frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 18 October 2024
DOI 10.3389/fcomp.2024.1427463

OPEN ACCESS

EDITED BY
Ilaria Tiddi,
VU Amsterdam, Netherlands

REVIEWED BY
Junxiang Chen,
Indiana University, United States
Alana Delaforce,
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

*CORRESPONDENCE
Catalina Gomez
✉ cgomezc1@jhu.edu
Mathias Unberath
✉ mathias@jhu.edu

†These authors have contributed equally to this work

RECEIVED 03 May 2024
ACCEPTED 26 September 2024
PUBLISHED 18 October 2024

CITATION
Gomez C, Yin J, Huang C-M and Unberath M (2024) How large language model-powered conversational agents influence decision making in domestic medical triage contexts. *Front. Comput. Sci.* 6:1427463. doi: 10.3389/fcomp.2024.1427463

How large language model-powered conversational agents influence decision making in domestic medical triage contexts

Catalina Gomez*†, Junjie Yin†, Chien-Ming Huang and Mathias Unberath*

Department of Computer Science, Johns Hopkins University, Baltimore, MD, United States

Introduction: Effective delivery of healthcare depends on timely and accurate triage decisions, directing patients to appropriate care pathways and reducing unnecessary visits. Artificial Intelligence (AI) solutions, particularly those based on Large Language Models (LLMs), may enable non-experts to make better triage decisions at home, thus easing the healthcare system's load. We investigate how LLM-powered conversational agents influence non-experts in making triage decisions, further studying different persona profiles embedded via prompting.

Methods: We designed a randomized experiment where participants first assessed patient symptom vignettes independently, then consulted one of the two agent profiles—rational or empathic—for advice, and finally revised their triage ratings. We used linear models to quantify the effect of the agent profile and confidence on the weight of advice. We examined changes in confidence and accuracy of triage decisions, along with participants' perceptions of the agents.

Results: In a study with 49 layperson participants, we found that persona profiles can be differentiated in LLM-powered conversational agents. However, these profiles did not significantly affect the weight of advice. Notably, less confident participants were more influenced by LLM advice, leading to larger adjustments to initial decisions. AI guidance improved alignment with correct triage levels and boosted confidence in participants' decisions.

Discussion: While LLM advice improves triage recommendations accuracy, confidence plays an important role in its adoption. Our findings raise design considerations for human-AI interfaces, highlighting two key aspects: encouraging appropriate alignment with LLMs' advice and ensuring that people are not easily swayed in situations of uncertainty.

KEYWORDS

human-AI interaction, decision-making, empirical studies, LLMs, triage

1 Introduction

Emerging Artificial Intelligence (AI) technologies promise transformational opportunities in healthcare delivery (Topol, 2019). Identifying the right care pathway for patients is crucial to reduce delays and unnecessary costs and visits. Patients face the challenge of assessing the type of care they need before approaching any health service. Empowering non-experts to make informed self-triage decisions from the comfort of their homes is key to mitigating the load on healthcare systems. This is evident in emergency departments where patients are further

triaged on arrival to determine the severity of their complaints, and the routine influx of people can compromise the ability to deliver critical care (Stanfield, 2015). The operation of the healthcare system and the quality of the services offered beyond emergency departments require proper management of finite resources and effective care planning, underscoring the universal need for precise and efficient patient assessment. AI-driven systems can efficiently analyze and interpret complex data, which can support decision-making processes involving triage outcomes. The application of AI in medical triage with imaging data has shown promise in radiology workflows, including diagnosis and treatment assessment in mammography (Rodriguez-Ruiz et al., 2019) and brain-related illnesses (Titano et al., 2018; O'neill et al., 2020), where time is particularly critical. However, these AI-based tools are targeted to healthcare professionals. More broadly, AI can be used to conduct the initial assessment of a patient seeking care, providing quick and reliable guidance on the urgency and type of care needed (van der Stigchel et al., 2023; Karlafti et al., 2023). The advanced conversational capabilities that Large Language Models (LLMs) bring open new opportunities in the context of medical triage, especially for non-experts to communicate effectively about their health concerns.

LLMs foster the design of chat interfaces by facilitating conversational interactions with users, potentially enhancing the experience over more traditional and rigid online symptom checkers. LLM-powered conversational agents can provide explanations and tailored answers to specific questions non-expert users may have, complementing them with intelligible information to make the right decisions. Recent research efforts mostly focus on testing the capabilities of LLMs in more specialized domains, including medicine (Johri et al., 2023; Brin et al., 2023) because of the broad range of possibilities for LLM-powered chatbots (Lee et al., 2023). However, there is limited understanding about how LLMs shape the user behavior, including their interpretations or actionable decisions under this novel interaction paradigm. A key area that recently has started to receive considerable attention is the design of conversational agents based on LLMs that allow for the integration of distinct personality traits. Imbuing LLMs with such traits has the potential to significantly transform user interactions, making them more engaging and relatable. While previous research (Safdari et al., 2023) has attempted to embed personality traits in LLMs, there is a scarcity of studies evaluating the accuracy with which LLM-generated personas reflect these traits. Furthermore, the perception of these personas by users has not been thoroughly investigated.

In this work, we focus on the application of LLMs as decision support systems through a conversational interface for the preliminary assessment of patients seeking medical care. Conducting evaluations directly with patients experiencing discomfort poses ethical and practical challenges, while simulations where participants act as patients lack realism and engagement. To address these issues, we designed scenarios where participants advise a third party describing a patient's symptoms. Through an online user study, we investigated the impact of LLM-powered conversational agents as a novel way of interaction on user behavior and decision-making quality in a medical triage context. By employing prompting engineering to create conversational agents representing different personas, we aimed to understand how these personalized AI interactions influence user perceptions and choices during medical care decision support tasks. We identified novel opportunities for medical triage support regarding improvements in participants' confidence and abilities to identify correct triage levels. However, the effect of the personality of the agent was not observed in our data sample. We found that self-reported confidence on decisions plays an important role on adopting AI's advice provided via conversations.

2 Related work

2.1 LLMs in conversational agents

Decision support systems (DSS) are interactive systems that assist humans in making decisions when numerous complex variables are involved. Several studies have evaluated the suitability of conversational agents for decision support systems (Jo et al., 2023; Fadhil and Schiavo, 2019; Xiao et al., 2023). The recent development of LLMs, such as ChatGPT[1] and Bloom (Scao et al., 2022), has provided a new way for users to interact with DSS using natural language as in a conversation. LLMs have demonstrated comparable, if not superior, responses in question-answering tasks, such as responding to complex clinical questions (Ayers et al., 2023; Hopkins et al., 2023). In parallel, Wei et al. (2024b) studied how generating a chain of thought—a series of intermediate reasoning steps—significantly improves the ability of LLMs to perform real life tasks. Further, Gupta et al. (2022) explored how LLMs affect trust in DSS and found that conversational interfaces were significantly more effective in gaining users' trust than traditional web interfaces. However, Johri et al. (2023) evaluated ChatGPT's performance in medical assessments and found that placing the conversational AI in an interactive setting reduces diagnostic accuracy. This finding is in line with another research suggesting that knowledge provided in conversation can negate the knowledge learned by the AI (Zuccon and Koopman, 2023).

