information for health conditions that were not captured by existing medical care [31]. But this technology also has raised concerns about information accuracy and potential bias [4, 5, 28].

Addressing certain health information needs can require access to specific context that users may not proactively include in health searches. For instance, for cause-of-symptoms intent (henceforth COSI) needs, where a person wants to understand what is causing a set of symptoms, laypeople may not know what to proactively share to help ascertain the possible causes. Within clinical contexts, healthcare professionals are trained in history taking, and tested by evaluations such as the objective structured clinical examination [19, 22].

Recent research has developed LLMs for augmenting clinicians in tasks like reasoning about possible diagnoses, including in obtaining relevant context for differential diagnoses [48, 68]. However, there is little understanding of how LLMs can optimally support laypeople in their online health information seeking (OHIS) journeys in non-clinical

*Both authors contributed equally.
†Corresponding authors: sayres@google.com, mikeshake@google.com

settings. Because laypeople lack clinical training, the types of questions posed to the LLM, decisions being made, and the appropriate behavior of the LLMs may differ substantially.

Our work examines how people interact conversationally with LLMs for their own health questions. We conducted 3 interview-based qualitative studies, and 1 quantitative study, to understand how participants engage with context-seeking in conversations about their own real-world health questions.

We make four key contributions:

(1) We provide qualitative evidence that proactive context-seeking by LLMs is valued by laypeople for their health questions, even if it delays an answer by several conversation turns.
(2) We describe a **Wayfinding Artificial Intelligence (Wayfinding AI)** which was iteratively designed based on qualitative feedback to engage in effective context-seeking, to help users obtain the most relevant information for their concern.
(3) We compare user perceptions of the Wayfinding AI to the baseline LLM from which it was trained; in a quantitative study (N=130) we show that the Wayfinding AI is significantly preferred over the baseline AI on helpfulness, relevance of questions asked, tailoring, and efficiency of answering health questions.
(4) We discuss design considerations for how LLMs should behave in conversational contexts, to help users better provide relevant context and to comprehend and trust the outputs.

2 Related Work

2.1 Online Health Information Seeking

Online health information seeking (OHIS) is a well-studied phenomenon in the human-computer interaction (HCI) literature. Prior work has explored OHIS in the context of different platforms including search engines and social media [13, 74], specific medical domains like mental health [51], reproductive health [16], rare disorders [9] or COVID-19 [17], as well as demographic factors like age, gender and ethnicity [2, 6, 16, 29, 30, 46]. Research has found that the framing of information, algorithmic transparency, and overall system design can influence important aspects of user perception including trust, comfort and anxiety [29, 71, 74]. Individuals' pursuit of health information is often driven by personal, emotional, and motivational factors and has been shown to take place within information ecosystems [51], rather than simply constituting a detached quest for facts [33, 37, 54]. In recent years, LLMs have become another popular starting point for OHIS journeys [27], and research has started probing for laypeople's attitudes towards LLMs versus other methods (like search engines) for their health queries [50]. Our work contributes to understanding how LLMs can optimally support laypeople in their OHIS journeys through proactive context-seeking behavior.

2.2 Expert-layperson conversations

When human experts in different domains converse with laypeople about their area of expertise, they adopt different strategies based on the layperson's questions and phrasing [23]: They will "zoom in" to gain context about specific questions when it is not initially provided, "zoom out" to clarify overall goals and explore other possible approaches; and "reframe" when question intent is ambiguous or relies on false assumptions.

In-person conversations, such as patient-provider discussions, can rely on a range of nonverbal cues to gauge understanding of content and level of concern [8]. As more conversations are done using text or other remote interfaces, experts have had to adapt conversational strategies to better assess these facets of communication [40, 41]. However, some of these strategies make conversation harder for laypeople, such as the use of "double-barrel" questions that are harder to answer concisely [40].

2.3 Clinician-AI conversations

Most clinical decision support AI has focused on reading of medical images [20, 66]. Use patterns include risks of over-reliance and under-reliance [15, 24, 55, 67]. Earlier applications tended to involve AI in narrowly-defined tasks, such as reading specific image modalities (X-rays, mammograms, retina images) for particular diseases [49, 60]. More recent work has examined how conversation may be used to support clinicians across a wider range of clinical domains, including differential diagnosis based on understanding complex information and history taking [48, 58, 68].

2.4 Layperson decision support with AI tools

Consumer-oriented, AI-powered apps are available in some countries for uses such as detecting common types of skin cancers [64]. Use of an app of this kind was associated with an increase in claims for skin malignancies in the health care system, suggesting greater detection (and potentially over-detection) of cancers [64]. AI tools to support understanding skin concerns based off smartphone images were helpful for laypeople [35, 59], could increase their ability to name a condition and may support patient-provider interactions [59]; but people tended to anchor on the example images shown by the app, suggesting the app design may strongly impact app benefits.

Many people use Internet search to find information about health topics online [69]. However, the risks and benefits of existing web search patterns are not well characterized. A study of Internet search for health information, diagnosis and triage on retrospective vignettes found that any benefit of Internet search was small, with a small fraction of respondents changing their diagnosis or triage decisions after Internet search [38].

2.5 Understanding conversational patterns with health chatbots

As more people have begun using conversational AI chatbots, they have started asking these tools their health questions [3, 21, 61]. Individual stories suggest these interactions may be highly valuable in some cases [31]; but may pose a range of risks [14, 18, 57]. Several studies suggest broad interest in health-related AI chatbots [3, 10, 52, 61]. Survey studies found a large proportion of respondents who either intended (78%, [61]) or recently had (9.9%, [3]) used ChatGPT for their health questions, with a highly significant effect of age (younger respondents using it more often). There has been a focus within the HCI literature on how laypeople have conversations with AI chatbots on their own health concerns. A study on chatbot-based symptom checkers identified a tension between a desires for efficient information access, and one for empathetic communication [73]. Several studies used a "Wizard-of-Oz" design, with human experts powering AI responses to understand layperson conversational behavior [39–42]. People chatting with clinical experts (dermatologists) disguised as a chatbot will exhibit different patterns of conversational turns compared to discussions with LLMs [42], with more messages overall, as well as more expressions of appreciation and question answering. Likewise, in a clinical pre-consultation setting, expressions of empathy and follow-up questions impacted users' perception of a chatbot's thoroughness and sincerity [41]. These suggest that the default conversational patterns of current LLMs may not sufficiently mirror the behaviors of human experts.

2.6 Need for understanding real-world usage patterns

Despite growing numbers of people conversing with AI chatbots, little is known about what successful interaction looks like, especially on health topics. Controlled studies on a set of hand-selected medical scenarios showed that, when

LLMs were given scenarios directly, they achieved high accuracy; but when laypeople were assigned to act out the scenarios with the same LLMs, performance at assessing conditions plummeted by more than half, to the same levels as without access to LLMs [5]. Similar results showed that LLMs would often "get lost" in multi-turn settings, exhibiting much lower performance when provided context across turns compared to up front [36]. The large gap in performance derived from difficulty of users in providing necessary context around each scenario. Analyses of LLM architectures suggests that human conversations may pose a particular challenge compared to other domains like software and mathematical text, due to the long-range attention across many conversation turns required [32]. Recent research has focused on developing frameworks to quantify aspects of conversation, such as the quality of clarifying questions [34, 43, 44].

These prior results point to conversation dynamics, and the need for context-seeking in particular, as key design factors influencing successful interaction with an AI. To our knowledge, there has not yet been a study of how laypeople use functional AIs (rather than Wizard-of-Oz simulations) on their own health problems. Do people experience the same challenges in providing necessary context to get relevant answers on their own concerns as on simulated scenarios [5]? And do their conversational patterns with AIs mirror those of humans presented as AIs [41, 42]? And are these prior insights actionable – does designing an AI to behave more like human experts lead to better conversations? This study sought to focus on understanding these real-life interaction patterns.

3 Methods

An overview of the 4 studies in this paper is provided in Table 1. The qualitative studies (Studies 1-3) and quantitative study (Study 4) are described in further detail below.

Study	**Key Research Questions**	**Methodology**	**Num of Participants**
Study 1	(1) How do people view AIs' context-seeking behaviors? (2) Is context-seeking valued even if it means deferring an answer for several conversation turns?	Interviews, remote usability session	10
Study 2	(1) How do people view a Wayfinding AI compared to a Baseline AI when using it for their health questions? (2) How is context-seeking perceived on models that tend to give longer responses by default?	Interviews, remote usability session	11
Study 3	(1) Does a user interface that explicitly separates answers from context seeking increase engagement with the AI? (2) What are user perspectives on different UI affordances for context seeking?	Interviews, remote usability session	12
Study 4	(1) Is a Wayfinding AI preferred over a Baseline AI for conversations on health questions? (2) How do conversation dynamics differ between Wayfinding AI and Baseline AIs?	Randomized blinded unmoderated survey study with remote chatbot use	130

Table 1. Summary of the studies described in this paper.

3.1 Ethics & Positionality Statement

This study was meant to focus on user experience, and participants were instructed not to consider information from the research tool as advice or to act on it. The protocol was reviewed and deemed exempt from Institutional Review Board (IRB) oversight by the Advarra IRB. The study was conducted in adherence to our organization's ethical, legal, and privacy standards for human subjects research.

3.2 AI prototypes

This work involved participants interacting with a series of AI prototypes to understand their own health questions.

