Taylor & Francis
Taylor & Francis Group

Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: www.tandfonline.com/journals/tbit20

The effect of agent persona on source-clicking, reliance and trust in generative conversational search and the moderating role of health literacy

Shatha Degachi , Evangelos Niforatos & Gerd Kortuem

To cite this article: Shatha Degachi , Evangelos Niforatos & Gerd Kortuem (06 Apr 2026): The effect of agent persona on source-clicking, reliance and trust in generative conversational search and the moderating role of health literacy, Behaviour & Information Technology, DOI: 10.1080/0144929X.2026.2638907

To link to this article: https://doi.org/10.1080/0144929X.2026.2638907

Published online: 06 Apr 2026.

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2026.2638907

The effect of agent persona on source-clicking, reliance and trust in generative conversational search and the moderating role of health literacy

Shatha Degachi, Evangelos Niforatos and Gerd Kortuem

Delft University of Technology, Delft, The Netherlands

ABSTRACT
The utilisation of digital health information is increasingly prevalent, and generative AI-based health information search is likely to become commonplace as well. Yet generative conversation search still has the potential to disseminate inaccurate or incomplete information. Calibrating user reliance on, and trust in, system responses to be more appropriate may mitigate some harms following from this. Indeed, past research shows that clicking on sources in conversational search can improve appropriate reliance, although low source click-through rates remain a challenge. This research explores the design of search agent personas to increase source-clicking rates and foster appropriate reliance and trust. Further, we investigate how health literacy variance moderates the relationship between persona and source-clicking, trust and reliance. Our results show that persona design is a promising direction for influencing source page use frequency, and that health literacy interacts with persona design to affect verification behaviour and perceived risk. This work contributes to the development of more verifiable generative conversational search systems in healthcare contexts.

ARTICLE HISTORY
Received 14 October 2025
Accepted 22 February 2026

KEYWORDS
Digital health; conversational search; appropriate trust; personality; verification behaviour

1. Introduction

The utilisation of digital health information is increasingly prevalent, with 58.15% of residents in the European Union having used the internet to find health information in 2024 (Eurostat 2022). With the introduction of Large Language Models (LLMs) (Brown et al. 2020; Radford et al. 2018) and subsequent rise in the commercial availability of conversational search tools,[1] generative AI-based health information search is likely to become commonplace as well. Like previous innovations in personal digital health, generative conversational search promises increased scalability, availability, anonymity and support to an overextended healthcare system (Crisp and Chen 2014; Crutzen et al. 2011; Luxton et al. 2011; Mohr, Benzer, and Young 2013; Versluis et al. 2022). It also delivers increased personalisation and usability compared to previous conversational agents and digital health tools (Freire, Wang, and Niforatos 2024).

Yet, generative conversational search still has the potential to disseminate inaccurate or incomplete information (Borji 2023; Ji et al. 2023; J. Li et al. 2024; Rawte, Sheth, and Das 2023; Shi et al. 2023). In the healthcare domain, this is especially concerning, as search agents may contribute to the spread of health misinformation (Weidinger et al. 2022). While digital health utilisation improves healthcare service utilisation (Eastin and Guinsler 2006; White and Horvitz 2014) and health outcomes (Jiang and Street 2017), exposure to misinformation negatively impacts the same (Lan, Mahmoud, and Franson 2024). Moreover, such inaccuracies may mislead users into unknowingly taking actions harmful to themselves or others (Metzger et al. 2024; Weidinger et al. 2022). Calibrating user trust in, and reliance on, system responses to reduce under- and over-trust may mitigate some of these concerns. When trust and reliance are appropriate, they are aligned with actual system capability and trustworthiness (Centeio Jorge et al. 2021; Lee and See 2004; O'Neill 2018), reducing the risks of dis- and mis-use (Parasuraman and Riley 1997) that arise from inappropriate trust.

HCI researchers have worked to address the problem of appropriate trust and reliance through many means, most often through explanations (Mehrotra et al. 2024). However, new research shows that even the mere presence of various explanation types can increase blind trust in, and overreliance on, AI tools (Bansal et al. 2021; Pafla, Larson, and Hancock 2024; Si et al. 2024; X. Wang and Yin 2021; Zhang, Vera Liao, and Bellamy

CONTACT Shatha Degachi c.degachi@tudelft.nl Mekelweg 5, 2628 CC Delft, Netherlands

2020). In response to these limitations, promising novel design directions are emerging. Indeed, recent research shows that clicking on sources in LLM responses (Kim et al. 2025) and verification-based explanations (Fok and Weld 2024) can improve appropriate reliance by empowering users to validate outputs. However, low source click-through rates remain a challenge (Kim et al. 2024, 2025; Narayanan Venkit et al. 2025), and many users may be missing key information needed to make reliance and trust judgements. One possible avenue for improving source-clicking (or source click-through rates) or verification behaviour could be the design of the search agent persona. The expression of personality in conversational agents through language features and other elements of interface design is well-researched (Harrison et al. 2019; Mairesse and Walker 2011; Ruane, Farrell, and Ventresque 2021; Smestad and Volden 2018). In healthcare, Nißen et al. (2022) found that affective bond with interpersonally distant agents (Peer vs Doctor) was moderated by age. Moreover, Biro, Linder, and Neyens (2023) found that agent persona qualifications (Nursing Student vs Doctor) impacted perceived usability. We build on these insights by investigating whether agent persona and personality can affect not only user experience but also verification behaviour and subsequently the appropriateness of reliance behaviour and trust beliefs.

However, trust and reliance judgments are influenced not only by system design but also by characteristics of the user and their environment, such as task topic expertise (Szymanski, Millecamp, and Verbert 2021; X. Wang and Yin 2021). As this work is part of the 'Our Smart Family Buddy' project, which aims to develop a digital intervention that facilitates a healthy lifestyle in young families in vulnerable circumstances, we specifically investigate how health literacy variance moderated the relationship between persona and source-clicking. Health literacy was further selected as an object of study as the cognitive demands of processing health information are inversely related to health literacy and message complexity (Meppelink et al. 2016; Squiers et al. 2012; Sweller 1994). Such cognitive load has been shown to impact user trust judgements in Human–AI collaboration (Ahmad et al. 2019; Khawaji et al. 2014; Zhou et al. 2017). Moreover, lower health literacy is associated with higher trust in information from social media and blogs, in addition to lower use of medical websites and less trust in information from specialists (Chen et al. 2018). Therefore, health literacy likely affects users' perception of, and interaction with, search agent personas. Figure 1 illustrates the hypothesised relationships between constructs studied in this paper. We therefore pose the following research question:

- **RQ1**: What effect does agent persona have on the appropriateness of user reliance on, and trust in, generative health conversational search?
  - **(a)**: What effect does health literacy have on the relationship above?

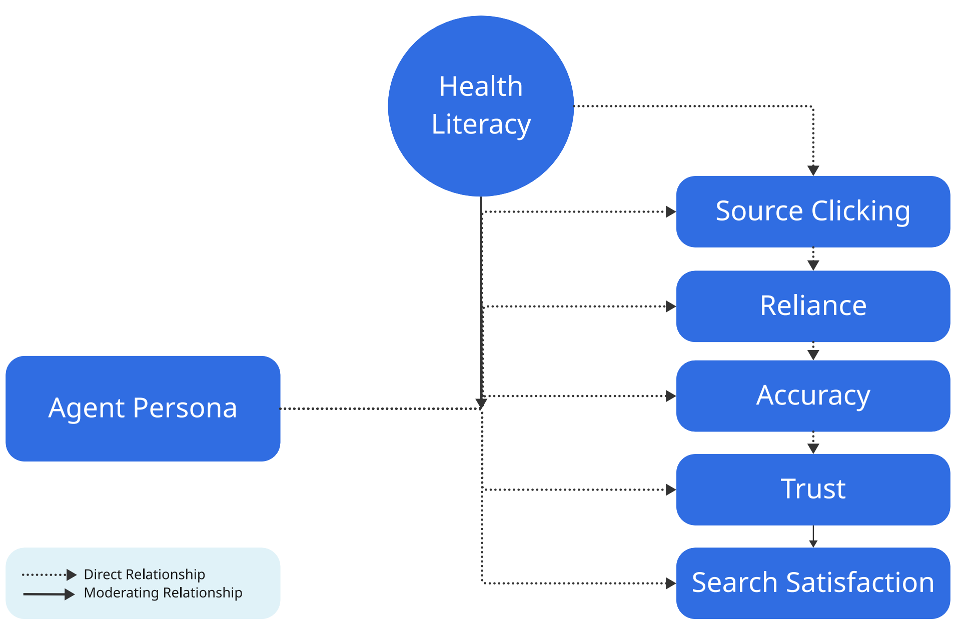


Figure 1. Diagram of hypothesised relationships between constructs under investigation. A flowchart connecting five constructs. Health Literacy is shown at the top, connected by a solid arrow to the dotted arrow between Agent Persona and Source-Clicking, Reliance and Trust.

To answer this question, our research explores the manipulation of search agent personas through domain-specific writing strategies to increase source click-through behaviour. We designed a between-subjects crowdsourced study with 57 participants, where participants interacted with one of two personas varied along the axes of warmth and formality, to complete a health information search task. Personas were designed based on writing strategies and language features described in the works of Ngai, Singh, and Yao (2022) and Nißen et al. (2022) on communication and conversational agents in digital health.

We contribute an understanding of persona design as a promising direction for influencing aspects of source-clicking, such as source page use frequency. Further, we uncover a need for more nuanced communication design in digital health contexts by surfacing interaction effects between health literacy, persona design, and attitude towards AI on source-clicking behaviour and perceived risk. We discuss potential directions for reducing friction in conversational search verification and rethinking the measurement of reliance and accuracy in digital health research. Overall, this work furthers the development of more acceptable and verifiable generative conversational search systems in healthcare contexts.

2. Related work

In this section, we cover the use of conversational search and LLMs in digital health, then unpack the possible harms of over-trust and reliance arising from their adoption. We further describe past research into the perception and implementation of conversational agent personality. Lastly, we define the concept of health literacy and show why it is relevant to understanding verification behaviour, reliance behaviour, and trust beliefs in this context.

2.1. LLMs and conversational search in digital health

The digitalisation of healthcare and health information communication over the past two decades has aimed to deliver increased scalability, availability, anonymity and support to an overextended healthcare system (Crisp and Chen 2014; Crutzen et al. 2011; Luxton et al. 2011; Mohr, Benzer, and Young 2013; Versluis et al. 2022). This digitalisation has occurred in many ways, from the utilisation of web engines for health information seeking to the development of diagnostic image analysis algorithms. Chatbots and conversational agents have been a large part of this trend (Bin Sawad et al. 2022; Kocaballi et al. 2019; Milne-Ives et al. 2020), with agents being used to support behaviour change, health monitoring, professional training, and other activities, including health information search. Bickmore et al. (2016) found that participants using conversational search were more satisfied with their search for clinical trials online than those using conventional search. Furthermore, they also found that conversational search enabled users with low health literacy to locate trials online, which they had previously been unable to do, thereby improving information accessibility. LLMs have now proliferated in healthcare, having been applied in diagnostic support, clinical documentation, medical research and more (L. Wang et al. 2024). Conversational generative search based on LLMs promises to be the next application domain to merge the improved accessibility of previous agents with the increased personalisation and usability of LLMs.