A chatbot-based symptom checker (CSC) is a particular integration of DSS and conversational AI, providing potential diagnoses through interactive interfaces by guiding users with a series of AI-driven questions (Montenegro et al., 2019). Previously (Cross et al., 2021), compared user's performance and experience in using traditional web search engines and traditional symptom checkers. The study reveals that symptom checkers are too constrained compared to search engines, which is useful for explorative hypothesis testing and differential diagnosis. CSCs enhance traditional symptom checkers by offering emotional support, tailored medical explanations, and flexibility previously lacking (You et al., 2023). For instance, a study found that interactive conversations with CSCs led to greater transparency and trust (Sun and Sundar, 2022). Tsai et al. (2021) enhanced a COVID-19 self-diagnosis system's transparency with three types of LLM-generated explanations, improving user experience by clarifying the model's reasoning and aiding in more informed decision-making. However, several studies have pointed out possible drawbacks of these conversations, specifically the issue of information overload

1 ChatGPT (2022). Available at: https://openai.com/blog/chatgpt.

(Fan et al., 2021; Ponnada, 2020), particularly in the context of no-experts users (Jiang et al., 2022). Although out-of-the-box LLMs may not be immediately effective in diagnosis and assessment, other research efforts have shown that these intelligent systems are favored by users for self-diagnosis purposes. Shahsavar et al. (2023) conducted a large-scale survey and found that 78.4% of the participants are willing to use ChatGPT for self-diagnosis and other health-related activities. In this work, we took a further step by evaluating actual interaction with a CSC powered by ChatGPT to better understand the outcomes of the human-AI interaction and human perceptions.

2.2 Personality traits in conversational agents

Personality has been a key focus of study as a precursor to human values (Parks-Leduc et al., 2015). Decades of research have demonstrated that personality traits are deeply embedded in human language (Goldberg, 1981; Digman and Takemoto-Chock, 1981). LLMs encapsulate extensive data on social, political, economic, and behavioral aspects and produce language that inherently reflects personality. Consequently, assessing and comparing the personality traits generated by LLMs offer potential benefits for ensuring LLM safety, responsible use, and alignment with ethical standards in its application. To date, efforts in this area have mainly concentrated on addressing specific harms, such as ensuring no explicit or implicit hate output is generated (Yuan et al., 2024), rather than exploring underlying behavioral tendencies of models.

People's perceptions of AI systems can significantly affect their behavior. Previous research in interactive systems and embodied agents has already shown that the mental structure people develop can influence the way they collaborate with an agent (Lee et al., 2010). Further, it has been found possible to identify a user's general behavior based on their initial engagement with these agents. LLMs, trained on extensive human data, exhibit synthetic personalities that can be easily noticed by the user. Safdari et al. (2023) proposed a framework to administer and validate LLMs' personality traits, demonstrating that personality assessments of instruction-finetuned LLMs are reliable and valid under specific prompts. Assigning LLMs with personality traits leads to more coherent responses and reduces harmful outputs, as indicated by Qian et al. (2018) and Safdari et al. (2023). Two personality traits stand out as the most prominent: personality trait related to emotion and personality trait related to reason. For instance, Gilad et al. (2021) studied the effects of warmth and competence perceptions on users' choice of an AI System and finds that a high-warmth system is generally preferred. Similarly, Sharma et al. (2023) has shown the importance of warmth and empathy in AI, with evidence that these qualities naturally enhance emotional support. Complementing this, frameworks have been developed to detect empathetic elements in AI dialogues (Sharma et al., 2020a; Pérez-Rosas et al., 2017).

Research into CSCs and personality traits in LLMs is still in its nascence. Specifically, prior works have mainly focused on assessing and evaluating LLMs personality traits, but the impact of LLMs' embedded personas on user behavior is not well-understood. This work seeks to harness LLMs' ability to emulate diverse personas to aid users in making triage-level decisions.

3 Methods

3.1 Study overview

Our research seeks to understand how LLM-powered conversational agent with rational and empathic persona profiles affects people's triage decisions.

We perform a between-subjects experiment, where we randomly assign participants to one of the two persona profiles that describes the conversational AI: rational agent or empathic agent. In addition to paralleling the experiments between two groups, a between-subjects experiment design avoids potential sequencing effects, where participants' behavior may be influenced by the order the personas are presented. After consenting to the study, participants consult with the AI characterized by the assigned persona profile. We adopt a sequential decision-making workflow where the agent's triage recommendation follows the participant's initial independent judgment, allowing for subsequent revisions (Tejeda et al., 2022). Participants first independently assess patient vignettes and will later make a final triage decision after interacting with the agent. This two-stage approach not only enables us to analyze how the AI's recommendations influence adjustments from their initial judgments but also to evaluate the enhancement in decision accuracy provided by the advice, using the initial decision accuracy as a benchmark or baseline accuracy in the triage task (Vodrahalli et al., 2022). Figure 1 depicts the experimental setup of the user study.

3.2 Shaping personality in LLMs: prompt design

We follow the guidelines for using ChatGPT's Application Programming Interface (API) to prompt the GPT-4 models. We start by specifying identities in the prompt using a template. An example is using different job identities for a set of health-related topics (Kojima et al., 2024; Wei et al., 2024a). In our case, we specified the role of a fellowship trained emergency physician (Johri et al., 2023). Importantly, modifications were made solely to the prompts input into the LLMs, with the underlying model remaining constant.

Taking cues from recent work on CSC (You et al., 2023), we incorporate two persona profiles—rational and empathic—in the conversation style by extending the prompt description. For the rational profile, we specify the characteristics in the prompts as being solution-oriented, having strong problem-solving skills, and being perceived as confident and knowledgeable. For the empathic profile, we describe the role that LLM is portraying as confidential, sensitive, and empathic and encourage it to be acknowledging, appreciating, and sympathizing following (Wei et al., 2024a) practices.

Then, we outline the task required of LLMs, comprising two main components: (1) The goal is to offer triage advice

