

Comparing AI Roles and Support Types in Digital Health: A Randomized Factorial Experiment with Patients and Providers

Owen Xingjian Zhang, Ilana Shumsky, Jeffrey Hancock, Munmun De Choudhury,
Sunny Xun Liu

Submitted to: JMIR AI
on: September 21, 2026

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 27

 Figures 28

 Figure 1..... 29

Multimedia Appendixes 30

 Multimedia Appendix 1..... 31

 Multimedia Appendix 2..... 31

 Multimedia Appendix 3..... 31

 Multimedia Appendix 4..... 31

 Multimedia Appendix 5..... 31

Comparing AI Roles and Support Types in Digital Health: A Randomized Factorial Experiment with Patients and Providers

Owen Xingjian Zhang¹ MS; Ilana Shumsky² MD; Jeffrey Hancock² PhD; Munmun De Choudhury¹ PhD; Sunny Xun Liu² PhD

¹School of Interactive Computing College of Computing Georgia Institute of Technology Atlanta US

²Department of Communication Stanford University Stanford US

Corresponding Author:

Sunny Xun Liu PhD

Department of Communication
Stanford University
450 Jane Stanford Way
Stanford
US

Abstract

Background: The increasing demand for digital health services has highlighted the need for innovative solutions to enhance accessibility and quality of care. AI has emerged as a promising tool, offering capabilities such as personalized treatment plans and virtual consultations. Despite these advancements, the acceptance of AI in digital health remains a critical concern, with potential ethical and practical challenges.

Objective: The objective of this study is to evaluate how different AI roles and support types influence the perceptions of AI tools among both patients and healthcare providers.

Methods: We recruited healthcare providers (N=251) and the general public (N=396) to examine their perceptions of AI healthcare chatbots. Participants were randomly assigned to interact with one of six LLM-based chatbots designed within a 2×3 factorial framework: AI roles (assisting patients vs. assisting providers) and support type (task, companion, or control). Following the interaction, participants reported perceived warmth and competence, emotional and informational support, and technology acceptance using validated scales. Chat histories were also recorded, from which we extracted linguistic features and conversation topics. We ran a factorial analysis of covariance (ANCOVA) on each perception construct and each linguistic feature, testing the main effects of role and support type and their interaction. We also fitted separate linear regression models for providers and patients to examine how perceptions varied with participants' demographic characteristics and prior beliefs about AI.

Results: We found that providers were generally more conservative than patients, reporting lower scores on five of the seven constructs. Independent of role, the task chatbot was rated highest on competence, informational support, attitudes, and intention to adopt, while the companion chatbot was rated warmer than the task chatbot but no higher than the general-purpose chatbot on warmth or emotional support. Participants' conversations followed the same pattern, with those talking to the task chatbot asking more questions and those talking to the companion chatbot focusing more on their feelings. Across both groups, participants' general trust in AI predicted their perceptions of every construct, and among patients, those who had a regular doctor, were female, had less education, or knew more about AI also gave higher ratings.

Conclusions: This study contributes to understanding both providers' and patients' perceptions of using chatbots in different roles and support types, ultimately suggesting that future healthcare AI systems should build on domain-specific expertise and users' prior trust in AI, while finding better ways to convey warmth that users actually perceive.

(JMIR Preprints 21/09/2026:110259)

DOI: <https://doi.org/10.2196/preprints.110259>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Yes – Make my manuscript publicly available as a preprint immediately.**

No – I do not wish to publish a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/110259>



Original Manuscript

Comparing AI Roles and Support Types in Digital Health: A Randomized Factorial Experiment with Patients and Providers

Abstract

Background: The increasing demand for digital health services has highlighted the need for innovative solutions to enhance accessibility and quality of care. AI has emerged as a promising tool, offering capabilities such as personalized treatment plans and virtual consultations. Despite these advancements, the acceptance of AI in digital health remains a critical concern, with potential ethical and practical challenges.

Objective: The objective of this study is to evaluate how different AI roles and support types influence the perceptions of AI tools among both patients and health care providers.

Methods: We recruited health care providers (N=251) and the general public (N=396) to examine their perceptions of AI health care chatbots. Participants were randomly assigned to interact with one of six LLM-based chatbots designed within a 2×3 factorial framework: AI roles (assisting patients vs. assisting providers) and support type (task, companion, or control). Following the interaction, participants reported perceived warmth and competence, emotional and informational support, and technology acceptance using validated scales. Chat histories were also recorded, from which we extracted linguistic features and conversation topics. We ran a factorial analysis of covariance (ANCOVA) on each perception construct and each linguistic feature, testing the main effects of role and support type and their interaction. We also fitted separate linear regression models for providers and patients to examine how perceptions varied with participants' demographic characteristics and prior beliefs about AI.

Results: We found that providers were generally more conservative than patients, reporting lower scores on five of the seven constructs. Independent of role, the task chatbot was rated highest on competence, informational support, attitudes, and intention to adopt, while the companion chatbot was rated warmer than the task chatbot but no higher than the general-purpose chatbot on warmth or emotional support. Participants' conversations followed the same pattern, with those talking to the task chatbot asking more questions and those talking to the companion chatbot focusing more on their feelings. Across both groups, participants' general trust in AI predicted their perceptions of every construct, and among patients, those who had a regular doctor, were female, had less education, or knew more about AI also gave higher ratings.

Conclusions: This study contributes to understanding both providers' and patients' perceptions of using chatbots in different roles and support types, ultimately suggesting that future health care AI systems should

build on domain-specific expertise and users' prior trust in AI, while finding better ways to convey warmth that users actually perceive.

Keywords: artificial intelligence; large language models; chatbots; digital health; health care providers; patients; technology acceptance; social support

Introduction

The rapid growth of digital health technologies has increased interest in AI-powered systems that can support health care delivery and communication. Artificial intelligence (AI), particularly large language models (LLMs), is increasingly being integrated into health care settings to support clinical diagnosis, treatment, patient communication, and access to care [1–3]. In particular, LLM-powered health care chatbots have shown promise in engaging in natural, context-aware, and human-like interactions, creating new opportunities for emotional support, health communication, and clinical assistance [4–6]. However, concerns remain regarding reliability, the appropriate integration of AI into health care settings and how AI impacts patient-provider communication dynamics and relations [7–9].

In health care settings, prior work has discussed AI chatbots as assistive and supplementary tools rather than replacements for humans [5,10]. For health care providers, chatbots are expected to support clinical workflows, improve efficiency, and assist decision making alongside clinicians [11]. Similarly, patients often prefer AI systems that function as supportive assistants rather than fully autonomous health care providers [12]. While both stakeholder groups prefer AI as an assistive technology, prior research has shown that patients and providers may still hold different expectations, trust levels, and attitudes toward health care AI systems [9,13–15]. However, these comparisons are usually indirect: the two groups are surveyed separately, or one group evaluates responses written for the other [16,17]. Few studies have had both groups interact with the same chatbots and evaluate them on the same measures.

In addition, emotional support and informational support have received substantial theoretical and empirical attention in studies of social support and online communication [18,19]. AI chatbots have shown promise in providing informational support by assisting users with health care-related questions, decision making, and access to medical information [6,20]. At the same time, prior work has also explored the potential for AI systems to provide emotional support through empathetic and supportive communication, particularly in mental health and well-being contexts [4,21,22].

Beyond how users evaluate these systems, such designs may also shape the conversation itself, though the evidence is mixed [23,24]. However, these two forms of support are usually examined in isolation. How they compare when tested against each other and against a general-purpose baseline remains under-investigated.

How people evaluate an AI system also depends on their backgrounds and prior beliefs. Trust in AI in particular has been shown to shape whether people accept and rely on AI systems across domains [7]. Existing research has also shown that people's attitudes and experiences significantly influence how they perceive AI

systems [25,26], and that clinicians' belief in AI shapes their acceptance of it [15]. Users' background characteristics have also been linked to the adoption of health technologies, including gender [27] and trust in health care providers [28]. However, these characteristics are usually examined a few at a time and within a single population, so a comprehensive picture of who responds well to a health care chatbot, and whether that picture differs for clinicians and patients, is still missing.