Wayfinding AI: The primary goal was to develop and test a Wayfinding AI to proactively seek out context of user's health queries, in order to provide a maximally relevant and accurate answer. The Wayfinding AI was developed using a combination of prompt tuning and reinforcement learning off of the baseline AI. Over the course of the 4 studies described in this paper (3 qualitative, 1 quantitative), we iteratively refined our Wayfinding AI to better address feedback from each qualitative study. For the final quantitative study, the Wayfinding AI configuration was frozen prior to commencing the study. We describe the full set of AIs across studies in supplemental Table S2.

Baseline AI: In Studies 2 and 4, we compared the Wayfinding AI to a Baseline AI. Our baseline AI utilized Gemini 2.5 Flash [12] off-the-shelf, without any additional tuning or system prompts specifically designed to engage in context-seeking. This AI served as a comparison baseline in our blinded, randomized Study 4.

AIs were presented to users as chatbots in a custom interface (Figure 1) which presented alternating user prompts and system responses. For one version of the interface, used in qualitative Study 3 and the quantitative Study 4, the interface used a 2-column design that surfaced the best answer using available information so far in a right-hand column, leaving only clarifying questions in the left (main) column with the user-AI conversation, in order to provide a clear affordance for context-seeking.

3.3 Qualitative studies (Studies 1-3)

Three rounds of qualitative studies took place from February to June 2025. We recruited a total of 33 US-based participants for 3 interview-based qualitative studies (10 in Study 1, 11 in Study 2, and 12 in Study 3). Participants ranged in age from 25-75; 17 female / 17 male. Participant demographics included 19 Caucasians, 6 Asian or Pacific Islanders, 4 Hispanics, and 4 Black or African Americans. The full list of participant information and their health questions are presented in Table 2.

All participants had a recent health-related query they were willing to discuss, and were open to interacting with one or more AIs for the study. Sessions lasted 1 hour each and had the following structure: First, participants discussed their question, what information sources they had used so far to understand it, and their current status of understanding the question. Second, participants engaged in a remote usability session while they interacted with an AI to better understand the topic of their question. Researchers probed on why participants provided the prompts they did; what their expectations were for AI responses; what information they hoped to obtain; and whether the AI's behavior matched their expectations. Third, participants used a second, alternate AI. The order of AIs tested by each participant was randomized across participants for each study. Finally, participants compared the two AIs and provided overall impressions and thoughts about their experience and what they learned.

The AIs studied varied between the three studies, based on the key research questions (see Table S2). Study 1 compared a Wayfinding AI that provided an immediate answer to a question to a modified version with a system

ID	Gender	Age	Race / Ethnicity	Health Question(s) Summary
P1	Female	47	Hispanic	Consistent and unexplained weight gain
P2	Female	26	Asian/Pac Is	Respiratory symptoms after travel
P4	Male	50	Cauc	Tinnitus and surgery
P5	Male	62	Asian/Pac Is	Rash diagnosis and treatment
P6	Female	59	Cauc	Cataract surgery
P7	Male	25	Cauc	Torso itchy when warm
P8	Male	46	Cauc	Medication's long-term side effects
P9	Female	36	AA	Long-lasting migraines and the causes
P10	Male	40	Cauc	Essential thrombocytosis
P11	Male	47	Cauc	Reverse or lessen tinnitus
P12	Male	31	Asian / Pac Is	Stress-induced dandruff
P13	Female	68	Asian / Pac Is	Blood pressure and sleep issues
P14	Male	65	Cauc	Natural blood pressure supplement
P15	Male	54	Asian / Pac Is	Rib pain
P16	Male	57	Asian / Pac Is	Daily headaches
P17	Female	33	Cauc	Child's fussiness and infertility concerns
P18	Female	41	Cauc	Perimenopause symptoms and Hormone replacement therapy
P19	Female	56	Hispanic	Osteoporosis
P20	Male	29	AA	Swelling causes
P21	Female	49	Cauc	Pain and sensitivity in leg
P22	Male	37	Cauc	Eosinophil Esophagitis: Swallowing issues
P23	Female	25	Hispanic	Frequent, sharp abdominal pain
P24	Male	38	Hispanic	Knee issues: clicking, locking, pain
P25	Male	61	AA	Suspected kidney stones and prevention tips
P26	Female	25	Cauc	Vertigo
P27	Male	45	Cauc	Lower back pain cause
P28	Female	60	Cauc	Knee pain, buckling, and exercise
P29	Female	53	Cauc	Cyst symptoms reoccurring after surgery
P30	Male	48	Cauc	Seed oils and gout flare-ups
P31	Female	75	Cauc	Acid reflux: symptoms and remedies
P32	Female	33	Cauc	Ovarian cysts
P33	Male	50	Cauc	Raynaud's syndrome and discolored feet

Table 2. Study 1-3 Participant Demographic Information.

prompt to ask at least 4 but no more than 10 clarifying questions before providing an answer. The prototypes in Study 1 were specific to cause-of-symptom-intent (COSI) questions, so for this study we recruited participants who specifically reported having a question involving understanding symptoms. Study 2 compared a tuned Wayfinding AI (that was designed to handle both COSI and non-COSI questions) to the base AI (without tuning). For this study, we recruited participants with a wider range of questions, including both cause-of-symptom as well as other intents (e.g., treatment or management for a known condition). Study 3 used a further refined Wayfinding AI, and compared two alternative UI arrangements, in which answers to questions were either placed in-line and above context-seeking questions (1-column layout) or in a separate side panel (2-column layout). For Study 3, we again focused on COSI questions, in order to focus our feedback on how the UI layouts facilitated context seeking for this common use case. The different UI layouts are illustrated in Figure 1.

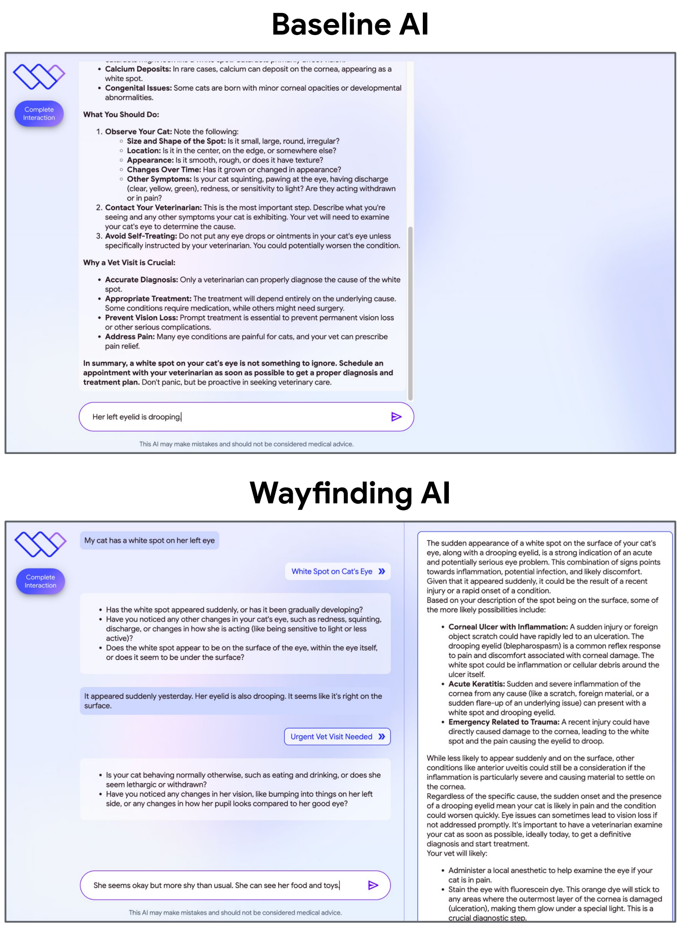

Fig. 1. Screenshots of the two AI chatbot designs used in Studies 3 and 4. Top shows a baseline, 1-column AI design. Bottom shows a 2-column design in which the current answer is shown in the right column, while clarifying questions are displayed underneath user prompts on the left.

3.4 Quantitative study (Study 4)

We assessed user satisfaction and preference between Wayfinding AI and Baseline AI in a quantitative survey-based study (Figure 2). We recruited US-based participants via a third party platform (Qualtrics). All participants were 21 years and older, were not health care professionals, and had a health-related question for which they were willing to interact with an AI. Mental-health-related queries were excluded from the study scope.

After answering standard demographic questions, participants were instructed to login to each AI using anonymous accounts set up specifically for this study; to have a conversation spending at least 3 minutes on their question; and

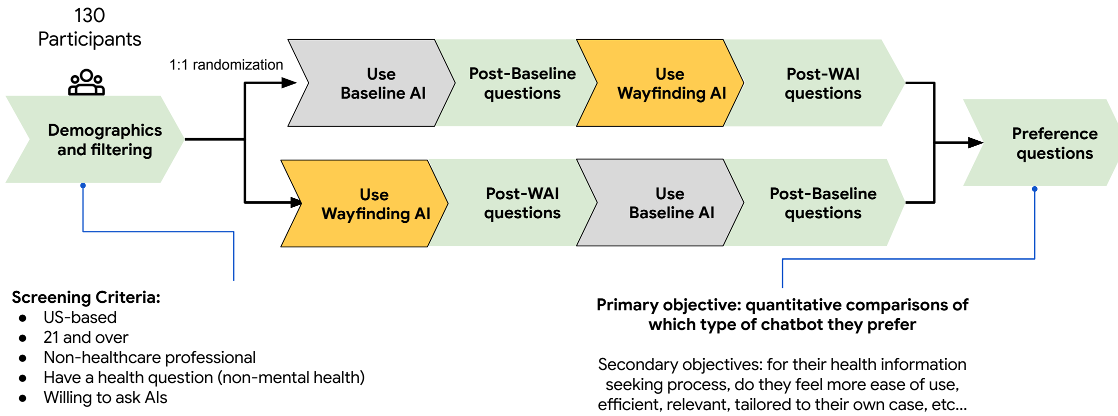

Fig. 2. Quantitative study design (Study 4). All participants used both AIs used in the study (Wayfinding AI and Baseline AI) to better understand their own health-related questions; but were randomly assigned in the order with which to use them. After using each AI, participants answered questions about their satisfaction with the experience; at the end of the survey, subjects explicitly compared the two AIs and expressed which they would prefer along a range of axes.

then to resume the survey. After interacting with each AI, participants answered questions about their satisfaction with the experience along 6 dimensions: Helpfulness, relevance of questions asked, tailoring to their situation, goal understanding, ease of use, and efficiency of getting useful information; and could provide open feedback about what they learned, as well as the option to upload their conversation with the AI. Sharing the conversation was not required to complete the survey. At the end of the study, participants were prompted to explicitly compare the two AIs and indicate which they would prefer in terms of each of the six dimensions above, and answered an additional question: "For a future topic, would you prefer the first or the second AI?". The order of AI exposure (Baseline AI first vs Wayfinding AI first) was randomized across participants.