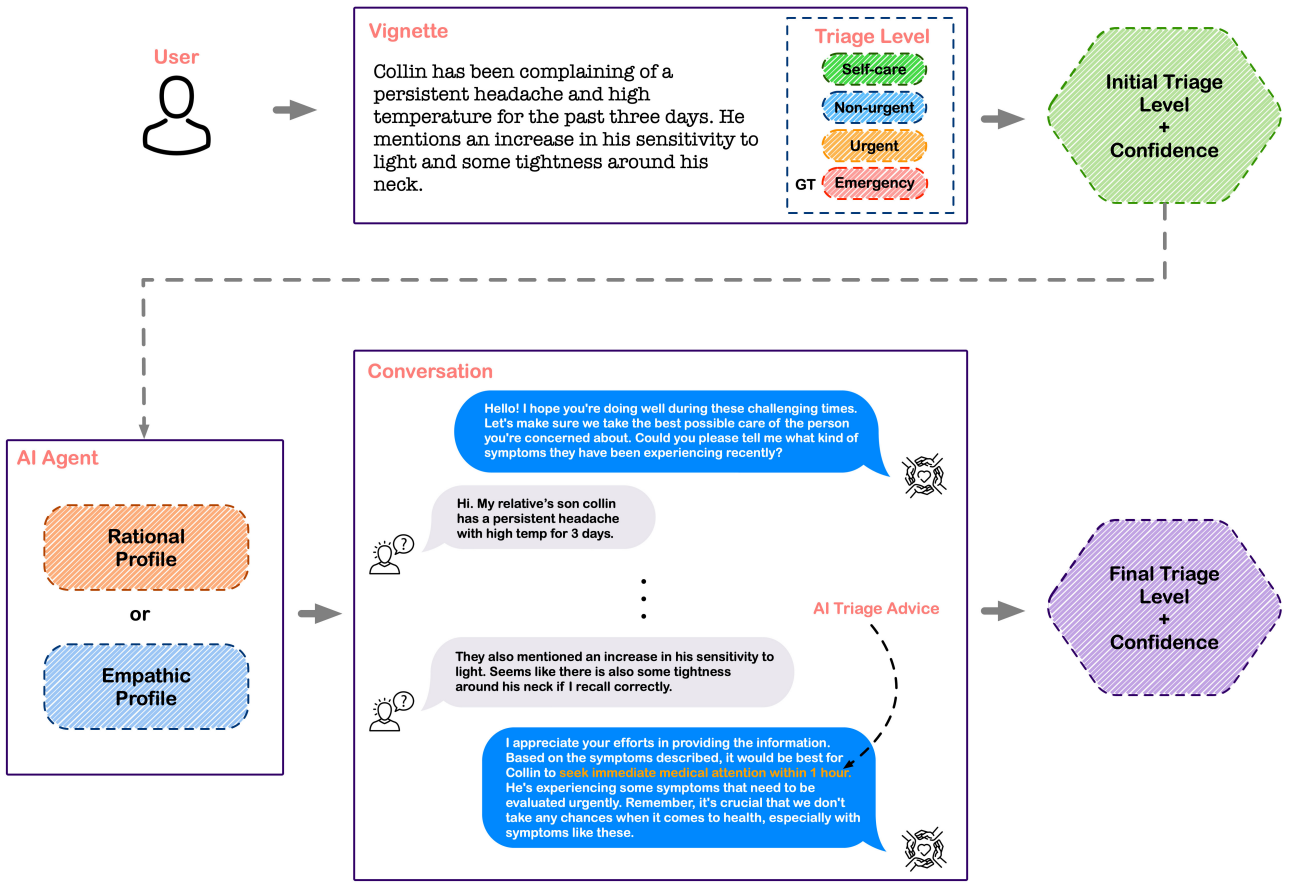

FIGURE 1
Experimental set-up: participants review a vignette where a third party is describing the symptoms of a patient that corresponds to one of the four triage levels. After providing an initial assessment of the triage level, participants interacted with the LLM-powered conversational agent that had either a rational or empathic profile. The conversation displays the interface used by participants to consult with the AI agent and receive triage advice. Lastly, participants provided their final triage decision and rated their confidence.

across four urgency levels; (2) The LLM must ask symptom-related questions to aid diagnosis. We add specific instructions for each persona type. For the rational persona, the LLM should adopt a rationale-based approach, providing justifications for each question's relevance to the triage recommendation. This aligns with studies advocating for increased transparency in symptom checkers through explanatory mechanisms (Tsai et al., 2021). For the empathic persona, the LLM should exhibit warmth and empathy, especially when posing questions.

To ensure consistent LLM behavior across participants, we established rules to structure the conversations. These rules were: (1) One question should be asked at a time, (2) A maximum of three questions per conversation is allowed, and 3) After making a triage decision, no further justifications should be provided.

We did not control the quality of the agent's triage advice since the inputs that participants provided during the conversations guided the agent's advice generation process. To further control the quality of the LLM's output, we leverage its one-shot learning capability (Brown et al., 2020) by providing a task demonstration at inference time. For more details about the prompts and the accuracy of the agent's triage advice, see Appendix A. Samples of the conversations showcasing the differing approaches and conversational styles of each persona profile are shown in Figure B.1 in Appendix B.

3.3 Triage scenarios construction

To construct the triage scenarios used in our study, we referenced the 45 standardized patient vignettes previously used to assess performance of symptom checkers (Semigran et al., 2015). These vignettes were identified from various clinical sources with the associated correct diagnosis. The description of the cases includes age, gender, symptoms, medical history data, etc. The cases in the vignettes span across the severity spectrum covering three categories of triage urgency: emergent care is required, non-emergent care is reasonable, and self care is sufficient.

From the initial set of vignettes, we created a more fined-grained version by further segmenting the diagnoses into four triage categories: requiring emergency care, urgent care, non-urgent care, and self-care appropriate. This refinement process is done with the help of clinicians. We refined a total of eight vignettes, ensuring equal representation across triage levels with two scenarios from each of the four triage categories, and adapted

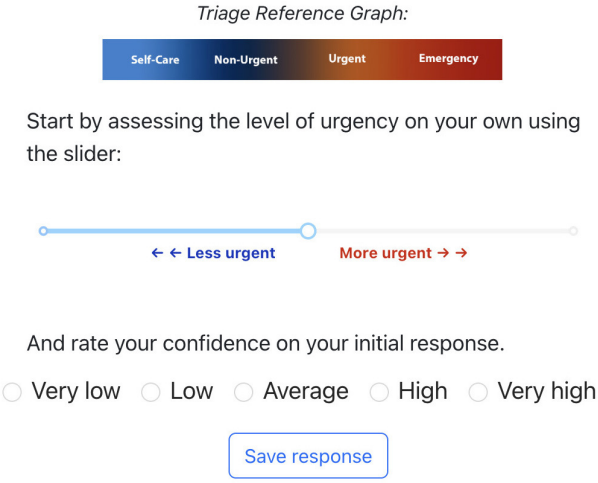

FIGURE 2
Interface for participants to rate the urgency level of each vignette. The triage reference bar was provided as a guide for people to adjust the slider.

them into immersive vignettes. Rather than listing generic patients' symptoms, immersive vignettes provided relatable context—for instance, a scenario might describe the patient as a friend's child, making the situation more personal and engaging for participants. The intention is to frame it in a way that invokes urgency, responsibility, and connection with the participants (Tilt, 2016). Although the immersive vignettes might not precisely reflect the participants' genuine information requirements, this kind of approach is by no means uncommon: numerous studies employ a similar method to enable control over the experiment conditions and comparison of results across participants (Li and Belkin, 2010; Kelly et al., 2015).

Following Cross et al. (2021) approach, we changed specialized medical terminology with more understandable layman expressions where suitable (e.g., "rhinorrhea" was replaced with "runny/stuffy nose", "abdominal pain" was replaced with "stomachache"). Three examples of the triage scenarios are provided in Table 1.

The experimental task required participants to provide a triage decision for the presented scenario before and after engaging with the LLM-powered conversational agent. To capture the full information about participants' decision making, we asked them to use a slider with a range from zero to one hundred (refer to Figure 2), in which the triage levels are represented by a set of ranges (0–25 = self-care, 25–50 = non-urgent, 50–75 = urgent, 75–100 = emergency).

3.4 Measures

3.4.1 Manipulation check

In order to perform manipulation check on LLM's persona profiles, we ask participants the following two question on 5-point Likert scales after interacting with the corresponding agent in each vignette: (1) *"I felt like the AI assistant was responsive and warm"* (2) *"I felt like the AI assistant was analytical and clearly explained its reasoning"*. These questions were informed by constructs used in previous research that evaluated personalized and mechanical conversations when patients make health-related decisions (Yun et al., 2021). In addition, we use a framework for characterizing the communication of empathy in text-based conversations (Cuadra et al., 2024; Sharma et al., 2020b). The publicly available empathy classifier is a Bidirectional Encoder Representations from Transformers (BERT) model trained on *post-response* pairs from the internet. The model outputs ratings for three communication mechanisms associated with empathy, namely, emotional reactions, interpretations, and explorations, with a score that indicates no communication (0), weak communication (1), and strong communication (2). We analyzed the conversations recorded from our participants by splitting each chat into post-response pairs, where the former corresponds to the participant's input and the latter to the agent's output. We add the scores across the three dimensions for each pair and then calculate an average score per conversation for each patient vignette, such that we have multiple measures per participant to conduct the statistical analysis.

