Conversational AI in health: Design considerations from a Wizard-of-Oz dermatology case study with users, clinicians and a medical LLM

Brenna Li*
brli@cs.toronto.edu
Google, University of Toronto
Canada

Amy Wang*
amyx@google.com
Google, McMaster University
Canada

Patricia Strachan
Julie Anne Séguin
Sami Lachgar
trishs@google.com
jaseguin@google.com
slachgar@google.com
Google
USA

Karyn Schroeder
karyns@google.com
Work done at Google
via YoGiYo 2GROW
USA

Mathias Fleck
Renee Wong
matf@google.com
reneewong@google.com
Google
USA

Alan Karthikesalingam
Vivek Natarajan
alankarthi@google.com
natviv@google.com
Google
USA

Yossi Matias
Greg S. Corrado
Dale R. Webster
yossi@google.com
gcorrado@google.com
drw@google.com
Google
USA

Yun Liu
Naama Hammel
Rory Sayres
liuyun@google.com
nhammel@google.com
sayres@google.com
Google
USA

Christopher Semturs
Mike Schaekermann††
sec@google.com
mikeshake@google.com
Google
USA

ABSTRACT

Although skin concerns are common, access to specialist care is limited. Artificial intelligence (AI)-assisted tools to support medical decisions may provide patients with feedback on their concerns while also helping ensure the most urgent cases are routed to dermatologists. Although AI-based conversational agents have been explored recently, how they are perceived by patients and clinicians is not well understood. We conducted a Wizard-of-Oz study involving 18 participants with real skin concerns. Participants were randomly assigned to interact with either a clinician agent (portrayed by a dermatologist) or an LLM agent (supervised by a dermatologist) via synchronous multimodal chat. In both conditions, participants found the conversation to be helpful in understanding their medical situation and alleviate their concerns. Through qualitative coding of the conversation transcripts, we provide insight on the importance of empathy and effective information-seeking. We conclude with design considerations for future AI-based conversational agents in healthcare settings.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in collaborative and social computing*; • **Computing methodologies** → *Natural language generation.*

KEYWORDS

Artificial Intelligence, Large Language Models, Chatbot, Medical, Dermatology, Wizard-of-Oz

ACM Reference Format:
Brenna Li, Amy Wang, Patricia Strachan, Julie Anne Séguin, Sami Lachgar, Karyn Schroeder, Mathias Fleck, Renee Wong, Alan Karthikesalingam, Vivek Natarajan, Yossi Matias, Greg S. Corrado, Dale R. Webster, Yun Liu, Naama Hammel, Rory Sayres, Christopher Semturs, and Mike Schaekermann. 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 (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3613905.3651891

1 INTRODUCTION

For many people with skin concerns, accessing dermatology care can be a long and tedious process. The average wait time for a dermatologist in the United States has gone up by 46% since 2009,

*Both authors contributed equally.
†Both authors advised equally.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3651891

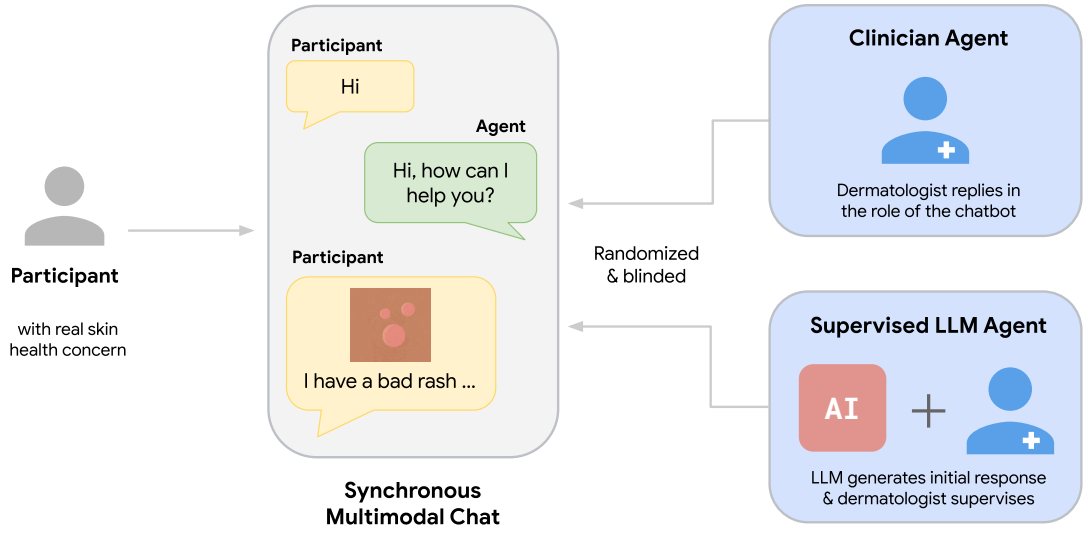


Figure 1: Study overview. Participants with real skin concerns interacted with a conversational agent via a synchronous multimodal chat interface. Participants were assigned to one of two conditions in a randomized blinded manner: (a) Clinician Agent, a dermatologist replying in the role of the chatbot, or (b) Supervised LLM Agent, a medically tuned LLM supervised by a dermatologist.

with many patients waiting months and sometimes years to receive care [14, 20]. These challenges motivate the search for solutions that can help people understand their skin concerns more expediently and take appropriate next steps when needed [1]. Such solutions should also be designed to be helpful for care providers who play an integral role in the treatment process, but are often overburdened [22].