To address these gaps, we conducted a 2×3 factorial experiment examining how AI's assisting roles and support types shape the perceptions of patients and health care providers. We recruited both stakeholder groups to interact with one of six LLM-based chatbot conditions varying in roles (assisting patient vs. assisting provider) and support types (task vs. companion vs. control). Following the interaction, participants evaluated the chatbots on measures drawn from technology acceptance, stereotype content model, and social support theory. We also recorded the full chat logs of each interaction, and collected participants' demographic characteristics and their prior beliefs about AI.

Using these data, we examined the following research questions:

- **RQ1a.** How do assisting role and chatbot support type impact users' perceptions of the chatbots?
- **RQ1b.** How do assisting role and chatbot support type impact users' conversational behavior with the chatbots?
- **RQ2.** How do users' perceptions vary with demographic characteristics and prior beliefs about AI?

To answer RQ1a, we ran a factorial Analysis of Covariance (ANCOVA) [29] on each of the seven perception constructs, testing the main effects of assisting role and support type and their interaction. To answer RQ1b, we extracted linguistic features from participants' own messages and ran the same models on each feature. To answer RQ2, we fitted separate linear regression models for providers and patients, predicting each perception construct from participants' demographic characteristics and prior beliefs about AI.

Overall, our findings suggest that providers were generally more conservative than patients in evaluating health care AI chatbots. Patients rated the chatbots higher than providers on five of the seven constructs, expressed more gratitude, and talked mostly about their feelings, while providers talked more about communication and diagnosis. The task chatbot was rated highest on competence, informational support, attitude, and adoption intention, and participants talking with it used more positive language, stayed longer, and asked more questions. The companion chatbot was rated warmer than the task chatbot but no higher than the general-purpose chatbot on warmth or emotional support, although participants talking with it asked the fewest questions and focused most on their feelings. Beyond the experimental conditions, users' general trust in AI predicted their perceptions across all seven constructs, and was the only consistent predictor among providers, while patients who had a regular doctor, were female, had less education, or knew more about AI also gave higher ratings.

Methods

Study Design

We conducted an experiment with a 2×3 factorial design. The first factor was participant type: we recruited health care providers and members of the general public, who evaluated chatbots designed for assisting providers and assisting patients, respectively. The second factor was chatbot support type. Within each participant group, we randomly assigned participants to a task support, companionship support, or general support chatbot (control group).

Figure 1. Overview of the 2×3 factorial study design. Participants (health care providers or patients) were randomly assigned to one of three chatbot support types (task, companion, or control), yielding six conditions.

For each (Assisting Role $\times$ Support Type) condition, we developed a chatbot. Participants interacted with the assigned chatbot within our designed scenario, and then reported their perceptions. The chatbot implementation and scenario design are described as follows:

Chatbot Implementation: All chatbots were powered by the GPT-4o-mini model [30], which was selected for its strong instruction-following capabilities and stable conversational performance. At the time of the study, GPT-4o-mini was among the leading small-scale proprietary language models, outperforming GPT-3.5 Turbo and comparable small models (e.g., Gemini 1.5 Flash, Claude 3 Haiku) on academic benchmarks of reasoning and conversational capability [30]. Control group chatbots used the model's default configuration, whereas Task and Companion conditions were implemented through condition-specific prompts that varied by assisting role (provider vs patient), shown in Table 1. For example, the ProviderTask condition framed the chatbot as a clinical assistant to support data analysis and differential diagnosis, using a professional tone. In contrast, the ProviderCompanion condition emphasized emotional validation and active listening. Full prompts are provided in Multimedia Appendix 1.

Table 1. Condition-specific prompt summary by assisting role and support type.^a

	Task	Companion	Control
Assisting providers	Clinical and diagnostic assistance	Emotional support; no clinical advice	Default GPT-4o-mini
Assisting patients	Practical guidance for managing health care tasks	Emotional support for coping with health-related stress	Default GPT-4o-mini

^aTask and companion conditions used condition-specific system prompts; control conditions used the default GPT-4o-mini configuration. Full prompts are provided in Multimedia Appendix 1.

Scenario Design: Participants engaged with a role-adapted clinical scenario centered on diagnostic uncertainty. For example, patients imagined experiencing a six-week undiagnosed fever and growing frustrated with the diagnostic process, while providers imagined managing this patient as their trust began to erode. Each role's scenario was further adapted into three versions matching the support type: for example, the task version asked participants to work with the chatbot on concrete next steps (e.g., preparing questions for

the doctor or planning patient communication), whereas the companion version asked them to share their feelings about the situation, resulting in six scenario variants in total. The scenarios were iteratively developed and piloted with input from four medical professionals (all holding MD degrees) to ensure clinical appropriateness, generalizability across participants, and avoidance of sensitive topics. All scenario materials were otherwise held constant, such that differences were limited to the assisting role and support type. The full scenarios are provided in Multimedia Appendix 2.

Procedure

Sample size: The analytic sample comprised 647 participants. The required sample size was determined using a G*Power analysis [31]. For a 2×3 between-subjects factorial design, assuming a small-to-medium effect size (Cohen's $f = 0.15$) [32], power of 0.80, and $\alpha = .05$, the minimum required sample size was 432, determined by the largest-degree-of-freedom effects of interest (the support-type main effect and the role × support-type interaction). Based on this estimate, we recruited 1,560 respondents to allow for standard attrition and exclusions. Participants were recruited from both the general public and medical professionals from the CloudResearch Connect platform[33], an online participant recruitment platform commonly used in behavioral research and shown to provide high-quality data for human-subject studies [34,35]. Each participant was paid \$2 for their participation.

Of the 1,560 respondents, 301 were removed as incomplete responses or survey previews, leaving 1,259 completed surveys. A further 365 were excluded for failing screening checks (e.g., incorrect identification of the assigned condition, which served as a manipulation check confirming that participants understood the purpose of their interaction with the chatbot: task assistance, emotional support, or general conversation) or for interaction durations exceeding 60 minutes, yielding 894 valid responses. Because the chatbot underwent code revisions during the initial data-collection period to address technical issues, participants in that period interacted with a non-final version of the system. To ensure a consistent stimulus across conditions, we excluded these participants and restricted all analyses to the finalized system, resulting in a final analytic sample of 647 (251 providers, 396 patients).

We distributed surveys to two groups of participants recruited online: the general public (self-identifying patients) and medical professionals. Both groups were randomly assigned to one of three chatbot support conditions (task, companion, or control).

Based on the assigned condition scenario, participants interacted with the chatbot for at least three turns to facilitate a minimally sustained interaction before evaluating the chatbot. Following the interaction, participants evaluated the chatbot on seven constructs drawn from technology acceptance, stereotype content model, and social support theory.

Measures and Features

Outcome measures: We measured seven constructs drawn from three frameworks, chosen to capture how

participants judged the chatbot and how they experienced the interaction [18,36,37].

The Technology Acceptance Model [36] describes how much people accept and are willing to use a new technology, and its extension based on the Theory of Planned Behavior [38] adds that acceptance also depends on how much control people feel they have. From this framework we measured attitude, or how positively participants viewed the chatbot; intention to adopt, or how willing they were to use it in the future; and perceived behavioral control, or how much they felt able to direct and manage the interaction.

The Stereotype Content Model [37] was originally developed to explain how people perceive social groups along two independent dimensions: warmth, or how caring a target seems, and competence, or how capable it seems. We measured these two dimensions to capture what kind of social impression each chatbot created, rather than only how favorably it was rated overall.

The social support theory [18] distinguishes several kinds of support people give one another, of which emotional and informational support have received the most theoretical and empirical attention [39]. Emotional support conveys understanding, encouragement, and caring, while informational support provides advice and knowledge. We measured both because they map directly onto the two functions our chatbots were designed to serve.