Question-answering systems built on LLMs already demonstrate impressive performance on health knowledge tasks (Q. Li, Li, and Li 2024; Seidel et al. 2024; Singhal et al. 2023; L. Wang et al. 2024). These systems use natural language processing to resolve single-shot queries; however, they lack the multi-turn interaction of conversational generative search, which allows for complex, memory-retaining, contextual dialogues with users (Mo et al. 2025; Radlinski and Craswell 2017). Moreover, much of the research on conversational generative search in healthcare has focused on the technical development and evaluation of such systems (Abbasian et al. 2024; Arias-Duart et al. 2025; Bedi et al. 2024), while the critical study of human factors in this context has not received as much attention. This work, therefore, focuses on the understudied human aspect of an emerging and increasingly popular technology in the high-risk healthcare domain.

2.2. Inappropriate trust and reliance in GenAI

LLMs introduce uncertainty into users' decision-making by introducing plausible errors into responses (Metzger et al. 2024; Weidinger et al. 2022). Under conditions of uncertainty, **trust** emerges as the belief that an agent will help users achieve their goal (Lee and See 2004). Trust in automation, and even in conversational agents, is well studied (Hoff and Bashir 2015; Lockey et al. 2021; Rheu et al. 2021); when trust and reliance are appropriate, they are aligned with actual system trustworthiness and capability (Centeio Jorge et al. 2021; Lee and See 2004; O'Neill 2018), and risk of system dis- and mis-use is reduced (Parasuraman and Riley

1997). Trust and reliance can be made more appropriate through a process of calibration; updating a trust stance by aligning perceptions of an actor's trustworthiness with its actual trustworthiness (de Visser et al. 2020). HCI researchers have tried to calibrate trust and reliance decisions, often through explanations or communication of limitations (Mehrotra et al. 2024); however, some issues have been observed in these approaches. Recent work shows that even the mere presence of many explanation types can increase blind, inappropriate, trust in, and overreliance on, AI tools (Bansal et al. 2021; Pafla, Larson, and Hancock 2024; Si et al. 2024; X. Wang and Yin 2021; Zhang, Vera Liao, and Bellamy 2020). Si et al. (2024) found that even when ChatGPT explanations were incorrect, users found them convincing, and that, being unfamiliar with the task topic, they chose to trust the agent by default. Moreover, Metzger et al. (2024) show that disclaimers on agents' limitations did not alter users' trust.

Nonetheless, some potentially effective design directions have been emerging recently in response to calls for research on trust calibration in the context of conversational agents (Desai et al. 2024). Though Metzger et al. (2024) found disclaimers to have limited effectiveness in impacting user trust, they did show that communication style (assertiveness) could be effectively leveraged to calibrate trust. Kim et al. (2024) and Hosking, Blunsom, and Bartolo (2024) similarly found that assertiveness impacted perceived uncertainty and perceived rate of error. This is because trust can be affected by aspects of the system beyond capability, such as interface design; these aspects may be warranted or unwarranted trustworthiness cues (Liao and Sundar 2022). In research on verification, Kim et al. (2025) found that users made more appropriate reliance decisions and were more accurate when they clicked on links in LLM outputs. Furthermore, Fok and Weld (2024) argue that explanations can only assist in appropriate trust and reliance decisions insofar as they enable easy verification of AI outcomes. Yet, verification behaviour and source-clicking in interaction with LLMs remain low (Kim et al. 2025; Narayanan Venkit et al. 2025). This disparity suggests room for improvement in the design of conversational interfaces, possibly through the use of communication style to enable and motivate verification and thus reduce inappropriate trust and reliance.

2.3. Agents with personality

Designing conversational agents to embody a given personality, persona or role has become a widespread practice among researchers and developers (Pradhan and Lazar 2021). Lessio and Morris (2020) define personality in conversational agents to mean '*a set of traits that is stable across situations and time and acts as a guiding influence on agent behaviour and interactions*' while a persona is '*a fictional character and can have a name, age, education or job, or even a defined backstory and personalities*' (Pradhan and Lazar 2021). Imbuing agents with personas or personality traits has been shown to impact many aspects of the user experience of conversational agents. In retail settings, agent personality can impact usability, engagement (Elsholz, Chamberlain, and Kruschwitz 2019) and even product ratings (Eun Rhee and Choi 2020). Personas and roles are similarly impactful in the healthcare domain. Nißen et al. (2022) found that younger participants had higher affective bonds with more interpersonally distant lifestyle intervention agents, such as experts, while older participants reported higher outcomes for closer peer-like agents. Biro, Linder, and Neyens (2023) found that users who interacted with a Nursing Student persona reported lower usability compared to a Doctor or Nurse persona. Additionally, chatbot language formality has been observed to affect users' perceived competence when disclosing sensitive health information and the quality of data they give Cox and Ooi (2022). In digital health communication research as a whole, there exist some contradictory findings. While some studies (Bernhardt and Felter 2004; Niu, Jeong, and Willoughby 2020) observed that users preferred, and were more influenced by, health information presented from a peer perspective over that of an expert, more recent research (Ngai, Singh, and Yao 2022) has found that social media posts utilising scientific language were more liked. This suggests there might be more nuances to be untangled in the presentation of digital health information.

Personality, persona or roles may be infused into agents through a number of means. In embodied or voice agents, voice and visual presentation can be designed to reflect specific characteristics (Luce Lupetti et al. 2023; Eun Rhee and Choi 2020). In text-only agents, both verbal and non-verbal cues are still available to designers. When designing agents' language, linguistic style such as verbosity, hedging or politeness words can reflect personality traits such as formality (Ruane, Farrell, and Ventresque 2021), extraversion (Mairesse and Walker 2011; Ruane, Farrell, and Ventresque 2021; Völkel et al. 2022), authoritativeness (Hosking, Blunsom, and Bartolo 2024; Kim et al. 2024; Metzger et al. 2024), scepticism (Wester et al. 2024) and neuroticism (Mairesse and Walker 2011). Designers may also utilise non-verbal cues such as agent proactivity (Meng, Lu, and Xu 2025) or emoji inclusion. Fadhil et al. (2018) found that in health coaching systems, mental wellbeing chatbots were rated more highly when emojis were included in their response, while the opposite was true for physical wellbeing agents. Lastly, an agent's role, whether as therapist, peer or friend, can be reflected through titles (e.g. Dr.), technical (or scientific) language (Biro, Linder, and Neyens 2023; Nißen et al. 2022) and the framing of information sources. Luce Lupetti et al. (2023) showed that when agents explicitly referred to healthcare professionals at a local hospital in their recommendations, they were less likely to be described as competent by users. This referencing behaviour may have cast the agent in a subservient role to the referenced practitioners in the users' perception and thus contributed to the understanding of the agent's persona.

Overall, this work builds on the previous research in digital health to investigate whether agent persona can impact not only user experience but also verification behaviour, creating a new design space for trust and reliance calibration study and safer LLM-based conversational agents.

2.4. Cognitive load and health literacy

Health literacy (HL) is a key to empowering individuals in managing their health. Low HL has been associated with poorer health outcomes (Berkman et al. 2011; Boren 2009), and worse utilisation of healthcare services, including the ability to interpret health messages (Berkman et al. 2011; Sorensen et al. 2012). Users with lower HL are more likely to trust health information from social media and blogs over medical websites or doctors (Chen et al. 2018). In 2015, across eight member states of the EU, 47% of the population had inadequate HL (Sørensen et al. 2015). Therefore, HCI researchers must develop a deeper understanding of how HL interacts with user behaviour to develop interventions which can meet the needs of lower HL groups (Zhou et al. 2017) and improve health equity.

Cognitive load, a more commonly studied concept in HCI (Hollender et al. 2010; Kosch et al. 2023), is a probable mechanism by which HL can affect user experience. HL is defined as '*the degree to which individuals have the ability to find, understand, and use information and services to inform health-related decisions and actions for themselves and others*'.[2] Given this emphasis on health information processing ability, HL is well aligned with cognitive load theory. Cognitive load theory posits that working memory has limited capacity, and design choices which do not account for this limitation may overload working memory and compromise learning (Sweller 1994, 2005). With lower health literacy, the need for cognition to process the same health messaging is higher (Meppelink et al. 2016; Squiers et al. 2012). If this need eclipses the available capacity, then information is not processed (Lang 2000). Moreover, HCI research indicates that trust judgements can be impacted by cognitive load in decision-making tasks, specifically suggesting that trust may be higher when cognitive demand is low (Ahmad et al. 2019; Khawaji et al. 2014; Zhou et al. 2017). These findings suggest that health literacy may be a key understudied influence on trust, reliance and verification behaviour in conversational generative search use. Indeed, the findings of Biro, Linder, and Neyens (2023) indicate that higher HL users were more trusting of conversational agent output. Therefore, in this work, we aim to untangle on the potential moderating effect of health literacy on user experience and behaviour in conversational agent use.

3. Hypotheses

Based on the past work unpacked in Section 2, we can see the potential for conversational agent persona and conversational style to shape not only user experience, but also user behaviour, and potentially verification behaviour specifically. Though we can also anticipate the effect of health literacy as understood through cognitive load theory moderating this relationship. We therefore break down our research question and formulate the following hypotheses.

H1 – **Source-clicking** Health messaging using technical (scientific) language was more effective and more liked (Biro, Linder, and Neyens 2023; Ngai, Singh, and Yao 2022), therefore, we expected users would perceive it as more credible and be more persuaded by the Scientific Persona. Moreover, we expected that users would be less motivated to verify its output. Indeed, since LLM output research has shown that assertiveness (Hosking, Blunsom, and Bartolo 2024; Kim et al. 2024; Metzger et al. 2024) can impact perceived uncertainty, error rate and trust, we expected the Scientific Persona to be perceived as confident and highly certain, hence demotivating verification. Thus we hypothesise that *the Scientific Persona will have a lower source click-through rate than the Friendly Persona.*

H2 – **Reliance** With reliance operationalised as agreement rate (Mehrotra et al. 2024), and following from the lack of verification behaviour previously hypothesised, we expected that appropriate reliance would suffer in the Scientific Persona. We therefore posit that *the Scientific Persona will have a higher agreement rate than the Friendly Persona.*

H3 – **Accuracy** Following from the hypothesised lack of appropriate reliance, we expected accuracy to suffer in the Scientific Persona. We hypothesise that *the Scientific Persona will result in lower accuracy than the Friendly Persona.*

H4 – **Trust** Trust, or perceived trustworthiness (Schlicker et al. 2022), is here operationalised as a multidimensional construct (Gulati, Sousa, and Lamas 2019). However, much of the past research did not delineate different dimensions of trust, thus, it is difficult to translate findings into the current hypothesis. Nonetheless, as scientific communication style and technical language use were more liked (Ngai, Singh, and Yao 2022) and more effective (Biro, Linder, and Neyens 2023) in past work, hence we expect the Scientific persona will be perceived as more competent and less risky. We also expect the increased warmth of the Friendly persona will correlate with increased perceived benevolence. We therefore posit that *the Scientific Persona will correlate with higher perceived competence, lower perceived benevolence, and lower perceived risk than the Friendly Persona.*

H5 – **Health Literacy** Given the current state of the literature, multiple hypothesis directions are possible. Despite a preference for peer and social media information sources in the Low HL group (Chen et al. 2018), and the negative correlation between trust and cognitive load in HCI (Ahmad et al. 2019; Khawaji et al. 2014; Zhou et al. 2017), we expected the effect of assertiveness (Hosking, Blunsom, and Bartolo 2024; Kim et al. 2024; Metzger et al. 2024) and technical language (Ngai, Singh, and Yao 2022) would be dominant in the search interaction, driving higher persuasiveness and agreement, thus demotivating verification, more so in this group than the High HL group. We favoured this direction, as the effects described in (Hosking, Blunsom, and Bartolo 2024; Kim et al. 2024; Metzger et al. 2024) were more specific to conversational agents. Therefore, we hypothesise:

- **H5.1**: Lower HL will correspond to a stronger drop in Source-Clicking.
- **H5.2**: Lower HL will correspond to a stronger rise in Reliance.
- **H5.3**: Lower HL will correspond to a stronger drop in Accuracy.
- **H5.4** Lower HL will correspond to a stronger rise in perceived competence as well as a stronger drop in perceived benevolence and perceived risk.

