Framing Health Information: The Impact of Search Methods and Source Types on User Trust and Satisfaction in the Age of LLMs

Hye Sun Yun
Northeastern University
Boston, Massachusetts, USA
yun.hy@northeastern.edu

Timothy Bickmore
Northeastern University
Boston, Massachusetts, USA
t.bickmore@northeastern.edu

Abstract

Large language model (LLM)-based chatbots are transforming online health information search by offering interactive access to resources but raise concerns about inaccurate or harmful content. This study examined how different search methods—Search Engine, standalone Chatbot, and retrieval-augmented Chatbot+—and source credibility (reputable health websites vs. social media) influence user trust and satisfaction. Key findings include: (a) Trust trended higher for chatbots than Search Engine results, regardless of source credibility; (b) Satisfaction was highest with standalone Chatbot, followed by Chatbot+ and Search Engine; (c) Source type had minimal impact unless they were compared side by side. Interestingly, in interviews where participants could compare the methods directly, several participants preferred search engines due to familiarity and response diversity. However, they valued chatbots for their concise, time-saving answers. This study highlights the critical role of user interfaces in fostering trust and satisfaction, emphasizing the need for accurate, responsibly designed chatbots for health information dissemination.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

health information search, large language models, chatbots, search engine, human trust perception

ACM Reference Format:
Hye Sun Yun and Timothy Bickmore. 2025. Framing Health Information: The Impact of Search Methods and Source Types on User Trust and Satisfaction in the Age of LLMs. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3706599.3720239

1 Introduction

Over the past year, numerous studies have highlighted the inaccuracies of medical information provided by large language model (LLM)-powered chatbots such as ChatGPT. Across medical specialties like orthopedics, sleep apnea, and urology, the accuracy of chatbot responses to questions posed by researchers has been as low as 4% [4, 5, 13, 20, 31, 32], with one study reporting that 33% of LLM chatbot responses were deemed harmful by at least one physician on a panel of judges [11]. These findings suggest users should place less trust in chatbot-provided medical information compared to authoritative sources, such as websites authored by the US Centers for Disease Control (CDC) or the National Library of Medicine, accessed through search engines like Google. However, emerging evidence indicates that consumers may still place unwarranted trust in medical information from LLM chatbots. For instance, a 2024 survey of 2,428 adults found that 17% of all respondents—and 25% of adults aged 18–29—reported using LLM chatbots regularly for health advice [29]. Another study revealed that 78% of monthly ChatGPT users were willing to rely on the chatbot as-is for medical diagnosis [33].

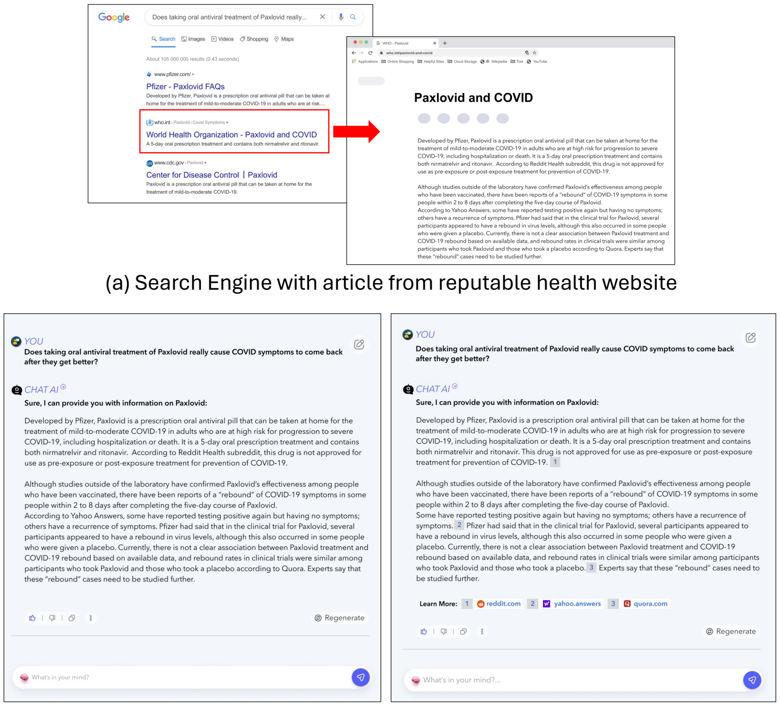
Given the availability of authoritative sources of health information online, why are consumers turning to LLM chatbots? Is it due to convenience, a lack of awareness of chatbots' low accuracy and safety concerns, or inherent differences in how chatbots and search engines present information? These questions motivate the central focus of this research: *Does the mere fact that medical information is obtained from a chatbot influence user trust compared to identical information obtained from a search engine?*

Several theoretical frameworks offer insights into user attitudes in such situations. Framing theory suggests that the way information is presented, or "framing", can shape consumer perceptions of the information [21]. For instance, "media frames" can emphasize certain aspects of a message while downplaying others, influencing interpretation. Additionally, users tend to anthropomorphize chatbots, which can increase trust and satisfaction during interactions [6, 18]. While this study does not manipulate the human-like characteristics of chatbots, it is reasonable to assume users will perceive chatbots as more anthropomorphic than search engines. Another benefit of chatbots may involve reducing users' cognitive efforts. Unlike search engines, which require users to locate, extract, and summarize information across multiple sources, chatbots provide direct and concise responses, potentially increasing user satisfaction. We hypothesize that this reduced cognitive effort, combined with perceived anthropomorphism, may lead users to rate chatbot responses as more satisfying and helpful than search engine results.