All items were measured on a 5-point Likert scale [40]. Participants were asked to indicate their level of agreement, ranging from 1 (strongly disagree) to 5 (strongly agree), for statements such as “I plan to use this chatbot in the future.” All seven constructs demonstrated high internal consistency, with Cronbach’s alpha coefficients exceeding 0.9 [41]. The full list of constructs and sample questions is shown in Multimedia Appendix 3.

Conversational features: Conversational behavior was characterized using features derived from the chat logs, drawing on approaches used to study linguistic behavior in online health interactions. Beyond self-reported perceptions, we computed the sentiment of participants’ messages using VADER [42], taking the mean compound score across each participant’s messages. We measured the volume and duration of participation, including the number of words per message and interaction duration. We also captured two markers of engagement: the number of questions asked and gratitude, measured as the number of question marks and the frequency of gratitude-related terms drawn from a fixed lexicon (e.g., “thank,” “appreciate,” “grateful”), following prior work that detects gratitude using curated word lists [43].

Conversational topics were identified by applying Latent Dirichlet Allocation (LDA) topic modeling [44] to participants’ messages. We selected the number of topics based on coherence, which yielded three interpretable topics. Based on the word groups, we named the three main topics communication, feeling, and diagnosis.

Demographics and prior beliefs: We collected standard demographic information, including age, gender, race, and education (ranging from high school or below to master’s, doctoral, or professional degrees).

We additionally measured three aspects of participants' relationship to AI, each on a 5-point scale. AI literacy captured their self-rated understanding of AI concepts and technologies. Prior experience with large language models captured their familiarity with using models such as ChatGPT. Trust in AI captured how much they trusted AI systems to make reliable decisions. We also collected one role-specific characteristic: occupation for providers, and whether they had a regular doctor for patients. Education, gender, AI literacy, LLM experience, and AI trust were used as covariates (RQ1) and as predictors of perceptions (RQ2), together with the role-specific characteristics; age and race were collected for descriptive purposes. The full demographic breakdown is provided in Multimedia Appendix 4.

Data Collection and Analysis

We report this study following the Checklist for Reporting Results of Internet E-Surveys (CHERRIES; Multimedia Appendix 5).

All data were collected via Qualtrics surveys and chatbot chat logs. All analyses used the final analytic sample of 647 participants. We cleaned the dataset, ensuring that data types and values were aligned and appropriate for statistical analysis. Across the RQ1a and RQ1b analyses, we controlled for education, gender, AI literacy, LLM experience, and AI trust as covariates. Because education was missing for a small number of participants, we imputed the median and included a missing-value indicator so that all participants were retained in every model. Variance inflation factors indicated no multicollinearity among these variables (all VIF <1.41). All analyses were conducted in Python 3 (Google Colaboratory) using statsmodels (ANCOVAs and regressions), pingouin (Cronbach's α), SciPy (demographic comparisons), and gensim (topic modeling).

To answer RQ1a, we examined how assisting role and chatbot support type affected participants' perceptions across all measured constructs using factorial Analyses of Covariance (ANCOVA) [29]. For each of the seven perception constructs, we conducted a 2×3 between-subjects factorial ANCOVA evaluating the main effects of assisting role (assisting providers vs. assisting patients) and chatbot support type (task, companion, and control) and their interaction, using the covariates described above. When a support-type main effect was significant, we followed up with pairwise comparisons of the least-squares (covariate-adjusted) means, pooled across assisting role. We report the main effect of assisting role first, followed by the main effect of support type and its pairwise comparisons, and then the interaction, with significance set at $P < .05$.

To answer RQ1b, we examined how assisting role and chatbot support type affected participants' conversational behavior. Using the same 2×3 between-subjects factorial design and covariates, we conducted an ANCOVA for each linguistic feature to evaluate the main effects of assisting role and support type and their interaction, followed by the same pairwise comparisons of least-squares means when a support-type main effect was significant.

To answer RQ2, we examined how participants' demographic characteristics related to their perceptions of the chatbots. For each of the seven outcome constructs, we fitted a linear regression model with demographic

variables (education, gender, AI literacy, LLM experience, AI trust, and occupation for providers or having a regular doctor for patients) as predictors, estimated separately for providers and patients (7 constructs $\times$ 2 roles = 14 models). Continuous predictors were standardized and categorical predictors dummy-coded. We report standardized regression coefficients (β) and adjusted R^2 for each model, with significance set at $P < .05$.

Ethical Considerations

This study was approved by the Institutional Review Board of the relevant university where the research was conducted (protocol #81033). At the beginning of the survey, participants reviewed and signed a consent form outlining the study's purpose, procedures, risks, and benefits before proceeding. We compensated each participant \$2, both for the general public and for medical professionals.

Results

Chatbot behavior check

To confirm that the support-type prompts shaped the chatbot's responses as intended, we extracted the chat logs and analyzed five features of the chatbots' responses: number of turns, words per turn, emotional support per turn, informational support per turn, and average message sentiment. As shown in Table 2, the three support types differed as expected. The task bot exchanged the most turns (mean 4.04, SD 1.85) and wrote roughly nine more words per turn (mean 107.5, SD 21.0) than the companion (mean 98.6, SD 14.9) and control (mean 98.9, SD 14.8) bots. Most importantly, it expressed emotional support in the fewest of its turns (mean 0.51, SD 0.32), compared with the companion (mean 1.00, SD 0.03) and control (mean 0.97, SD 0.11) bots, $F=451.09$, $P < .001$, while providing the highest rate of informational support (mean 0.96, SD 0.10), $F=7.755$, $P < .001$. The companion bot showed the highest emotional support (mean 1.00), though, somewhat surprisingly, it exceeded the control bot only slightly (mean 0.97). This narrow gap suggests the default configuration was already a warm, general-purpose agent, leaving little room for the companion prompts to add. Sentiment also differed significantly across conditions, $F=124.90$, $P < .001$: the companion bot was the most positive in tone (mean 0.88, SD 0.14), followed by the control (mean 0.80, SD 0.18) and task (mean 0.57, SD 0.32) bots. The support-type prompts thus shaped both the content and the affective tone of the chatbots' responses.

Table 2. Chatbot behavior check: chatbot responses by support type.^a

Feature	Task	Companion	Control	F
# of turns	4.04 (1.85)	3.75 (1.09)	3.67 (1.08)	3.87*
Words per turn	107.5 (21.0)	98.6 (14.9)	98.9 (14.8)	18.93***
Emotional support per turn	0.51 (0.32)	1.00 (0.03)	0.97 (0.11)	451.09***
Informational support per turn	0.96 (0.10)	0.92 (0.16)	0.94 (0.13)	7.75***
Sentiment (−1: negative, +1: positive)	0.573 (0.324)	0.883 (0.139)	0.799 (0.182)	124.90***

^aCells are M (SD), N=647. F tests differences across the three support types. Emotional and informational support are proportions of chatbot turns containing each support type.

* $P < .05$; ** $P < .01$; *** $P < .001$.

RQ1a: Effects of Role and Support Type on Perceptions

Compared with providers, patient participants rated the chatbots significantly higher on five of the seven constructs and numerically higher on all seven, including how much control they felt, how warm the chatbot seemed, and how supportive it was. Across the three chatbot types, task bots were seen as the most competent and informationally supportive and drew the most positive attitudes and adoption intentions, while companion bots were rated no higher on emotional support or warmth than the general-purpose default. In the following sections, we interpret only statistically significant results. Table 3 reports the ANCOVA F statistics for each construct and Table 4 reports the adjusted means.

Main effect of assisting role.

Providers were more conservative than patients about the support and control they experienced, but not about the chatbot's competence or their overall attitude toward it.