4. Materials and methods

To address our **RQ1**, we design a between-subject crowdsourced study with 57 participants. Based on writing strategies and language features described in health communication literature (Ngai, Singh, and Yao 2022; Nißen et al. 2022), we created two simulated conversations labelled (1) Scientific and (2) Friendly. Participants were randomly assigned to interact with one of the two conversations. Participants relied on the simulated conversation to complete a search task on the topic of *Coffee Intake and Heart Health*. The study lasted on average 17 min and 53 s.

To assess whether our persona design was effective and perceived as intended, a between-subject crowdsourced study was also conducted with 70 participants. Here, the participants were also randomly assigned to interact with one of the two simulated conversations designed and asked to rate them on scientificness and friendliness. The median task duration was 8 min, 25 s.

4.1. Designing the stimuli

Two simulated conversations were created to ensure participants in each group were exposed to identical stimuli and to enhance the internal validity of this work. The simulated responses were based on the work of Ngai, Singh, and Yao (2022) and Nißen et al. (2022) on online health information communication and chatbot design. Therefore, the Scientific Persona responses exhibited the following characteristics: '*describing actions; quoting sentences from public figures; attributing information to credible-sounding sources; including medical experts, doctors/nurses, scientific studies, and legal documents; using jargon, terminology, and/or statistics*' (Ngai, Singh, and Yao 2022). On the other hand, the Friendly Persona responses exhibited: '*first or second-person address form (e.g. we should listen, you must act…); author visibility such as sharing personal experiences and feelings; and use of informal expressions (e.g. using sentence fragments, questions, contractions, emojis, swear words)*' (Ngai, Singh, and Yao 2022). We also incorporated interpersonal closeness cues from social role theory, so that the Friendly (peer-like) Persona responses included emojis and humour (Nißen et al. 2022). For example, in the Friendly condition, the agent replies: '*Although, how you prepare your coffee also matters – filtered coffee is generally less likely to mess up your cholesterol levels than boiled or unfiltered coffee, it's tastier too!*'. In contrast, in the Scientific condition, '*The impact varies by individual and preparation method, with filtered coffee*

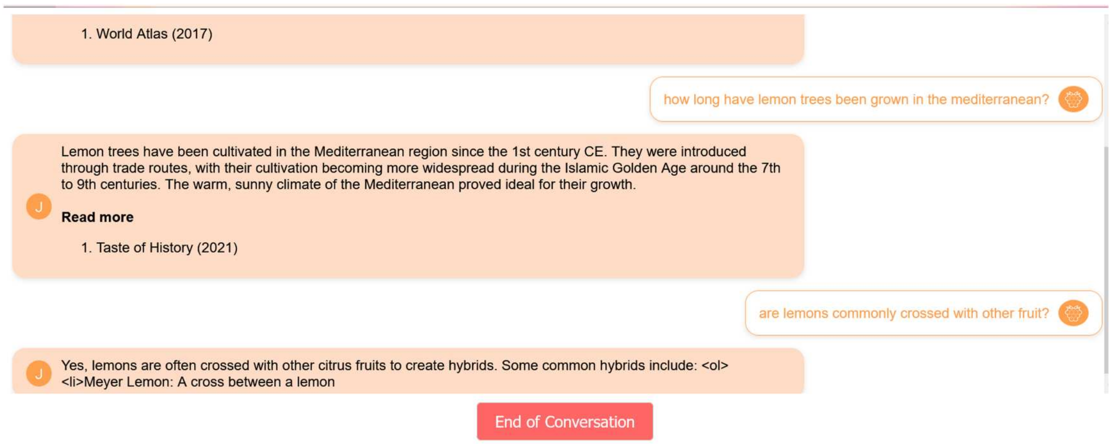


Figure 2. Screenshot of tutorial interface showing two turns between user and search agent. A chat messaging interface screenshot. The agent message bubbles have an orange background with an icon showing the letter J. The conversation is about lemon agriculture. A red button at the bottom says End of Conversation.

being less atherogenic than boiled coffee'. See Appendix 1 for the full text of the simulated conversations.

The conversations consisted of three turns, the average turn count in generative conversation interactions (Zhao et al. 2024). The conversations were initially generated using a retrieval augmented generation architecture (RAG) (Shuster et al. 2021) built on *GPT4.o*[3] by OpenAI then edited to align more closely with the writing strategies described in the literature, as well as to introduce factual errors into the responses. The conversations aimed to simulate a realistic accuracy level for an RAG agent in the health domain; thus we set the accuracy at 70% (Seidel et al. 2024; Wu et al. 2024) by introducing a factual error into one of the three turns. User input was the same across all conversations, while the agent's responses varied in style but communicated the same facts.

An interactive interface was built using HTML and JavaScript to host the conversations and allow participants to engage with them. This interface can be seen in Figure 3. To increase the sense of ownership over the conversation and identification with the simulated user, participants could 'send' input messages to the 'AI assistant' via button-click. Responses from the simulated agent appeared after a 'loading' pause, with a typing animation to emulate live conversational agent interaction. Each turn in the interface contained two source links: (1) a link to a page emulating a scientific journal where (relevant) snippets from a journal article were displayed, (2) a link to a page emulating a popular media page where

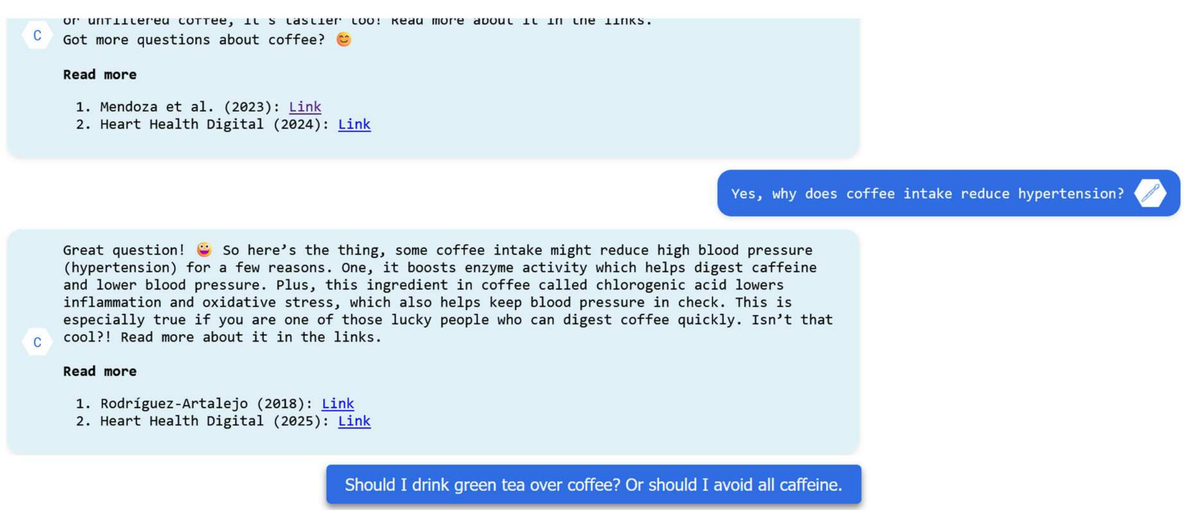


Figure 3. Screenshot of the simulated friendly persona conversation interface, showing one turn between user and search agent. Another chat messaging interface screenshot. The agent message bubbles have a blue background with an icon showing the letter C. The conversation is about coffee intake. A blue button at the bottom says: Should I drink green tea over coffee? Or should I avoid all caffeine.

(relevant) snippets from an online blog or article were displayed (see Appendix 2). Both personas explicitly referred to the links in the response (e.g. 'Read more about it in the links') (Luce Lupetti et al. 2023). We chose these sources to reduce noisy data caused by linking to live webpages and because popular generative search tools, when using links, also often provide a diverse group of sources. During the manipulation check, links and factual errors were removed, as they were irrelevant to the participants' task.

A 'tutorial' interface was also created to familiarise the participant with the interaction and reduce the friction and novelty effect during the study task. This interface functioned the same as the task interface but is designed to look like a different agent, with a different name, different colours, fonts, user avatars and (non-healthcare-related) discussion topics (Figure 2).

4.2. Manipulation checks

To evaluate our stimuli, we designed a between-subjects study. 70 participants were recruited via the crowdsourcing platform Prolific[4] and compensated at the recommended rate of 9 credits per hour. Participants were screened on English language fluency, country of residence (such that they were residents in the EU or UK), whether they had used an AI chatbot before, and task approval rating (≥ 95%).

Participants were randomly assigned to one of two groups: the Scientific Persona Group or the Friendly Persona Group. In each group, participants reviewed a pre-generated conversation between a simulated user and their assigned persona search agent. As an attention check, participants were asked to summarise the conversation in a copy-paste-disabled text box (150–200 characters). Participants then scored the agent on two 11-point scales –Informal-Scientific and Cold-Friendly. Participants were also asked to select which role best suited the presented agent: Health Expert, Friend, Journalist or Health Online Influencer.

4.3. Pilot

We conducted a small-scale offline pilot with five participants. Based on this pilot, we acquired a more grounded estimate of the study duration, assessed task clarity, and resolved any technical issues with the survey platform.

4.4. Participants

63 participants were recruited via the crowdsourcing platform, Prolific. Six participants were removed from the final dataset due to failed attention checks and/or our inability to verify that their actions were logged correctly, resulting in 57 total participants included over the course of July 2025. Participants were pre-screened and excluded from the full study if they were too familiar with the search task topic (Coffee Intake and Heart Health), had never used an AI chatbot before or had participated in the intervention validation study. Participants were also screened on English language fluency, country of residence (such that they were residents in the EU or UK, in accordance with ethics board recommendations), and task approval rating (≥ 95%).

4.5. Measures

We combine click-through behaviour, agreement and self-reported user experience metrics to capture user reliance behaviour and trust beliefs in generative conversational search. We also capture user decision-reasoning via open-text questions.

4.5.1. Control variables

Attitude towards AI Measured using the 5-item ATAI scale developed by Sindermann et al. (2021). The ATAI uses an 11-point Likert scale to capture negative and positive dimensions of attitude towards AI, which are here scored and analysed separately. Positive items include '*Artificial intelligence will benefit humankind*' while negative items include '*I fear artificial intelligence*'. We capture participants' attitudes to understand the extent to which their pre-existing beliefs about AI systems explain the variance observed in our data, as compared to our independent variables, and thus the extent to which our interventions are impactful.

Health Literacy (HL) Measured by the (short) European Health Literacy Survey Questionnaire, a 16-item survey (Pelikan and Ganahl 2017; Pelikan, Straßmayr, and Ganahl 2020). The EU-HLSQ16 uses a four-point scale per item from *Very Easy (4)* to *Very Difficult (1)*, with higher overall score indicating higher health literacy. The scale includes items such as '*How easy would you say it is to use information the doctor gives you to make decisions about your illness?*'. We collected health literacy to investigate its role as a possible moderator in the relationship between agent persona and reliance behaviour and trust beliefs. As lower health literacy means a higher cognitive load for users making trust judgements in health information seeking contexts (Meppelink et al. 2016; Squiers et al. 2012; Sweller 1994), and cognitive load moderates the appropriateness of trust (Ahmad et al. 2019; Khawaji et al. 2014; Zhou et al. 2017), understanding the role of health literacy is then key to interpreting user–agent interactions.