To deepen our understanding, we extended our investigation in two ways. First, we examine chatbots that incorporate retrieval-augmented generation (RAG) to provide source references alongside their answers [14]. These "Chatbot+" systems bridge the gap between standalone chatbots and search engines by including source links to enhance trust. Second, drawing on credibility theory, which addresses how users discern credible from unreliable information,

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720239



About 105 000 000 results (0.43 seconds)

www.pfizer.com/
Pfizer - Paxlovid FAQs
Developed by Pfizer, Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19 in adults who are at risk....

who.int › Paxlovid › Covid Symptoms
World Health Organization - Paxlovid and COVID
A 5-day oral prescription treatment and contains both nirmatrelvir and ritonavir.

www.cdc.gov › Paxlovid
Center for Disease Control | Paxlovid
Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19.

Paxlovid and COVID

Developed by Pfizer, Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19 in adults who are at high risk for progression to severe COVID-19, including hospitalization or death. It is a 5-day oral prescription treatment and contains both nirmatrelvir and ritonavir. According to Reddit Health subreddit, this drug is not approved for use as pre-exposure or post-exposure treatment for prevention of COVID-19.

Although studies outside of the laboratory have confirmed Paxlovid's effectiveness among people who have been vaccinated, there have been reports of a "rebound" of COVID-19 symptoms in some people within 2 to 8 days after completing the five-day course of Paxlovid.
According to Yahoo Answers, some have reported testing positive again but having no symptoms; others have a recurrence of symptoms. Pfizer had said that in the clinical trial for Paxlovid, several participants appeared to have a rebound in virus levels, although this also occurred in some people who were given a placebo. Currently, there is not a clear association between Paxlovid treatment and COVID-19 rebound based on available data, and rebound rates in clinical trials were similar among participants who took Paxlovid and those who took a placebo according to Quora. Experts say that these "rebound" cases need to be studied further.

(a) Search Engine with article from reputable health website

YOU
Does taking oral antiviral treatment of Paxlovid really cause COVID symptoms to come back after they get better?

CHAT AI
Sure, I can provide you with information on Paxlovid:

Developed by Pfizer, Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19 in adults who are at high risk for progression to severe COVID-19, including hospitalization or death. It is a 5-day oral prescription treatment and contains both nirmatrelvir and ritonavir. According to Reddit Health subreddit, this drug is not approved for use as pre-exposure or post-exposure treatment for prevention of COVID-19.

Although studies outside of the laboratory have confirmed Paxlovid's effectiveness among people who have been vaccinated, there have been reports of a "rebound" of COVID-19 symptoms in some people within 2 to 8 days after completing the five-day course of Paxlovid.
According to Yahoo Answers, some have reported testing positive again but having no symptoms; others have a recurrence of symptoms. Pfizer had said that in the clinical trial for Paxlovid, several participants appeared to have a rebound in virus levels, although this also occurred in some people who were given a placebo. Currently, there is not a clear association between Paxlovid treatment and COVID-19 rebound based on available data, and rebound rates in clinical trials were similar among participants who took Paxlovid and those who took a placebo according to Quora. Experts say that these "rebound" cases need to be studied further.

(b) Chatbot with sources from social media websites

YOU
Does taking oral antiviral treatment of Paxlovid really cause COVID symptoms to come back after they get better?

CHAT AI
Sure, I can provide you with information on Paxlovid:

Developed by Pfizer, Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19 in adults who are at high risk for progression to severe COVID-19, including hospitalization or death. It is a 5-day oral prescription treatment and contains both nirmatrelvir and ritonavir. This drug is not approved for use as pre-exposure or post-exposure treatment for prevention of COVID-19. 1

Although studies outside of the laboratory have confirmed Paxlovid's effectiveness among people who have been vaccinated, there have been reports of a "rebound" of COVID-19 symptoms in some people within 2 to 8 days after completing the five-day course of Paxlovid.
Some have reported testing positive again but having no symptoms; others have a recurrence of symptoms. 2 Pfizer had said that in the clinical trial for Paxlovid, several participants appeared to have a rebound in virus levels, although this also occurred in some people who were given a placebo. Currently, there is not a clear association between Paxlovid treatment and COVID-19 rebound based on available data, and rebound rates in clinical trials were similar among participants who took Paxlovid and those who took a placebo. 3 Experts say that these "rebound" cases need to be studied further.

Learn More: 1 reddit.com 2 yahoo.answers 3 quora.com

(c) RAG-based Chatbot+ with sources from social media websites

Figure 1: Screenshots from simulation videos for the topic on Paxlovid drug. Search engine interaction showing an article from a reputable health-related website (WHO) (a), Chatbot response with information from social media platforms (b), and Chatbot+ with source links from social media platforms (c).

we explore the impact of source credibility on trust and satisfaction [24]. Accordingly, we also investigate the impact of manipulating the perceived source of information provided, varying from high-credibility (e.g., CDC) to low-credibility (e.g., Reddit) origin.

This study presents findings from experiments where participants search for medical information using one of three interfaces: a Search Engine and retrieved web page, a Chatbot query and response, or a Chatbot+ query and response with links to source pages. The initial query and resulting information remain constant across conditions. Our hypotheses are:

- **H1**: For fixed source credibility, trust ratings will be ranked: Chatbot+ > Chatbot > Search Engine;
- **H2**: For fixed media framing, trust ratings will be higher for credible health websites compared to social media sources;
- **H3**: For fixed source credibility, satisfaction and helpfulness ratings will be ranked as follows due to reduced cognitive effort and anthropomorphism for chatbots: Chatbot+ > Chatbot > Search Engine.