3.4.2 Objective measures

- ***Weight of advice:*** First, we want to understand the influence of the LLM-powered agent's advice on triage decisions. For this, we take cues from recent research (Panigutti et al., 2022) and employ the weight of advice (WoA), a common metric used in psychology to capture the extent to which the algorithmic suggestion (with or without AI) affects the participant's estimate (Yaniv and Foster, 1997). The WoA is described as:

$$\text{WoA} = \frac{\text{Decision}_{\text{post}} - \text{Decision}_{\text{pre}}}{\text{Advice}_{\text{AI}} - \text{Decision}_{\text{pre}}} \quad (1)$$

where $\text{Decision}_{\text{pre}}$ and $\text{Decision}_{\text{post}}$ are the participant's decision before and after engaging with conversational AI, respectively, and $\text{Advice}_{\text{AI}}$ is the advice generated by the LLM that is presented to the user. Given that such advice is categorized into four distinct triage levels, each defined by a specific interval (0–25, 25–50, 50–75, 75–100), we convert these categories into numerical values by calculating the midpoint of their respective intervals. We calculate the WoA by taking the magnitude of both the numerator and the denominator as outlined in related work (Panigutti et al., 2022). A positive WoA value indicates that a participant adjusted their decision after considering the LLM-powered agent's advice. The closer the value is to one, the more significant the influence of the agent's advice on the participant's final response. A value of zero is observed when there is no change in the participant's response. In this metric, we choose to make the discrete triage advice from the AI agent continuous to capture more fine-grained differences when a participant makes their decision, given that within each triage level there is still a hierarchy of urgency depending on the nature of the symptom.

TABLE 1 Example triage scenarios provided to users.

#	Triage scenarios	Diagnosis	Triage
1	Your friend texted you that his sister's 4-year-old son Timmy, who is visiting for a weekend trip, has started feeling unwell. Timmy has been dealing with a tummy ache and diarrhea, which started to show some blood traces after the first day...	Hemolytic uremic syndrome	Emergency
2	Your friend Sarah brought her 5-month-old son, Liam, to your place for a weekend getaway. During the stay, Sarah noticed that Liam seems to have some trouble passing stool...Liam didn't have any trouble passing stool when he was born, and there was no history of excessive vomiting, bloating or other health issues.	Constipation	Non-urgent
3	You invited your brother-in-law for a weekend dinner and he cannot make it. His 12-year-old daughter Laura is experiencing a somewhat troublesome skin condition. Over the phone, he said that Laura has been dealing with areas of particularly dry and itchy skin, localized mainly to the insides of her elbows, behind her knees, and on her ankle...	Eczema	Self-care

Each scenario is labeled with ground truth diagnosis and fine-grained triage level.

- ***Change in decision:*** Since our study measures triage decisions in numerical form as opposed to in category, we are able to examine more closely how AI influences the decisions of the participants. Specifically, we define two kinds of changes: *change in decision within the category* and *change in decision between categories*. Changes within the category suggest that the participant changed their decision after interacting with the LLM-powered conversational agent, but the change is within the range that represents the triage level (e.g., participant change their decision from 0 to 15, $0,15 \in$ Self-care). While changes within the category do not change the action corresponding to that triage category, they serve as a useful measure to better inform us about the participant's perception of triage urgency. Changes between the category suggest that the participant changed their triage level from one to another (e.g., a participant changes their decision from 0 to 45, $0 \in$ Self-care, $45 \in$ Non-Urgent).
- ***Task performance:*** Task performance is measured as the accuracy of triage decisions across all category levels. To determine each participant's overall accuracy, we first classified each triage decision as accurate or inaccurate by comparing it to the vignette's ground truth. This process involved converting participants' numerical responses into corresponding triage levels using the four predefined intervals. Subsequently, we calculated accuracy as the proportion of trials correctly identified out of five. Accuracy of triage advice from the agents can be calculated using this approach, i.e., as the proportion of vignettes in which the agent provided the correct triage level. Decisions made before interacting with the LLM-powered conversational agent form the pre-AI or baseline accuracy. Similarly, decisions after the conversation form the post-AI accuracy. For a more comprehensive report of participants' and agents' performance, we calculate precision and recall independently for each triage category (and then take the average) by combining triage outcomes as true positives, false positives, or false negatives within a specific group: all triage advice generated from interactions with the rational or empathic agent, and all triage decisions from participants grouped by the agent profile or the stage of decision making (pre- or post-AI).

Subjective measures: To investigate changes in participants' experience across conditions, we include subjective measures to gauge the differences in the perceived interactions with the the LLM-powered conversational agents using 5-point Likert scales.

- ***Quality:*** To probe for participants' perception about the quality and effectiveness of AI's advice, we ask them the following two questions: (1). *"I like the recommendations provided by the AI system."* (2). *"The system provides good recommendations for me."*
- ***Trust:*** To measure participant's trust on the AI agent, we ask them the following two questions (Körber, 2019): (1). *"I feel like the AI system can be trusted."* (2). *"I am convinced by the recommendations that the system provided to me."*
- ***Satisfaction:*** To measure participant's satisfaction on the AI agent, we ask them the following two questions (Tsai, 2019): (1). *"Overall, I am satisfied with the system."* (2). *"I feel like I will use this system again."*

3.5 Procedure

Participants, after providing their demographic information including age, gender, education level, and familiarity with AI, were briefed on the main task involving five patient scenarios for triage. Four of these scenarios were randomly selected from our pool of eight vignettes, with one scenario chosen from each of the four triage levels. These four were fixed for all participants. The fifth scenario was randomly selected from the remaining four vignettes. This approach ensured that each participant evaluated two cases from one triage level and one case from each of the other three levels. The tasks were completed in random order. The two experimental groups differed only in how the LLMs responded based on assigned personas. Participants made initial and final triage recommendations, rating their confidence on a 5-point Likert scale for each one. After completing the five vignettes, participants responded to a questionnaire evaluating their experience with the agent, including perceived quality, trust, and satisfaction. The study concluded with a rating of their prior exposure to ChatGPT. Additional details and questionnaires can be found in the Appendix A. This user study was approved by our institutional review board.

3.6 Participants

We recruited 60 participants via online distribution of the study within the U.S. The study was prompted via local university channels and Reddit groups. The study required participants to be above the age of 18 and to be proficient in English. Participants were informed that the study would take ∼30 min, and upon completion, they could sign up to receive a $20 gift card. After considering agreement to participate in the study through informed consent, 52 participants completed the study. A summary of the demographic information is shown in Table 5 in Appendix A. For the data analysis, we included 49 valid participants after filtering data samples in which participants' inputs to the conversations were empty as this may indicate lack of task attention.

3.7 Statistical analysis

In our statistical analysis of continuous variables with one observation per participant, we initially applied independent samples t-tests. For continuous variables with repeated measures within participants and one observation per participant and condition, we utilized a two-way mixed model Analysis of Variance (ANOVA) with the agent profile as the between subjects effect and the time point as the within subjects effect (before or after the interactions). In cases of continuous variables with repeated observations per participant (between agents), a mixed effects linear regression was applied to include a covariate (participants' initial confidence). For categorical and binary response variables to capture decision changes, we used logistic regressions. For Likert scale data (ordinal data) concerning multiple observations per condition, we employed an Aligned Rank Transform mixed model ANOVA. Additionally, for Likert scale data with one observation per participant and independent samples, such as the subjective measures, we used the non-parametric Mann-Whitney Test. We assessed the assumptions of normality and homogeneity of variance of the parametric tests using the Shapiro-Wilk and Levene's tests, respectively. The non-parametric alternative of the test was used if any of the assumptions was not satisfied. We followed Cohen's guidelines (Cohen, 1988) on effect sizes and considered $\eta_p^2 = 0.01$ a small effect size, $\eta_p^2 = 0.06$ a medium effect size, and $\eta_p^2 = 0.14$ a large effect size. For Cohen's index, 0.2 is considered small, 0.5 medium, and 0.8 a large effect. Participant ID was included as a random effect to address repeated measurements and individual variability when needed. Statistical significance was set at $\alpha < 0.05$.