3.5 Analysis

3.5.1 Qualitative study analysis. All qualitative interview sessions (Studies 1-3) were recorded. Recordings were transcribed using a text-to-speech medical transcription model. Transcripts were analyzed using reflexive thematic analysis [11]. In addition, transcripts of conversations with each AI were analyzed to understand turn structure. Conversations were broken into a set of alternating user and AI turns; user turns were manually coded to determine whether responses directly answered clarifying questions asked by each AI.

3.5.2 Quantitative study analysis. Quantitative data analyses were performed using Python 3.0 and the SciPy package. Confidence intervals on measurements were computed using bootstrapping with 1,000 iterations. Paired comparisons on survey responses between AIs used the Wilcoxon signed-rank test for ordinal responses, and paired two-sided T tests for interval responses. Significance on comparison tests used Bonferroni correction for multiple comparisons across different aspects of user satisfaction and/or preference.

Participants in the quantitative study could optionally upload their conversations with each AI along with their survey responses. Conversations were screened for personally-identifiable information by Qualtrics before sharing with researchers. A total of 119 conversations for the Wayfinding AI and 112 for the Baseline AI were shared. Each conversation was broken into turns where user prompts and AI responses were analyzed. We prompted Gemini 2.5 Pro [12] to classify user prompts and AI responses in multiple ways for analysis. Initial user prompts were classified as reflecting cause-of-symptom intent (COSI; e.g. "a cough that won't go away") or non-cause-of-symptom intent (NCOSI; e.g. "I am curious to know about keto"), to understand how specific the benefits of the Wayfinding AI might be to COSI. All user prompts were also classified in one of 11 turn types, in order to understand conversational dynamics. The classification scheme was derived by internal review and discussion among researchers. Finally, AI responses were classified as containing clarifying questions (e.g. "Have you ever had chickenpox or the shingles vaccine previously?"), open-ended questions ("Could you tell me a little about what you hope to achieve with your workout plan?") or no questions. All prompts used to classify conversational turns are included in the Appendix.

4 Results

4.1 Study 1: Perceived value of context-seeking

Participants in Study 1 consistently found value in an AI that delayed answers for 4 or more turns to collect context around questions. 8 of the 10 participants stated that they preferred to use a prototype that only asked clarifying questions for 4 turns before providing an answer (AI 1b) over a prototype that provided immediate answers before context-seeking (AI 1a). Participants called out the value of context-seeking in establishing a conversational interaction that felt more human (P2: "I feel good. I feel like I'm having real conversations with someone, as opposed to me trying to figure out the level of information that I need to give ... that's one of my main issues, sometimes." P7: "This one feels more like the way it would work if you talk to a doctor. Conversational ... it does make make me feel a little more confident that it wants to know more before jumping right into an answer.") Conversations took between 2 to 6 turns for AI 1a, and between 4 and 16 turns for AI 1b. 9 of 10 participants directly answered one or more clarifying questions from 1a, and all 10 answered clarifying questions from 1b.

While participants found context-seeking useful for understanding possible matching conditions for symptoms, they also expressed other needs related to their health questions that the initial Wayfinding AI prototypes did not serve as well. Many participants wanted to understand management options for conditions; since the initial AIs in this study focused on cause-of-symptom intent, it took more prompting to get useful information about management. Therefore we sought to refine our Wayfinding AI to one that could more flexibly provide context-seeking for both COSI and non-COSI needs. Because non-COSI needs may inherently involve less zooming-in, we opted to focus on testing Wayfinding AIs that surfaced immediate answers in addition to context-seeking.

4.2 Study 2: Tension between longer answers and context-seeking

Study 2 involved 11 participants, who had both COSI and non-COSI questions. Participants found AI 2a (with context-seeking behavior) to be helpful for both use cases. Overall, 6 participants preferred 2a, 3 preferred 2b, and 1 had no preference. Participants who preferred AI 2b cited its context-seeking behavior, which helped them arrive at more relevant results. Participants preferring the baseline AI generally did so because they expressed that their needs were different and less likely to require "zooming in", and they found the baseline AI responses to be comprehensive. These participants tended to have short conversations, with only 1 or 2 turns.

We noted an unexpected behavior in Study 2 compared to Study 1: only 1 of 10 participants in Study 2 directly answered clarifying questions asked by AI 2a (whereas all or nearly all participants did so for 1a and 1b). More often, participants would either pivot to related but distinct questions (e.g. P12: "What is the best diet for healthy scalp and hair?" → "Does lack of sleep lead to vitamin deficiencies?"); proactively zoom in to learn more details of a suggested topic (e.g. P18: "what are recommendations for treatment of perimenopause?" → "What are the risks and benefits of HRT in perimenopause?") or correct the model when its questions seemed to have an inaccurate premise (e.g. P19: "I'm a 56-year-old woman. Be more specific for what I need to do").

When prompted, participants indicated that this was because they did not always notice the questions, or perceive them as essential for getting relevant answers. (P11: "They're just questions for me to ask myself and to think about… It just ends with some rhetorical questions to consider.") AI 2b tended to produce long answers (longer than 1a or 2a), and the clarifying questions were not as consistently placed, sometimes appearing at the end of an answer, but sometimes in the middle. Participants were more likely to report scanning these longer answers for relevant information, whereas participants in Study 1 more often read the whole answers. However, when probed about answer length, participants tended to report that they valued the longer and more comprehensive answers. (P12: "I do quite like this ... it looks like it's trying more to not just give standard answers.")

Due to the dual feedback of (1) desire for immediate, detailed answers and (2) challenges with consistent context-seeking, we explored UI approaches to provide affordances for both.

4.3 Study 3: Separating answers from context-seeking

In Study 3, 12 participants used a refined Wayfinding AI that surfaced clarifying questions consistently at the end of answers, presented in two different user interface (UI) layouts. AI 3a placed answers directly above clarifying questions in the main column, while AI 3b placed them on the right-hand side in a two-column format. AI 3b had two variants, where the initial answer was either collapsed or expanded by default. The answer could be expanded by clicking on a clickable button (whose text provided a short summary of the response) in the main column. Participants would test one of the variants randomly, but be shown the second variant to get feedback on their preferences for expanded vs. collapsed behavior, as well as between 3a and 3b.

Participants were split on preference between 1-column and 2-column layouts. 6 preferred 3b, 4 preferred 3a, and 1 had no preference. Participants who preferred 3b cited how separating out clarifying questions encouraged a more naturalistic conversation (P27: "These are logical questions. They're on topic. And I could see how they would help the AI learn more about my problem to give me better responses.") However, participants uniformly preferred the expanded variant, with some participants not initially recognizing that the answers could be expanded. Participants who preferred the 1-column view cited the simpler mental model associated with it (P31: "Right off, I like this version better than the side one ... You want to ask a question, then you want the answer down below. It's just less confusing.").

4.4 Study 4: Wayfinding AI preferred over baseline AI

Based on the qualitative studies, we developed AI 4a, a Wayfinding AI that provided answers to health questions as well as context-seeking in separate columns (with answers always visible in an expanded state). We compared this Wayfinding AI to a baseline AI (4b) in a quantitative study with US-based participants, who tested both models on their own health questions.

Of the 130 participants, 59% were 50 years of age or younger and 44% identified as female. The participant sample included various racial/ethnic groups (8% Asian, 10% Black or African American, 12% Hispanic, Latino or Spanish Origin, 72% White). Slightly more than half (61%) of participants had earned a Bachelor's degree or a higher qualification. The majority (73%) reported feeling extremely confident using mobile phone apps by themselves, and a smaller portion

(61%) reported the same level of confidence filling out medical forms by themselves. Slightly less than half (43%) said they used AI chatbots at least several times a week, and 15% reported never having used an AI chatbot before. A detailed breakdown of participant characteristics is provided in Table S3.

When asked about their satisfaction with each app separately, participants reported higher satisfaction with the Wayfinding AI on the dimensions of helpfulness, relevance of questions asked, tailoring of responses to the participants' situations, understanding of user goals, and efficiency of getting an answer (Figure 3).