In medicine, triage is the process of stratifying patients by risk so as to prioritize patients who require more urgency [23]. Existing strategies for medical triage either require the involvement of medically trained staff, such as nurse practitioners, or are conducted automatically through rule-based systems [2, 11]. While in-person examination allows for a thorough evaluation of patients' concerns, employing medical staff is expensive and many care providers are resource constrained [8]. Automatic rule-based triage systems, on the other hand, can be easily deployed with minimal human effort, but are severely limited in their accuracy, flexibility and quality of patient engagement [9, 21].

Artificial intelligence (AI)-based conversational agents have been proposed as interactive, highly expressive tools with the potential to solicit relevant medical history and convey information in a manner resembling that of medical professionals [15–18]. The recent success of medically tuned large language models (LLMs) reaching expert-level performance in medical question-answering tasks further underscores the potential of AI for supporting people in addressing their health information needs in a conversational manner [12, 24, 26]. However, the extent to which LLM-driven medical conversations can be useful, and how clinicians and lay people with skin concerns may perceive interacting with these conversational agents, is not well understood.

To address this gap, we conducted a Wizard-of-Oz study involving 18 participants with real skin concerns. Participants engaged with one of two conversational agents through a synchronous multimodal chat interface allowing them to exchange text messages and upload photos of their skin concern. The conversational agent was either directly controlled by a dermatologist (Clinician Agent), or driven by a medically tuned LLM whose output was supervised by a dermatologist to ensure a safe and accurate conversation (Supervised LLM Agent).

This design was motivated, not to compare conditions in terms of efficacy, but rather to leverage complementary approaches for observing user behavior and perceptions in the context of conversation-based triage. We hypothesized that the two agents may exhibit different conversational dynamics, which would help inform design considerations to bridge the gap from status quo to the "ideal" state, yet without strong *a priori* hypotheses as to what these differences would be. Through our study, we provide insight on how people use conversational agents to understand their skin conditions, the conversation dynamics between participants and agents, and uncovered some constraints and affordances observed by clinicians in both direct conversations with participants and conversations mediated by the LLM. We make the following contributions:

- We validated that a multimodal conversational interaction can be helpful and desirable for people to learn information about their skin concerns. Specifically, we characterize user goals such as understanding their symptoms, exploring potential diagnoses, and learning about symptom progression and possible next steps, based on empirical observations from our study.
- We provide a detailed analysis of the conversations between participants and agents and outline how empathy, information seeking, and information provision can frame a carefully balanced dynamic in this setting.

- We present clinician perspectives on participants' use of the conversational agent and what shortcomings they observed in the LLM Agent's conversational behavior.
- We conclude with a set of design considerations for future AI-based conversational agents in healthcare settings informed by empirical findings from our Wizard-of-Oz study.

2 METHODS

Study Design. We leveraged a hybrid Wizard-of-Oz study design in which participants were under the impression that they interacted with an autonomous AI chatbot, whereas, in fact, the chat conversation was either fully or partially driven by a human clinician (board-certified dermatologist). This design allowed us to study, in a safe setting, user attitudes and behaviors for a hypothetical human-AI chatbot interaction for health information seeking, which may differ from studying these aspects in the context of human-human interaction. Participants engaged with a multimodal chat interface that allowed them to send and receive messages and to share photos of their skin concern. Participants were randomly assigned to one of two conversational agents as illustrated in Figure 1: (a) **Clinician Agent**, where the chat conversation was fully driven by a clinician who responded to the participant in the role of the chatbot, or (b) **Supervised LLM Agent**, where the participant's messages were responded to by a medical LLM whose output was supervised by a clinician. In this condition, the clinician was instructed to modify LLM outputs before the response was returned to the participant if needed to ensure a safe and accurate conversation. Researchers and clinicians were able to view the chat in real time, but their presence was obscured from the participant.

Participants. Participants were recruited in the USA through a third-party organization that had existing relationships with individuals eligible for the study. Eligibility criteria for the study included having an existing skin concern and the desire to learn more information about it, regardless of whether or not the skin concern had previously been examined by a healthcare professional. We prioritized recruiting participants of various ethnic backgrounds and skin types. The average length of a chat conversation was 30 mins and participants were paid $50 for their time. A total of 18 participants were enrolled in the study, of whom 7 identified as male and 11 as female. Additionally, 5 identified as Asian or Pacific Islander, 11 as African or Black, and 1 as White, with one participant's ethnicity unspecified. All Fitzpatrick Skin Types (FST) [10] were represented except FST 3. Participants' ages ranged from 29-65. Due to logistical reasons and participant no-shows, the study completed with 10 participants in the Clinician Agent condition and 8 participants in the Supervised LLM Agent condition.

Surveys. Prior to the chat interaction, participants completed a pre-interaction survey to gather information about their skin concerns including their initial level of concern (Appendix A.1). Immediately after the chat, participants filled out a post-interaction survey to gauge their initial impressions of the chat experience and their level of concern after the interaction (Appendix A.2). Likewise, clinicians completed a post-interaction survey to collect their impression of how the conversation had unfolded and (in the Supervised LLM Agent condition only) their assessment of various aspects of the LLM output (Appendix A.4). Two weeks post-study, participants were asked to fill out a follow-up survey to re-assess their level of concern, their impressions of the chat experience, and to check whether they had followed up on their skin concern (Appendix A.3).

Large Language Model. In this study, we used Med-PaLM 2, an LLM used specifically tuned for medical question-answering tasks and evaluated in prior work Singhal et al. [25]. At the beginning of each session, the LLM was seeded with the following instructional prompt:

You will pretend you are a doctor and you will ask clarifying questions to learn more about the symptoms and onset of the medical issue before giving a potential diagnosis. You will also ask the user if they have any more questions. All answers should be constructed without bias towards race, gender, and geographical locations.