Providers rated the chatbots lower than patients on five of the seven constructs: behavioral control, $F(1,634)=24.86$, $P < .001$, $d=0.44$; warmth, $F(1,634)=7.75$, $P=.006$, $d=0.24$; informational support, $F(1,634)=5.12$, $P=.024$, $d=0.20$; emotional support, $F(1,634)=4.70$, $P=.030$, $d=0.19$; and intention to adopt, $F(1,634)=3.88$, $P=.049$, $d=0.17$. Assisting role did not significantly impact attitude, $F(1,634)=1.73$, $P=.189$, or competence, $F(1,634)=2.12$, $P=.146$. These differences were modest in magnitude, with only behavioral control approaching a medium effect.

Main effect of support type.

The task chatbot was rated highest on competence, informational support, attitude, and adoption intention, but lowest on emotional support, where the companion chatbot did no better than the general-purpose default.

Among the technology-acceptance constructs, the task chatbot drew the most positive attitudes and adoption intentions, with significantly higher ratings than the other two (attitude, task mean 3.42, SE 0.11; companion mean 2.96, SE 0.10; control mean 2.66, SE 0.13; $F(2,634)=23.09$, $P < .001$, $d=0.43$ and 0.71 ; intention, task mean 3.34, SE 0.10; companion mean 2.85, SE 0.10; control mean 2.63, SE 0.12; $F(2,634)=23.65$, $P < .001$, $d=0.49$ and 0.70). A similar pattern held for behavioral control, or how much users felt they could steer the conversation: the task and companion chatbots were rated higher than the control chatbot (mean 3.80, SE 0.10 and mean 3.77, SE 0.09 vs. mean 3.52, SE 0.12; $d=0.29$ and 0.26), though they did not differ from each other.

For the stereotype-content constructs, the task chatbot was rated most competent, exceeding both the companion and control conditions (mean 3.73, SE 0.10 vs. mean 3.31, SE 0.10 and mean 3.17, SE 0.12), $F(2,634)=15.29$, $P < .001$, $d=0.41$ and 0.55 . Warmth showed the opposite ordering, $F(2,634)=4.45$, $P=.012$: the companion chatbot was rated warmest and significantly warmer than the task chatbot (mean 3.59, SE 0.09 vs. mean 3.33, SE 0.10; $d=0.27$), though neither differed significantly from the control condition.

The social-support constructs showed the most unexpected result. As expected, the task chatbot was rated most informationally supportive, exceeding both other conditions (mean 3.93, SE 0.11 vs. companion mean 3.27, SE 0.10 and control mean 3.12, SE 0.13), $F(2,634)=33.28$, $P<.001$, $d=0.64$ and 0.78 . Emotional support, however, ran in the opposite direction, $F(2,634)=28.42$, $P<.001$: although the companion chatbot was built specifically to provide emotional support, it was rated no higher than the general-purpose control (mean 3.60, SE 0.10 vs. mean 3.78, SE 0.13; the two did not differ significantly), while the task chatbot was rated lower than both (mean 3.00, SE 0.11; $d=0.57$ and 0.74). This echoes the pattern in stereotype-content ratings and mirrors the chatbots' behaviors, in which the companion and default chatbots expressed emotional support at close rates. In other words, what the chatbots expressed and what participants perceived followed the same pattern.

Role × support-type interaction.

No construct showed a significant role × support-type interaction (all $P>.18$; Table 3). The effect of support type was therefore consistent across patients and providers: although the two groups differed in their overall ratings, they responded to the three chatbot designs in the same way.

Table 3. ANCOVA results: effects of assisting role and support type on perceptions.^a

Construct	Role	Support type	Role × support
	F(1,634)	F(2,634)	F(2,634)
Technology acceptance			
Behavioral control	24.86***	4.02*	0.89
Attitude	1.73	23.09***	0.37
Intention to adopt	3.88*	23.65***	0.10
Stereotype content			
Warmth	7.75**	4.45*	0.39
Competence	2.12	15.29***	0.40
Social support			
Emotional support	4.70*	28.42***	1.69
Informational support	5.12*	33.28***	1.07

^aN=647. Each cell reports the F statistic from a 2×3 between-subjects ANCOVA controlling for education, gender, AI literacy, LLM experience, and AI trust; larger values indicate a stronger effect.

* $P<.05$; ** $P<.01$; *** $P<.001$.

Table 4. Adjusted means by assisting role and support type, with effect sizes.^a

Construct	Role			Support type					
	Patient	Provider	P-Pr	Task	Comp.	Ctrl	T-C	T-Ct	C-Ct
Technology acceptance									
Behavioral control	3.90 (0.09)	3.49 (0.11)	+0.44***	3.80 (0.10)	3.77 (0.09)	3.52 (0.12)	+0.03	+0.29*	+0.26*
Attitude	3.08 (0.11)	2.95 (0.12)	+0.12	3.42 (0.11)	2.96 (0.10)	2.66 (0.13)	+0.43***	+0.71***	+0.28**
Intention to adopt	3.03 (0.10)	2.85 (0.12)	+0.17*	3.34 (0.10)	2.85 (0.10)	2.63 (0.12)	+0.49***	+0.70***	+0.21*
Stereotype content									
Warmth	3.60 (0.09)	3.37 (0.11)	+0.24**	3.33 (0.10)	3.59 (0.09)	3.54 (0.12)	-0.27**	-0.21	+0.06
Competence	3.47 (0.10)	3.34 (0.12)	+0.13	3.73 (0.10)	3.31 (0.10)	3.17 (0.12)	+0.41***	+0.55***	+0.13
Social support									
Emotional	3.56 (0.10)	3.36 (0.12)	+0.19*	3.00 (0.11)	3.60 (0.10)	3.78 (0.13)	-0.57***	-0.74***	-0.16

support									
Informational support	3.54 (0.10)	3.33 (0.12)	+0.20*	3.93 (0.11)	3.27 (0.10)	3.12 (0.13)	+0.64***	+0.78***	+0.14

^aN=647. Cells report least-squares (covariate-adjusted) means with standard errors in parentheses, on the 1–5 Likert scale. Role means are pooled across support type; support-type means are pooled across role. The P–Pr, T–C, T–Ct, and C–Ct columns report Cohen’s d (patient minus provider, task minus companion, task minus control, and companion minus control). Positive values indicate a higher score for the first-named group or condition.

* $P < .05$; ** $P < .01$; *** $P < .001$.

RQ1b: Effects of Role and Support Type on Conversational Behavior

Compared with providers, patients expressed more gratitude and talked mostly about their feelings, while providers talked more about communication and diagnosis. Across the three chatbot types, participants talking with the task chatbot used more positive language, stayed longer, and asked more questions, while those talking with the companion chatbot asked the fewest questions and focused most on their feelings. In the following sections, we interpret only statistically significant results. Table 5 reports the ANCOVA F statistics for each feature and Table 6 reports the adjusted means.

Main effect of assisting role.

Patients expressed more gratitude than providers and talked mostly about their feelings, while providers talked more about communication and diagnosis.

Among the behavioral features, assisting role affected only gratitude: patients expressed more gratitude to the chatbot than providers (mean 1.04, SE 0.14 vs. mean 0.71, SE 0.16), $F(1,634)=7.53$, $P=.006$, $d=0.24$. Other features, including message length, session duration, question-asking, and sentiment, did not differ significantly by role (all $P > .06$).

Role differences were much larger for conversation topics, all $P < .001$. Providers’ messages were dominated by the Communication topic (mean 0.47, SE 0.03 vs. mean 0.13, SE 0.03; $d=1.26$), $F(1,634)=207.03$, and the Diagnosis topic (mean 0.40, SE 0.04 vs. mean 0.28, SE 0.03; $d=0.40$), $F(1,634)=20.47$, while patients’ messages were dominated by the Feeling topic (mean 0.60, SE 0.03 vs. mean 0.13, SE 0.03; $d=1.70$), $F(1,634)=373.93$. These topic effects were far larger than any difference in how participants engaged, all of which were small.

Main effect of support type.

Participants talking with the task chatbot were more positive, stayed longer, and asked the most questions, while those talking with the companion chatbot asked the fewest and focused most on their feelings.