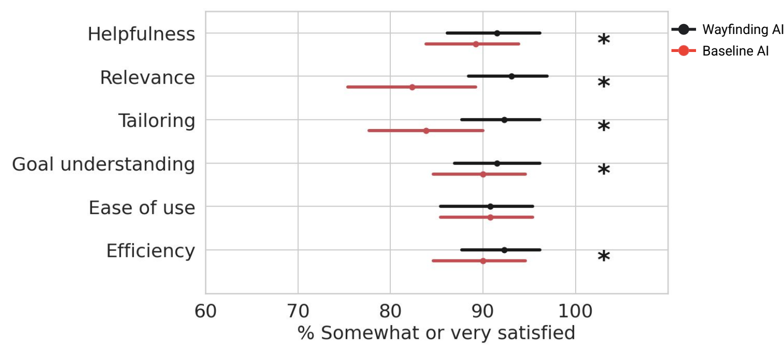

Fig. 3. Summary of satisfaction measures for Wayfinding AI and Baseline AI. Values reflect the percent of respondents who report being somewhat or very satisfied with each agent. Error bars represent 95% confidence intervals determined by bootstrap. Asterisks indicate significant difference by a Wilcoxon test, at $p < 0.05$, with Bonferroni correction for multiple comparisons.

When asked to explicitly compare the two AIs they tested, participants significantly preferred the Wayfinding AI on the dimensions of helpfulness, relevance of the questions asked, degree of interactions tailored to the participant's situation, and goal understanding (Figure 4). Responses were neutral for ease of use, efficiency, help in achieving goal and which AI the user would prefer for a future search.

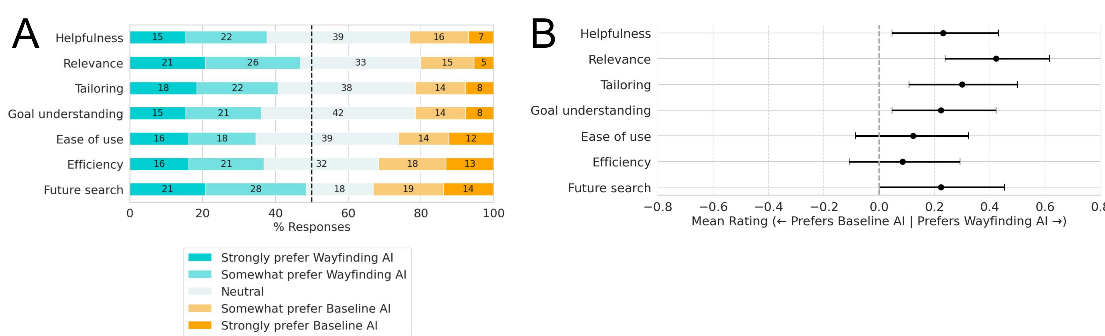

Fig. 4. Summary of side-by-side comparison between Wayfinding AI and Baseline AI. **A**, distributions of survey responses for each preference question; **B**, Summary of preference distributions. Values reflect the arithmetic mean of Likert scale responses on a scale of [-2, +2]; negative values indicate preference for Baseline AI, positive values indicate preference for Wayfinding AI. Error bars represent 95% confidence intervals determined by bootstrap.

4.4.1 Different conversation dynamics with the Wayfinding AI. We observed marked differences in the structure of conversations between the two AIs in Study 4. The initial responses of Baseline AI rarely asked questions to users to proactively gain context on the users' concerns; and when it did it sometimes asked more open-ended questions. (E.g., when prompted for help with workout planning, it completed its response with "Let me know what you're aiming for, and we can build from there!") By comparison, the Wayfinding AI asked specific clarifying questions for 96% of conversations (Table 3). For instance, a participant hoping to understand a recent test result was asked "Have you had any recent infections, surgeries or injuries?" (Table S1).

Questions asked in first agent response	Wayfinding AI	Baseline AI
No questions	4.2%	90.1%
Open-ended question(s)	8.4%	1.8%
Clarifying question(s)	87.7%	8.0%

Table 3. Breakdown of first responses by each AI in Study 4 based on the type of questions it asked of users. The visualization for question types for 13 turns are presented in Appendix Figure S3.

Compared to the Baseline AI, conversations with the Wayfinding AI tended to be longer (Mean 4.51 turns for Wayfinding AI vs. 3.69 turns for Baseline AI; $p = 0.025$, paired T test). This difference was driven by substantially longer conversations for COSI questions ($p < 0.001$), while non-COSI questions were not significantly different in length ($p = 0.105$; Table 4).

	Number of conversation turns, Wayfinding AI	**Number of conversation turns, Baseline AI**	**N, Wayfinding AI**	**N, Baseline AI**	**Difference in conversation length, p (Paired T test)**
All data	4.51 [3.99, 5.03]	3.69 [3.22, 4.15]	119	112	0.0252
Cause-of-symptom-intent topic (COSI)	4.96 [4.32, 5.61]	3.29 [2.79, 3.78]	84	73	0.0009
Non-cause-of-symptom-intent topic (Non-COSI)	3.43 [2.67, 4.19]	4.44 [3.52, 5.35]	35	39	0.1050

Table 4. Summary of conversation lengths, for all conversations, conversations with a cause-of-symptom intent (COSI), and conversations with a non-cause-of-symptom intent (Non-COSI). Metrics include the point estimate followed by the 95% confidence interval determined by bootstrap. Note that N reflects the number of participants (out of 130 total) who opted to upload a copy of their conversation after interacting with the Wayfinding AI and Baseline AI respectively. Conversation upload was an optional step in the study flow.

We classified each user prompt across turns in their conversations (details in Appendix). The patterns of conversations were sharply different between AIs (Figure 5). With the Wayfinding AI, users engaged in a much higher rate of responding to AI responses with clarifying information (e.g. "I enjoy walking" when asked what type of activities the user enjoyed, for the question of how often to work out; or "I have painful itching" when asked whether a rash has any symptoms.). Conversely, when using Baseline AI, users were much more likely to elaborate, meaning they volunteered information that was not directly in response to the AI response (e.g. "what sleep medications are non habit forming?" when asking about insomnia treatments and being provided a set of recommendations); and were relatively more likely to pivot to different topics (e.g. "Can you reverse type 2 diabetes?" after initially asking "What is the prevnar vaccine? What does it protect you from?") and/or end the conversation earlier.

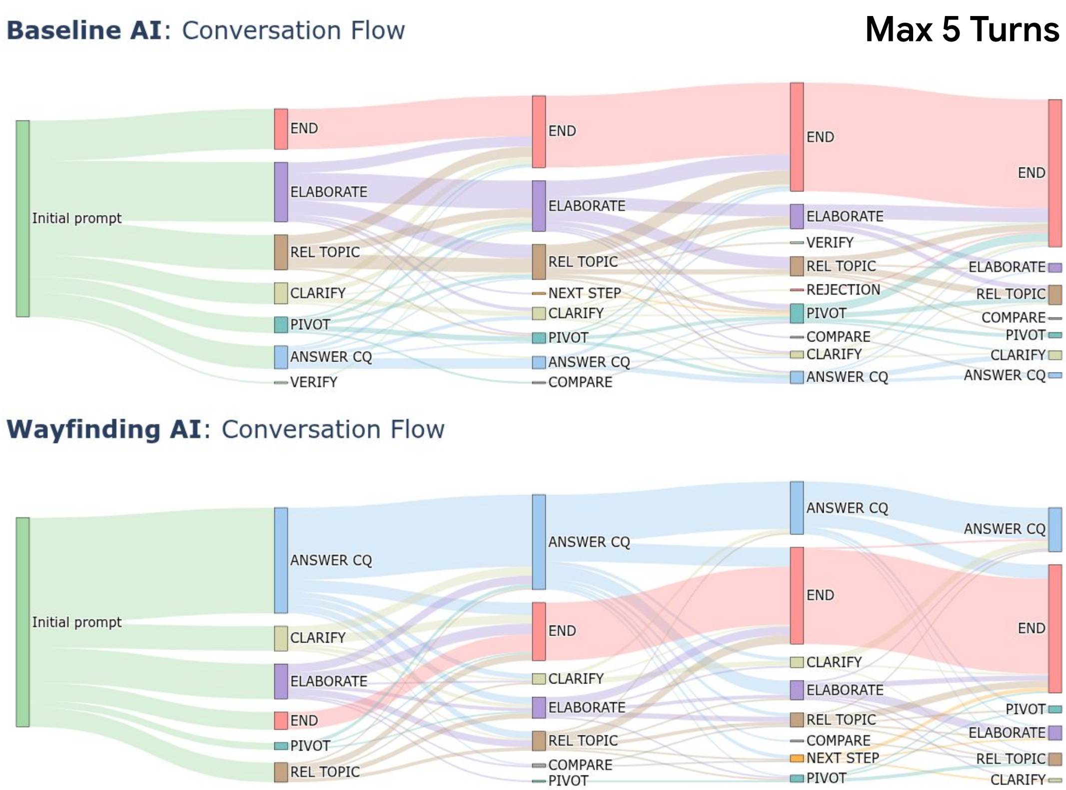

Fig. 5. Sankey flow diagrams illustrating distributions of user prompt types across conversations in Study 4. Top, summary of all conversations with Baseline AI. Bottom, summary of all conversations with Wayfinding AI. Nodes indicate different user prompt types. Thickness of lines connecting two nodes indicates relative proportion of conversations with the first prompt type followed by the second prompt type. The first 5 turns are shown. The abbreviations describing each turn type are described in more detail in the Appendix. Abbreviations: ANSWER_CQ: Answering clarifying questions asked by the AI; ELABORATION: Request for AI to elaborate on an aspect of its response; REL_TOPIC: Related topic exploration; CLARIFY: user provides refinement or clarification of their need without prompting; PIVOT: User pivots to a new topic; COMPARE: user asks AI to compare two entities related to the question; NEXT_STEP: User asks about a task-related next step; VERIFY: User expresses uncertainty or seeks confirmation of information provided. Details on how labels were applied to conversations are in the Appendix.