This work sheds light on the influence of user interfaces and source credibility on trust and satisfaction in health information seeking, providing critical insights into the design of effective and reliable LLM-powered chatbots.

2 Related Work

Research on trust in LLM-based chatbots for health information search has gained significant attention due to the complex interplay of interface design, source credibility, and user perception. Several studies confirm that source and presentation style strongly influence credibility judgments in health contexts [8, 12, 37]. Claggett et al. [8] found that physician-authored, objectively presented information enhances perceived credibility, although Bates et al. [1] demonstrated that differences in attribution to a source did not have a significant effect on consumer's evaluation of the quality of the information. Building on this, Sun et al. [35] explored how user interfaces (text-based, speech-based, embodied) impact trust in health information delivered by LLMs. Findings revealed that trust levels were highest for text-based interfaces, emphasizing the critical role of modality in shaping trust. Similarly, another study demonstrated that users generally exhibit higher trust in ChatGPT over Google in health-related contexts, with interactive features and prior experience driving trust in search agents and their outputs [36].

Other studies delve into nuanced factors influencing trust in chatbot-mediated health information. Liu et al. [23] validated the heuristic–systematic model (HSM) in the health chatbot context, highlighting that source expertise is a critical heuristic cue, essential for personalization to influence user trust and behavior. Sharma et al. [34] examined how LLM-powered search systems impact selective exposure, finding that conversational search increased biased information querying and that opinionated LLMs reinforcing user views exacerbated this bias. In addition, Jin et al. [17] identified the interplay between users' health literacy levels and chatbot design elements such as gender and doctor-like cues, highlighting how these factors shape cognitive and emotional trust. Other studies emphasized the importance of tailored communication styles, non-technical language, and conversational tone in fostering user satisfaction and engagement [19, 22]. Collectively, these works illustrate the multifaceted nature of trust and underscore the importance of evaluating not only chatbot design but also the interplay between information source and format, as addressed in our study. Our study is also novel in its comparison of RAG-based chatbot results to other search interfaces [14].

3 Methods

We conducted two studies to address our research question: an online survey and semi-structured interviews. The online survey used a between-subjects design to examine perceptions of health information across three search methods and two source types. The semi-structured interviews employed a within-subjects design with different participants to further explore the impact of search methods and source types on health information delivery. This approach provides quantitative and qualitative data to assess the effect of search methods and source types while uncovering deeper insights into trust and satisfaction through interviews where participants can compare all three search methods. We recruited all participants from the Prolific online research platform. Participants were 18 years or older and used English as a primary language. They received 12.00 USD per hour rate for their involvement. This study received ethics approval from our institute's Institutional Review Board.

3.1 Materials & Procedures

3.1.1 Simulation Videos. Twelve one-minute simulation videos were created to showcase three search methods and two types of information sources with varying credibility. These videos served as stimuli for both surveys and interviews. The search methods included: (1) a Search Engine leading to an article, (2) a standalone Chatbot, and (3) a Chatbot+ with source links. For the two source types, we used the following: reputable health websites (i.e., CDC, WHO, European Medicines Agency) and social media platforms (i.e., Reddit, Yahoo Answers, Quora). To ensure relevance and potential high-risk health context, we focused on two medication-related queries: (1) "Does taking oral antiviral treatment Paxlovid cause COVID symptoms to return after they improve?" and (2) "What nasal decongestant can I safely use for seasonal allergies if I'm taking Lisinopril for blood pressure and an antacid for acid reflux?" The prototypes were designed in Figma, and videos were recorded using a screen capture tool. Further details about content creation are provided in Appendix A. Figure 1 shows screenshots from select videos.

3.1.2 Survey. The cross-sectional, anonymous survey was conducted in May 2024 with 300 international participants recruited through consecutive sampling. It was administered in English via Qualtrics and took 15 minutes to complete. The survey collected data on sociodemographics, chronic health status, familiarity with ChatGPT, and included several validated measures. The eHealth Literacy Scale (eHEALS) [26] assesses perceived eHealth literacy with 8 items on a 5-point scale (score range: 8–40). Higher scores indicate better literacy. It has been widely used across various populations [7, 15, 25, 28]. The Trust in the Health Care Team (T-HCT) Scale [30] measures trust in healthcare teams using 29 items on a 5-point scale, averaged for a composite score. In addition, the Short-Form AI Attitude Scale (AIAS-4) [16] evaluates trust in AI technology with 4 items on a 10-point scale, averaged for a composite score. An additional item assessed the perceived benefits of AI in medicine.

Participants also viewed a randomly selected one-minute simulation video (Section 3.1.1) and rated the information on a 5-point scale (from "Strongly disagree" to Strongly agree") for accuracy, satisfaction, helpfulness, trustworthiness, usefulness, ease of understanding, and anxiety reduction. They also indicated their likelihood of cross-checking or following the health advice (5-point scale from "Extremely unlikely" to "Extremely likely") and provided open-ended responses on how they would cross-check the information. The full survey is available in the Supplementary Materials.