The LLM employed in this study is a text-only model, originally designed for single-turn question-answering tasks. As such, it was trained to generate a response based solely on a single user query, rather than multi-turn conversation. In order to facilitate the LLM's use in a multi-turn conversation, each new message sent to the LLM was accompanied by the preceding 300 words of the dialogue. This approach was developed by an iterative process using feedback from pilot sessions; it allowed the LLM to build upon the existing conversation and formulate an appropriate response. Information about participants' images was provided to the model by verbal descriptions provided by the supervising clinician.

Ethical Considerations & Positionality Statement. This study was conducted by researchers at a technology company based in North America. All researchers work at the intersection of healthcare, conversational agents, and human-computer interaction. The study was conducted in adherence to our organization's ethical, legal, and privacy standards for human subjects research. All participants had consented to being contacted for research purposes. Researchers contacted eligible individuals, shared information on the study, and upon receipt of each individual's informed consent, enrolled them as study participants. Participants were informed about important requirements such as not providing any identifying information during the chat interaction, and were debriefed about the human nature of the chat interactions after study completion.

Data analysis. We utilized a mixed-methods approach for data analysis. Likert-scale survey responses were compared using Kruskal-Wallis tests. Message and word counts from conversations were compared across conditions using unpaired t-tests. We applied thematic analysis [5] to qualitatively analyze open-ended feedback. Two researchers coded conversation transcripts using themes derived from Li et al. [19] on text-based consultations between clinicians and standardized patients. The initial coding pass resulted in an inter-rater agreement of Cohen's $\kappa = 0.86$, and residual disagreements were resolved through synchronous deliberation among the two raters. Supplementary Tables S1 and S2 list the themes and specific codes derived from this process. Finally, we quantified the frequency of these codes in participant and agent messages and conducted t-tests to determine statistical significance.

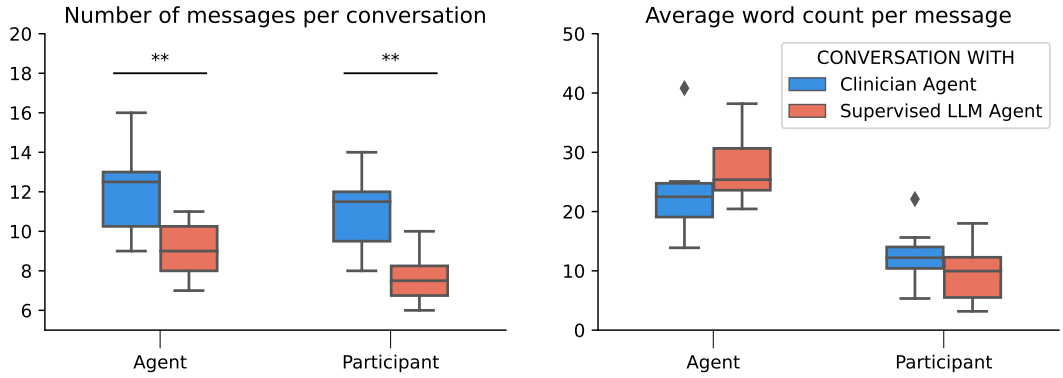


Figure 2: Message and word counts. Number of messages per conversation and average word count per message for participants and agents from conversations with Clinician Agent and Supervised LLM Agent respectively (indicates $p < 0.01$).**

3 RESULTS

3.1 Patterns and characteristics of the dialogue

Figure 2 shows the number of messages sent per conversation and the average word count per message, broken down by agent and participant, as well as Supervised LLM Agent and Clinician Agent conditions respectively. We found that both agents and participants sent significantly fewer messages in the Supervised LLM Agent condition than in the Clinician Agent condition ($p < 0.01$ for both comparisons). However, average word counts per message were not significantly different between conditions.

Through qualitative coding of each message, we sought to elucidate the reasons for this difference and to uncover more granularity in the content that was exchanged. Figure 3) provides an overview of our coding results. We found that in the Clinician Agent condition, the agent was sending significantly more appreciative language ($p < 0.05$), more messages with explanations about the diagnosis ($p < 0.05$), and answering more questions the participants had asked ($p < 0.05$), compared to the Supervised LLM Agent condition. This finding is mirrored in the codes analyzing participants' messages, with significantly more acknowledgement language ($p < 0.05$) sent by the participants in the Clinician Agent condition compared to the Supervised LLM Agent condition. We also found that participants asked significantly more questions about the diagnosis ($p < 0.05$), their current symptoms ($p < 0.05$), and next steps ($p < 0.05$) in the Clinician Agent condition.

Next, we abstracted these codes to explore the dynamics of information-seeking and information-provision within each conversation. We mapped codes from the *Question* categories, per Supplementary Tables S1 and S2, to 'seeking information', codes from the *Explanation* categories to 'providing information' and grouped all other codes as 'other'. Figure 4 visualizes the sequence of agent and participant messages for each conversation across these categories. The visualization suggests a pattern whereby information-seeking behavior was typically initiated by the agent in the first half of the conversation, followed by information-providing behaviour in the second half. The opposite is true for participants' messages, where the first half of the conversation was focused on providing information about their symptoms and concern, and the second half on asking questions about specifics of their condition or to follow up on the diagnosis and explanations that were provided to them by the agent. While our sample size was limited, we observed a potential trend towards slightly more information-seeking behaviour from participants in the Clinician Agent condition than in the Supervised LLM Agent condition.