5 Discussion

In this work, we provide qualitative and quantitative evidence across four studies demonstrating that people value AIs that explicitly engage in context-seeking to better understand their health questions. Such “Wayfinding” AIs result in longer conversations with more turns, that are preferred over conversations with standard AIs, and are perceived as being more relevant to the person’s individual circumstances.

These findings suggest that design considerations in AI behavior and interface can dramatically impact the quality of health information obtained. This also has implications for how AIs may be developed and assessed, and how this technology may impact access to and quality of health information.

5.1 AI design can affect conversational dynamics

We show that design choices in AIs can result in significant user-perceived differences in the quality of conversations. Developing an AI that proactively engages in context-seeking results in more relevant information being provided, which is overall viewed by participants as more helpful and tailored to their goals and circumstances (Figure 4). Conversations are also longer and more detailed (Table 4 and Figure 5).

Design choices include several aspects of AI interface and behavior, which may interact. AI response length can often be long, and length is known to influence human judgments of answer quality [45, 47, 65, 72]. We saw in Study 2 that this length also interacts with context-seeking: Participants mostly ignored the clarifying questions the Wayfinding AI from that study asked, in large part because of long responses that did not consistently place clarifying questions in the same location. AIs that have been further trained to ask questions in a consistent format, and UI affordances to highlight clarifying questions, can restore user engagement with context-seeking (as seen in Studies 3 and 4).

Proactive context-seeking results in behavior that is more in line with human experts, who when conversing with laypeople on topics will often "zoom in" to better understand details [23] often with targeted clarifying questions [8, 41, 42]. Conversational patterns between laypeople and either an LLM or a clinician presented as an LLM have been shown to systematically differ in terms of how information is provided, and non-need related communication such as empathy [42]. Further development of LLMs may aid in other conversational dynamics that human experts tend to engage in, such as "zooming out" to understand wider goals and "reframing" when a layperson is operating on an unsupported premise [23].

5.2 Measuring the value of context-seeking

Our work focuses on laypeople using AI to understand their own real-world health questions and sharing their impressions of their experience. This approach is distinct from most current approaches for assessing AI performance on health topics. Existing benchmarks focus primarily on closed-form question sets, such as those given to medical students [53, 62]. Evaluations of model performance also assess how well they answer open-ended questions, including questions commonly asked online [7, 56, 63]; but these have tended to focus on single-turn responses [26]. More recent work has started exploring multi-turn evaluation, either by simulating interactive conversations [43] or providing clinical review of responses for one turn in a conversation [1].

These automated approaches to measuring health information may overestimate the ability of models to help users, by assuming they know what relevant context to share [5]. Work assessing context-seeking in LLMs found that directly prompting LLMs to ask clarifying questions unexpectedly degraded performance [43], suggesting both that the development of effective context-seeking is non-trivial, and that existing benchmarks may not fully assess the value of perceived relevance and tailoring to users. More broadly, most evaluation approaches rely on assessments of interactions by human raters, by rubrics designed by human raters, or by automated evaluation methods. The extent to which these approaches result in better alignment with the preferences of the people asking the questions (versus a rater assessing the conversation after-the-fact) is unclear.

5.3 Limitations & Future Work

This work has some limitations. Our study focused on text-only AIs; many real-world health questions may be multi-modal, involving images, audio, or video. The set of participants was also limited to the US and these findings may not reflect sentiments from people outside the US. Participants conversed with AIs twice, as part of a study; this behavior may not precisely match behavior if participants were using either AI when they first had their question. The AIs tested here varied both in context-seeking behavior and in UI presentation (motivated by the findings of Studies 2 and 3 which suggested the UI may aid in context seeking by separating out clarifying questions). Our findings therefore reflect the aggregate effects of both UI and model behavior. Future research will be needed to better understand the separate contributions of UI and model behavior, and investigate context-seeking in a multi-modal context, and with a broader sample of participants.

6 Conclusions

AI technology, in distilling a large knowledge corpus and tailoring it to individual users, may greatly broaden layperson access to expertise. Within the health domain, this has the promise to expand access to high-quality medical information. Yet it also poses many risks, including that people will be misinformed, overwhelmed, or confused. The value people extract from this technology, and their ability to apply it to their own circumstances, will depend acutely on a range of design considerations. There is an urgency in elucidating these principles, and applying it to develop accurate and accessible tools. Our work contributes to this effort by demonstrating the impact of context-seeking on conversational health AIs. By encouraging proactive collection of information, the pattern and quality of interactions with an AI can vary significantly. The resulting conversations are viewed as more helpful, relevant, and tailored to each individual's circumstances. Future work is needed to understand how these different patterns of information-seeking may inform decisions around seeking care, and ultimately whether they can promote better health outcomes.

Acknowledgments

We thank Laura Vardoulakis, Tiffany Guo, and Meredith Ringel Morris for their thoughtful reviews of the manuscript.

References

[1] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775* (2025).

[2] Laima Augustaitis, Leland A. Merrill, Kristi E Gamarel, and Oliver L. Haimson. 2021. Online Transgender Health Information Seeking: Facilitators, Barriers, and Future Directions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 205, 14 pages. doi:10.1145/3411764.3445091

[3] Julie Ayre, Erin Cvejic, and Kirsten J McCaffery. 2025. Use of ChatGPT to obtain health information in Australia, 2024: insights from a nationally representative survey. *Medical Journal of Australia* 222, 4 (2025), 210–212.

[4] Boris Babic, Sara Gerke, Theodoros Evgeniou, and I Glenn Cohen. 2021. Direct-to-consumer medical machine learning and artificial intelligence applications. *Nature Machine Intelligence* 3, 4 (2021), 283–287.

[5] Andrew M Bean, Rebecca Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera, Sara Hincapié Monsalve, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, et al. 2025. Clinical knowledge in LLMs does not translate to human interactions. *arXiv preprint arXiv:2504.18919* (2025).

[6] Shirley Ann Becker. 2004. A study of web usability for older adults seeking online health resources. *ACM Trans. Comput.-Hum. Interact.* 11, 4 (Dec. 2004), 387–406. doi:10.1145/1035575.1035578

[7] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. 2025. MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. *arXiv preprint arXiv:2505.23802* (2025).

[8] Rainer Bromme, Regina Jucks, and Anne Runde. 2005. Barriers and biases in computer-mediated expert-layperson-communication: An overview and insights into the field of medical advice. *Barriers and biases in computer-mediated knowledge communication: And how they may be overcome* (2005), 89–118.

[9] Ka I Chan, Siying Hu, Yuntao Wang, Xuhai Xu, Zhicong Lu, and Yuanchun Shi. 2025. The Odyssey Journey: Top-Tier Medical Resource Seeking for Specialized Disorder in China. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1048, 18 pages. doi:10.1145/3706598.3713315

[10] I-Chiu Chang, Yi-Syuan Shih, and Kuang-Ming Kuo. 2022. Why would you use medical chatbots? interview and survey. *International Journal of Medical Informatics* 165 (2022), 104827.

[11] Victoria Clarke and Virginia Braun. 2014. Thematic analysis. In *Encyclopedia of critical psychology*. Springer, 1947–1952.

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[13] Munmun De Choudhury, Meredith Ringel Morris, and Ryen W. White. 2014. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1365–1376. doi:10.1145/2556288.2557214

[14] Julian De Freitas and I Glenn Cohen. 2024. The health risks of generative AI-based wellness apps. *Nature medicine* 30, 5 (2024), 1269–1275.

[15] Henrik HJ Detjen, Lars Densky, Niklas von Kalckreuth, and Marvin Kopka. 2025. Who is Trusted for a Second Opinion? Comparing Collective Advice from a Medical AI and Physicians in Biopsy Decisions After Mammography Screening. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.

[16] Umama Dewan, Cora Sula, and Nora Mcdonald. 2024. Teen Reproductive Health Information Seeking and Sharing Post-Roe. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 814, 12 pages. doi:10.1145/3613904.3641934

[17] Cinzia Di Novi, Matija Kovacic, and Cristina Elisa Orso. 2024. Online health information seeking behavior, healthcare access, and health status during exceptional times. *Journal of Economic Behavior & Organization* 220 (2024), 675–690.

[18] Audrey Eichenberger, Stephen Thielke, and Adam Van Buskirk. 2025. A Case of Bromism Influenced by Use of Artificial Intelligence. *Annals of Internal Medicine: Clinical Cases* 4, 8 (2025), e241260.

[19] George Libman Engel and William L Morgan. 1973. Interviewing the patient. *(No Title)* (1973).

[20] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature medicine* 25, 1 (2019), 24–29.

[21] Xiangmin Fan, Daren Chao, Zhan Zhang, Dakuo Wang, Xiaohua Li, and Feng Tian. 2021. Utilization of self-diagnosis health chatbots in real-world settings: case study. *Journal of medical Internet research* 23, 1 (2021), e19928.

[22] Moshe Y Flugelman. 2021. History-taking revisited: Simple techniques to foster patient collaboration, improve data attainment, and establish trust with the patient. *GMS Journal for Medical Education* 38, 6 (2021), Doc109.

[23] Beverly Freeman, Roma Ruparel, and Laura M Vardoulakis. 2025. Zoom in, Zoom out, Reframe: Domain Experts' Strategies for Addressing Non-Experts' Complex Questions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[24] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 31.

[25] Rachel S Goodman, J Randall Patrinely, Cosby A Stone, Eli Zimmerman, Rebecca R Donald, Sam S Chang, Sean T Berkowitz, Avni P Finn, Eiman Jahangir, Elizabeth A Scoville, et al. 2023. Accuracy and reliability of chatbot responses to physician questions. *JAMA network open* 6, 10 (2023), e2336483–e2336483.