4 Results

From the valid data samples of 49 participants, the distribution over the experimental groups was $n = 21$ in the rational-based agent and $n = 28$ in the empathic-based agent, and a total of 245 data samples.

4.1 Manipulation check: can participants perceive different agent profiles as they provide triage recommendations?

We determined whether participants perceived distinct traits in different persona profiles. This involved evaluating their perception of the agent's rational-related and empathy-related behavior. To assess participants' overall perception of the agent they interacted with, we calculated separate averages for ratings of rational-related and empathy-related behaviors across the five trials. Specifically, we measured the agent's profile impact on these perceptions using independent t-tests. After validating normality and homogeneity of variance ($p > 0.05$), we found that participants gave higher rational-related ratings to the rational agent than to the empathic agent [$t_{(46)} = 2.26, p = 0.028$], with an estimated score difference of 0.42 points ($M = 4.13, SD = 0.59$ for the rational agent and $M = 3.71, SD = 0.73$ for the empathic agent). The effect size, as measured by Cohen's d, was $d = 0.63$, indicating a medium effect. For empathy ratings, the homogeneity of variance was satisfied ($p > 0.05$) while normality was not ($p < 0.05$). Therefore, we utilized the non-parametric alternative of an independent t-test. The empathic agent scored slightly higher ($M = 4.21, SD = 0.65$) than the rational agent ($M = 3.99, SD = 0.65$), but the difference was not significant according to the Mann-Whitney test ($W = 225, p = 0.160$). The analysis of conversations recorded under each agent profile using the empathy classifier suggests that the agent's profile significantly affected empathy ratings. Since both the normality and homogeneity of variance were not satisfied, a Mann-Whitney test ($W = 0, p < 0.001$) showed that on average, empathy scores were higher in conversations with the empathic agent ($M = 2.85, SD = 0.15$) than the rational one ($M = 1.81, SD = 0.09$) with a large effect size ($r = 0.85$). Table 6 in Appendix B presents some examples of conversation fragments and the empathy-related ratings.

4.2 How are participants influenced by the agent profile when making triage recommendations?

4.2.1 Confidence on triage recommendations

We compared confidence ratings given by participants for triage recommendations before and after interacting with agents of different profiles, analyzing each trial's measurements individually. Since confidence was measured using a Likert scale, we applied an Aligned Rank Transform (ART) ANOVA, a non-parametric alternative to traditional factorial ANOVA, better suited for this type of data (Smith-Renner et al., 2020). We defined the agent profile as the between subjects effect and the time point (before or after the agent's advice) as the within subjects effect, and included their interaction effect as well. No significant effect of the agent profile on confidence was found [$F_{(1,47)} = 0.04, p = 0.845$], with average ratings being 3.70 ($SD = 0.87$) for the rational and 3.74 ($SD = 0.85$) for the empathic agents. However, the time point significantly affected confidence [$F_{(1,47)} = 7.37, p = 0.009, \eta_p^2 = 0.14$], resulting in higher confidence ratings after participants

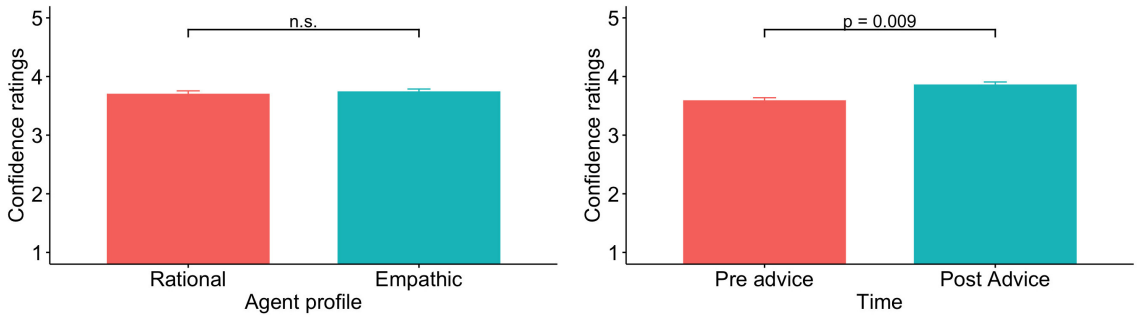

FIGURE 3
Bar plots of confidence across the two persona profiles and before and after interacting with the agent.

TABLE 2 Linear mixed model for confidence in triage recommendations and weight of advice.

Outcome	Predictor	Estimate	Standard error	t	p-value
Weight of advice	Intercept	1.25	0.20	6.42	<0.001
	Agent profile (empathic)	0.02	0.09	0.25	0.803
	Confidence rating	−0.18	0.05	−3.55	<0.001

Random effects: $\sigma^2 = 0.372$, $N_{ID} = 49$, observations= 226, marginal $R^2 = 0.06$, conditional $R^2 = 0.135$.

interacted with the agent ($M = 3.85, SD = 0.84$) compared to their initial confidence perception ($M = 3.58, SD = 0.85$), regardless of its profile. Figure 3 illustrates these confidence changes. There was no significant interaction effect [$F_{(1,47)} = 0.05, p = 0.817$].

4.2.2 Weight of advice

We examined how the agent profile delivering triage advice to participants affected their triage decisions using the WoA metric across trials without averaging. To handle repeated observations per participant and explore how initial confidence ratings impacted the WoA, we employed a mixed effects linear regression. This model included two predictors: the agent profile and initial confidence, with participants treated as a random effect. We identified and removed outliers using the interquartile range to improve the model fit (226/245 trials). Table 2 summarizes the results of the linear regression with the WoA as the outcome and the two predictors. The effect of the agent profile on the WoA was statistically non-significant ($p = 0.803$). On average, WoA values were >0.5 for both agents: 0.62 ($SD = 0.69$) for the rational agent and 0.63 ($SD = 0.62$) for the empathic one. Figure 4 shows the WoA distribution separated by the agent profile. The negative coefficient for the initial confidence factor [$-0.18, 95\%CI(-0.27, -0.08)$] and its significance suggest that higher WoA values were generally associated with lower confidence ratings reported by participants in their initial triage response ($p =< 0.001$). In trials where participants altered their initial assessments (WoA$\neq$ 0), we noted a similar pattern in the Change in Decision model. The initial confidence had a negative coefficient, and the odds ratio for changes across different triage levels, compared to within the same category, was less than one. However, this effect was close to significant. More detailed results are presented in Appendix C.

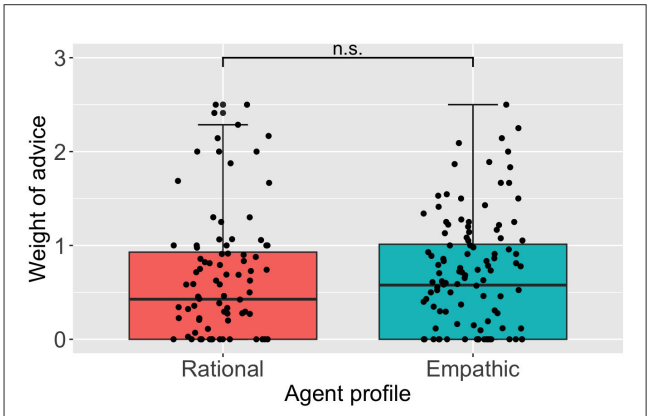

FIGURE 4
Box and whisker plot for the weight of advice values across the two persona profiles. Values closer to zero mean a lower influence of the advice.