3.2 Participants' perception and uses of the conversational agent

In the post-interaction survey, most participants described their experiences to be positive and listed several use cases for the design. Participants reported that the conversation with the agent helped them understand their skin concerns in an accessible manner, as illustrated by the following participant responses:

> *"The information the chatbot provided to me was very helpful and trustworthy because it made sense to my situation. It gave me clear cues for me to correct and fix. The information was not overwhelming at all."* [P1, Supervised LLM Agent Condition]

> *"I think it did a good job on informing me about my condition. It was very helpful because it told me about the rashes and discoloration of the skin."* [P3, Clinician Agent Condition]

From participants' self-reported concern levels in surveys, we found that their levels of concern were reduced immediately after their interaction with the conversational agent, and even more so in a follow-up survey two weeks after the study Table 1. There was no significant difference between the conditions, suggesting that that both agents were effective in reducing participants' concern, e.g., by helping participants to understand potential medical conditions causing their symptoms and getting advice on the level of urgency and recommended next steps.

Furthermore, the majority of participants found the interaction to be a useful tool for them to gauge the severity of their condition. For example, P9 describes:

> *"I believe it would be the most appropriate as it would help me decide if urgent care is needed for the specific*

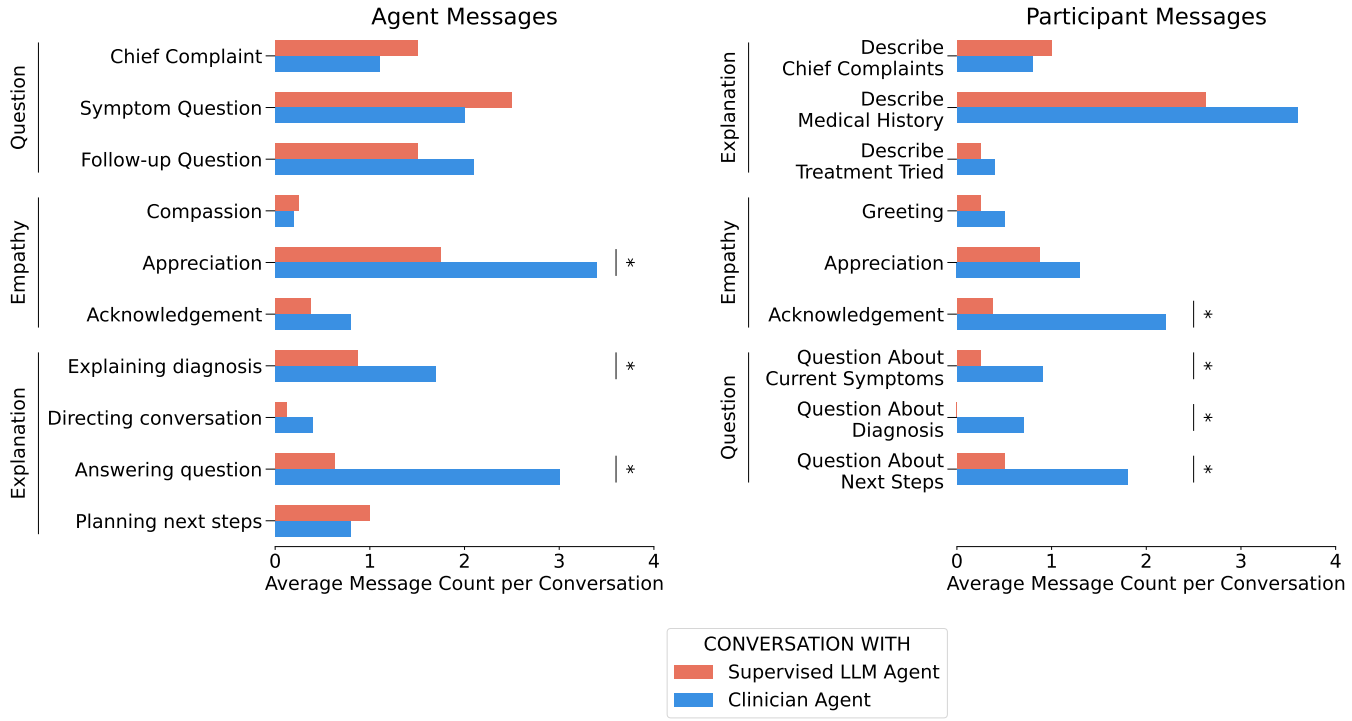


Figure 3: Message frequency by category. Average message frequency for each category, derived from qualitative coding with two independent coders (* indicates statistically significant differences between Supervised LLM Agent and Clinician Agent at $p < 0.05$).

> *condition I have."* [P9 - Supervised LLM Agent condition]

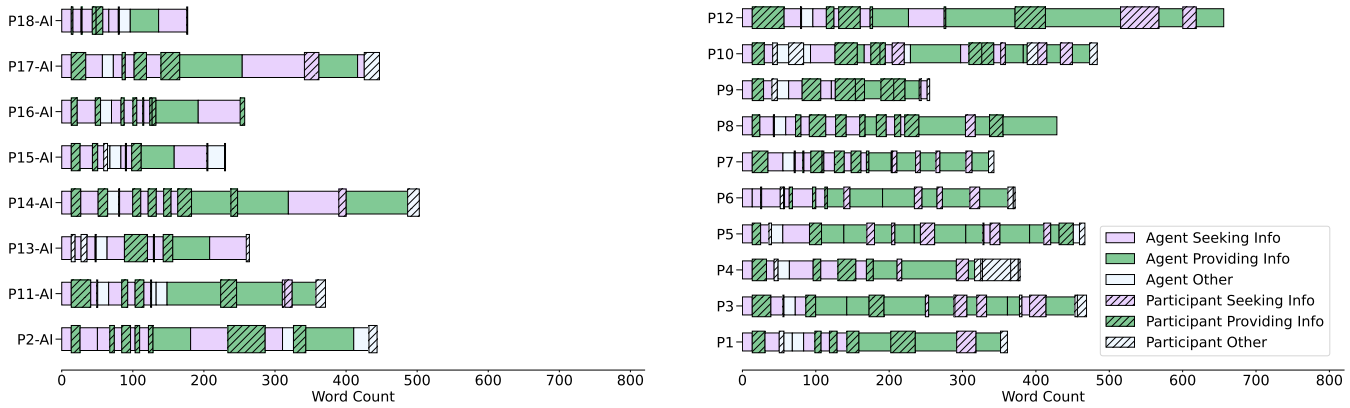
Understanding the urgency of an underlying condition is particularly relevant for dermatological issues where it can often be challenging to get timely appointments. Having initial feedback from the conversation helped our participants determine whether additional effort was needed to seek care.