[26] Yuexing Hao, Jason Holmes, Jared Hobson, Alexandra Bennett, Elizabeth L McKone, Daniel K Ebner, David M Routman, Satomi Shiraishi, Samir H Patel, Nathan Y Yu, et al. 2025. Retrospective Comparative Analysis of Prostate Cancer In-Basket Messages: Responses From Closed-Domain Large Language Models Versus Clinical Teams. *Mayo Clinic Proceedings: Digital Health* 3, 1 (2025), 100198.

[27] Yuexing Hao, Zeyu Liu, Robert N Riter, and Saleh Kalantari. 2024. Advancing Patient-Centered Shared Decision-Making with AI Systems for Older Adult Cancer Patients. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[28] Yuexing Hao, Zhiwen Qiu, Jason Holmes, Corinna E Löckenhoff, Wei Liu, Marzyeh Ghassemi, and Saleh Kalantari. 2025. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *npj Digital Medicine* 8, 1 (2025), 450.

[29] Christina N. Harrington and Lisa Egede. 2023. Trust, Comfort and Relatability: Understanding Black Older Adults' Perceptions of Chatbot Design for Health Information Seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 120, 18 pages. doi:10.1145/3544548.3580719

[30] Christina N. Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*

(New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 604, 15 pages. doi:10.1145/3491102.3501995

[31] Courtney Hofmann and Holly Gilmer. 2025. *Partners in Diagnosis: ChatGPT, a Mother's Intuition, and a Doctor's Expertise with Courtney Hofmann.* https://ai-podcast.nejm.org/e/partners-in-diagnosis-chatgpt-a-mother-s-intuition-and-a-doctor-s-expertise-with-courtney-hofman-and-dr-holly-gilmer/ NEJM AI Podcast.

[32] Toshish Jawale, Chaitanya Animesh, Sekhar Vallath, Kartik Talamadupula, and Larry Heck. 2024. Are human conversations special? a large language model perspective. *arXiv preprint arXiv:2403.05045* (2024).

[33] Xiaoyun Jia, Yan Pang, and Liangni Sally Liu. 2021. Online health information seeking behavior: a systematic review. In *Healthcare*, Vol. 9. MDPI, 1740.

[34] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

[35] Justin D Krogue, Rory Sayres, Jay Hartford, Amit Talreja, Pinal Bavishi, Natalie Salaets, Kimberley Raiford, Jay Nayar, Rajan Patel, Yossi Matias, et al. 2024. Searching for Dermatology Information Online using Images vs Text: a Randomized Study. *medRxiv* (2024), 2024–10.

[36] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120* (2025).

[37] Sylvie D Lambert and Carmen G Loiselle. 2007. Health information—seeking behavior. *Qualitative health research* 17, 8 (2007), 1006–1019.

[38] David M Levine and Ateev Mehrotra. 2021. Assessment of diagnosis and triage in validated case vignettes among nonphysicians before and after internet search. *JAMA network open* 4, 3 (2021), e213287–e213287.

[39] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N Truong, and Alex Mariakakis. 2024. Beyond the Waiting Room: Patient's Perspectives on the Conversational Nuances of Pre-Consultation Chatbots. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

[40] Brenna Li, Tetyana Skoropad, Puneet Seth, Mohit Jain, Khai Truong, and Alex Mariakakis. 2023. Constraints and Workarounds to Support Clinical Consultations in Synchronous Text-based Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[41] Brenna Li, Saba Tauseef, Khai N Truong, and Alex Mariakakis. 2025. A Comparative Analysis of Information Gathering by Chatbots, Questionnaires, and Humans in Clinical Pre-Consultation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[42] Brenna Li, Amy Wang, Patricia Strachan, Julie Anne Séguin, Sami Lachgar, Karyn C Schroeder, Mathias S Fleck, Renee Wong, Alan Karthikesalingam, Vivek Natarajan, et al. 2024. Conversational AI in health: Design considerations from a Wizard-of-Oz dermatology case study with users, clinicians and a medical LLM. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–10.

[43] Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems* 37 (2024), 28858–28888.

[44] Shuyue Stella Li, Jimin Mun, Faeze Brahman, Pedram Hosseini, Bryceton G. Thomas, Jessica M. Sin, Bing Ren, Jonathan S. Ilgen, Yulia Tsvetkov, and Maarten Sap. 2025. ALFA: Aligning LLMs to Ask Good Questions A Case Study in Clinical Reasoning. arXiv:2502.14860 [cs.CL] https://arxiv.org/abs/2502.14860

[45] Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. 2024. Does style matter? disentangling style and substance in chatbot arena. *LMSYS Blog* (2024).

[46] Q. Vera Liao and Wai-Tat Fu. 2014. Age differences in credibility judgments of online health information. *ACM Trans. Comput.-Hum. Interact.* 21, 1, Article 2 (Feb. 2014), 23 pages. doi:10.1145/2534410

[47] Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, et al. 2022. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *arXiv preprint arXiv:2212.07981* (2022).

[48] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2025. Towards accurate differential diagnosis with large language models. *Nature* (2025), 1–7.

[49] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[50] Tamir Mendel, Nina Singh, Devin M Mann, Batia Wiesenfeld, and Oded Nov. 2025. Laypeople's Use of and Attitudes Toward Large Language Models and Search Engines for Health Queries: Survey Study. *J Med Internet Res* 27 (13 Feb 2025), e64290. doi:10.2196/64290

[51] Ashlee Milton, Juan F. Maestre, Abhishek Roy, Rebecca Umbach, and Stevie Chancellor. 2024. Seeking in Cycles: How Users Leverage Personal Information Ecosystems to Find Mental Health Information. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 540, 16 pages. doi:10.1145/3613904.3641894

[52] Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital health* 5 (2019), 2055207619871808.

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

[54] Simmi Oberoi, Neha Chaudhary, Siriesha Patnaik, and Amarjit Singh. 2016. Understanding health seeking behavior. *Journal of family medicine and primary care* 5, 2 (2016), 463–464.

[55] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[56] Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine* 30, 12 (2024), 3590–3600.

[57] Laura Reiley. 2025. What My Daughter Told ChatGPT Before She Took Her Life. *The New York Times* (18 Aug. 2025). https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html

[58] Khaled Saab, Jan Freyberg, Chunjong Park, Tim Strother, Yong Cheng, Wei-Hung Weng, David GT Barrett, David Stutz, Nenad Tomasev, Anil Palepu, et al. 2025. Advancing Conversational Diagnostic AI with Multimodal Reasoning. *arXiv preprint arXiv:2505.04653* (2025).

[59] Rory Sayres, Anna Devon-Sand, Mike Schaekermann, Margaret Ann Smith, Patricia Strachan, Grace Hong, Trinh Nguyen, Doris Wong, Naama Hammel, Dawn Siegel, et al. 2025. Navigating Skin Concerns with AI: A Human-Centered Investigation of a Dermatology App in a Diverse Community. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[60] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, et al. 2019. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* 126, 4 (2019), 552–564.

[61] Yeganeh Shahsavar and Avishek Choudhury. 2023. The role of AI chatbots in healthcare: a study on user intentions to utilize ChatGPT for self-diagnosis. *JMIR Preprints* (2023).

[62] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[63] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine* 31, 3 (2025), 943–950.

[64] Anna M Smak Gregoor, Tobias E Sangers, Lytske J Bakker, Loes Hollestein, Carin A Uyl-de Groot, Tamar Nijsten, and Marlies Wakkee. 2023. An artificial intelligence based app for skin cancer detection evaluated in a population based setting. *NPJ digital medicine* 6, 1 (2023), 90.

[65] Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. 21–29.

[66] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine* 3, 1 (2020), 17.

[67] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature medicine* 26, 8 (2020), 1229–1234.

[68] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature* (2025), 1–9.

[69] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–6.

[70] Xun Wang and Robin A Cohen. 2023. *Health information technology use among adults: United States, July–December 2022*. US Department of Health and Human Services, Centers for Disease Control and Prevention.

[71] Yuheng Wu, Yujie Dong, Yi Mou, and Ki Joon Kim. 2025. How the Algorithmic Transparency of Search Engines Influences Health Anxiety: The Mediating Effects of Trust in Online Health Information Search. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 307, 10 pages. doi:10.1145/3706598.3713199

[72] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. *arXiv preprint arXiv:2305.18201* (2023).

[73] Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. 2023. Beyond self-diagnosis: how a chatbot-based symptom checker should respond. *ACM Transactions on Computer-Human Interaction* 30, 4 (2023), 1–44.

[74] Hye Sun Yun and Timothy Bickmore. 2025. Framing Health Information: The Impact of Search Methods and Source Types on User Trust and Satisfaction in the Age of LLMs. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 286, 7 pages. doi:10.1145/3706599.3720239

A Sample Chat History

Turn	Speaker	Conversation Context
1	User	> Normal acceptable range of platelets
2	Response	Hello! To help me understand your situation, I need to ask some questions. The more information you provide, the better I can pinpoint the cause of your symptoms. I will also provide my current best guess after each response, along with recommended next steps. ... To help me understand why you're asking about platelet ranges, could you tell me: 1. Have you had a recent blood test that showed your platelet count? If so, what was the result? 2. Are you experiencing any symptoms that might be related to a low or high platelet count (e.g., easy bruising, bleeding, fatigue)? 3. Do you have any known medical conditions or are you taking any medications that could affect your platelet count?
	User	> Yes, platelet count was 642,000.
3	Response	Okay, a platelet count of 642,000 is above the normal range. This is referred to as thrombocytosis. Here are some possible causes, ranked by how well they might explain an isolated high platelet count: ... **Questions to improve my answer:** 1. **Have you had any recent infections, surgeries, or injuries?** 2. **Do you have any known inflammatory conditions (e.g., rheumatoid arthritis, inflammatory bowel disease) or iron deficiency?** 3. **Are you experiencing any unusual bleeding or bruising, or any symptoms suggestive of blood clots?**
	User	> No recent infections, surgeries, or injuries.