4.3 Does the interaction with different agents affect participants' abilities to identify the correct triage level?

We tested whether participants' accuracy in identifying the correct triage level was influenced by the agent profile and if the agent's triage advice resulted in accuracy improvements. Thus, we measured the effect of the agent profile, time point (before or after the agent's advice), and their interaction on the accuracy of the triage levels using a two-way mixed model ANOVA. Because the normality assumption was not met ($p < 0.05$), we calculated these effects using the non-parametric alternative ART ANOVA. We found a non-significant effect of the agent profile on the accuracy of triage recommendations [$F_{(1,47)} = 0.44, p = 0.509$]. Overall,

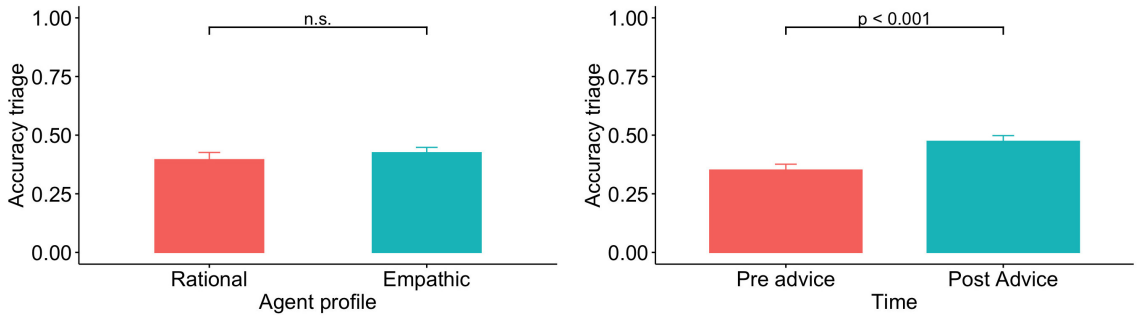

FIGURE 5
Bar plots of triage accuracy across the two persona profiles and before and after interacting with the agent.

participants showed a significantly higher accuracy of triage levels after they received the agent's advice ($M = 0.47, SD = 0.17$) compared to participants' average accuracy before interacting with the agent ($M = 0.35, SD = 0.18$), [$F_{(1,47)} = 23.46, p < 0.001, \eta^2 = 0.33$]. Figure 5 presents the accuracy of triage decisions. When looking at the accuracy before interacting with the agent separated by the agent profile, participants' average accuracy was 0.32 ($SD = 0.18$) and 0.37 ($SD = 0.17$) in the rational and empathic agent groups, respectively. The average accuracy changed to 0.47 ($SD = 0.19$) and 0.48 ($SD = 0.16$) after participants interacted with the rational and empathic agents, respectively. However, these differences across agents are non-significant as suggested by the interaction term [$F_{(1,47)} = 0.00, p = 0.982$]. Table 9 in Appendix D.2 presents precision and recall values of participants' triage decisions across different agents and before and after the interaction. Metrics at the category level and their average are improved after receiving advice from the agent. We also quantified the agents' performance and present evaluation metrics for the triage advice provided by agents with different profiles in Table 8 Appendix D.1.

4.4 Does the interaction with different agent profiles affect the quality, trust and satisfaction perceived during the interaction?

We examined whether participants' ratings of quality of advice, trust, and satisfaction with the agent were affected by the profile of the agent they interacted with. Table 3 presents the descriptive statistics and results of the Mann-Whitney tests. We could only find non-significant differences in subjective ratings between the rational and empathic agents.

TABLE 3 Descriptive statistics (mean and standard deviation) and Mann-Whitney U test results for the subjective measures.

Metric	Rational-based agent	Empathic-based agent	Test result
Quality	4.00 (0.81)	3.84 (0.82)	$W = 321.5, p = 0.557$
Trust	3.55 (0.82)	3.52 (0.84)	$W = 304, p = 0.842$
Satisfaction	3.76 (0.83)	3.75 (0.94)	$W = 294, p = 1.00$

5 Discussion

We empirically evaluated the influence of LLM-powered conversational agents on human decision making in a medical triage scenario, focusing on how design choices of the AI's persona profile and advice affect the user experience.

Our study suggests that LLM-powered conversational agent personas can be integrated into LLMs providing triage advice. Participants could differentiate that the rational agent explained its reasoning and was more analytical than the empathic one, but the perception of empathy-related attributes like responsiveness and warmth did not significantly differ between agents. Despite distinct personality traits and different conversational styles, which was captured by the empathy classifier, the empathic agent's use of empathetic expressions and the rational agent's analytical behavior did not lead to a marked difference in perceived empathy. Even though the empathic agent included expressions such as "*I'm sorry to hear about ...*" or "*It's challenging to see anyone go through such discomfort.*" during the conversations and the rational one did not, the latter's personality was not inherently unempathetic. Hence, participants might have rated the rational agent higher in empathy than anticipated, blurring the distinction with the empathic agent.

The effect of the persona profile did not result in statistically significant differences in our dependent variables. Our results suggest that conversational style—specifically between rational and empathetic personalities—might not substantially affect user experience in terms of advice-taking trends, and perceived satisfaction, likability, and usefulness, aligning with similar findings in other studies (You et al., 2023). This could be attributed to the subtlety of LLMs personality perception and interpretation, indicating that minor variations in agent profiles may not significantly influence users' behavioral responses or experience perception. Previous research on CSC has explored personality traits, but not with the complexity of LLMs as in our study (You et al., 2023; Tsai et al., 2021). The intricate nature of LLMs makes controlling outputs more challenging, potentially blurring conversational styles. Furthermore, our interaction context, which

involved users providing advice, might have led them to focus more on content than on the agent's personality traits.

After participants interacted with the LLM-powered agent, regardless of the persona profile, they showed an improvement in their accuracy in identifying the correct triage level. On average, the AI-recommended triage levels were more accurate (rational agent: $M = 0.55, SD = 0.19$, empathic agent: $M = 0.50, SD = 0.19$) than the participants' initial decisions, indicating potential benefits from their interaction with AI in refining their choices. However, the agents failed to predict the self-care category, while participants were able to identify a few such cases. This category appeared to be the most challenging to correctly triage. The quality of algorithmic recommendations can influence how people interpret and incorporate them (Yin et al., 2019). Our analysis of the agents' triage advice performance showed no evidence that persona profiles affected the advice quality globally, potentially explaining why the persona profile did not impact the user experience overall. Because the AI's recommendations present flaws, people need to judge when to accept or reject its advice, particularly to avoid unintended harm. Therefore, implementing mechanisms to calibrate trust appropriately is essential in enabling users to make well-informed decisions (Nourani et al., 2020). Augmenting standard LLM frameworks with targeted medical domain data can enhance their capabilities for clinical decision making (Zakka et al., 2024; McDuff et al., 2023). In this study, we did not employ fine-tuning techniques on the LLMs; instead, we exclusively leveraged refined prompting strategies to guide the behavior of the models. Furthermore, the low baseline accuracy of participants suggests that determining the appropriate urgency of care is a challenging task. Even though our evidence showed that participants improved their accuracy and confidence after interacting with the AI agent, the average performance across all four triage categories remained below fifty percent. For a more proper assessment of performance at each triage level, participants would need exposure to more cases, which will extend the duration and workload of the study since each case requires a separate conversation with the AI agent.