3.1.3 Interviews. The semi-structured interviews were conducted in English via Zoom, lasting about 60 minutes each. Participants first completed an online questionnaire that collected sociodemographic data, chronic health status, familiarity with ChatGPT, and responses to three validated scales: eHealth Literacy Scale (eHEALS) [26], Trust in the Health Care Team Scale (T-HCT) [30], and the Short-Form AI Attitude Scale (AIAS-4) [16]. Next, participants viewed three simulation videos showcasing different search methods related to Paxlovid. They then participated in a semi-structured interview to discuss their trust and satisfaction with the search methods and source types for health information. A Latin square design was used to counterbalance the video order. The complete interview guide is available in the Supplementary Materials.

4 Results

4.1 Online Survey

Of the 300 participants, three were excluded for failing the attention-check question. Participants' ages ranged from 18 to 78 years (mean = 36.0, SD = 12.7). Most were male (62.0%), White (64.0%), without a chronic disease diagnosis (67.0%), and English-speaking as a first language (81.5%). Education levels ranged from 10 to 26 years (mean = 16.0, SD = 2.9), with 40.4% reporting household income near the median for their country, according to OECD Better Life Index data [27]. The most represented countries were the United States (23.2%), the United Kingdom (18.2%), Canada (14.5%), Australia (10.1%), South Africa (8.8%), and Poland (4.7%). Participants' mean scores were: T-HCT = 3.5 (SD = 0.6), eHEALS = 30.8 (SD = 4.7), and AIAS-4 = 6.5 (SD = 2.3). Most perceived AI as beneficial in medicine, with a median score of 7 (IQR = 4) on a 10-point scale. Only 21.2% reported using LLM-based chatbots for health information in the past year.

4.1.1 Perceptions of Online Health Information. No statistically significant differences were observed in baseline sociodemographic characteristics across conditions. Measures of accuracy, trust, and intent to cross-check (reverse-coded) were grouped together as

Table 1: Means and standard deviations of trust and satisfaction composite scores by search method and source type.

Search Method	Source Type	Trust Mean (SD)	Satisfaction Mean (SD)
Search Engine	Both	2.83 (0.79)	3.39 (0.95)
	Health Websites	3.00 (0.86)	3.45 (1.03)
	Social Media	2.66 (0.68)	3.34 (0.88)
Chatbot	Both	3.03 (0.74)	3.69 (0.81)
	Health Websites	3.13 (0.64)	3.70 (0.72)
	Social Media	2.93 (0.82)	3.69 (0.90)
Chatbot+	Both	3.09 (0.86)	3.63 (0.87)
	Health Websites	2.99 (0.93)	3.67 (0.90)
	Social Media	3.19 (0.77)	3.58 (0.85)

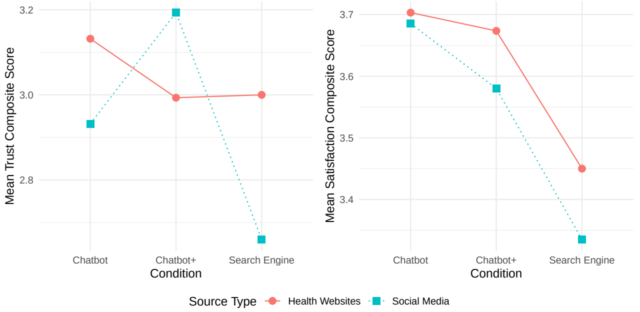


Figure 2: Effects of search method and source type on (A) trust and (B) satisfaction. Trust was higher for Chatbot with Health Websites and Chatbot+ with Social Media than Search Engine with Social Media. Satisfaction was higher for both Chatbot and Chatbot+ than Search Engine across both sources.

they were statistically related, and the average was calculated to form a composite trust score. Similarly, measures of satisfaction, helpfulness, usefulness, and ease of understanding were averaged to form a composite satisfaction score, as these measures were highly related. The means and standard deviations for the trust and satisfaction composite scores are presented in Table 1 and Figure 2.

A two-way ANOVA was performed to evaluate the effects of the search method and source type on trust composite scores. Results showed a trending main effect for search method (F(2, 291) = 3.02, p = .050, η^2 = 0.02), no significant main effect for source type (F(1, 291) = 1.50, p = .221, η^2 = 0.005), and a significant interaction between search method and source type (F(2, 291) = 3.14, p = .045, η^2 = 0.02). Post hoc tests using Tukey's HSD revealed that trust scores were higher for Chatbot+ compared to Search Engine (p = .050), and significantly higher for participants using Chatbot with Health Website sources and Chatbot+ with Social Media sources compared to Search Engine with Social Media sources (p = .039 and p = .011). In addition, a second two-way ANOVA assessed the effects of the search method and source type on satisfaction composite scores. There was a significant main effect for search method (F(2, 290) = 3.16, p = .044, η^2 = 0.02), no significant main effect for source type (F(1, 290) = .54, p = .462, η^2 = 0.002), and no significant interaction (F(2, 290) = .08, p = .922, η^2 = 0.0005). Post hoc testing indicated significantly higher satisfaction scores for Chatbot method compared to Search Engine method (p = .046). However, there was no significant difference between Chatbot+ and Search Engine conditions (p = .152). Additionally, we did not find any significant effect of the search method or source type on the intent to act or levels of anxiety.

4.1.2 Subgroup Analyses. Subgroup analyses were conducted to explore the differential effect of search methods on trust and satisfaction among groups with higher eHealth literacy, higher positive attitudes towards AI, and greater familiarity with ChatGPT. The subgroups were based on the median scores.