4.5.2. Dependent variables

Source-Clicking Recorded using logs and event trackers embedded into the simulated conversation interface. We capture source-clicks as the mechanism through which reliance behaviour and trust beliefs may be intervened on, due to the connection between appropriate reliance and source-clicking established in past research, where increased cross-referencing with source reduced overreliance on agents (Kim et al. 2025). We capture this variable using two measures labelled: (1) 'LinksClicked', which captures unique links the user clicked on, and (2) 'SourcesChecked', which captures times the user loaded the source page. Therefore, a user could click on any given link only once, but could load the same source page multiple times by navigating back and forth between the search window and source page.

Reliance Operationalised and measured as the rate of agreement with the simulated conversation (Mehrotra et al. 2024). Participants may over-rely on the system when they agree with an incorrect response, and under-rely when they disagree with a correct one. Three true-or-false questions, one per conversational turn, are designed based on the task topic to capture reliance. Questions were written to be of a similar difficulty level. Questions can be seen in Appendix 3.

Accuracy Operationalised and measured as the rate of correct responses to the aforementioned three true-or-false questions (Appendix 3) posed post information exposure (Mehrotra et al. 2024).

Trust or Perceived Trustworthiness (Schlicker et al. 2022). Perceived Trustworthiness is measured using the 12-item 5-point Likert scale developed by Gulati, Sousa, and Lamas (2019), which captures trust along four dimensions: perceived benevolence, competence, risk and trust, using three items per dimension. For example, in benevolence, '*I believe that [the agent] will act in my best interest*'. Items for all dimensions except risk are positively framed, with higher scores indicating higher overall trust. We expand on past work into uncertainty communication and reliance behaviour (Kim et al. 2024, 2025) by also investigating trust beliefs in conversational search. Thus we can uncover implications for future trust calibration research.

Search Satisfaction Measured by a single item 11-point scale: '*Rate your satisfaction with this AI-assisted search session*'. Satisfaction is a popular metric for assessing the usability of information retrieval tools (Liu et al. 2018). Though not directly related to our RQ, we use it here to understand potential future design implications of our interventions.

4.6. Tasks

Participants completed one search task on the topic of *Coffee Intake and Heart Health*, relying on a simulated conversation with linked website sources for information. After reviewing the conversation and possibly the sources, the participants answered three true-or-false questions, one for each conversational turn. The search task was untimed to allow the users to read the conversation at their pace and explore the linked sources in the responses.

We chose the topic of Coffee and Heart Health, as the topic of nutrition is of interest to many people (Vrabič Dežman 2024), and coffee intake is a common habit (Mendoza et al. 2023), yet nutrition topics can be misinformation-laden (Colditz, Woods, and Primack 2018; Denniss, Lindberg, and McNaughton 2023). Therefore, we believed it would be an engaging subject for participants, but sufficiently arcane such that they would not know the answers to the questions on the topic before the study.

4.7. Procedure

4.7.1. Research design

We designed a between-subjects repeated measures study where participants were randomly assigned to one of two groups: Scientific Persona Group or Friendly Persona Group, as illustrated in Figure 4. Participants were recruited via the crowdsourcing platform Prolific[5] and compensated at the recommended rate of 9 credits per hour. After consenting to participation in the study, participants filled in a series of survey questions on their (1) Attitude Towards AI, (2) their Experience with, and Expectations of, Chatbots, and (3) their Knowledge of the Task Topic, as well as general Health Literacy. Participants interacted with a tutorial example interface on a non-health-related topic to familiarise themselves with the system before the search task. In each group, participants reviewed a pre-generated conversation between a simulated user and their assigned persona search agent.

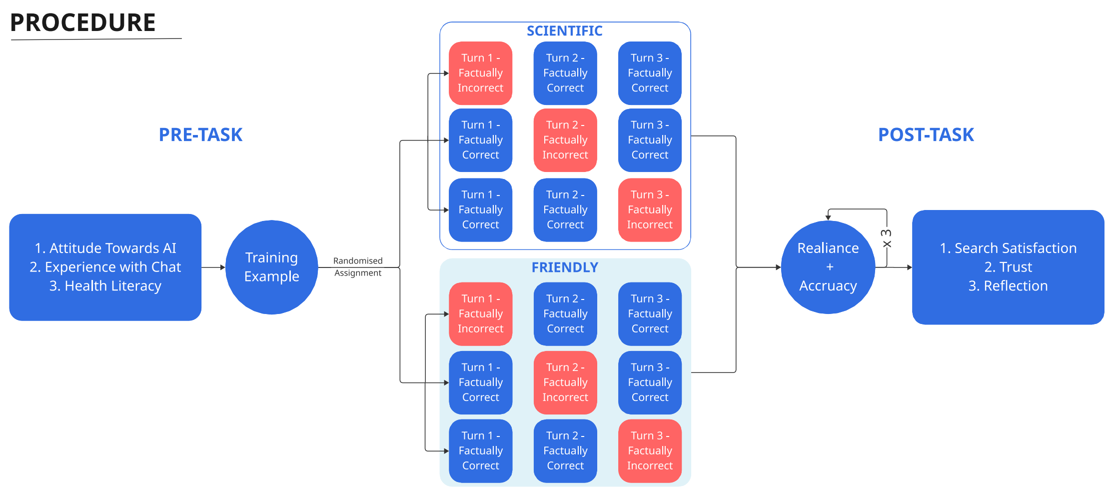


Figure 4. Flowchart capturing study procedure workflow. A flowchart captures the study procedure workflow. The pre-task phase collects data on participants' attitudes towards AI, experience with chat systems, and health literacy. The procedure phase randomly assigns users into one of two conditions: (1) scientific or (2) friendly persona, each with three turns where LLM output is either correct or incorrect. The post-task phase measures search satisfaction, trust and reflection.

They did so by clicking a button to input the pre-determined user message into a chat and reading the agent response, at each turn, for three turns, where one turn contained a factually inaccurate statement. In contrast, the two source links contained accurate information. As an attention check, participants were asked to summarise the conversation reviewed in 150–200 characters. After summarising the conversation, users answered three true-or-false questions and explained their answers. To introduce a sense of risk necessary for trust formation (Mehrotra et al. 2024), participants were informed that they would receive a bonus payment of 0.03 credits per True or False question answered correctly. Next, they were asked to rate and explain their search satisfaction and trust in the agent. Lastly, they were asked to reflect on their source-clicking behaviour in past conversational search. The order of factual inaccuracies' appearance in the conversation was counterbalanced such that participants may experience it in the first, second or third turn.

This procedure was approved by the Human Research Ethics Committee at Delft University of Technology and pre-registered on Open Science Framework at: 10.17605/OSF.IO/D36EG.

4.7.2. Data analysis

For the collected quantitative data, we first performed Shapiro–Wilk tests to estimate the distribution of our data. For normally distributed variables (i.e. Trust), we used t-tests and linear regression to investigate related hypotheses. For non-normal count variables (i.e. Source-Clicking), we used log-linear (or Poisson) regression analysis. Poisson regression is well suited for count variables such as the number of visits (Coxe, West, and Aiken 2009; Hutchinson and Holtman 2005; Imrey 2000). Meanwhile, for proportion variables (i.e. Reliance, Accuracy), we used Mann–Whitney tests and binomial logistic regression as these methods were best suited for proportions (Bewick, Cheek, and Ball 2005; Harrell 2015; Laerd Statistics 2018). All statistical analyses were performed using JASP (JASP Team 2025).

Qualitative data, such as user reasoning, collected through open-text questions, were analysed through a combination of content (Downe-Wamboldt 1992; Harwood and Garry 2003) and thematic analysis (Braun and Clarke 2021; Clarke and Braun 2013), supplemented with natural language processing (NLP) models. Guetterman et al. (2018) demonstrated that NLP methods can facilitate the analysis of qualitative data, resulting in faster theme discovery and more comprehensive results. We used SentenceBERT models[6] and semi-supervised (constrained) K-means clustering[7] to group user responses into the maximum number of clusters possible while maintaining a cluster size of more than three items. Given our use of NLP tools, our thematic analysis approach then combined inductive and deductive approaches to the data. The first author performed selective coding, code group generation and theme generation. The main researchers were involved in theme review and refinement.

Statement of Positionality: The lead researcher on this work is situated within the field of human–computer trust and trust calibration, aiming to forward a more nuanced understanding of trust. Further, the authors of this work are embedded in the domain of LLM adoption and acceptance. These perspectives naturally colour our orientation towards our participants and our data.

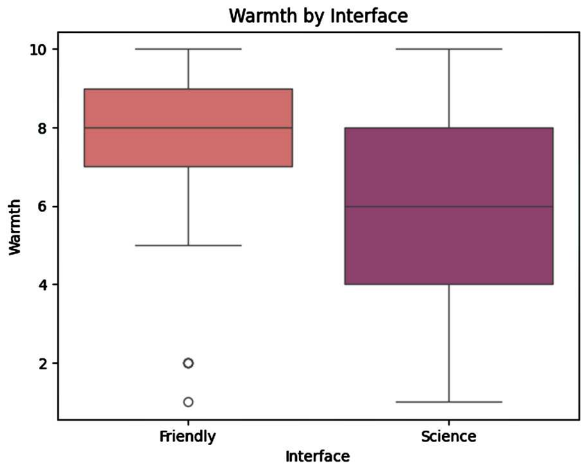


Figure 5. Box plot comparing the two personas on the cold-friendly scale. A boxplot comparing perceived friendliness by persona. The Friendly persona has a higher mean perceived friendliness and a tighter distribution of values.

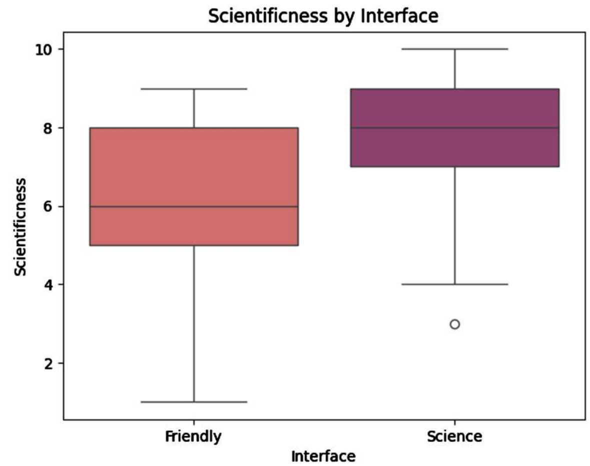


Figure 6. Box plot comparing the two personas on the informal-scientific scale. A boxplot comparing perceived scientificness by persona. The Scientific persona has a higher mean perceived scientificness and a tighter distribution of values.

5. Results

5.1. Manipulation checks

35 participants were randomly assigned to each condition. Initially, Mann–Whitney U analysis showed that while perceived Friendliness was significantly greater in the Friendly persona ($U = 971$, $p<0.001$), perceived Scientificness was only slightly greater in the Scientific persona ($U = 492.5$, $p = 0.211$). Therefore, the simulated conversations were reworked to introduce more statistical figures into the Scientific responses, while reducing medical jargon and introducing humour into the Friendly responses (Nißen et al. 2022). The personas were then retested using the same procedure with an additional 70 participants (who had not participated in the previous test). Mann–Whitney U analysis showed that perceived friendliness was significantly higher in the Friendly persona ($U = 380$, $p = 0.006$), and perceived scientificness was significantly higher in the Scientific persona ($U = 322.5$, $p<0.001$). Thus the stimuli were determined to be valid and reliable. Figures 6 and 5 illustrate these distributions.

Further exploratory analysis using Pearson's Chi-Square shows a significant association between the designed stimuli and selected agent role ($\chi = 15.847$, $p = 0.001$, $V = 0.429$). For example, participants labelled the Scientific persona as a Health Expert in 28 out of 42 instances, compared to only 11 out of 44 for the Friendly persona. Similarly, participants labelled the Friendly persona as a Health Online Influencer in 21 instances, compared to 7 for the Scientific persona.