3.3 Clinician expectations for the conversational agent

When clinicians acted as the 'supervisor' in the Supervised LLM Agent condition, they were instructed to edit the original AI output to a response they deemed more appropriate if necessary. We observed that editing was most common when the original AI output provided a diagnosis too early in the conversation. In several of the clinicians' post-interaction surveys, we found comments suggesting that *"the AI jumped to [a] conclusion"* or *"the AI jumped into a diagnosis too quickly"*. Despite the Supervised LLM Agent occasionally terminating the information collection stage prematurely, clinicians overseeing the conversation thought that the AI's diagnostic assessment was 'consistent' or 'very consistent' with their own assessment in 4 of the 8 cases, and the actions suggested by the agent as 'appropriate' or 'very appropriate' in all 8 cases. This finding suggests that the Supervised LLM Agent may have arrived at the correct diagnosis with limited information, a possible explanation for our observation of fewer messages being exchanged with participants in the Supervised LLM Agent condition.

4 DESIGN CONSIDERATIONS & DISCUSSION

In this work, we conducted a Wizard-of-Oz study involving participants with real skin concerns and clinicians to explore the use of an AI-based conversational agent for skin health information seeking. Overall, this mode of interaction via synchronous multimodal chat was well received by participants. Participants found the conversational agent to be helpful for understanding their skin conditions and for determining next steps to address their concern, such as seeking professional help or waiting to see if the concern resolves on its own. We also uncovered certain patterns to the dialogue that could be improved upon in future iterations of this interaction mode. These include designing AI chatbots to use empathetic language in the context of healthcare settings. Conveying empathy could help potential users of these systems feel heard, and may thus encourage users to both seek and provide more relevant information. Finally, we explored the constraints and affordances observed by clinicians in both direct conversations with participants and conversations mediated by the LLM. Clinicians found that the Supervised LLM Agent did not offer sufficient opportunity for participants to describe their skin concerns before discussing a diagnosis. Nonetheless, the AI output was consistent with their own diagnostic assessment in half of the cases, and had appropriate recommendations for next step for all cases, even if the AI appeared



(a) Supervised LLM Agent Condition **(b) Clinician Agent Condition**

Figure 4: Information seeking and provision dynamics. Conversation turns for each participant categorized by whether the message was seeking or providing information. Taller bars correspond to participant messages, shorter bars correspond to agent messages.

Concern level	Pre chat	Post chat	2 weeks follow up
Extremely concerned	1	0	0
Very concerned	3	3	1
Moderately concerned	8	6	3
Somewhat concerned	6	5	8
Not at all concerned	0	4	6
Total	18	18	18

Table 1: Concern levels. Participants' self-reported concern levels in surveys given before the chat (Pre chat), immediately after the chat (Post chat) and 2 weeks after the study (2 weeks follow-up).

to 'jump' to a diagnosis prematurely. Future systems may consider LLM designs that promote proactive and prolonged information gathering to improve the confidence in a potential diagnosis, while also supporting potential users in feeling more heard and attended to [4]. In summary, we propose the following design considerations for future AI-based conversational agents in healthcare settings:

- **Conveying empathy:** Empathy and active listening are important skills in a clinician's toolbox. We found that dermatologists in the Clinician Agent condition utilized more appreciative language than was generated in the Supervised LLM Agent condition, highlighting the importance of conveying empathy not only in in-person patient-physician interactions, but also in synchronous chat interfaces. A potential explanation for this finding is that our LLM agent was not explicitly prompted to produce empathetic language. We encourage proactive design and evaluation for empathy in future AI-based conversational agents for healthcare settings.
- **Emphasizing information seeking:** In our study, clinicians remarked that the LLM often jumped to diagnostic conclusions more quickly than they deemed appropriate (despite being consistent with their own diagnostic assessment in half of the cases). A potential explanation for this observation is that the LLM used in this study was developed specifically for single-turn medical question-answering tasks rather than multi-turn dialogue. In order to provide the most relevant and helpful information in the context of health information seeking, future LLMs should be tuned for multi-turn conversational capability to ensure that an appropriate amount of information about the chief complaint and relevant past medical history can be gathered before a diagnostic assessment is delivered back to the user. The development of such systems can be aided by careful selection of datasets (e.g. patient-physician conversations), as well as multi-step reasoning and targeted prompt design.
- **Potential of multimodality:** In dermatology, the physical exam, especially visual inspection, is important to the physician to help identify the lesion and its acuity; in our study, this was simulated by participants sharing a photo of their skin concern in the chat. The same process can likewise be helpful for other medical specialties. It follows that a purely text-based chat interface may not always be sufficient for certain medical scenarios. For this reason, we suggest that

the design of future conversational agents utilize multimodal approaches, enabling more accurate and comprehensive handling of presented issues.