Continued on next page

Table S1 – continued from previous page

Turn	Speaker	Conversation Context
4	Response	Okay, that helps narrow things down. The absence of recent infections, surgeries, or injuries makes reactive thrombocytosis less likely but doesn't rule it out entirely... Given your elevated platelet count of 642,000 and the absence of recent triggers, here's an updated list of possible causes: ... **Questions to improve my answer:** 1. **Do you have any known inflammatory conditions (e.g., rheumatoid arthritis, inflammatory bowel disease) or iron deficiency?** 2. **Are you experiencing any unusual bleeding or bruising, or any symptoms suggestive of blood clots?** 3. **Have you had a complete blood count (CBC) done recently, or just the platelet count? If a CBC was done, what were your red blood cell and white blood cell counts?**
	User	> No known inflammatory conditions.
5	Response	Okay, that's helpful. Knowing that you don't have any *known* inflammatory conditions helps to slightly decrease the likelihood of reactive thrombocytosis due to a diagnosed inflammatory disorder... ... **Questions to improve my answer:** 1. **Are you experiencing any unusual bleeding or bruising, or any symptoms suggestive of blood clots?** 2. **Have you had a complete blood count (CBC) done recently, or just the platelet count? If a CBC was done, what were your red blood cell and white blood cell counts?** 3. **Are you taking any medications or supplements (including over-the-counter)?**
	User	> Taking hydroxurea once a day.
6	Chatbot	Okay, that's a very important piece of information. ... **Questions to improve my answer:** 1. **Have you been officially diagnosed with Essential Thrombocythemia (ET) or another Myeloproliferative Neoplasm (MPN)?** 2. **When was your last appointment with your hematologist, and what did they say about your platelet count at that time?** 3. **Are you experiencing any new or worsening symptoms since starting hydroxyurea?**
	User	> At my last appointment with my hematologist, platelet count was 425,000.

Continued on next page

Table S1 – continued from previous page

Turn	Speaker	Conversation Context
7	Chatbot	Okay, that's helpful context. Knowing that your platelet count was 425,000 at your last hematologist appointment, and is now 642,000, suggests a recent upward trend despite being on hydroxyurea. This reinforces the need to investigate the cause of the increase. ... **To provide more targeted information, it would be helpful to know:** 1. **Have you been officially diagnosed with Essential Thrombocythemia (ET) or another Myeloproliferative Neoplasm (MPN)?** 2. **Are you experiencing any new or worsening symptoms since starting hydroxyurea, or any symptoms related to blood clots or bleeding?** 3. **When was your last appointment with your hematologist, besides the one where your platelet count was 425,000? Was a complete blood count performed and were there any other abnormal results?**
	User	> My platelet count was 642,000, now it's gone down to 425,000.

Table S1. **Sample Multi-turn Chat History (P10).** This example demonstrates an interaction with a deferred-answer prototype. The blue highlighting indicates the AI's question that the user's answer directly corresponds to.

B Supplemental Methods

B.1 Agents used throughout the study

Across the 4 studies in this paper, we tested a range of conversational agents that were iteratively updated in terms of both model behavior and user interface. The full set of variants we tested is described in Table S2. For the first study, we also include an illustration of the different conversational dynamics between the Immediate response behavior used by Agent A1 (as well as all other agents in later studies) and the Delayed-answer behavior used by Agent A2.

AI	Brief description	AI Details	Study Examining AI
1a	Context-seeking + Immediate answer	Initial model prompted to aid cause-of-symptom-intent (COSI) questions. Provides set of matching conditions followed by clarifying questions	Study 1 (qualitative)
1b	Context-seeking + Delayed answer	Variant of AI, with system prompt to only ask clarifying questions for at least 4 (but no more than 10) turns, in order to make the set of matching conditions as relevant as possibly	Study 1 (qualitative)
2a	Generalized context-seeking	Refined version of A1, provides initial answer but tuned to address both COSI and NCOSI questions	Study 2 (qualitative)
2b	Baseline model for 2a	Base model for Wayfinding AI 2a, without any additional tuning or system prompts to engage in context-seeking	Study 2 (qualitative)
3a	Wayfinding AI candidate, 1-column	Refined version of 2a, with clarifying questions consistently formatted at end of each prompt and presented underneath user prompts in a single column	Study 3 (qualitative)
3b	Wayfinding AI candidate, 2-column	Refined version of 2a, in a 2 column format: clarifying questions underneath user prompts in left column, answers in right column. Answers are initially collapsed and can be expanded by users selecting expandable chips in left column.	Study 3 (qualitative)
4a	Wayfinding AI candidate	Refined version of 3b, using a 2-column format but where answers are expanded by default and always immediately visible	Study 4 (quantitative)
4b	Baseline model for 4a	Baseline model for Wayfinding AI 4a, without any additional tuning or system prompts to engage in context-seeking	Study 4 (quantitative)

Table S2. Summary of different AI versions tested in this study.

A. Immediate-Answer Prototype

User: Woman horizontal line on abdomen

Assistant Response: Hello! To help me understand your situation, I need to ask some questions. The more information you provide, the better I can pinpoint the cause of your symptoms. I will also provide my current best guess after each response, along with recommended next steps.

Best-matching Possible Causes So Far:

1. **Linea Alba/Nigra:** This is a common condition where a vertical line appears on the abdomen. During pregnancy, it darkens due to hormonal changes and is then called linea nigra.
2. **Surgical Scar:** A horizontal line could be the result of a previous abdominal surgery.
3. **Stretch Marks (Striae):** These can appear as horizontal lines, often due to rapid changes in skin tension, such as during pregnancy or weight fluctuations.

Urgent Medical Attention: None of the information suggests an immediate need for urgent medical attention. However, if the line is accompanied by severe pain, swelling, or other concerning symptoms, seek medical care promptly.

Questions to Improve My Answer:

1. Can you describe the line in more detail? Is it raised, discolored, indented, or flat?
2. Are you pregnant or have you recently been pregnant?
3. Do you have any other symptoms, such as pain, itching, or skin changes in the area around the line?

User: I recently gave birth to my second child; she's about 6 months old now. I've had some pain in my lower back lately.

Assistant Response: Okay, that information is helpful! Knowing that you recently gave birth and are experiencing lower back pain helps narrow down the possibilities.

B. Deferred-Answer Prototype

User: My eyes are watery watching TV in the dark.

Assistant Response: Okay, I understand. Watery eyes while watching TV in the dark can be caused by a few different things. To help me narrow down the possibilities, could you tell me, do your eyes feel dry, gritty, or irritated in addition to being watery?

User: They feel kind of gritty.

Assistant Response: Thanks. Do you notice this grittiness and watering happening at other times, or is it mostly when you're watching TV in the dark?

User: Mostly just then. I do a lot of reading, and it doesn't feel so bad. Maybe a little dry.

Assistant Response: Okay. When you're watching TV in the dark, do you find yourself blinking less often than usual?

User: I guess I hadn't noticed, sorry!

Assistant Response: No problem at all! Do you wear contact lenses?

User: No

Assistant Response: Okay, thanks. Here are some possible causes for your watery, gritty eyes when watching TV in the dark:

Dry Eye Syndrome:

- **Explanation:** Dry eye occurs when your eyes don't produce enough tears or the tears are of poor quality...

Fig. S1. Illustration of differences in agent behavior tested in Study 1. Left, AI 1a provided immediate answers to responses, with context-seeking clarifying questions at the bottom of each response. Right, AI 1b deferred answers for 4 or more conversation turns, instead only asking clarifying questions. During this section of the session, the researcher engaged in a remote moderated usability session with the participant, understanding whether the two behaviors met or differed from their expectations for a conversational AI.

B.2 Conversation classification

We used a large language model (Gemini 2.5 Pro) to classify conversations, as well as individual conversational turns.

B.2.1 Cause-of-symptom intent classification. For identifying cause-of-symptom-intent from non-cause-of-symptom-intent, we used the following prompt:

```
You are an expert health assistant.

Your task is to analyze a conversation between a User and an AI Assistant,
and determine if it reflects an intent related to discovering the cause of
somebody's health symptoms (either the user or someone else).

The conversation will be provided below. User's comments are provided after
the string "USER:", and the Assistant's replies are after the string
"ASSISTANT:". Sometimes, the Assistant response will include markup like
"(RIGHT)", "(LEFT)", or "(CHIP)"; this reflects formatting of the answer
in an interface. Ignore it for determining the intent of the conversation.

Return 'COSI' if the conversation has cause of symptom intent, and 'NCOSI' if
it does not. If the intent can not be reasonably determined, return 'Unsure'.
Do not return any response other than 'COSI', 'NCOSI', or 'Unsure'.

CONVERSATION:
```

B.2.2 User turn type classification. For classifying conversation turn types, we used the following prompt:

```
You are an expert in analyzing conversational user behavior. You will be provided with a
user's previous turn query, assistant's response and the user's subsequent response if the
user responded. Your task is to classify the user's response to the assistant's previous
turn in a multi-turn conversation. Categorize the user response into **one and only one** of
the following categories:

1. RESPONSE_WITH_INFO - User responds with information solicited by the assistant.
    *   Signals: Responding with facts that were asked by the assistant.
    *   Examples: "my height is 5ft", "no I do not currently take any medications", "this
    has been happening for the last 6 months"

2. REJECTION - User indicates the response is incorrect, irrelevant, misleading, or
fundamentally flawed. Includes direct disagreement, challenges to facts, and statements of
misinformation.
    *   Signals: "That's wrong," "This is incorrect," "You're lying," "I don't agree,"
    "That's not what I asked," challenges with evidence, skeptical questions implying
    disbelief if the intent is clearly rejection.
```

    *   Examples: "That's not the right answer." "This is incorrect, it was 1945 not 1946." "You're confusing two different things." "Are you sure that's correct? Because I found X."