Confidence is positively influenced by the intervention of the LLM-powered conversational agent, leading to more assured decisions post-interactions. On average, self-reported confidence was below four on the Likert scale, regardless the agent profile and even after the intervention, indicating an opportunity to enhance how confident participants feel after interacting with the agent. Besides, our findings indicate that participants' confidence in their initial triage assessment impacts the weight of AI advice. Higher initial confidence correlates with lower weight of advice values. Users with higher reported confidence levels might feel more certain and therefore make smaller changes regardless of the agent persona profile, while less confident users are more susceptible to being influenced by the agent's advice. This pattern has been pointed out as troublesome in the context of conversational generative AI because these provide answers to users' questions upon request when they are dealing with uncertainty (Kidd and Birhane, 2023). The significance of a control variable highlights two aspects: (1) its importance in understanding the influence of LLMs' advice and (2) the need to design interactions where LLMs aim not simply to persuade users, but to better inform them with correct information.

In interpreting our results, we acknowledge that our study presents some limitations. First, while the simulated scenarios were suitable for our objectives, they may not fully reflect the complexities and stress of real medical situations, which can notably affect AI interactions. Second, our participant sample predominantly consisted of individuals under 25 years old, female, and with a decent familiarity with AI and ChatGPT. This demographic may not accurately represent the typical end users of conversational agents for medical triage. Third, since we did not impose any limit on participants' input format during their interaction with LLMs, some participants provided empty or monosyllable responses mostly. In any of these cases, we cannot differentiate participants who were not engaged during the experiment. Lastly, the structured nature of our conversations, resulting from our specific prompting strategy, could have constrained the expression of the predefined personalities and conversation dynamics.

To conclude, the potential of LLM with its more sophisticated capabilities to enhance domestic healthcare triage systems requires careful design and evaluation of human interactions for a successful implementation. Our findings indicate that while LLM-powered conversational agents can positively influence user confidence and alignment with correct decisions, the variation in agent personalities does not significantly impact decision-making processes. This suggests a complex interplay between LLM-powered AI conversational styles and human decision-making, highlighting the need for further exploration in this domain.

Data availability statement

The datasets presented in this article are not readily available because in the current IRB protocol, permission has not been granted for data collected in the study to be made publicly available. Requests to access the datasets should be directed to cgomezc1@jhu.edu. Access will be granted on condition of approval from the IRB.

Ethics statement

The studies involving humans were approved by Johns Hopkins University Homewood Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin because participants were provided with the informed consent on a website since the study was conducted online. Upon acceptance, they clicked on a button that directed them to the main study.

Author contributions

CG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. JY: Conceptualization,

Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. C-MH: Conceptualization, Investigation, Supervision, Writing – review & editing. MU: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was in part supported by internal funds of Johns Hopkins University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2024.1427463/full#supplementary-material

References

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Int. Med.* 183, 589–596. doi: 10.1001/jamainternmed.2023.1838

Brin, D., Sorin, V., Vaid, A., Soroush, A., Glicksberg, B., Charney, A., et al. (2023). Comparing ChatGPT and GPT-4 performance in usmle soft skill assessments. *Sci. Rep.* 13:16492. doi: 10.1038/s41598-023-43436-9

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20* (Red Hook, NY: Curran Associates Inc.).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Cross, S., Mourad, A., Zuccon, G., and Koopman, B. (2021). "Search engines vs. symptom checkers: A comparison of their effectiveness for online health advice," in *Proceedings of the Web Conference 2021, WWW '21* (New York, NY: Association for Computing Machinery), 206–216. doi: 10.1145/3442381.3450140

Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., et al. (2024). "The illusion of empathy? notes on displays of emotion in human-computer interaction," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24* (New York, NY: Association for Computing Machinery), 1–18. doi: 10.1145/3613904.3642336

Digman, J. M., and Takemoto-Chock, N. K. (1981). Factors in the natural language of personality: re-analysis, comparison, and interpretation of six major studies. *Multivariate Behav. Res.* 16, 149–170. doi: 10.1207/s15327906mbr1602_2

Fadhil, A., and Schiavo, G. (2019). Designing for health chatbots. *arXiv* [preprint]. doi: 10.48550/arXiv.1902.09022

Fan, X., Chao, D., Zhang, Z., Wang, D., Li, X., and Tian, F. (2021). Utilization of self-diagnosis health chatbots in real-world settings: case study. *J. Med. Int. Res.* 23:e19928. doi: 10.2196/19928

Gilad, Z., Amir, O., and Levontin, L. (2021). "The effects of warmth and competence perceptions on users' choice of an AI system," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21* (New York, NY: Association for Computing Machinery), 1–13. doi: 10.1145/3411764.3446863

Goldberg, L. R. (1981). Language and individual differences: the search for universals in personality lexicons. *Rev. Person. Soc. Psychol.* 2, 141–165.

Gupta, A., Basu, D., Ghantasala, R., Qiu, S., and Gadiraju, U. (2022). To trust or not to trust: How a conversational interface affects trust in a decision support system. *Proc. ACM Web Conf.* 2022, 3531–3540. doi: 10.1145/3485447.3512248

Hopkins, A. M., Logan, J. M., Kichenadasse, G., and Sorich, M. J. (2023). Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spect.* 7:pkad010. doi: 10.1093/jncics/pkad010

Jiang, J., Kahai, S., and Yang, M. (2022). Who needs explanation and when? Juggling explainable ai and user epistemic uncertainty. *Int. J. Hum. Comp. Stud.* 165:102839. doi: 10.1016/j.ijhcs.2022.102839

Jo, E., Epstein, D. A., Jung, H., and Kim, Y.-H. (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, CHI '23* (New York, NY: Association for Computing Machinery), 1–16. doi: 10.1145/3544548.3581503

Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Cai, Z. R., et al. (2023). Testing the limits of language models: a conversational framework for medical AI assessment. *medRxiv* [preprint]. doi: 10.1101/2023.09.12.23295399

Karlafti, E., Anagnostis, A., Simou, T., Kollatou, A. S., Paramythiotis, D., Kaiafa, G., et al. (2023). Support systems of clinical decisions in the triage of the emergency department using artificial intelligence: the efficiency to support triage. *Acta Med. Lituanica* 30, 19–25. doi: 10.15388/Amed.2023.30.1.2

Kelly, D., Arguello, J., Edwards, A., and Wu, W.-C. (2015). "Development and evaluation of search tasks for iir experiments using a cognitive complexity framework," in *Proceedings of the 2015 International Conference on the Theory of Information Retrieval, ICTIR '15* (New York, NY: Association for Computing Machinery), 101–110. doi: 10.1145/2808194.2809465

Kidd, C., and Birhane, A. (2023). How ai can distort human beliefs. *Science* 380, 1222–1223. doi: 10.1126/science.adi0248

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2024). "Large language models are zeroshot reasoners," in *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22* (Red Hook, NY Curran Associates Inc.).

Körber, M. (2019). "Theoretical considerations and development of a questionnaire to measure trust in automation," in *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20* (Springer), 13–30.

Lee, M. K., Kiesler, S., and Forlizzi, J. (2010). "Receptionist or information kiosk: how do people talk with a robot?," in *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work, CSCW '10* (New York, NY: Association for Computing Machinery), 31–40. doi: 10.1145/1718918.1718927

Lee, P., Bubeck, S., and Petro, J. (2023). Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New Engl. J. Med.* 388, 1233–1239. doi: 10.1056/NEJMsr2214184

Li, Y., and Belkin, N. J. (2010). An exploration of the relationships between work task and interactive information search behavior. *J. Am. Soc. Inf. Sci. Technol.* 61, 1771–1789. doi: 10.1002/asi.21359

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., et al. (2023). Towards accurate differential diagnosis with large language models. *arXiv* [preprint]. doi: 10.48550/arXiv.2312.00164

Montenegro, J. L. Z., da Costa, C. A., and da Rosa Righi, R. (2019). Survey of conversational agents in health. *Exp. Syst. Appl.* 129, 56–67. doi: 10.1016/j.eswa.2019.03.054

Nourani, M., King, J., and Ragan, E. (2020). "The role of domain expertise in user trust and the impact of first impressions with intelligent systems," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Volume 8*, 112–121.