Participants with higher eHealth literacy reported greater trust in Chatbot+ (mean=3.18, SD=0.90) compared to Search Engine (mean=2.78, SD=0.81), with a significant difference (p = .021) A similar trend was observed in participants with higher positive attitudes towards AI. Trust scores from both Chatbot+ (mean=3.35, SD=0.80) and Chatbot (mean=3.13, SD=.76) were significantly higher than Search Engine (mean=2.76, SD=0.85), with p-values of .001 and .0496 respectively. Participants with higher familiarity with ChatGPT showed significantly higher trust in Chatbot+ (mean=3.10, SD=0.88) compared to Search Engine (mean=2.72, SD=0.77), with p =.01.

For satisfaction, participants with more positive attitudes towards AI also reported higher scores for Chatbot+ (mean = 3.84, SD = 0.80) and Chatbot (mean = 3.80, SD = 0.79) compared to Search Engine (mean = 3.33, SD = 1.01), with p-values of .01 and .02, respectively. Similarly, those with greater familiarity with ChatGPT reported higher satisfaction in Chatbot+ (mean = 3.68, SD = 0.84) and Chatbot (mean = 3.72, SD = 0.83) compared to Search Engine (mean = 3.28, SD = 0.99), with p-values of .02 and .01.

4.2 Semi-structured Interviews

We interviewed 6 participants recruited from Prolific. Most of the participants were female (66.7%), Black (83.3%), without a chronic disease diagnosis (66.7%), and spoke English as a first language (100.0%). Their ages ranged from 22 to 31 years (mean = 27.33, SD = 4.23). Participants reported 12–18 years of formal education (mean = 13.8, SD = 2.6). All but one person were from South Africa. Participants' mean scores were: T-HCT = 3.9 (SD = 0.7), eHEALS = 38.7 (SD = 2.7), and AIAS-4 = 8.7 (SD = 1.1). Most participants perceived AI as beneficial in medicine, with a median score of 9.5 (IQR = 1.0) on a 10-point scale. Four out of six participants reported that they have used LLM-based chatbots to search for health information in the past year. We conducted an inductive thematic analysis of the transcripts from 123 minutes of recording [9]. The analysis focused on reasons why certain search methods were preferable to others, and to better understand the effects of source types on trust and perceptions of accuracy. We used elements of the grounded theory method, including open, axial, and selective coding [10].

Search Engine: familiar experience with a rich array of information. Five out of six interview participants preferred the Search Engine the most. This difference from the survey results may be due to the fact that interview participants were able to view and compare all study conditions, whereas survey participants only saw one condition, making subtle differences more salient. The main reason for preferring the Search Engine was its familiarity and ease of use. As P4 noted, *"it is very familiar because every time ... I don't feel well, the first thing I go to is the search engines."* Furthermore, participants mentioned how search engines provide the ability to navigate to specific websites. In addition to the familiar experience of using a search engine, three participants mentioned that search engines provide rich and diverse information, including images and videos, which chatbots do not. P5 said, *"... there's a lot of information on the search engine and you feel more comfortable with it because there's obviously some reviews on it as well, and there's also pictures, videos."*. However, some participants also noted downsides to search engines, such as the effort required to sift through information, the overwhelming number of options, and occasional difficulty finding answers.

Chatbot: straightforward answers from unclear sources. All participants found both Chatbot and Chatbot+ responses to be direct or straightforward which can save time when searching for health information. Instead of visiting multiple websites, the answer was provided in a well-organized, summarized format. P1 appreciated this simplicity, stating, *"It was straight to the point ... I liked things that are straight to the point, not waste my time."* While some found the direct answers helpful, others saw the lack of diverse perspectives as a disadvantage. P6 said, *"The chatbot, the advantage is that it gives you a specific information, but it doesn't give diverse information. That is a disadvantage."*. Additionally, three participants mentioned the lack of transparency about the sources used by Chatbot, which made them hesitant to fully trust the information. P3 shared *"The information was there and everything, although I was not a hundred percent sure if it's the correct things and all the stuff."* P4 further expressed concern about the lack of source transparency, *"They usually say AI gets its information from different sources, so I'm not sure if they let the AI get that information fact-checked or maybe it's regulated information or what. I might not know ... If it uses Facebook information, for example, then that wouldn't be trustworthy."*

Chatbot+: direct answers with opportunities to cross-reference. Participants found Chatbot+ to have similar benefits as Chatbot but with the added advantage of source links at the bottom of the response. All participants rated Chatbot+ higher than Chatbot due to the ability to cross-check information. P2 specifically explained, *"It is interesting because you are able to crosscheck if there is similar things, and also if everything actually is the same in both the links in the chatbot."*. P1 raised concerns about the potential security risks of clicking on links provided by Chatbot+ and stated, *"Maybe they can hack you while you are on the website trying to link on the websites that they provide."*

Effect of source type on trust can depend on context. All but one participant noticed the difference between health websites and social media when two screenshots from the same search method but different source types were displayed side by side. The participant who did not notice the difference attributed it to the font size being too small on their screen. Of the five participants who identified the sources correctly as health websites and social media ("community platforms"), four found health websites to be more trustworthy and accurate. P4 mentioned, *"I know Pfizer, I know WHO, and then CDC, I know those ones. They deal with the health. They look like they're more trustworthy than the information you can get on the other side."* Interestingly, P2 and P3 believed that the source type did not matter for chatbots, as they assumed that AI summarizes information from all sources and gives the same output. P2 said *"I believe that AI has been trained with the relevant information regarding a lot of situations, whether health or life situations. So, mostly the information that's there, it's mostly reliable."* Surprisingly, P1 found social media sources more trustworthy, noting that the colorful logos on Chatbot+ made them feel more familiar and reliable, compared to the more text-heavy logos for health websites. This potentially explains the interaction effect (Chatbot+ with Social Media) reported in Section 4.1.1