3.VERIFICATION_VALIDATION - User expresses uncertainty or seeks confirmation of the information provided, **without asserting it's wrong**. Asks for evidence, justification, or source information to **assess reliability**. This category applies when the user is unsure and seeks to confirm rather than directly disagreeing.

    *   Signals: "Are you sure?", "How do you know?", "Can you provide a source?", "What's the basis for that?" **without directly contradicting the assistant**.
    *   Examples: "Are you sure about that number? Can you show me where you got that from?", "Where did you get that information?", "What makes that source reliable? Is it peer-reviewed?"

4. REFINEMENT_CLARIFICATION - User provides more details to clarify the original query, narrow the scope, or add constraints to improve relevance. The focus is on making the assistant's subsequent response *more useful*, not questioning its validity.

    *   Signals: Adding keywords, specifying criteria, excluding terms, defining terms.
    *   Examples: "I meant Windows 8 **on a tablet**." "Exclude results from Wikipedia." "What about vegan options?"

5. ELABORATION - User requests more detail on a specific aspect of the response that the assistant already provided. The user isn't questioning the response itself, just wanting more information *on the topics already introduced*. This also includes cases when the user wants to see a different viewpoint or consider another factor, alternative solutions **to the problem or situation that the assistant has addressed**.

    *   Signals: "Tell me more about...", "Explain X in more detail," "Give me an example of...", "What are the benefits/drawbacks **of what you just said**?"
    *   Examples: "Explain how that works **that you just mentioned**." "Can you give me some real-world examples **of this approach**?" "What are the specific advantages of this approach **you're describing**?"

6. RELATED_TOPIC_EXPLORATION - User asks about related but not directly implied topics, seeking a broader understanding **that goes beyond the original scope AND beyond the details already offered by the assistant**. This is a "tangential but related" question.

    *   Signals: Questions about related concepts, tangential connections, and similar ideas **not already discussed**.

    *   Examples: (After travel planning) "What are some good bars nearby **the hotel you recommended**?" (After discussing a historical event) "What were the social consequences of that **war you just discussed**?"

7. COMPARATIVE - User wants a comparison between different options **presented in the response**. *This only applies if the assistant has already offered multiple options*.

* Signals: How does X compare to Y?, Pros and cons, which is better for me, compare and contrast.
* Examples: "What are the pros and cons of **each of those options**?", "How does option A compare to option B? **You mentioned both**", "Which one is better for X purpose? **You gave me 3 options**"

8. TASK_ORIENTED_FOLLOWUP - User asks for assistance with a task that logically follows from the initial response. This is about getting the assistant to *do something* based on the information it provided, rather than just asking questions.

* Signals: Requests for automation, generation, translation, coding examples, shopping lists, etc.
* Examples: (After finding a recipe) "Can you create a shopping list **for that recipe**?" (After finding information on a software library) "Can you show me a code example **using that library**?"

9. PIVOT - User abruptly changes the topic to something unrelated to the previous discussion.

* Signals: Introducing a completely new subject with no obvious connection to the prior conversation.
* Examples: (After discussing travel) "By the way, do you know anything about quantum physics?"

10. CONVERSATION_CLOSURE - User signals the end of the conversation, with no need for followup

* Signals: Signaling end of conversation by saying thank you, or stop or related phrases.
* Examples: "thank you!", "thanks", "ty", "that answered my question", "that's all", etc.

11. NO_RESPONSE - the user did not respond to the assistant's message.

Instructions:

1. You are given a previous turn user query, assistant response, and next turn user response (None if the user did not respond/speak).
2. Determine the type of the next turn user response based on the categories above. Note that the categories include cases where the user responded with statements, as well as cases where the user responded with a question or questions. Based on the type of response, determine the best category.

```
3.  If the user did not respond, output NO_RESPONSE
4.  Explain your reasoning in one or two sentences, justifying why you chose a specific
category. This should refer to the definitions provided.
5.  Choose only one category that best describes the user response behavior to the assistant.

Output your response in the format:

Reason: {{reason}}
ResponseType: {{user response type}}

**Input:**
```

B.2.3 AI question-type classification. For classifying AI responses as containing clarifying questions, open-ended questions, or no questions, we used the following prompt:

```
You are an expert conversation analyst. Your task is to classify the assistant's
response provided below.

# ASSISTANT RESPONSE
{assistant_response}

# TASK
Analyze the "ASSISTANT RESPONSE" above. Choose ONLY ONE of the following categories that best describes the
Assistant's Response:
- 'Clarifying Question': The assistant asked for more detail.
- 'Open-ended Question': The assistant asked a broad, engaging question.
- 'No Question': The assistant did not ask a follow-up question.

Your output must be ONLY the category name (e.g., 'Clarifying Question').
```

C Supplemental Results

Breakdown	# of Participants	Percent (%)
Age Range		
21 to 30	19	14.62
31 to 40	23	17.69
41 to 50	35	26.92
51 to 60	34	26.15
61 to 70	7	5.38
71 or over	1	0.77
n/a	11	8.46
Gender		
Man	73	56.15
Woman	57	43.85
Racial or ethnic groups		
Asian	11	8.46
Black or African American	13	10.00
Hispanic, Latino or Spanish Origin	16	12.31
White	93	71.54
Completed highest degree		
Less than a high school degree	1	0.77
High school diploma or the equivalent (for example: GED)	12	9.23
Some college credit, no degree	31	23.85
Associates degree (for example: AA, AS)	7	5.38
Bachelor's degree (for example: BA, BS)	50	38.46
Master's degree (for example: MA, MS, MEng, MEd, MSW, MBA)	24	18.46
Doctorate degree (for example: PhD, EdD)	5	3.85
Confidence in filling out medical forms by myself (without asking for help from someone else)		
Extremely confident	79	60.77
Very confident	29	22.31
Moderately confident	16	12.31
Slightly confident	6	4.62
Confidence in using mobile phone apps by myself (without asking for help from someone else)		
Extremely confident	95	73.08
Very confident	21	16.15
Moderately confident	10	7.69
Slightly confident	3	2.31
Not at all confident	1	0.77
Frequency in using AI chatbots (e.g. ChatGPT, Claude, Gemini)		
Most days	24	18.46
Several times a week	32	24.62
Several times a month	24	18.46
Several times a year	31	23.85
Never. I have never used an AI chatbot or don't know what it is.	19	14.62
Total	130	100

Table S3. Participant metadata summary for Study 4.

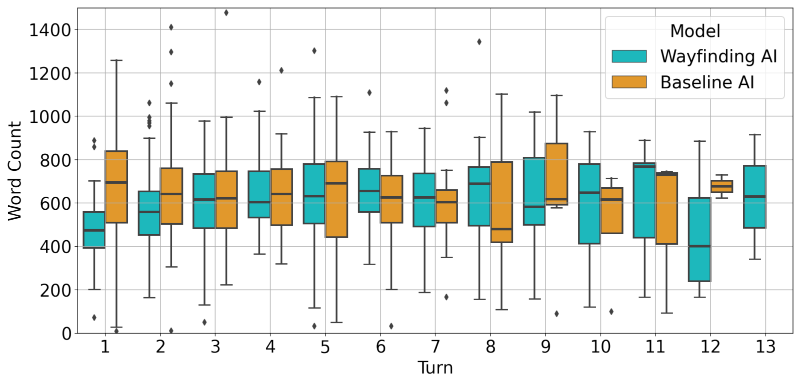

Fig. S2. Comparisons of each turn's word counts between Wayfinding AI and Baseline AI.

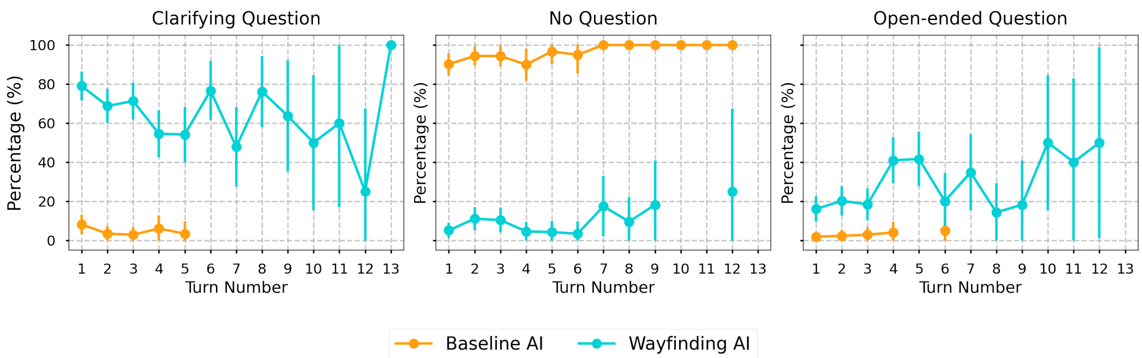

Fig. S3. Rates of AI responses containing Clarifying questions (left), No questions (center), and Open-ended questions (right) over conversational turns, for the Baseline AI (orange) and Wayfinding AI (blue).