5.2. Hypothesis testing

We first describe the sample characteristics in our study. Next, we study the effect of persona on source-clicking behaviour, reliance, accuracy, trust and satisfaction while examining the moderating effect of health literacy (HL) on these relationships. Lastly, we report on user reasoning on True or False questions and their reflections on trust and verification behaviour.

After testing for normality, the collected control and dependent variable groups were also tested for multicollinearity. No significant correlations between independent variables in the dependent group (e.g. the different dimensions of trust were correlated, but trust and reliance were not). However, Pearson's r testing did show a significant, but weak, positive correlation between the acceptance of AI and HL variables in the control group ($\rho = 0.264$, $p = 0.047$), see Figure 7, and thus regression analysis could proceed as planned (Shrestha 2020).

5.2.1. Sample characteristics

Our sample characteristic distribution can be seen in Table 1. Based on the HL groupings described by Mekhail et al. (2022), our sample included 22 participants in the High HL group, 18 in the Middle HL group and 17 in the Low HL group.

5.2.2. Effect on source-clicking behaviour

Of 57 participants, 38 (66.7%) did not click on any link. The median frequency of source page checking was also 0. With four participants using source pages more than

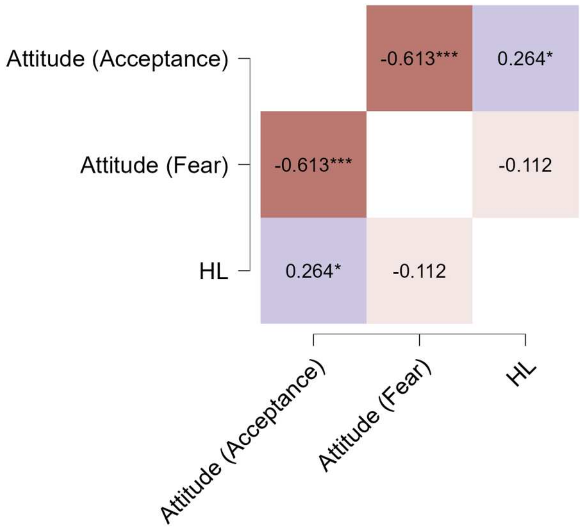


Figure 7. Heatmap showing correlations within the control variables group. Heatmap showing correlations within study control variables. Acceptance of AI and Health Literacy have a weak positive correlation (0.264). Colour (blue-red) is used to indicate the direction of the correlation.

20 times, the mean frequency of source checking was 3.561 with an $SD = 7.646$. To test **H1** and **H5.1**, we perform general linear regression assuming a Poisson distribution, on 'LinksClicked' and 'SourcesChecked'. The overall regression models were significantly predictive of (1) LinksClicked with $df = 39, X^2 = 75.474, p < .001$ and (2) SourcesChecked with $df = 39, X^2 = 279.959, p < .001$. Persona alone was not a significant predictor of LinksClicked, but did predict SourcesChecked ($\beta = -71.188, z = -2.711, p = 0.007$). **This indicates that users checked the source pages more often in the 'Friendly' condition, and thus H1 partially holds**. Moreover, the results showed significant two-way and three-way interactions between persona design, HL and attitude towards AI. Tables 2 and 3 list the significant coefficients in these regressions. **Figures 8 and 9 show that while the participants in the Low HL group did click on links less often than the High HL group in the Friendly condition, the reverse was true in the Scientific condition. Thus H5.1 partially holds**.

Table 1. Participant characteristics.

Characteristic	n	%
Gender		
Other	1	1.8%
Female	18	31.59%
Male	38	66.67%
Age		
18–24	4	7.02%
25–34	22	38.60%
35–44	10	17.54%
45–54	14	24.56%
55–64	6	10.53%
65+	1	1.8%
Education		
High School Graduate	18	31.59%
Some Higher Education	4	7.02%
2-year Degree	3	5.26%
4-year Degree	23	40.35%
Post-Graduate Degree	9	15.79%
Conversational Agent Use		
Daily	15	26.32%
3–5 Times a Week	20	35.09%
Weekly	10	17.54%
Less than Once a Week	12	21.05%

Table 2. General linear regression model coefficients for LinksClicked.

	Estimate	Standard Error	z	p
(Intercept)	−1.246	2.309	−0.540	0.590
HL (Mid)	8.301	3.139	2.645	0.008
Attitude (Acceptance)	5.314	2.658	2.000	0.046
Attitude (Fear)	−7.629	2.825	−2.701	0.007
Persona (Scientific) * HL (Low)	56.956	28.683	1.986	0.047
Persona (Scientific) * Attitude (Fear)	32.677	16.261	2.010	0.044
HL (Mid) * Attitude (Acceptance)	−12.829	3.706	−3.462	< .001
Persona (Scientific) * HL (Low) * Attitude (Fear)	−39.911	18.490	−2.158	0.031
Persona (Scientific) * HL (Mid) * Attitude (Fear)	−40.612	17.377	−2.337	0.019

5.2.3. Effect on reliance behaviour and trust beliefs

Mann–Whitney independent means comparison tests on **H2** and **H3**, along with Binomial Logistic regression for **H5.2** and **H5.3**. Testing did not indicate that Reliance was significantly higher, nor Accuracy significantly lower, in the Scientific Persona. Moreover, this

Table 3. General linear regression model coefficients for SourcesChecked.

	Estimate	Standard Error	z	p
(Intercept)	−0.465	1.645	−0.282	0.778
Persona (Scientific)	−71.188	26.257	−2.711	0.007
HL (Mid)	4.410	1.976	2.232	0.026
Attitude (Acceptance)	7.272	1.917	3.793	<.001
Attitude (Fear)	−11.917	2.018	−5.904	<.001
Persona (Scientific) * HL (Low)	82.029	26.772	3.064	0.002
Persona (Scientific) * HL (Mid)	85.577	26.424	3.239	0.001
Persona (Scientific) * Attitude (Fear)	52.365	15.618	3.353	<.001
Persona (Scientific) * Attitude (Acceptance)	65.098	25.694	2.534	0.011
HL (Mid) * Attitude (Acceptance)	−9.190	2.262	−4.064	<.001
HL (Low) * Attitude (Fear)	18.335	4.283	4.281	<.001
HL (Mid) * Attitude (Fear)	9.032	2.278	3.966	<.001
Persona (Scientific) * HL (Low) * Attitude (Acceptance)	−78.189	26.332	−2.969	0.003
Persona (Scientific) * HL (Mid) * Attitude (Acceptance)	−77.379	25.829	−2.996	0.003
Persona (Scientific) * HL (Low) * Attitude (Fear)	−59.938	16.121	−3.718	<.001
Persona (Scientific) * HL (Mid) * Attitude (Fear)	−69.291	16.155	−4.289	<.001

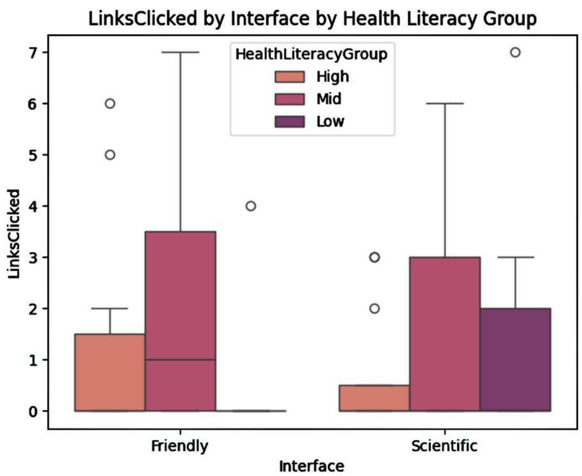


Figure 8. Box plot comparing LinkedClicked by persona and health literacy group. Box plot comparing the number of links clicked by persona and health literacy group. For both the Friendly and Scientific interfaces, the Mid HL group clicks links most often. The Low HL group clicks on links more frequently in the Scientific persona condition.

testing did not reveal any significant two-way or three-way interactions of our control variables on reliance or accuracy. **Therefore, H2, H3, H5.2 and H5.3 were not supported.**

Independent means t-test comparison also revealed no significant difference in any trust dimension by Persona alone. **Hence, H4 was also not supported.** In an **exploratory** one sample Wilcoxon signed-rank-test (since trust dimensions were non-normal when not split by Persona), perceived competence ($M_F = 11.45$,

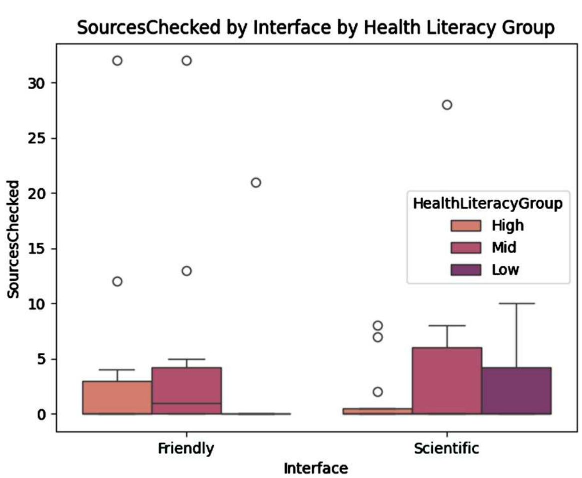


Figure 9. Box plot comparing SourcesChecked by persona and health literacy. Box plot comparing the number of sources checked by persona and health literacy group. For both the Friendly and Scientific interfaces, the Mid HL group checks the source pages most frequently. The Low HL group checks the source pages more frequently in the Scientific persona condition.

Table 4. Linear regression model coefficients for perceived risk.

	Unstandardised	Standard Error	t	p
(Intercept)	1.665	3.285	0.507	0.615
Attitude (Acceptance)	8.767	3.251	2.697	0.010
Persona (Scientific) * HL (Low)	−25.760	11.174	−2.305	0.027
Persona (Scientific) * Attitude (Acceptance)	−21.926	9.933	−2.207	0.033
Persona (Scientific) * HL (Low) * Attitude (Acceptance)	24.428	11.835	2.064	0.046

$M_S = 11.39$), benevolence ($M_F = 10.10$, $M_S = 10.64$) and trust ($M_F = 9.59$, $M_S = 9.93$) were significantly greater than the neutral point of 9 ($V_{comp} = 1424.50$, $p_{comp} < 0.001$, $V_{ben} = 1154.50, p_{ben} < 0.001$, $V_{trust} = 820.50, p_{trust} = 0.008$), while perceived risk ($M_F = 8.45$, $M_S = 7.89$) was significantly lower ($V_{risk} = 85.00, p_{risk} < 0.001$), indicating uniformly high trust in both personas.

Nonetheless, when testing **H5.4**, the linear regression model was significantly predictive of perceived risk ($R^2 = 0.514$, $RMSE = 2.017$, $F(17, 39) = 2.45$, $p = 0.011$) and indicated several significant two-way and three-way interaction effects on Persona. Further, a significant main effect of Acceptance of AI can be observed ($t = 2.697$, $p = 0.010$). See Table 4 for all significant interfactions. Figure 10 shows that the Medium HL group perceived lower risk in the Scientific Persona ($M_F = 9.17, M_S = 7.67$), while the Low HL group showed the opposite ($M_F = 8.14, M_S = 9.10$). The High HL group showed the lowest perceived risk regardless of persona ($M_F = 7.80$, $M_S = 7.00$). Trust, perceived competence and perceived benevolence, the remaining

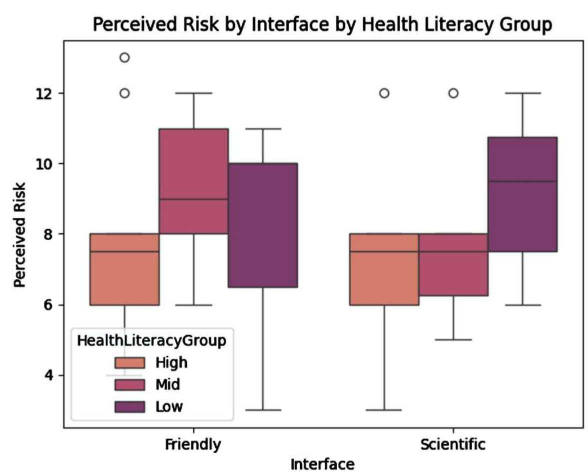


Figure 10. Box plot comparing perceived risk by persona by health literacy group. Box plot comparing perceived risk by persona and HL group. For the Friendly interface, the Mid HL group has the highest mean perceived risk. For the Scientific interface, the Low HL group has the highest mean perceived risk.