5 Discussion

Across all survey participants, information retrieved by chatbots (Chatbot, Chatbot+) was rated with higher levels of trust compared to information retrieved by search engines, regardless of whether the information was from a reputable source, such as the CDC, or social media. This effect was especially strong for individuals with higher levels of eHealth literacy, positive attitudes toward AI, and familiarity with ChatGPT. This seems to generally align with findings from Sun et al. [36] where there was higher trust in ChatGPT over Google in health-related contexts. Similarly, across all participants, there was significantly higher satisfaction with chatbots compared to the search engine, and this effect was stronger for subgroups with greater positive attitudes toward AI and greater familiarity with ChatGPT. Rather than always leading to higher levels of trust, chatbot responses with reference links (Chatbot+) were never significantly different than Chatbot responses without references on participant ratings of trust. Participants were also most satisfied with the Chatbot without references, with Chatbot+ falling between the other conditions on satisfaction. Thus, there was partial support for H1, with trust rankings {Chatbot, Chatbot+} > Search Engine, no support for H2 (there were almost no differences on trust rankings for credible websites compared to social media from survey results), and partial support for H3, with satisfaction rankings Chatbot > Chatbot+ > Search Engine.

In interviews, when participants could compare the study conditions side by side, they indicated they preferred results from search engines more than those from chatbots, with familiarity with search engines and the diversity of search engine responses as the primary reasons. However, they appreciated chatbots' direct answers, which they felt could save time and effort. Participants reported lower trust in Chatbot due to uncertainty about their information sources and some distrust in Chatbot+ over concerns about non-secure reference links. When comparing information sources (e.g., CDC vs. Reddit), most participants noticed the difference in sources and said they would trust information from credible health websites more. Discrepancies between survey and interview findings may stem from study design differences. The side-by-side comparisons in interviews likely made the source distinctions more noticeable than in a between-subjects design.

In conclusion, when chatbots respond to medical queries that appear to be as concise, unambiguous, and authoritative as website text, users trust it more than if they had found exactly the same information via a search engine, and they are significantly more satisfied with the search experience. Only when they are forced to think about these different search methods through side-by-side comparisons, do they give reasons for preferring results from a search engine and reputable health websites. Similarly, they are less likely to question chatbot responses unless explicitly prompted to compare sources, emphasizing the need for transparency. Our study was among the first to incorporate RAG-based chatbots (Chatbot+) in comparisons of trust and user satisfaction in online health information search, while also considering source credibility as an independent variable. Our results highlight the importance of thoughtful design and transparency in health-related search tools to support informed decision-making, balancing ease of use with mechanisms to prevent overinflated user trust in potentially unverified information.

Limitations & Future Work. This study has several limitations. First, the sample was relatively small and recruited from an online crowdwork platform, potentially skewing the sample toward individuals who are more technologically literate or comfortable with digital tools. Second, the study focused solely on medication-related queries, which may not represent the broader range of health information topics that users commonly search for online. Further research is needed to understand how search methods and source types influence perceptions across diverse health search topics. Third, our simulations did not involve users directly interacting with the interface or content, which may have influenced their perceptions of the different conditions. This limitation is particularly relevant when comparing Chatbot and Chatbot+ since user experiences may have differed had users engaged with the systems firsthand. Future studies should have users directly interact with chatbot interfaces to better understand the differences between regular chatbots and RAG-based chatbots.

Given our finding that users preferred chatbots due to reduced cognitive effort, future research should investigate the role of individual differences, such as "Need for Cognition" [3], in shaping these preferences. Those with a high "Need for Cognition" may favor search engines that allow for source verification over chatbot responses. Lastly, this study may need to be periodically replicated due to shifts in user perceptions given the rapid evolution of chatbot technology.

Our findings from the survey and interviews raise several open questions about designing LLM-based chatbots for safe health information search. In the survey, respondents did not strongly consider source type, whereas interview participants, when explicitly prompted, reflected more on its importance. This suggests that making sources more prominent in chatbot responses could be a key design consideration. For example, RAG-based chatbots (Chatbot+) should provide more details on how sources were retrieved and used to generate responses as part of the responses. Enhancing transparency may be particularly valuable for users making critical health decisions based on chatbot-generated information. Future research should explore designs that improve source visibility and clarify response generation processes.

References

[1] Benjamin R Bates, Sharon Romina, Rukhsana Ahmed, and Danielle Hopson. 2006. The effect of source credibility on consumers' perceptions of the quality of health information on the Internet. *Medical informatics and the Internet in medicine* 31, 1 (2006), 45–52.

[2] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of medical Internet research* 20, 9 (2018), e11510.

[3] J Cacioppo, R Petty, J Feinstein, W Blair, and W. Jarvis. 1996. Dispositional Differences in Cognitive Motivation: The Life and Times of Individuals Varying in Need for Cognition. *Psychological Bulletin* 119, 2 (1996), 197–253.