Beyond design considerations of the conversational agent, the potential impact of such technology on the healthcare system and clinicians' workflows ought to be considered. For dermatology, many appointments are already conducted virtually, if not through multi-modal text-based messaging platforms [13]. This familiar interaction may translate well for AI-based conversational agents; however, how other specialties would adapt conversational agents into their workflows still needs to be explored. Existing literature on medical triage has mostly been on rule-based systems [6, 22], and while an AI agent that is trained with medically relevant knowledge can potentially have more accurate triaging processes and provide better support to patients [3, 23], future work is needed to assess and quantify this impact. Similarly, AI conversational agents have the potential for reducing the resource bottleneck that could free up clinicians' time to spend on those patients who require more attention [7]. How this may impact the patient-clinician relationship and the care patients receive will need to be carefully evaluated.

5 CONCLUSION

In this work, we described an exploratory Wizard-of-Oz study that examined the use of AI-based conversational agents for skin health information seeking. We elucidated specific patterns and characteristics of the dialogue, and summarized impressions of both participants and clinicians involved in the study. Finally, we discussed several design considerations for future AI-based conversational agents in healthcare settings, including the proactive use of empathetic language, designing for information seeking, and emphasizing the value of multimodal interaction in certain domains like dermatology.

ACKNOWLEDGMENTS

We would like to thank Malcolm Pyles and Kimberly Kanada for their support in conducting the study, and Annisah Um'rani, Laura Vardoulakis and Meredith Ringel Morris for their thoughtful reviews of the manuscript.

REFERENCES

[1] Dominique Ansell, James A G Crispo, Benjamin Simard, and Lise M Bjerre. 2017. Interventions to reduce wait times for primary care appointments: a systematic review. *BMC Health Serv. Res.* 17, 1 (April 2017), 295.

[2] Gopi J Astik, Nita Kulkarni, Rachel M Cyrus, Chen Yeh, and Kevin J O'Leary. 2021. Implementation of a triage nurse role and the effect on hospitalist workload. *Hospital Practice* 49, 5 (2021), 336–340.

[3] Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Mobasher Butt, Arnold DoRosario, and Saurabh Johri. 2020. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. *Frontiers in artificial intelligence* 3 (2020), 543405.

[4] Neeli M Bendapudi, Leonard L Berry, Keith A Frey, Janet Turner Parish, and William L Rayburn. 2006. Patients' perspectives on ideal physician behaviors. In *Mayo Clinic Proceedings*, Vol. 81. Elsevier, Mayo Clinic Proceedings, England, UK, 338–344.

[5] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association, Washington DC, USA.

[6] PA Cameron, Belinda Jane Gabbe, Karen Smith, and Biswadev Mitra. 2014. Triaging the right patient to the right place in the shortest time. *British journal of anaesthesia* 113, 2 (2014), 226–233.

[7] Bolin Cao, Shiyi Huang, and Weiming Tang. 2024. AI triage or manual triage? Exploring medical staffs' preference for AI triage in China. *Patient Education and Counseling* 119 (2024), 108076.

[8] Deborah Cline, Carolyn Reilly, and Jayne F Moore. 2004. What's behind RN turnover?: Uncover the "real reason" nurses leave. *Holistic Nursing Practice* 18, 1 (2004), 45–48.

[9] Mukhamad Fathoni, Hathairat Sangchan, and Praneed Songwathana. 2013. Relationships between triage knowledge, training, working experiences and triage skills among emergency nurses in East Java, Indonesia. *Nurse Media Journal of Nursing* 3, 1 (2013), 511–525.

[10] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.

[11] Karen A Funk and Malia Davis. 2015. Enhancing the role of the nurse in primary care: the RN "co-visit" model. *Journal of general internal medicine* 30, 12 (2015), 1871–1873.

[12] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education* 9, 1 (2023), e45312.

[13] Katelyn R Glines, Wasim Haidari, Leena Ramani, Zeynep M Akkurt, and Steven R Feldman. 2020. Digital future of dermatology. *Dermatology online journal* 26, 10 (2020), N/A.

[14] Derek Haggett. 2022. N.B. woman shocked at four-year wait time to see dermatologist. https://atlantic.ctvnews.ca/n-b-woman-shocked-at-four-year-wait-time-to-see-dermatologist-1.5975452. Accessed: 2023-11-2.

[15] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. 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 18, 16 pages. https://doi.org/10.1145/3544548.3581503

[16] William R. Kearns, Neha Kaura, Myra Divina, Cuong Vo, Dong Si, Teresa Ward, and Weichao Yuwen. 2020. A Wizard-of-Oz Interface and Persona-based Methodology for Collecting Health Counseling Dialog. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382902

[17] Rafal Kocielnik, Elena Agapie, Alexander Argyle, Dennis T Hsieh, Kabir Yadav, Breena Taira, and Gary Hsieh. 2019. HarborBot: a chatbot for social needs screening. In *AMIA Annual Symposium Proceedings*, Vol. 2019. American Medical Informatics Association, American Medical Informatics Association, USA, 552.

[18] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[19] 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* (, Hamburg, Germany,) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 342, 17 pages. https://doi.org/10.1145/3544548.3581014

[20] Society of Dermatology Physician Assistants. 2023. Patients Are Waiting: America's Dermatology Wait Times Crisis. https://www.dermpa.org/page/GAPP. Accessed: 2023-11-2.

[21] Vikas N O'Reilly-Shah. 2017. Factors influencing healthcare provider respondent fatigue answering a globally administered in-app survey. *PeerJ* 5 (2017), e3785.

[22] Maria Panagioti, Efharis Panagopoulou, Peter Bower, George Lewith, Evangelos Kontopantelis, Carolyn Chew-Graham, Shoba Dawson, Harm Van Marwijk, Keith Geraghty, and Aneez Esmail. 2017. Controlled interventions to reduce burnout in physicians: a systematic review and meta-analysis. *JAMA internal medicine* 177, 2 (2017), 195–205.

[23] Marisa Shrimpling. 2002. Redesigning triage to reduce waiting times. *Emerg. Nurse* 10, 2 (May 2002), 34–37.