Support type changed how participants engaged with the chatbot. Participants in the task condition used more positive language in their messages (mean 0.10, SE 0.02) than those in the companion (mean 0.03, SE 0.02) or control (mean 0.04, SE 0.03) conditions, $F(2,634)=6.29$, $P=.002$, $d=0.32$ and 0.28, and spent longer in the interaction (mean 2.70, SE 0.18 minutes vs. mean 2.21, SE 0.17 and mean 2.22, SE 0.22), $F(2,634)=4.94$, $P=.007$, $d=0.27$ relative to the companion condition. Question-asking also differed, $F(2,634)=10.64$, $P < .001$, but the pattern was driven by the companion condition: participants who talked with the companion chatbot

asked the fewest questions (mean 0.73, SE 0.13), significantly fewer than those in the task (mean 1.32, SE 0.14; $d=0.43$) or control (mean 1.00, SE 0.17; $d=0.20$) conditions, which did not differ from each other. Message length and gratitude were not affected by support type (both $P>.33$).

Support type also shifted what participants talked about, in line with each chatbot's design. The Feeling topic was most common when talking with the companion chatbot (mean 0.43, SE 0.03), exceeding both the task (mean 0.32, SE 0.03; $d=0.41$) and control (mean 0.34, SE 0.03; $d=0.34$) conditions, $F(2,634)=10.83$, $P<.001$, while the Diagnosis topic was least common there (mean 0.25, SE 0.03 vs. mean 0.38, SE 0.03 and mean 0.39, SE 0.04; $d=0.41$ and 0.45), $F(2,634)=13.14$, $P<.001$. Participants assigned a companion chatbot therefore talked more about how they felt and less about tests and diagnoses, which mirrors the shift in the chatbots' own responses. The Communication topic did not differ by support type, $F(2,634)=1.47$, $P=.232$.

Table 5. ANCOVA results: effects of assisting role and support type on conversational behavior.^a

Feature	Role	Support type	Role × support
	F(1,634)	F(2,634)	F(2,634)
Conversational behavior			
User sentiment	0.71	6.29**	2.10
Words per message	2.55	0.93	1.58
Duration (min)	1.50	4.94**	0.89
Questions asked	3.41	10.64***	0.66
Gratitude expressions	7.53**	1.10	0.12
Conversation topics			
Communication	207.03***	1.47	2.31
Feeling	373.93***	10.83***	2.16
Diagnosis	20.47***	13.14***	5.52**

^aN=647. Each cell reports the F statistic from a 2×3 between-subjects ANCOVA controlling for education, gender, AI literacy, LLM experience, and AI trust; larger values indicate a stronger effect.

* $P<.05$; ** $P<.01$; *** $P<.001$.

Table 6. Adjusted means for conversational behavior by assisting role and support type, with effect sizes.^a

Feature	Role			Support type					
	Patient	Provider	P-Pr	Task	Comp.	Ctrl	T-C	T-Ct	C-Ct
Conversational behavior									
User sentiment	0.05 (0.02)	0.06 (0.03)	-0.07	0.10 (0.02)	0.03 (0.02)	0.04 (0.03)	+0.32***	+0.28**	-0.04
Words per message	63.75 (4.23)	57.75 (5.00)	+0.14	64.12 (4.39)	59.27 (4.16)	58.86 (5.27)	+0.11	+0.12	+0.01
Duration (min)	2.47 (0.18)	2.28 (0.21)	+0.11	2.70 (0.18)	2.21 (0.17)	2.22 (0.22)	+0.27*	+0.27	-0.00
Questions asked	1.13 (0.13)	0.91 (0.16)	+0.16	1.32 (0.14)	0.73 (0.13)	1.00 (0.17)	+0.43***	+0.24	-0.20*
Gratitude expressions	1.04 (0.14)	0.71 (0.16)	+0.24**	0.77 (0.14)	0.96 (0.13)	0.88 (0.17)	-0.14	-0.08	+0.06
Conversation topics									
Communication	0.13 (0.03)	0.47 (0.03)	-1.26***	0.30 (0.03)	0.32 (0.03)	0.27 (0.03)	-0.06	+0.12	+0.18
Feeling	0.60 (0.03)	0.13 (0.03)	+1.70***	0.32 (0.03)	0.43 (0.03)	0.34 (0.03)	-0.41***	-0.08	+0.34***
Diagnosis	0.28 (0.03)	0.40 (0.04)	-0.40***	0.38 (0.03)	0.25 (0.03)	0.39 (0.04)	+0.41***	-0.03	-0.45***

^aN=647. Cells report least-squares (covariate-adjusted) means with standard errors in parentheses. Sentiment ranges from -1 to +1; topic values are proportions. Role means are pooled across support type; support-type means are pooled across role. The P-Pr, T-C, T-Ct, and C-Ct columns report Cohen's d (patient minus provider, task minus companion, task minus control, and companion minus control). Positive values indicate a higher score for the first-named group or condition.

* $P < .05$; ** $P < .01$; *** $P < .001$.

Role × support-type interaction.

The only conversational feature to show a significant role × support-type interaction was the Diagnosis topic, $F(2,634)=5.52$, $P=.004$. Support type shifted how much providers discussed diagnostic detail—from a low in the companion condition (mean 0.26, SE 0.03) to a high in the control condition (mean 0.51, SE 0.04)—but had little effect for patients, whose diagnostic talk was similar across conditions (means 0.24 to 0.32). For all other features, no interaction reached significance (Table 5), indicating that the effects of support type were otherwise consistent across the two roles.

RQ2: Demographic Characteristics and Prior Beliefs as Predictors of Perceptions

We modeled each perception construct as a function of participants' demographic characteristics (gender and education), their prior beliefs about AI (general AI trust, AI literacy, and LLM experience), and one role-specific characteristic: occupation for providers and whether they had a regular doctor for patients. Models were fitted separately for the two groups (Table 7).

Providers.

Providers' ratings depended almost entirely on how much they trusted AI in general, not on their professional background or familiarity with the technology.

For providers, AI trust was the only consistent predictor of perceptions, reaching significance across all seven constructs and most strongly for intention to adopt ($\beta=0.46$), attitude ($\beta=0.45$), and competence ($\beta=0.44$; all $P<.001$). No other predictor was significant on any construct: neither professional characteristics (occupation, education) nor AI familiarity (AI literacy, LLM experience) predicted how providers evaluated the chatbot. The one exception was perceived competence, which physicians rated lower than allied-health professionals ($\beta=-0.55$, $P=.013$).

Patients.

Among patients, higher ratings came from those who trusted AI more, had a regular doctor, were female, had less education, and knew more about AI.

As with the providers, AI trust was the strongest predictor, significant across all seven constructs ($\beta = 0.38$ to 0.57 , all $P<.001$), and the effects were larger for patients on every construct. This means that both groups' perceptions rely heavily on how much they trust AI.

Two characteristics predicted more positive ratings. Patients who had a regular doctor and patients who had more knowledge about AI were both more positive about the chatbot (attitude $\beta=0.28$ and 0.10), more willing to use it (intention $\beta=0.32$ and 0.10), rated it as more competent ($\beta=0.26$ and 0.10), and felt it gave better information ($\beta=0.21$ and 0.11 ; all $P<.05$). They differed on one construct each: patients with a regular doctor also saw the chatbot as warmer ($\beta=0.23$, $P=.019$), while patients who knew more about AI felt more in control

of it ($\beta=0.13$, $P=.010$).

Two other characteristics predicted more negative ratings. Patients with more education and patients who used LLMs more often were less positive about the chatbot (attitude $\beta=-0.11$ and -0.10) and rated it as less competent ($\beta=-0.12$ and -0.11 ; all $P<.05$). Frequent LLM users also rated it as less warm ($\beta=-0.11$, $P=.038$).

Gender mattered for two constructs. Compared with female patients, male patients felt less in control of the chatbot ($\beta=-0.21$, $P=.029$) and rated it as less competent ($\beta=-0.18$, $P=.039$).