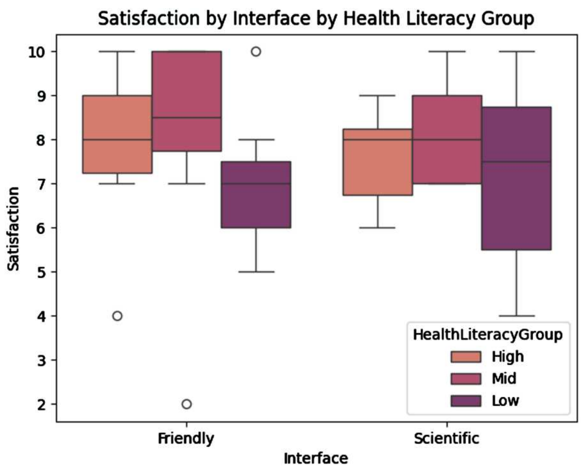


Figure 11. Box plot comparing satisfaction by persona and health literacy. A boxplot comparing satisfaction by persona and HL group. The High- and Mid HL groups have slightly higher mean satisfaction with the agent than the Low HL group.

dimensions of trust, did not show a significant regression. **Thus H5.4 is not supported, as the direction of the effect on perceived risk is reversed.**

5.2.4. Effect on satisfaction

Mann–Whitney and ANCOVA analyses showed no significant effect of Persona on Satisfaction, nor any significant interaction effects from health literacy or attitude towards AI. The mean Satisfaction was 7.79 in the Friendly group and 7.57 in the Scientific group. Figure 11 illustrates this distribution by persona and health literacy group.

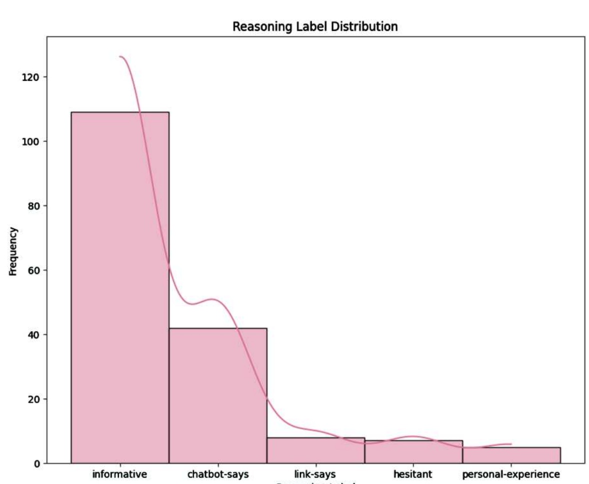


Figure 12. Bar plot illustrating frequency of reasoning given for reliance. Bar plot illustrating the frequency of reasoning given for reliance. The informative category has the highest frequency with more than 100 instances, followed by chatbot-says (40+), link-says (<20), hesitant and personal experience.

5.3. User reasoning and reflection

5.3.1. Reasoning

The reasoning behind user responses provided by participants was tagged with one of five, mutually exclusive, labels using content analysis. Figure 12 shows that the most frequent method of justification was quoting facts about the topic as supporting evidence with no attribution (here labelled as information), followed by referring to the chatbot as an authority on the topic.

Interestingly, some reasoning revealed that even when the chatbot provided participants with the correct answer, participants sometimes made deduction errors that led to them ultimately selecting the incorrect response, e.g. '*The AI answer specified headache and [fatigue] as symptoms, but these could cause dizziness, so I marked [it] as true*'. The reverse is also true, such that users inadvertently provided the correct answer, even when relying only on the incorrect chatbot '*anything over 600 mg is considered to increase the risk of heart disease and risk of high blood pressure, so 500 mg is near that, so I said true*'.

5.3.2. Reflections on trust and verification behaviour

To enable semi-supervised learning, a subset (36.84%) of the reflections on trust provided was tagged with one of six labels that represent a key theme: *confirmation-bias, uses-sources, too-friendly, distrust, reliable, inaccessible*. The remainder of the dataset was left unlabelled. This dataset was then used with Sentence-BERT embedding and semi-supervised clustering to create five clusters, which were labelled with a theme, as seen below.

- Reflection on Coda's Informativeness (Cluster 0): In this theme, participants attributed their satisfaction with the agent to its comprehensiveness, clarity and relevance. Nonetheless, some participants found the agent to be lacking in 'intelligence' beyond its informativeness and were less satisfied.
- Largely Satisfied but Cautious in Digital Health (Cluster 1): Here, participants expressed a high level of trust in the agent, but also some reservations. These reservations were related to the high-risk nature of the digital health context or the users' relative inexperience in the topic.
- Reflection on Coda's Friendliness (Cluster 2): Participants were in two minds concerning the agent's friendliness; some found the informality off-putting and inappropriate, while others found it human-like and 'nice'.

- Reflections on Coda as an AI agent (Cluster 3): In this theme, participants relate their trust and distrust of our agent to their thoughts on AI tools in general. This theme highlights how expectations of, and experience with, AI tools colour interactions with new tools, as one user states: '*I was satisfied with Coda because it did what I expect an AI assistant to be capable of. I did not trust it 100%, because you should never do that. AI (wrongly) think they're always 100% correct*'.
- Reflections on Verification Behaviour (Cluster 4): Participants highlighted multiple modes of interaction with the linked sources. In some cases, they trust the agent because information in the links or independent sources confirms its response. However, other participants found the inclusion of sources to improve their trust even without utilising them, as one participant mentioned, '*Yes, I was satisfied since it also pasted source links, however, I did not double check it. I would have to double check it to feel completely satisfied*'.

Further thematic analysis resulted in the production of 37 codes, including the initial six labels. The most commonly used codes were: satisfaction ($N = 23$), source-use ($N = 22$) and trust ($N = 21$). From those codes, two additional themes were derived, namely:

- The Effect of Confirmation Bias: These reflections show how aligning with user pre-conceived notions of health issues engenders higher trust, even without verification; '*there is no reason for me not to trust Coda. The information I already knew matched with the information provided by it*'.
- Reflections on Inaccessibility: In this theme, some participants discuss having difficulty understanding the agent due to the increased complexity in the Scientific persona, and therefore having difficulty making accurate trust judgments.

The reflections on verification behaviour were similarly treated, with 38.6% of responses tagged with one of four labels: *always, in-the-past-not-now, never, now-not-always.* The resulting five clusters were labelled with a theme as seen below.

- Reflections on the Purpose of Sources (Cluster 0): In this theme, users stated that they were most likely to interact with a link if they were interested in diving deeper into a specific topic.
- Difficulties Interacting with Links in Study Conditions (Cluster 1): Here, participants highlighted some issues that arose from the crowdsourced and online nature of the study that interfered with link interaction, such as moving between tabs or technical issues.
- Always/Never, Rules for Engaging with Links (Cluster 2): This cluster groups participants who established some extreme guidelines for their interactions with links in chatbots. Participants described themselves as either never interacting with links or always utilising them. One participant says: '*Yes, I always try and do a cursory search about the information presented to me, it pays to not blindly trust AI as it can hallucinate answers*'.
- In-Task vs Past Behaviour (Cluster 3): In this theme, users reflected on their use of links in the task and in the past. In some cases, users felt they often used links in the past, but didn't need to do so for this task. In others, they felt they checked links more often in the task than they normally would, one participant says: '*I rarely go to the source links when using AI chat functions, I prefer instead to rely on the summary answers. In this instance, though, I did view some of the links to read more details*'.
- Cluster 4: Items in cluster four did not result in a new theme, as they could be described using the previous four themes. For example, one user states '*No, I didn't. To be honest, I used AI tools for quick responses and convenience-unless I was doing really deep research, I don't think I could be bothered to use the links*', which falls under the established theme of 'Reflections on the Purpose of Sources'.

Further thematic analysis resulted in the production of 34 codes, including the initial four labels. The most commonly used codes were: in-the-past-not-now ($N = 16$), never ($N = 13$) and info-verification ($N = 13$). From these codes, an additional theme was derived, namely:

- Reluctance to Navigate Away from the Agent: This theme complements participant reflections on the purpose of links by surfacing their prescriptive purpose of conversational agents. Specifically, many participants expect these agents to be their sole source of information. As one user stated: '*I've never used links included in a[n] AI chat session. The reason I use an AI chatbot is to get information in one place. If I wanted to go to differ[ent] webpage, I'd use Google*'.

Moreover, this analysis revealed two additional purposes for source-clicking according to participants: information verification and source tracing. These purposes align with the initial assumption of this study, showing that participants are indeed interested in using links to verify information correctness and source trustworthiness.

6. Discussion

This research aimed to answer the research question *What effect does agent persona have on the appropriateness of user reliance on, and trust in, generative health conversational search?* and the sub-question *(a): What effect does health literacy have on the relationship above?* Our results revealed:

- **Direct effects on the underlying mechanism through which we aimed to affect reliance and trust; verification behaviour.**
- **Two- and three-way interaction effects of persona design, health literacy and attitude towards AI on verification behaviour.**
- **Two- and three-way interaction effects of persona design, health literacy and attitude towards AI on perceived risk.**
- **No difference in search satisfaction by persona.**

Nonetheless, questions about reliance, accuracy and other dimensions of trust in relation to persona and HL remain open.

6.1. Promising directions: persona design for verification behaviour

By partially supporting **H1**, our results indicate that conversational agent persona design is a promising approach to increasing user verification behaviour without compromising user satisfaction. In this study, we can observe the significant change in source-clicking elements such as source page views and, when accounting for HL, link clicks as the search agent persona is varied. Therefore, this study contributes to a growing body of evidence (Biro, Linder, and Neyens 2023; Nißen et al. 2022) on the importance of intentional persona design in the digital health domain. We further underscore the importance of accounting for the target user group's HL in verification behaviour research and conversational search design.

Indeed, while our hypothesis (**H1**) was true for the High HL group, the reverse was the case for the Low HL group. The Low HL group clicked on links more often in the Scientific Persona, not the Friendly Persona. Several explanations for this observation are possible. First, the High HL group may have been more influenced by the perceived formality or authority of the persona than the Low HL group. As previous research (Hosking, Blunsom, and Bartolo 2024; Kim et al. 2024; Metzger et al. 2024) shows, LLM output assertiveness can impact perceived uncertainty, error rate and trust; this aspect of the Scientific Persona may have unexpectedly been perceived more acutely in the higher HL group with their existing higher levels of trust in medical institutions and experts (Chen et al. 2018). Second, the Low HL group may have struggled with understanding the response more in the Scientific persona condition than the HL group, and thus was more motivated to investigate the topic further. User reflection on their experience with the agent indeed suggests that some found the Scientific agent inaccessible. This explanation aligns with past HCI research (Ahmad et al. 2019; Khawaji et al. 2014; Zhou et al. 2017) connecting higher cognitive load to decreased trust, though trust ratings here were unchanged, it seems verification behaviour was. These observations give rise to an interesting tension between system accessibility and safety (as driven by verification) as design goals.