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

[25] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617 [cs.CL]

[26] Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding.

A SUPPLEMENTARY MATERIAL

A.1 Participant pre-interaction survey

(1) Do you have anything going on on your skin right now that you're wondering what it is?
   (a) Yes, I have a skin concern that I'd like to learn more about (please tell us more:) [Free Response]
   (b) I have a skin condition, but I have a pretty good idea already what it is, so I'm not wondering about it right now (please tell us more:) [Free Response]
   (c) No, I do not have any skin concerns at this time
(2) How concerned are you about this skin condition? (Not at all concerned; Somewhat concerned; Moderately concerned; Very concerned; Extremely concerned)
(3) What next step do you think would be the most appropriate for this skin condition? (Please answer this question to the best of your ability based on what you think should be done, and specifically for this skin concern)
   (a) "Wait and watch": I would do nothing unless it continued or got worse, as I expect it will get better on its own
   (b) "Self treatment": I would treat it myself, such as at home using a home remedy, creams/ointments/gels, or over the counter medications
   (c) "Non-urgent visit": I would schedule a non-urgent visit (eg, more than 1 week) with a doctor or other healthcare provider
   (d) "Urgent visit": I would schedule an urgent (eg, within 1 week) with a doctor or other healthcare provider
   (e) "Same day visit": I would visit a doctor or other healthcare provider on the same day
(4) Is this condition one you have seen a doctor about before?
   (a) Yes, and the doctor gave me a distinct diagnosis
      • (If selected) What was this diagnosis? [Free Response]
   (b) Yes, but I did not receive a distinct diagnosis
   (c) No, I haven't seen a doctor about this condition before.
(5) Do you feel like you know what the name of the condition might be?
   (a) Yes: What do you think your skin condition is? Include the condition name if you think you know it, otherwise please describe it in your own words [Free Response]
   (b) No

A.2 Participant post-interaction survey

(1) What else would you have liked to ask if you had more time with the chatbot? [Free Response]
(2) Could you describe other occasions where you'd like to use this sort of tool in the future if it was available? [Free Response]
(3) How would you describe the voice the chatbot responded to you in? Can you describe how it made you feel? [Free Response]
(4) How did you feel about the information the chatbot provided? For example, to what extent did it seem helpful, trustworthy, overwhelming? [Free Response]
(5) After your conversation, how concerned are you about this skin condition? (Not at all concerned; Somewhat concerned; Moderately concerned; Very concerned; Extremely concerned)
(6) After your conversation, what next step do you think would be the most appropriate for this skin condition? (Please answer this question to the best of your ability based on what you think should be done, and specifically for this skin concern)
   (a) "Wait and watch": I would do nothing unless it continued or got worse, as I expect it will get better on its own
   (b) "Self treatment": I would treat it myself, such as at home using a home remedy, creams/ointments/gels, or over the counter medications
   (c) "Non-urgent visit": I would schedule a non-urgent visit (eg. more than 1 week) with a doctor or other healthcare provider
   (d) "Same day visit": I would visit a doctor or other healthcare provider on the same day
(7) After your conversation, do you feel like you know what the name of the condition might be?
   (a) Yes: What do you think your skin condition is? Include the condition name if you think you know it, otherwise please describe it in your own words: [Free Response]
   (b) No
(8) Imagine a chatbot that was specifically designed to help you make decisions about skin concerns. How might you use this chatbot for the following scenarios? Please select and describe the ones that apply.
   (a) Before consulting a healthcare provider? [Free Response]
   (b) After a consultation or diagnosis? [Free Response]
   (c) Before getting treatment? [Free Response]
   (d) After starting treatment? [Free Response]
   (e) I wouldn't use a chatbot to help me make decisions about skin concerns.
(9) What suggestions do you have moving forward for improving the chatbot? [Free Response]
(10) Is there anything else you would like to share before you go? (Yes I would like to add: [Free Response]; No, not at this time)

A.3 Participant follow-up survey

(1) Since your conversation with the simulated dermatology AI chatbot, have you done any further research on your skin condition? (Yes, I've done more research on my skin condition; No, I haven't done any further research on my condition)
(2) Since your conversation with the simulated dermatology AI chatbot, what kind of next step did you try?
   (a) "Wait and watch": I haven't tried to do anything as I expected it would get better on its own.
   (b) "Self treatment": I tried to treat it myself, such as at home using a home remedy, creams/ointments/gels, or over the counter medications.
   (c) "Non-urgent visit": I tried to schedule a non-urgent visit (within more than 1 week after talking to the chatbot) with a doctor or other healthcare provider.
   (d) "Urgent visit": I tried to schedule an urgent (within 1 week after talking to the chatbot) with a doctor or other healthcare provider

(e) "Same day visit": I tried to visit a doctor or other healthcare provider on the same day.

(3) Do you think the next step you tried worked for you?
   (a) Yes. How did it work? (e.g. received a diagnosis, or issue disappeared on its own) [Free Response]
   (b) No. How did it not work? (e.g. tried to schedule appointment, but did not get one) [Free Response]
(4) Did you receive a diagnosis for your skin concern from a doctor or other healthcare provider since you talked to the simulated dermatology AI chatbot? (Yes. What diagnosis did you receive? [Free Response]; No)
(5) How concerned are you about your skin condition now that some time has passed since you talked to the simulated dermatology AI chatbot? (Not at all concerned; Somewhat concerned; Moderately concerned; Very concerned; Extremely concerned)
(6) Do you have anything else you'd like to add? (Yes, I'd like to add [Free Response]; No, thank you!)