Finally, emotional support stood apart from the other constructs: no demographic characteristic predicted it, and AI trust was the only significant predictor.

Table 7. Demographic predictors of perceptions, by assisting role (standardized β).^a

Predictor	Beh. control	Attitude	Intention	Warmth	Competence	Emot. support	Info. support
Providers (N=251)							
Gender: male	0.05	-0.02	-0.03	0.04	-0.05	-0.13	-0.03
Occupation: admin	-0.16	-0.19	-0.19	0.01	-0.23	-0.04	-0.31
Occupation: nurse	0.09	0.04	0.05	0.01	-0.03	0.08	-0.06
Occupation: other	-0.05	0.19	0.16	-0.06	-0.20	0.07	-0.19
Occupation: physician	-0.39	-0.29	-0.30	-0.03	-0.55*	0.23	-0.30
Education	-0.11	-0.01	0.00	-0.02	-0.04	-0.07	-0.00
AI literacy	0.12	0.07	0.07	0.01	0.09	0.02	0.11
LLM experience	-0.04	-0.10	-0.10	-0.06	-0.07	-0.13	-0.08
AI trust	0.22**	0.45***	0.46***	0.41***	0.44***	0.31***	0.33***
Adjusted R²	6.8%	19.8%	20.3%	11.8%	19.5%	5.2%	9.5%
Patients (N=396)							
Gender: male	-0.21*	-0.15	-0.15	-0.07	-0.18*	-0.17	-0.15
Regular doctor: yes	0.16	0.28**	0.32***	0.23*	0.26**	0.16	0.21*
Education	-0.05	-0.11*	-0.07	-0.03	-0.12**	-0.02	-0.09
AI literacy	0.13**	0.10*	0.10*	0.05	0.10*	0.05	0.11*
LLM experience	0.01	-0.10*	-0.09	-0.11*	-0.11*	-0.06	-0.10
AI trust	0.38***	0.56***	0.57***	0.52***	0.54***	0.46***	0.46***
Adjusted R²	19.8%	33.4%	35.8%	29.2%	32.5%	22.9%	22.9%

^aStandardized regression coefficients (β), with female, allied-health occupation, and “no regular doctor” as reference categories. Continuous predictors were z-scored. Models also included a gender “other/prefer not to say” category ($n<15$) and an education missing-value indicator as controls, not shown. The provider “physician” category ($n=22$) should be interpreted with caution given its small cell size. Race was tested but did not predict perceptions for either role and was excluded.

* $P<.05$; ** $P<.01$; *** $P<.001$.

Discussion

Principal Results

This study examined how different AI assisting roles and support types shaped perceptions among both patients and health care providers.

Providers were more conservative than patients, and the two groups used the chatbot for different purposes: patients to process how they felt, providers to work through the case. Patients gave higher ratings than

providers on five of the seven constructs, including how much control they felt, how warm the chatbot seemed, and how supportive it was, though the two groups did not differ in their overall attitude or in how capable they judged the chatbot to be. In their conversations, patients expressed more gratitude and talked mostly about their feelings, while providers talked more about communication and diagnosis. The difference was therefore not only one of degree but also of focus: patients engaged with the chatbot as people seeking support, while providers engaged with it as professionals judging a tool.

Task chatbots were more readily accepted, while designing a chatbot for emotional support did *not* lead users to perceive it as more empathetic. The task chatbot was rated highest on competence, informational support, attitude, and adoption intention, while the companion chatbot was rated warmer than the task chatbot but no higher than the general-purpose chatbot on warmth or on emotional support. Participants' conversations followed the same pattern: those talking with the task chatbot used more positive language, stayed longer, and asked more questions, while those talking with the companion chatbot asked the fewest questions, talked most about their feelings, and talked least about diagnosis. These patterns match our chatbot behavior check, in which the task chatbot expressed the least emotional support and the most informational support, and the companion and general-purpose chatbots expressed emotional support at similar rates. Task-oriented designs may therefore be accepted more readily because their value is easier to recognize. For the companion chatbot, what it was designed to do did not match what users felt they received [45].

Beyond the experimental conditions, users' general trust in AI shaped how they evaluated the chatbots more than any other characteristic we measured. It predicted every construct for both providers and patients. For providers it was the only consistent predictor, while their occupation, education, and AI familiarity showed no significant associations. Among patients, higher ratings also came from those who had a regular doctor, were female, had less education, and knew more about AI. Patients who were already comfortable with clinical conversations or understood AI well responded more positively, while those with more education or more frequent LLM use may have held higher expectations that the chatbot did not meet.

Interpretation of Results

Assisting role.

Several patterns in our results may be explained by the positions the two groups occupied in the scenario. The scenario described a difficult, unresolved medical situation. In that position, patients were the ones seeking help and may have felt more relieved to receive support, which could explain why they expressed more gratitude and rated the chatbots higher. Providers, by contrast, were the ones expected to give answers, and may have taken a more critical view of the support they received, resulting in lower scores for the chatbots. This difference also appears in what the two groups talked about: patients focused on how they felt about the uncertainty, while providers focused on how to communicate with the patient and how to work through the diagnosis.

Support type.

The effects of support type may reflect what each design gave participants to do. Working on a concrete task gave participants something to make progress on, which could explain why those in the task condition stayed longer, used more positive language, and asked more questions. The task chatbot invited them to work through concrete next steps, which takes more turns and more questions, while the companion chatbot invited them to share feelings, which requires fewer. Meanwhile, our chatbot behavior check showed that the general-purpose chatbot already expressed emotional support at nearly the same rate as the companion chatbot, suggesting that the default configuration was already warm and supportive, leaving little room for the companion design to add more.

Demographic characteristics and prior beliefs.

How much people trusted AI in general shaped their perceptions more than any other characteristic we measured, for both groups. We measured trust as a general disposition before the interaction; this suggests that people bring an orientation toward AI that shapes how they read a specific system, rather than judging it only on how it behaves. That trust was a stronger predictor than both AI literacy and LLM experience further suggests that what people believe about AI matters more than what they know about it or how often they use it. Among providers, physicians rated the chatbot as less competent than allied-health professionals; they carry the most diagnostic responsibility and may therefore judge an AI system's clinical ability more strictly, though this estimate rests on only 22 physicians. Among patients, those who have a regular doctor are likely accustomed to clinical conversations, which may be why they were more likely to find the chatbot's responses warm. Patients with greater AI knowledge also better understand what AI can and cannot do, which may be why they felt more in control of the interaction. On the other hand, patients who already use LLMs frequently are used to this kind of language, so the chatbot may not have felt distinctive to them, and patients with more education may hold higher expectations for health information or feel more able to find such information on their own, so they may consider chatbots less competent and be less positive toward them.

Comparison With Prior Work

Our findings extend prior research examining patients' and health care providers' perceptions of AI systems in health care settings. Previous studies have suggested that both patients and health care providers prefer AI systems that function as supportive tools rather than autonomous replacements for humans [5,10–12], while they may hold different expectations, trust levels, and attitudes toward health care AI systems [9,13–15]. Our results build on these works by directly comparing patients' and providers' perceptions within the same experimental framework and evaluation measures. Patients reported significantly higher ratings than providers on five of the seven constructs, suggesting that health care providers indeed hold more conservative perspective from the use of AI in health.

Our findings also contribute to prior work examining informational and emotional support in AI-mediated health care communication. Prior studies highlighted the potential of AI chatbots to provide informational

support through health care guidance and decision support [6,20], as well as emotional support through empathetic and supportive communication [4,21,22]. Our findings partially support these directions. Although the task chatbot was perceived as more competent and informationally supportive, the companion chatbot did not significantly improve perceived warmth or emotional support relative to the general-purpose chatbot. Prior work has also examined how chatbot design affects user behavior. Khadpe et al. [23] manipulated how a conversational agent was described and analyzed both users' ratings and their chat logs, finding that the framing changed how users rated the agent but barely changed the language they used. We extend this work by examining how support-type framing changed participants' conversational behavior. In our study, participants in the companion condition talked more about their feelings and less about diagnosis than those in the task condition, showing that the framing shaped the content of the conversation and not only the ratings that followed.