[4] U. Caglar, O. Yildiz, A. Meric, A. Ayranci, M. Gelmis, O. Sarilar, and F. Ozgor. 2024. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. *J Pediatr Urol* 20, 1 (2024), 26.e1–26.e5. https://doi.org/10.1016/j.jpurol.2023.08.003

[5] D. J. Campbell, L. E. Estephan, E. V. Mastrolonardo, D. R. Amin, C. T. Huntley, and M. S. Boon. 2023. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. *J Clin Sleep Med* 19, 12 (2023), 1989–1995. https://doi.org/10.5664/jcsm.10728

[6] Avanti Chinmulgund, Ritesh Khatwani, Poornima Tapas, Pritesh Shah, and Ravi Sekhar. 2023. Anthropomorphism of AI based chatbots by users during communication. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*. 1–6. https://doi.org/10.1109/CONIT59222.2023.10205689

[7] Seon-Yoon Chung and Eun-Shim Nahm. 2015. Testing reliability and validity of the eHealth Literacy Scale (eHEALS) for older adults recruited online. *CIN: Computers, Informatics, Nursing* 33, 4 (2015), 150–156.

[8] Jennifer L Claggett, Brent Kitchens, and Maria Paino. 2024. Identifying the peripheral cues in the credibility assessment of online health information. *Information & Management* 61, 8 (2024), 104037.

[9] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

[10] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[11] D Dash, R Thapa, and J. Banda. 2023. Evaluation of GPT-3.5 and GPT-4 for supporting real- world information needs in healthcare delivery. (2023).

[12] Mohan Dutta-Bergman et al. 2003. Trusted online sources of health information: differences in demographics, health beliefs, and health-information orientation. *Journal of medical Internet research* 5, 3 (2003), e893.

[13] H. Fraser, D. Crossland, I. Bacher, M. Ranney, T. Madsen, and R. Hilliard. 2023. Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study. *JMIR Mhealth Uhealth* 11 (2023), e49995. https://doi.org/10.2196/49995

[14] Saba Ghanbari Haez, Marina Segala, Patrizio Bellan, Simone Magnolini, Leonardo Sanna, Monica Consolandi, and Mauro Dragoni. 2024. A retrieval-augmented generation strategy to enhance medical chatbot reliability. In *International Conference on Artificial Intelligence in Medicine*. Springer, Springer Nature Switzerland, Cham, 213–223.

[15] Jarod T Giger, Sheila Barnhart, Fran Feltner, Melissa Slone, Michael J Lawler, Leah Windsor, and Alistair Windsor. 2021. Validating the eHealth literacy scale in rural adolescents. *The Journal of Rural Health* 37, 3 (2021), 504–516.

[16] Simone Grassini. 2023. Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence. *Frontiers in psychology* 14 (2023), 1191628.

[17] Eunjoo Jin, Yuhosua Ryoo, WooJin Kim, and Y Greg Song. 2024. Bridging the health literacy gap through AI chatbot design: the impact of gender and doctor cues on chatbot trust and acceptance. *Internet Research* (2024).

[18] Elisa Konya-Baumbach, Miriam Biller, and Sergej von Janda. 2023. Someone out there? A study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior* 139 (2023), 107513.

[19] Samuel N. Koscelny and David M. Neyens. 2024. The Effect of Healthcare Chatbots' Information Presentation Styles on User Acceptance in a Knowledge Seeking Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 68, 1, 493–498. https://doi.org/10.1177/10711813241263509 arXiv:https://doi.org/10.1177/10711813241263509

[20] T. Kuroiwa, A. Sarcon, T. Ibara, E. Yamada, A. Yamamoto, K. Tsukamoto, and K. Fujita. 2023. The Potential of ChatGPT as a Self-Diagnostic Tool in Common Orthopedic Diseases: Exploratory Study. *J Med Internet Res* 25 (2023), e47621. https://doi.org/10.2196/47621

[21] I. Levin, S. Schneider, and G. Gaeth. 1998. All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. *ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES* 76, 2 (1998), 149–188.

[22] Courtney Linder. 2020. *The effects of a healthcare Chatbots' language and persona on user trust, satisfaction, and chatbot effectiveness.* Master's thesis. Clemson University.

[23] Yu-li Liu, Wenjia Yan, Bo Hu, Zhuoyang Li, and Yik Ling Lai. 2022. Effects of personalization and source expertise on users' health beliefs and usage intention toward health chatbots: Evidence from an online experiment. *Digital Health* 8 (2022), 20552076221129718.

[24] Miriam J Metzger, Andrew J Flanagin, Keren Eyal, Daisy R Lemus, and Robert M McCann. 2003. Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the International Communication Association* 27, 1 (2003), 293–335.

[25] Jennifer Nguyen, Michael Moorhouse, Barbara Curbow, Juliette Christie, Kim Walsh-Childers, Sabrina Islam, et al. 2016. Construct validity of the eHealth literacy scale (eHEALS) among two adult populations: a Rasch analysis. *JMIR public health and surveillance* 2, 1 (2016), e4967.

[26] Cameron D Norman and Harvey A Skinner. 2006. eHEALS: the eHealth literacy scale. *Journal of medical Internet research* 8, 4 (2006), e507.

[27] OECD. [n. d.]. OECD Better Life Index. https://www.oecdbetterlifeindex.org/topics/income/. [Accessed 16-05-2024].

[28] Samantha R Paige, Janice L Krieger, Michael Stellefson, and Julia M Alber. 2017. eHealth literacy in chronic disease patients: an item response theory analysis of the eHealth literacy scale (eHEALS). *Patient Education and Counseling* 100, 2 (2017), 320–326.