Overall, these findings suggest that agents' personas can be dynamically varied throughout the interaction to reflect system confidence and encourage appropriate verification. With the Low HL target group in mind, designers and developers can create warmer, informal responses when system confidence is high and colder, scientific responses when confidence is low. Of course, these aspects of information presentation must be balanced with other design considerations, such as accessibility, persona consistency (Medhi Thies et al. 2017; Smestad and Volden 2018) and system effectiveness (Biro, Linder, and Neyens 2023).

6.2. Lost in translation? capturing reliance and accuracy in complex tasks

While previous research (Kim et al. 2025) found appropriate reliance and accuracy improved with increased source-clicking, we did not see a similar effect in this work. One explanation for this observation is the complexity of the task in this study. Previous decision-making studies have used more binary, clear-cut, question-answering tasks such as '*Is it illegal to collect rainwater in Colorado?*' and '*Do more than two-thirds of South America's population live in Brazil?*', while we tasked users with information retrieval and comprehension from a more fact-dense, medical response emulating real-world health information communication. In fact, we can see from the reasoning participants gave in the study that even when exposed to correct information, users still made deduction errors that resulted in incorrect answers and vice versa.

These findings suggest we may need to rethink how reliance and accuracy are captured in digital health research. In this work, qualitative user reflections offered an explanation for user reliance rates as well as valuable insight into agent utilisation, with far fewer users referring to the links in their response reasoning than the conversational agent. Therefore, a mixed-method approach with increased focus on qualitative insights may best serve future work. HCI research has increasingly adopted mixed-method approaches as a representation of the multidisciplinary nature of the field, and frameworks are emerging to support researchers' understanding of the trade-offs in different methodology combinations within this approach (van Turnhout et al. 2014). Likewise, health research has also begun to address the complexity of healthcare inquiry through the use of mixed methods (O'Cathain, Murphy, and Nicholl 2007; Rana and Chimoriya 2025). This work offers support for the use of mixed-method approaches to build a more complete picture of the user experience within conversational search, and may be enhanced in the future by the inclusion of conversational log data and think-aloud methods.

6.3. Inaccurate risk assessment: perceived risk and over-trust

The uncovered moderating effect of HL on the relationship between persona and perceived risk, along with the weak correlation between HL and acceptance of AI, further highlights the importance of understanding the target group's literacy levels when designing digital health systems. The lower perceived risk in the High HL group was unexpected, given this group's preference for information from medical experts (Chen et al. 2018), but could be explained by the higher acceptance of AI. Other research has observed a similar trend, where users with higher health literacy were significantly more likely to trust an educational healthcare chatbot (Biro, Linder, and Neyens 2023). The Low HL group reporting higher perceived risk in the Scientific persona condition reversed our **H5.4**, also revealing an unexpected finding. It is possible this group found the search agent to be riskier due to its complexity and inaccessibility, which then drove higher verification. Similar to Luce Lupetti et al. (2023), reflecting on our results also suggests that limiting user trust, or at least increasing perceived risk, in conversational agents may be more beneficial for creating safer and more verifiable digital health tools.

Exploratory analysis of trust dimensions in this work showed a uniformly high trust in our conversational search agent. We can term this inappropriate trust (Mehrotra et al. 2024), i.e. trust which is out of sync with actual agent trustworthiness, or more specifically, over-trust, since the agent communicated factual errors. Future work may therefore need to focus on design for the mitigation of over-trust in digital health conversational agents. Past research has already observed instances of over-trust and over-reliance in LLM output in other domains (Kim et al. 2025; Łajewska et al. 2024; Eleni Spatharioti et al. 2023). In the healthcare setting, evaluation of LLM-based conversational agents has shown that the errors which agents make in responding to health inquiries may render reliance on them risky and harmful to end users (Howard, Hope, and Gerada 2023; Shiferaw et al. 2024). User reflection in this study and other work on LLM search outputs (Narayanan Venkit et al. 2025) also showed that for many users, the inclusion of links increased trust without motivating link-use, with link credibility being assumed. Thus reducing verification friction may be the only way to enable accurate trust judgments. It may be most effective for designers to further reduce the friction in link use by including snippets, images and more attention-grabbing links. Indeed, Narayanan Venkit et al. (2025)'s interviews suggest that for some participants, including hover-over snippets increased the usability of sources, uncovering a potentially impactful novel design direction.

7. Limitations

Several limitations arose in this study. First, though we aimed to introduce the sense of risk necessary for trust formation through the use of bonus payments for correct responses, this addition may still be insufficient to motivate realistic engagement with the system. Crowdsource workers can often be busy, possibly compromising data quality (Ikeda and Hoashi 2017); however, this is a trait that workers share with some health information seekers too, such as young parents. Therefore, we can still interpret our results through that lens. Moreover, crowdsource workers are likely more digitally literate than the average user, which may colour their opinions about AI tools. Of course, the adoption of AI tools has been a popular media discussion topic in recent years (Vrabič Dežman 2024; W. Wang, Downey, and Yang 2023) and thus even digitally inexperienced users are likely to have been exposed to some information on the topic. Our smaller sample size may additionally hinder the generalisability of our findings in relation to trust, which is not repeatedly measured unlike source-clicking and reliance. Lastly, we acknowledge that it is difficult to capture trust beyond perceived trustworthiness in these short-term online studies. Trust building is a long process with many points for feedback, calibration and repair (Hoff and Bashir 2015); however, logistical constraints make the longitudinal observation of trust formation difficult. In a similar vein, we can see that attitude towards AI shows significant two-way and three-way interactions with health literacy and persona design. This aspect of user attitude may be overly influential in the first interactions with a new tool, as they have not yet had the history of interaction needed to calibrate their trust in the new tool specifically. Nonetheless, we hope these results offer initial insights into trust and perceived trustworthiness in health information conversational search agents.

8. Future work

As mentioned in our discussion, this work further emphasises the necessity of studying over-trust in generative conversational search, specifically in the digital health domain. Future work should aim to not only mitigate over-trust but also understand the antecedents of user reliance behaviour and trust beliefs. Additionally, future work may build on this study to investigate how friction in verification behaviour can be reduced through response and link design. Some promising design directions are discussed here, such as hover-over snippets, but much room for exploratory research remains. Lastly, the domain of trust and reliance in conversational search would benefit from more longitudinal, in-the-field studies, to allow us to understand the evolution of these behaviours and beliefs over time and potentially uncover interaction points where trust calibration is most effective.

9. Conclusion

In this paper, we aimed to reduce the possible harms arising from the generative conversation search's potential to disseminate inaccurate or incomplete information through trust and reliance calibration. We used source-clicking as the mechanism through which to impact reliance behaviour and trust beliefs. We explored the design of search agent personas to increase source click-through behaviour and hopefully improve appropriate reliance and trust, while also investigating the moderating effects of health literacy. Our results show that persona design is a promising direction for influencing source-clicking, and that health literacy interacts with persona design to affect verification behaviour and perceived risk. Based on these results, we discussed potential design directions for conversational search agents, as well as future research directions in the fields of trust and reliance in HCI.

Notes

1. blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing – last access March 11, 2026. openai.com/index/introducing-chatgpt-search – last access March 11, 2026. cloud.google.com/blog/products/ai-machine-learning/vertex-ai-io-announcements – last access March 11, 2026.
2. nnlm.gov/guides/intro-health-literacy.
3. https://openai.com/index/hello-gpt-4o/–last accessed on March 11, 2026.
4. prolific.com – last access March 11, 2026.
5. prolific.com – last access March 11, 2026.
6. sbert.net – last access March 11, 2026.
7. github.com/Behrouz-Babaki/COP-Kmeans – last access March 11, 2026.

Acknowledgments

We gratefully acknowledge the crowdworkers and other stakeholders who have contributed to this research. The authors report there are no competing interests to declare.

Author contributions

CRediT: **Shatha Degachi:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing; **Evangelos Niforatos:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing; **Gerd Kortuem:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is part of the larger Our Smart Family Buddy project. The Our Smart Family Buddy program and this work are made possible by financial support from the De Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), under Grant KICH1.GZ01.20.016.

References

Abbasian, Mahyar, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, et al. 2024. "Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI." *Npj Digital Medicine* 7 (1): 82. https://doi.org/10.1038/s41746-024-01074-z.

Ahmad, Muneeb Imtiaz, Jasmin Bernotat, Katrin Lohan, and Friederike Eyssel. 2019. "Trust and Cognitive Load During Human–Robot Interaction." arXiv:1909.05160 [cs]. https://doi.org/10.48550/arXiv.1909.05160.

Arias-Duart, Anna, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. 2025. "Automatic Evaluation of Healthcare LLMs Beyond Question-Answering." arXiv:2502.06666 [cs]. https://doi.org/10.48550/arXiv.2502.06666.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 1–16. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445717.

Bedi, Suhana, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, et al. 2024. "A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs)." 2024.04.15.24305869 pages. Preprint: medRxiv. https://doi.org/10.1101/2024.04.15.24305869.

Berkman, Nancy D., Stacey L. Sheridan, Katrina E. Donahue, David J. Halpern, and Karen Crotty. 2011. "Low Health Literacy and Health Outcomes: An Updated Systematic Review." *Annals of Internal Medicine* 155 (2): 97–107. https://doi.org/10.7326/0003-4819-155-2-201107190-00005.

Bernhardt, Jay M., and Elizabeth M. Felter. 2004. "Online Pediatric Information Seeking among Mothers of Young Children: Results from a Qualitative Study Using Focus Groups." *Journal of Medical Internet Research* 6 (1): e36. https://doi.org/10.2196/jmir.6.1.e7.

Bewick, Viv, Liz Cheek, and Jonathan Ball. 2005. "Statistics Review 14: Logistic Regression." *Critical Care* 9 (1): 112–118. https://doi.org/10.1186/cc3045.

Bickmore, Timothy W., Dina Utami, Robin Matsuyama, and Michael K. Paasche-Orlow. 2016. "Improving Access to Online Health Information with Conversational Agents: A Randomized Controlled Experiment." *Journal of Medical Internet Research* 18 (1): e1. https://doi.org/10.2196/jmir.5239.

Bin Sawad, Abdullah, Bhuva Narayan, Ahlam Alnefaie, Ashwaq Maqbool, Indra Mckie, Jemma Smith, Berkan Yuksel, Deepak Puthal, Mukesh Prasad, and A. Baki Kocaballi. 2022. "A Systematic Review on Healthcare Artificial Intelligent Conversational Agents for Chronic Conditions." *Sensors* 22 (77): 2625. https://doi.org/10.3390/s22072625.

Biro, Joshua, Courtney Linder, and David Neyens. 2023. "The Effects of a Health Care Chatbot's Complexity and Persona on User Trust, Perceived Usability, and Effectiveness: Mixed Methods Study." *JMIR Human Factors* 10:e41017. https://doi.org/10.2196/41017.

Boren, Suzanne Austin. 2009. "A Review of Health Literacy and Diabetes: Opportunities for Technology." *Journal of Diabetes Science and Technology* 3 (1): 202–209. https://doi.org/10.1177/193229680900300124.

Borji, Ali. 2023. "A Categorical Archive of ChatGPT Failures." arXiv:2302.03494 [cs]. https://doi.org/10.48550/arXiv.2302.03494.

Braun, V., and V. Clarke. 2021. "Thematic Analysis: A Practical Guide [eBook version]." In *SAGE moradi H, vaezi A. lessons learned from Korea: COVID-19 Pandemic*, Vol. 41, 873–4.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." In *Proceedings of the 34th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada) (NIPS '20)*, Article 159, 25 pages. Red Hook, NY: Curran Associates Inc..