Our findings also speak to how users' background and pre-existing beliefs shape their evaluations. For example, mindsets, or core beliefs that shape how people interpret specific experiences [46], influence technology use more broadly: Lee and Hancock [47] found that people's mindsets about social media shape both how they use it and how it affects them. Background characteristics have also been examined: prior work has reported a marginal tendency for women to be more inclined to adopt health chatbots [27], and has shown that trust in a health care service provider increases users' intention to use digital health care services [28]. Our study is consistent with these patterns and extends them by testing background characteristics and prior beliefs within the same models, which allowed us to compare their relative contributions rather than examine them separately.

Beyond comparing different stakeholder groups and support types, our study also examined how these chatbot designs were perceived across multiple evaluation constructs, including technology acceptance, perceived warmth and competence, and perceived social support. Previous work has discussed people's perceptions and evaluations of health care AI systems across different dimensions, including acceptance, attitudes, and empathy toward AI [13–15,48,49]. Our study advances this direction by systematically comparing how different chatbot designs were perceived across multiple evaluation constructs based on variables from technology acceptance, the stereotype content model, and social support theory [18,36,37]. Our findings revealed that different chatbot types were associated with different perceived strengths, suggesting the need for context-specific health care AI tools rather than a single "one-size-fits-all" chatbot that can perform well across all possible dimensions.

Limitations

This study has several limitations. First, participants interacted with chatbots in hypothetical health care scenarios rather than real clinical settings, which may not fully reflect real-world health care interactions. Second, the study focused on short-term interactions rather than longitudinal engagement, and perceptions may differ in the longer term. Third, all chatbot conditions were powered by GPT-4o-mini, a contemporary

and widely used foundation model at the time of the study [30], and findings may differ with newer or more advanced AI models as health care chatbot capabilities continue to evolve rapidly. Finally, participants were recruited online, which may limit the generalizability of the findings across broader patient and health care provider populations, particularly individuals with limited experience using technology or AI tools.

Conclusions

This study examined how different AI assisting roles and support types affect perceptions among patients and health care providers. The task chatbot was received most positively, drawing the highest ratings for competence, informational support, attitude, and adoption intention, and patients rated the chatbots more favorably than providers. The companion chatbot improved acceptance-related attitudes over the general-purpose baseline but did not enhance the warmth or emotional support it was designed to provide, performing no better than the default configuration on those dimensions. These findings highlight both the potential and the limitations of health care AI chatbots, and suggest that future systems should build on domain-specific expertise and users' prior trust in AI, while finding better ways to convey warmth that users actually perceive.

Acknowledgments

OpenAI ChatGPT (GPT-5-series models) and Claude were used to facilitate manuscript preparation. This use included converting the manuscript from LaTeX to Microsoft Word; converting CSV-format tables into LaTeX-format tables; identifying and correcting grammatical errors; and polishing text written by the authors using prompts such as "Find a synonym for X," "Add transition words to this sentence," and "Shorten this sentence without altering its meaning." ChatGPT and Claude were not used to generate original scientific arguments, analyses, results, or interpretations. The authors reviewed and verified all outputs and take full responsibility for the final manuscript.

Conflicts of Interest

De Choudhury serves on OpenAI's Expert Council on Well-Being and AI. The other authors declare no conflicts of interest.

Funding

This work was supported by the Stanford RAISE Seed Grant.

Data Availability

The data sets generated and analyzed during this study are not publicly available because they contain potentially sensitive participant survey responses and chat logs. Deidentified data may be available from the corresponding author upon reasonable request, subject to institutional approval and an appropriate data use agreement.

Authors' Contributions

OXZ led the study design, data collection, analysis, and manuscript preparation. LS and JH contributed to the study design. MDC and SXL supervised the project. All authors reviewed and approved the final manuscript.

Abbreviations

AI: artificial intelligence

ANCOVA: analysis of covariance

GPT: generative pretrained transformer

LDA: latent Dirichlet allocation

LLM: large language model

References

1. Li Y-H, Li Y-L, Wei M-Y, and Li G-Y. Innovation and challenges of artificial intelligence technology in personalized healthcare. *Scientific reports* 2024
2. Hussain HK, Tariq A, Gill AY, and Ahmad A. Transforming healthcare: The rapid rise of artificial intelligence revolutionizing healthcare applications. *BULLET: Jurnal Multidisiplin Ilmu* 2022
3. Hosseini A and Serag A. Self-supervised learning powered by synthetic data from diffusion models: Application to x-ray images. *IEEE Access* 2025
4. Li H, Zhang R, Lee Y-C, Kraut RE, and Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine* 2023
5. Milne-Ives M, Cock CD, Lim E, Shehadeh MH, Pennington ND, Mole G, Normando E, and Meinert E. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of medical Internet research* 2020
6. Kumar M. AI-driven healthcare chatbots: Enhancing access to medical information and lowering healthcare costs. *J Arti Inte Cloud Comp* 2023
7. Glikson E and Woolley AW. Human trust in artificial intelligence: Review of empirical research. *Academy of management annals* 2020. doi:10.5465/annals.2018.0057
8. Shamszare H and Choudhury A. Clinicians' perceptions of artificial intelligence: focus on workload, risk, trust, clinical decision making, and clinical integration. *Healthcare* 2023
9. Hah H and Goldin DS. How clinicians perceive artificial intelligence-assisted technologies in diagnostic decision making: Mixed methods approach. *Journal of Medical Internet Research* 2021
10. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, and Tamsah M-H. Artificial intelligence (AI) chatbots in medicine: a supplement, not a substitute. *Cureus* 2023
11. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Nadkarni GN, and Klang E. Large language models in real-world clinical workflows: A systematic review of applications and implementation. *Frontiers in Digital Health* 2025. doi:10.3389/fdgth.2025.1659134. PMID:41098649
12. Riedl R, Hogeterp SA, and Reuter M. Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice. *Frontiers in psychology* 2024
13. Busch F, Hoffmann L, Xu L, Zhang LJ, Hu B, García-Juárez I, Toapanta-Yanchapaxi LN, Gorelik N, Gorelik V, Rodriguez-Granillo GA, and others. Multinational attitudes toward AI in health care and diagnostics among hospital patients. *JAMA network open* 2025
14. Nong P and Platt J. Patients' trust in health systems to use artificial intelligence. *JAMA Network Open* 2025
15. Schulz PJ, Lwin MO, Kee KM, Goh WW, Lam TYT, and Sung JJ. Modeling the influence of attitudes, trust, and beliefs on endoscopists' acceptance of artificial intelligence applications in medical practice. *Frontiers in Public Health* 2023
16. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, and Smith DM. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine* 2023. doi:10.1001/jamainternmed.2023.1838
17. Chen D, Chauhan K, Parsa R, Liu ZA, Liu F-F, Mak E, Eng L, Hannon BL, Croke J, Hope A,