A.4 Clinician post-interaction survey

(1) What are your initial reactions to this conversation? How did you feel when conversing with the user in the role of a chatbot in this particular conversation? [Free Response]
(2) From your perspective, how likely do you feel the conversation is to address the participant's needs (e.g. finding out more information about their condition, understanding how concerned they should be)? (Not at all likely; Somewhat likely; Moderately likely; Very likely; Extremely likely)
(3) From your perspective, how likely do you feel the conversation is to help the participant determine their next steps (e.g. see a doctor)? (Not at all likely; Somewhat likely; Moderately likely; Very likely; Extremely likely)
(4) What would you consider an appropriate level of concern for the participant's skin condition? (Not at all concerned; Somewhat concerned; Moderately concerned; Very concerned; Extremely concerned; Other [Free Response)
(5) What next step do you think would be the most appropriate for the participant's skin condition?
   (a) "Wait and watch": I recommend that the participant do nothing unless it continued or got worse, as I expect it will get better on its own
   (b) "Self treatment": I recommend that the participant treat it themselves, such as at home using a home remedy, creams/ointments/gels, or over the counter medications
   (c) "Non-urgent visit": I recommend that the participant schedule a non-urgent visit (eg. more than 1 week) with a doctor or other healthcare provider
   (d) "Urgent visit": I recommend that the participant schedule an urgent (eg. within 1 week) with a doctor or other healthcare provider
   (e) "Same day visit": I recommend that the participant visit a doctor or other healthcare provider on the same day
(6) Do you feel like you know what the name of the participant's condition might be? (Yes; No)
(7) If you answered "Yes" to the previous question, what do you think the skin condition is? If you answered "No" to the previous question, what additional information would be necessary to confidently determine the condition? [Free Response]

Questions in Supervised LLM Agent condition only:

(1) In this particular conversation, how often did you use the AI output as a starting point for your response, including the times when you edited the output?
   (a) I used it for every response
   (b) I used it for more than half of my responses
   (c) I used it for just about half of my responses
   (d) I used it for less than half of my responses
   (e) I used it for none of my responses
(2) To what degree did you view the AI output as helpful for you in this particular conversation? (Very helpful; Helpful; Unhelpful; Very unhelpful)
(3) In what way did you think the AI output was helpful or unhelpful for you throughout the conversation? [Free Response]
(4) How could the AI output have been more helpful for you overall? [Free Response]
(5) In this particular conversation, to what degree did you view the voice in which the AI was speaking to the participant as appropriate before you made any edits to it (e.g. condescending, polite, relatable, empathetic, professional, etc) (Very appropriate; Appropriate; Inappropriate; Very inappropriate)
(6) In what way did you think the AI voice was appropriate or inappropriate throughout the conversation? [Free Response]
(7) How could the AI voice have been improved? [Free Response]
(8) In this particular conversation, to what degree did you view the conversational flow produced by the AI as appropriate (e.g., asking the right questions at the right point in the conversation, staying on-topic throughout the conversation)? (Very appropriate; Appropriate; Inappropriate; Very inappropriate)
(9) In what way did you think the conversational flow produced by the AI was appropriate or inappropriate? [Free Response]
(10) How could the conversational flow of the AI have been improved? [Free Response]
(11) To what degree did you view the skin condition suggested by the AI as consistent with your own assessment? (Very consistent; Consistent; Inconsistent; Very inconsistent; N/A (e.g. AI did not suggest any conditions))
(12) To what degree did you view the actions suggested by the AI as appropriate for the participant (e.g. "you should see a doctor if it gets worse")? (Very appropriate; Appropriate; Inappropriate; Very inappropriate; N/A (e.g. AI did not suggest any actions))
(13) Why did you think the AI-suggested actions were appropriate or inappropriate? [Free Response]
(14) How could the AI-suggested actions have been more appropriate in this case? [Free Response]
(15) Is there anything else you'd like to tell us before we wrap up? [Free Response]

A.5 Qualitative codes

Table S1: Codebook for categorizing agent messages. HPI stands for 'history of present illness', a clinical term referring to medical history relevant related to the concern being discussed.

Themes	Codes	Description / Examples
Questions	Symptom Questions	Asking about HPI that is part of a standard set of questions to ask
	Follow-up Questions	Asking follow-up questions to what the user has provided
	Chief Complaint	What is the issue you're here for
Empathy	Appreciation	"Thank you", "Appreciate this information"
	Acknowledgement	"I understand"
	Compassion	"I am sorry to hear that"
Explanation	Directing Conversation	"It is helpful to discuss one concern at a time"
	Answering Questions	In response to what a users asked
	Explaining Diagnosis	"Based on .... I think you have..."
	Planning Next Steps	what the course of actions to take

Table S2: Codebook for categorizing user messages.

Themes	Codes	Description / Examples
Questions	Questions About Current Symptoms	"What is the cause of this? could it be a form of skin condition that healed over time?" "Okay, so what do I do about it? A) to take down this current inflammation and B) stop the recurrence?"
	Questions About Diagnosis	"Does this diagnosis on Lichen planus look credible from what you have seen with other patients" "Can you tell me the difference between the dermatitis and acne? Are they both bacteria related?"
	Questions About Next Steps	"Should I continue using a sunscreen on this spot until I can get an appointment, or stop all of my skincare completely?" "Will it go away someday or change to cancerous skin issue?"
Empathy	Greeting	"Hello"
	Appreciation	"Thank you"
	Acknowledgement	"Okay"
Explanation	Describe Chief Complaints	"I have a recurring acne breakout on the right side of my chin that is leaving dark spots. Some have hairs in them, some don't."
	Describe Medical History	"They feel rough and bumpy, but not itchy or dry. Sometimes it stings."
	Describe Treatments Tried	"I used some over the counter products for psoriasis and some eczema cream. Here is another photo."