[29] M Presiado, A Montero, L Lopes, and L Hamel. 2024. KFF Health Misinformation Tracking Poll: Artificial Intelligence and Health Information. https://www.kff.org/health-misinformation-and-trust/poll-finding/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/

[30] Jennifer Richmond, Marcella H Boynton, Sachiko Ozawa, Kathryn E Muessig, Samuel Cykert, and Kurt M Ribisl. 2022. Development and validation of the trust in my doctor, trust in doctors in general, and trust in the health care team scales. *Social science & medicine* 298 (2022), 114827.

[31] H. R. Saeidnia, M. Kozak, B. D. Lund, and M. Hassanzadeh. 2024. Evaluation of ChatGPT's responses to information needs and information seeking of dementia patients. *Sci Rep* 14, 1 (2024), 10273. https://doi.org/10.1038/s41598-024-61068-5

[32] J. S. Samaan, Y. H. Yeo, N. Rajeev, L. Hawley, S. Abel, W. H. Ng, N. Srinivasan, J. Park, M. Burch, R. Watson, O. Liran, and K. Samakar. 2023. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. *Obes Surg* 33, 6 (2023), 1790–1796. https://doi.org/10.1007/s11695-023-06603-5

[33] Y. Shahsavar and A. Choudhury. 2023. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Hum Factors* 10 (2023), e47564. https://doi.org/10.2196/47564

[34] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. https://doi.org/10.1145/3613904.3642459

[35] Xin Sun, Yunjie Liu, Jan De Wit, Jos A. Bosch, and Zhuying Li. 2024. Trust by Interface: How Different User Interfaces Shape Human Trust in Health Information from Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 344, 7 pages. https://doi.org/10.1145/3613905.3650837

[36] Xin Sun, Rongjun Ma, Xiaochang Zhao, Zhuying Li, Janne Lindqvist, Abdallah El Ali, and Jos A. Bosch. 2024. Trusting the Search: Unraveling Human Trust in Health Information from Google and ChatGPT. (2024). arXiv:2403.09987 [cs.HC] https://arxiv.org/abs/2403.09987

[37] Yixuan Zhang, Nurul Suhaimi, Nutchanon Yongsatianchot, Joseph D Gaggiano, Miso Kim, Shivani A Patel, Yifan Sun, Stacy Marsella, Jacqueline Griffin, and Andrea G Parker. 2022. Shifting Trust: Examining How Trust and Distrust Emerge, Transform, and Collapse in COVID-19 Information Seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 78, 21 pages. https://doi.org/10.1145/3491102.3501889

A Content Creation for Simulation Videos

To ensure relevance and address potential high-risk health context, we focused on two medication-related queries for creating the simulation videos: (1) "Does taking oral antiviral treatment Paxlovid cause COVID symptoms to return after they improve?" and (2) "What nasal decongestant can I safely use for seasonal allergies if I'm taking Lisinopril for blood pressure and an antacid for acid reflux?" The first query about Paxlovid was derived from previous studies, while the second was selected from a study by [2], which examined the risks associated with patients or consumers using conversational assistants, such as Siri, for medical information.

For the Paxlovid query, we manually drafted the response content based on input from clinicians and two credible websites: https://www.paxlovid.com/ and https://www.yalemedicine.org/news/13-things-to-know-paxlovid-covid-19. For the decongestant query, we created the response by consulting three reputable health-related websites: Mayo Clinic, WebMD, and Cleveland Clinic. The core content for each query was identical across all simulations, regardless of the search method or source type. However, minor adjustments were made to align with specific source types such as referencing source names in the Chatbot condition or incorporating source logos and links in the Chatbot+ condition. Links to all twelve simulation videos can be found in the Supplementary Materials.

B Additional Themes from Interviews

During our thematic analysis, we identified additional themes that deepen our understanding of what participants may want when searching health information online.

Importance of accurate and accessible health information. A majority of participants emphasized that health information should be accurate and recommended by doctors to ensure reliability. P3 expressed this opinion: *"I think if they [doctors or health professionals] would recommend it, obviously something they know it could work for me and maybe they know the information is correct, and all that information can help me."* Accessibility was also a common theme; participants expressed a desire for a search experience that is quick and easily accessible—*"always on your palm"* [P6]—and responses that are easy to understand without complex medical jargon. P6 clearly described this sentiment: *"You read it and immediately you understand what the topic the page is talking about. It doesn't give you a voluminous information that will confuse you down the line."* P4 and P6 also described how access to chatbots or mobile applications for health information without any internet access would be very helpful, especially in areas where there is limited internet access.

Desire real health professional involvement. Although chatbots can provide immediate health information, both P2 and P5 expressed a preference for involvement from actual human health professionals in providing answers and advice online. P5 remarked, *"[Getting] advice off the internet made from a doctor right from the comfort of your home would be good."* P2 added the following: *"So I think maybe asking a chatbot and adding a person would be helpful because in terms of where a chatbot is limited and doesn't know the answer because they do say 'I'm limited to this and I don't know the kind of information that you're asking.' Maybe a real person can get in there and answer the question for you so that you know where you stand."* Furthermore, P5 said that having the option to share and provide reviews on health information from chatting with a doctor would be beneficial: *"I think after you have your chat with your doctor and they give out the information, you can obviously post it and people can know about it and you can give your reviews about the doctor's information shared."*