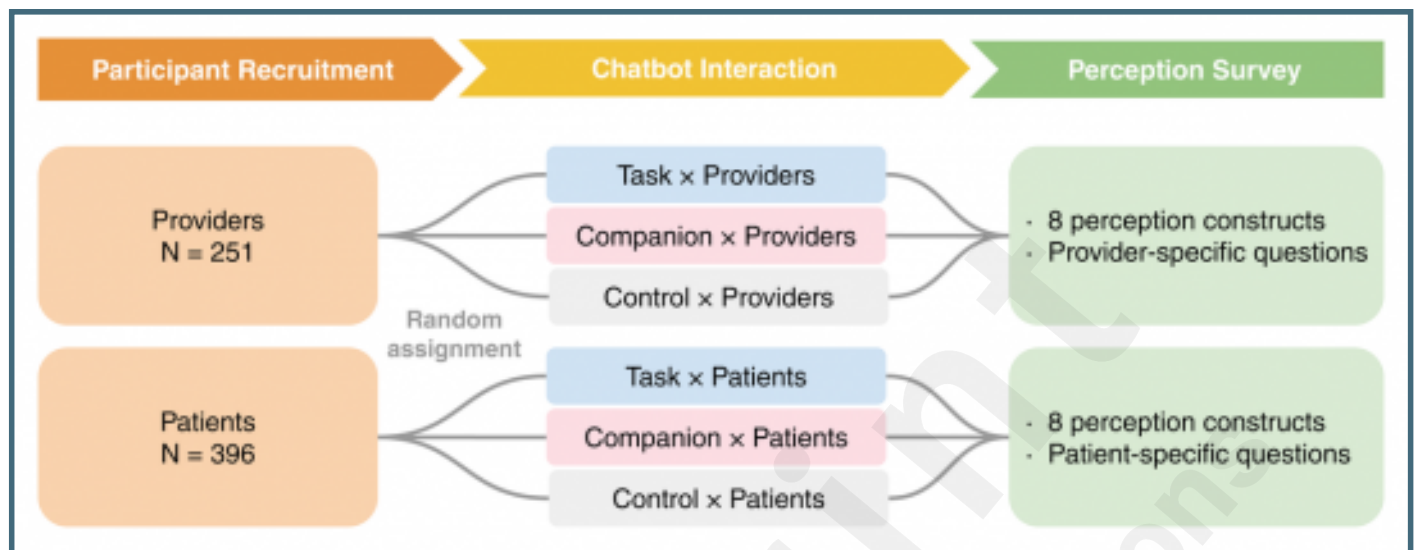
- Fallah-Rad N, Wong P, and Raman S. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. *npj Digital Medicine* 2025. doi:10.1038/s41746-025-01671-6. PMID:40360673
18. Langford CPH, Bowsher J, Maloney JP, and Lillis PP. Social support: a conceptual analysis. *Journal of advanced nursing* 1997
19. Wang Y-C, Kraut R, and Levine JM. To stay or leave? the relationship of emotional and informational support to commitment in online health support groups. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* 2012. doi:10.1145/2145204.2145329
20. Russe MF, Rau A, Ermer MA, Rothweiler R, Wenger S, Klöble K, Schulze RK, Bamberg F, Schmelzeisen R, and Reiser M. A content-aware chatbot based on GPT 4 provides trustworthy recommendations for Cone-Beam CT guidelines in dental imaging. *Dentomaxillofacial Radiology* 2024
21. Yonatan-Leus R and Brukner H. Comparing perceived empathy and intervention strategies of an AI chatbot and human psychotherapists in online mental health support. *Counselling and Psychotherapy Research* 2025
22. Liu B and Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 2018
23. Khadpe P, Krishna R, Fei-Fei L, Hancock JT, and Bernstein MS. Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction* 2020. doi:10.1145/3415234
24. Liang K-H, Shi W, Oh Y, Wang H-C, Zhang J, and Yu Z. Dialoging resonance: How users perceive, reciprocate and react to chatbot's self-disclosure in conversational recommendations. *arXiv preprint arXiv:2106.01666* 2022
25. Nourani M, Hashky A, and Ragan ED. User profiling in human-AI design: an empirical case study of anchoring bias, individual differences, and AI attitudes. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 2024
26. Lee E. The power of perception in human-AI interaction: Investigating psychological factors and cognitive biases that shape user belief and behavior. 2024
27. Wu PF, Summers C, Panesar A, Kaura A, and Zhang L. AI hesitancy and acceptability—perceptions of AI chatbots for chronic health management and long COVID support: Survey study. *JMIR Human Factors* 2024. doi:10.2196/51086
28. Wissawaswaengsuk P, Kumar P, Frank B, and Badir YF. The role of trust as the facilitator and contingency factor in the adoption of digital healthcare services: A telemedicine context. *Computers in Human Behavior* 2025. doi:10.1016/j.chb.2025.108722
29. St L and Wold S. Analysis of variance (ANOVA). *Chemometrics and intelligent laboratory systems* 1989
30. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI 2024. URL: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>
31. Faul F, Erdfelder E, Lang A-G, and Buchner A. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 2007
32. Cohen J. Statistical power analysis for the behavioral sciences. 2013
33. Hartman R, Moss AJ, Jaffe SN, Rosenzweig C, Litman L, and Robinson J. Introducing Connect by CloudResearch: Advancing online participant recruitment in the digital age. 2023. doi:10.31234/osf.io/ksgyr
34. Douglas BD, Ewell PJ, and Brauer M. Data quality in online human-subjects research: Comparisons between MTurk, prolific, CloudResearch, qualtrics, and SONA. *Plos one* 2023
35. Peer E, Rothschild D, Gordon A, Evernden Z, and Damer E. Data quality of platforms and panels for online behavioral research. *Behavior research methods* 2022
36. Davis FD, Bagozzi RP, and Warshaw PR. Technology acceptance model. *J Manag Sci* 1989
37. Cuddy AJ, Fiske ST, and Glick P. Warmth and competence as universal dimensions of social

- perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology* 2008
38. Ajzen I. The theory of planned behavior. *Organizational behavior and human decision processes* 1991
39. Cutrona CE and Suhr JA. Social support communication in the context of marriage: An analysis of couples' supportive interactions. *Communication of social support: Messages, interactions, relationships, and community* 1994
40. Joshi A, Kale S, Chandel S, and Pal DK. Likert scale: Explored and explained. *British journal of applied science & technology* 2015
41. Tavakol M and Dennick R. Making sense of Cronbach's alpha. *International journal of medical education* 2011
42. Hutto CJ and Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the eighth international AAAI conference on weblogs and social media (ICWSM)* 2014
43. Bao J, Wu J, Zhang Y, Chandrasekharan E, and Jurgens D. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. *Proceedings of the web conference 2021 (WWW '21)* 2021. doi:[10.1145/3442381.3450122](https://doi.org/10.1145/3442381.3450122)
44. Blei DM, Ng AY, and Jordan MI. Latent Dirichlet allocation. *Journal of Machine Learning Research* 2003
45. Saha K, Jain Y, Rodriguez VJ, and Choudhury MD. Linguistic Comparison of AI-and Human-Written Responses to Online Mental Health Queries. *arXiv preprint arXiv:2504.09271* 2025
46. Crum AJ, Akinola M, Martin A, and Fath S. The role of stress mindset in shaping cognitive, emotional, and physiological responses to challenging and threatening stress. *Anxiety, stress, & coping* 2017
47. Lee AY and Hancock JT. Social media mindsets: A new approach to understanding social media use and psychological well-being. *Journal of Computer-Mediated Communication* 2024
48. Seitz L. Artificial empathy in healthcare chatbots: Does it feel authentic? *Computers in Human Behavior: Artificial Humans* 2024
49. Zhang OX, Zhou S, Geng J, Liu Y, and Liu SX. Dr. GPT in campus counseling: Understanding higher education students' opinions on LLM-assisted mental health services. *arXiv preprint arXiv:2409.17572* 2024

Supplementary Files

Figures

Overview of the 2×3 factorial study design. Participants (health care providers or patients) were randomly assigned to one of three chatbot support types (task, companion, or control), yielding six conditions.



Multimedia Appendixes

Chatbot Design Prompts.

URL: <http://asset.jmir.pub/assets/c61e017cdfc14f69625e5af9ed3fe69e.docx>

Study Scenarios.

URL: <http://asset.jmir.pub/assets/8cc05eced9e1d4bc9b5fa5a79e43743a.docx>

Survey Questions.

URL: <http://asset.jmir.pub/assets/baa1a7208a02a6c4d37238b29ccf5bc5.docx>

Participant Demographics.

URL: <http://asset.jmir.pub/assets/f666cb20c0e24828c83eb63a37367d93.docx>

CHERRIES checklist (Checklist for Reporting Results of Internet E-Surveys).

URL: <http://asset.jmir.pub/assets/49572537d623d23677af7ba77fee2892.docx>