O'neill, T. J., Xi, Y., Stehel, E., Browning, T., Ng, Y. S., Baker, C., et al. (2020). Active reprioritization of the reading worklist using artificial intelligence has a beneficial effect on the turnaround time for interpretation of head CT with intracranial hemorrhage. *Radiology* 3:e200024. doi: 10.1148/ryai.2020200024

Panigutti, C., Beretta, A., Giannotti, F., and Pedreschi, D. (2022). "Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22* (New York, NY: Association for Computing Machinery). doi: 10.1145/3491102.3502104

Parks-Leduc, L., Feldman, G., and Bardi, A. (2015). Personality traits and personal values: a meta-analysis. *Person. Soc. Psychol. Rev.* 19, 3–29. doi: 10.1177/1088868314538548

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., and An, L. (2017). "Understanding and predicting empathic behavior in counseling therapy," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds R. Barzilay and M. -Y. Kan (Vancouver, BC: Association for Computational Linguistics), 1426–1435. Available at: https://aclanthology.org/P17-1131

Ponnada, S. (2020). "Reimagining the covid-19 digital experience: the value of user empowerment and accessibility in risk communication," in *Proceedings of the 38th ACM International Conference on Design of Communication, SIGDOC '20* (New York, NY: Association for Computing Machinery), 1–3. doi: 10.1145/3380851.3418619

Qian, Q., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018). "Assigning personality/profile to a chatting machine for coherent conversation generation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI '18* (AAAI Press), 4279–4285.

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Teuwen, J., Broeders, M., Gennaro, G., et al. (2019). Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur. Radiol.* 29, 4825–4832. doi: 10.1007/s00330-019-06186-9

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., et al. (2023). Personality traits in large language models. *arXiv* [preprint]. doi: 10.21203/rs.3.rs-3296728/v1

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, B., et al. (2022). Bloom: a 176b-parameter open-access multilingual language model. *arXiv* [preprint]. doi: 10.48550/arXiv.2211.05100

Semigran, H. L., Linder, J. A., Gidengil, C., and Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 351:h3480. doi: 10.1136/bmj.h3480

Shahsavar, Y., and Choudhury, A. (2023). User intentions to use chatgpt for self-diagnosis and health-related purposes: cross-sectional survey study. *JMIR Hum. Fact.* 10:e47564. doi: 10.2196/47564

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., and Althoff, T. (2023). Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* 5, 46–57. doi: 10.1038/s42256-022-00593-2

Sharma, A., Miner, A. S., Atkins, D. C., and Althoff, T. (2020a). A computational approach to understanding empathy expressed in text-based mental health support. *arXiv* [preprint]. doi: 10.18653/v1/2020.emnlp-main.425

Sharma, A., Miner, A. S., Atkins, D. C., and Althoff, T. (2020b). "A computational approach to understanding empathy expressed in text-based mental health support," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, eds B. Webber, T. Cohn, Y. He, and Y. Liu (Association for Computational Linguistics), 5263–5276. Available at: https://aclanthology.org/2020.emnlp-main.425

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., et al. (2020). "No explainability without accountability: an empirical study of explanations and feedback in interactive ML," in *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems, CHI '20* (New York, NY: Association for Computing Machinery), 1–13. doi: 10.1145/3313831.3376624

Stanfield, L. M. (2015). Clinical decision making in triage: an integrative review. *J. Emerg. Nurs.* 41, 396–403. doi: 10.1016/j.jen.2015.02.003

Sun, Y., and Sundar, S. S. (2022). "Exploring the effects of interactive dialogue in improving user control for explainable online symptom checkers," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22* (New York, NY: Association for Computing Machinery), 1–7. doi: 10.1145/3491101.3519668

Tejeda, H., Kumar, A., Smyth, P., and Steyvers, M. (2022). AI-assisted decision-making: a cognitive modeling approach to infer latent reliance strategies. *Comp. Brain Behav.* 5, 491–508. doi: 10.1007/s42113-022-00157-y

Tilt, C. A. (2016). Corporate social responsibility research: the importance of context. *Int. J. Corp. Soc. Respons.* 1, 1–9. doi: 10.1186/s40991-016-0003-7

Titano, J. J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., et al. (2018). Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* 24, 1337–1341. doi: 10.1038/s41591-018-0147-y

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7

Tsai, C.-H. (2019). *Controllability and Explainability in a Hybrid Social Recommender System* (PhD thesis). Pittsburgh, PA: University of Pittsburgh.

Tsai, C.-H., You, Y., Gui, X., Kou, Y., and Carroll, J. M. (2021). "Exploring and promoting diagnostic transparency and explainability in online symptom checkers," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21* (New York, NY: Association for Computing Machinery), 1–17. doi: 10.1145/3411764.3445101

van der Stigchel, B., van den Bosch, K., van Diggelen, J., and Haselager, P. (2023). Intelligent decision support in medical triage: are people robust to biased advice? *J. Public Health* 45:fdad005. doi: 10.1093/pubmed/fdad005

Vodrahalli, K., Daneshjou, R., Gerstenberg, T., and Zou, J. (2022). "Do humans trust advice more if it comes from AI? An analysis of human-AI interactions," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, AIES '22* (New York, NY: Association for Computing Machinery), 763–777. doi: 10.1145/3514094.3534150

Wei, J., Kim, S., Jung, H., and Kim, Y.-H. (2024a). Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction*. 8, 1–35. doi: 10.1145/3637364

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2024b). "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22* (Red Hook, NY: Curran Associates Inc.).

Xiao, Z., Liao, Q. V., Zhou, M., Grandison, T., and Li, Y. (2023). "Powering an ai chatbot with expert sourcing to support credible health information access," in *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23* (New York, NY: Association for Computing Machinery), 2–18. doi: 10.1145/3581641.3584031

Yaniv, I., and Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *J. Behav. Decis. Mak.* 10, 21–32. doi: 10.1002/(SICI)1099-0771(199703)10:1<21::AID-BDM243>3.0.CO;2-G

Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems, CHI '19* (New York, NY: Association for Computing Machinery), 1–12. doi: 10.1145/3290605.3300509

You, Y., Tsai, C.-H., Li, Y., Ma, F., Heron, C., and Gui, X. (2023). Beyond self-diagnosis: how a chatbot-based symptom checker should respond. *ACM Trans. Comput. Hum. Interact.* 30. doi: 10.1145/3589959

Yuan, L., Chen, Y., Cui, G., Gao, H., Zou, F., Cheng, X., et al. (2024). "Revisiting out-of-distribution robustness in NLP: benchmark, analysis, and LLMS evaluations," in *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23* (Red Hook, NY: Curran Associates Inc.).

Yun, J. H., Lee, E.-J., and Kim, D. H. (2021). Behavioral and neural evidence on consumer responses to human doctors and medical artificial intelligence. *Psychol. Market.* 38, 610–625. doi: 10.1002/mar.21445

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., et al. (2024). Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI* 1:AIoa2300068. doi: 10.1056/AIoa2300068

Zuccon, G., and Koopman, B. (2023). Dr ChatGPT, tell me what i want to hear: how prompt knowledge impacts health answer correctness. *arXiv* [preprint]. doi: 10.48550/arXiv.2302.13793