Chen, Xuewei, Jennifer L. Hay, Erika A. Waters, Marc T. Kiviniemi, Caitlin Biddle, Elizabeth Schofield, Yuelin Li, Kimberly Kaphingst, and Heather Orom. 2018. "Health Literacy and Use and Trust in Health Information." *Journal of Health Communication* 23 (8): 724–734. https://doi.org/10.1080/10810730.2018.1511658.

Clarke, Victoria, and Virginia Braun. 2013. *Successful Qualitative Research: A Practical Guide for Beginners*. London, UK: Sage Publications Ltd.

Colditz, Jason B., Michelle S. Woods, and Brian A. Primack. 2018. "Adolescents Seeking Online Health Information: Topics, Approaches, and Challenges." In *Technology and Adolescent Mental Health*, edited by Megan A. Moreno and Ana Radovic, 21–35. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-69638-6_2.

Cox, Samuel Rhys, and Wei Tsang Ooi. 2022. "Does Chatbot Language Formality Affect Users' Self-disclosure?" In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22)*, 1–13. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3543829.3543831.

Coxe, Stefany, Stephen G. West, and Leona S. Aiken. 2009. "The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives." *Journal of Personality Assessment* 91 (2): 121–136. https://doi.org/10.1080/00223890802634175.

Crisp, Nigel, and Lincoln Chen. 2014. "Global Supply of Health Professionals." *New England Journal of Medicine* 370 (10): 950–957. https://doi.org/10.1056/NEJMra1111610.

Crutzen, Rik, Gjalt-Jorn Y. Peters, Sarah Dias Portugal, Erwin M. Fisser, and Jorne J. Grolleman. 2011. "An Artificially Intelligent Chat Agent That Answers Adolescents' Questions Related to Sex, Drugs, and Alcohol: An Exploratory Study." *Journal of Adolescent Health* 48 (5): 514–519. https://doi.org/10.1016/j.jadohealth.2010.09.002.

Denniss, Emily, Rebecca Lindberg, and Sarah A. McNaughton. 2023. "Quality and Accuracy of Online Nutrition-Related Information: A Systematic Review of Content Analysis Studies." *Public Health Nutrition* 26 (7): 1345–1357. https://doi.org/10.1017/S1368980023000873.

Desai, Smit, Christina Wei, Jaisie Sin, Mateusz Dubiel, Nima Zargham, Shashank Ahire, Martin Porcheron, et al. 2024. "CUICHI 2024: Building Trust in CUIs-from Design to Deployment." arXiv:2401.13970 [cs]. https://doi.org/10.1145/3613905.3636287.

de Visser, Ewart J., Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A.

Neerincx. 2020. “Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams.” *International Journal of Social Robotics* 12 (2): 459–478. https://doi.org/10.1007/s12369-019-00596-x.

Downe-Wamboldt, Barbara. 1992. “Content Analysis: Method, Applications, and Issues.” *Health Care for Women International* 13 (3): 313–21. https://doi.org/10.1080/07399339209516006.

Eastin, Matthew S., and Natalie M. Guinsler. 2006. “Worried and Wired: Effects of Health Anxiety on Information-Seeking and Health Care Utilization Behaviors.” *CyberPsychology & Behavior* 9 (4): 494–498. https://doi.org/10.1089/cpb.2006.9.494.

Elsholz, Ela, Jon Chamberlain, and Udo Kruschwitz. 2019. “Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement.” In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR ’19)*, 301–305. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3295750.3298956.

Eurostat. 2022. “Individuals Using the Internet for Seeking Health-Related Information.” https://doi.org/10.2908/TIN00101.

Fadhil, Ahmed, Gianluca Schiavo, Yunlong Wang, and Bereket A. Yilma. 2018. “The Effect of Emojis When Interacting with Conversational Interface Assisted Health Coaching System.” In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (New York, NY) (PervasiveHealth ’18)*, 378–383. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3240925.3240965.

Fok, Raymond, and Daniel S. Weld. 2024. “In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-advised Decision Making.” *AI Magazine* 45 (3): 317–332. https://doi.org/10.1002/aaai.12182.

Freire, Samuel Kernan, Chaofan Wang, and Evangelos Niforatos. 2024. “Conversational Assistants in Knowledge-Intensive Contexts: An Evaluation of LLM-Versus Intent-Based Systems.” arXiv:2402.04955 [cs]. https://doi.org/10.48550/arXiv.2402.04955.

Guetterman, Timothy C., Tammy Chang, Melissa DeJonckheere, Tanmay Basu, Elizabeth Scruggs, and V. G. Vinod Vydiswaran. 2018. “Augmenting Qualitative Text Analysis with Natural Language Processing: Methodological Study.” *Journal of Medical Internet Research* 20 (6): e231. https://doi.org/10.2196/jmir.9702.

Gulati, Siddharth, Sonia Sousa, and David Lamas. 2019. “Design, Development and Evaluation of a Human–Computer Trust Scale.” *Behaviour & Information Technology* 38 (10): 1004–1015. https://doi.org/10.1080/0144929X.2019.1656779.

Harrell, Frank E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7.

Harrison, Vrindavan, Lena Reed, Shereen Oraby, and Marilyn A. Walker. 2019. “Maximizing Stylistic Control and Semantic Accuracy in NLG: Personality Variation and Discourse Contrast.” In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*. https://doi.org/10.48550/arXiv.1907.09527.

Harwood, Tracy G., and Tony Garry. 2003. “An Overview of Content Analysis.” *The Marketing Review* 3 (4): 479–498. https://doi.org/10.1362/146934703771910080.

Hoff, Kevin Anthony, and Masooda Bashir. 2015. “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (3): 407–434. https://doi.org/10.1177/0018720814547570.

Hollender, Nina, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. “Integrating Cognitive Load Theory and Concepts of Human–Computer Interaction.” *Computers in Human Behavior* 26 (6): 1278–1288. https://doi.org/10.1016/j.chb.2010.05.031.

Hosking, Tom, Phil Blunsom, and Max Bartolo. 2024. “Human Feedback Is Not Gold Standard.” arXiv:2309.16349. https://doi.org/10.48550/arXiv.2309.16349.

Howard, Alex, William Hope, and Alessandro Gerada. 2023. “ChatGPT and Antimicrobial Advice: The End of the Consulting Infection Doctor?” *The Lancet Infectious Diseases* 23 (4): 405–406. https://doi.org/10.1016/S1473-3099(23)00113-5.

Hutchinson, M. Katherine, and Matthew C. Holtman. 2005. “Analysis of Count Data Using Poisson Regression.” *Research in Nursing & Health* 28 (5): 408–418. https://doi.org/10.1002/nur.20093.

Ikeda, Kazushi, and Keiichiro Hoashi. 2017. “Crowdsourcing GO: Effect of Worker Situation on Mobile Crowdsourcing Performance.” In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI ’17)*, 1142–1153. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3025453.3025917.

Imrey, Peter B. 2000. *14 -- Poisson Regression, Logistic Regression, and Loglinear Models for Random Counts*, 391–437. San Diego: Academic Press. https://doi.org/10.1016/B978-012691360-6/50015-X.

JASP Team. 2025. “JASP (Version 0.95.1) [Computer Software].” https://jasp-stats.org/.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. “Survey of Hallucination in Natural Language Generation.” *ACM Computing Surveys* 55 (12): 248:1–248:38. https://doi.org/10.1145/3571730.

Jiang, Shaohai, and Richard L. Street. 2017. “Pathway Linking Internet Health Information Seeking to Better Health: A Moderated Mediation Study.” *Health Communication* 32 (8): 1024–1031. https://doi.org/10.1080/10410236.2016.1196514.

Jorge, Carolina Centeio, Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. “Trust Should Correspond to Trustworthiness: A Formalization of Appropriate Mutual Trust in Human-Agent Teams.” In *Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST2021): Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS2021)* 3022, 12 pages.

Khawaji, Ahmad, Fang Chen, Jianlong Zhou, and Nadine Marcus. 2014. “Trust and Cognitive Load in the Text-Chat Environment: The Role of Mouse Movement.” In *Proceedings of the 26th Australian Computer–Human*

Interaction Conference on Designing Futures: The Future of Design (OzCHI '14), 324–327. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2686612.2686661.

Kim, Sunnie S. Y., Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. ""I'm Not Sure, But…": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 822–835. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3630106.3658941.

Kim, Sunnie S. Y., Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. "Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies." arXiv:2502.08554 [cs]. https://doi.org/10.1145/3706598.3714020.

Kocaballi, Ahmet Baki, Shlomo Berkovsky, Juan C. Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. 2019. "The Personalization of Conversational Agents in Health Care: Systematic Review." *Journal of Medical Internet Research* 21 (11): e15360. https://doi.org/10.2196/15360.

Kosch, Thomas, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and PawełW. Woźniak. 2023. "A Survey on Measuring Cognitive Workload in Human-Computer Interaction." *ACM Computing Surveys* 55 (13s): 283:1–283:39. https://doi.org/10.1145/3582272.

Laerd Statistics. 2018. "Binomial Logistic Regression Using SPSS Statistics." https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php.

Łajewska, Weronika, Damiano Spina, Johanne Trippas, and Krisztian Balog. 2024. "Explainability for Transparent Conversational Information-Seeking." Preprint. https://doi.org/10.1145/3626772.3657768.

Lan, Shih-Han, Sara Mahmoud, and Kari L. Franson. 2024. "A Narrative Review on the Impact of Online Health Misinformation on Patients' Behavior and Communication." *American Journal of Health Behavior* 48 (2): 564–572. https://doi.org/10.5993/AJHB.48.2.25.

Lang, Annie. 2000. "The Limited Capacity Model of Mediated Message Processing." *Journal of Communication* 50 (1): 46–70. https://doi.org/10.1111/j.1460-2466.2000.tb02833.x.

Lee, John D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50.30392.

Lessio, Nadine, and Alexis Morris. 2020. "Toward Design Archetypes for Conversational Agent Personality." In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3221–3228. Toronto, ON: IEEE Press. https://doi.org/10.1109/SMC42975.2020.9283254.

Li, Jin, Yiyan Deng, Qi Sun, Junjie Zhu, Yu Tian, Jingsong Li, and Tingting Zhu. 2024. "Benchmarking Large Language Models in Evidence-Based Medicine." *IEEE Journal of Biomedical and Health Informatics* 29 (9): 1–14. https://doi.org/10.1109/JBHI.2024.3483816.

Li, Qing, Lei Li, and Yu Li. 2024. "Developing ChatGPT for Biology and Medicine: A Complete Review of Biomedical Question Answering." *Biophysics Reports* 10 (3): 152–171. https://doi.org/10.52601/bpr.2024.240004.

Liao, Q. Vera, and S. Shyam Sundar. 2022. "Designing for Responsible Trust in AI Systems: A Communication Perspective." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (Seoul, Republic of Korea) (FAccT '22)*, 1257–1268. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533182.

Liu, Mengyang, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. ""Satisfaction with Failure" or "Unsatisfied Success": Investigating the Relationship between Search Success and User Satisfaction." In *Proceedings of the 2018 World Wide Web Conference (Lyon, France) (WWW '18)*, 1533–1542. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186065.

Lockey, Steven, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. 2021. "A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions." In *Proceedings of the Annual Hawaii International Conference on System Sciences*, 10 pages. Hawai'i, USA: Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.664.

Lupetti, Maria Luce, Emma Hagens, Willem Van Der Maden, Régine Steegers-Theunissen, and Melek Rousian. 2023. "Trustworthy Embodied Conversational Agents for Healthcare: A Design Exploration of Embodied Conversational Agents for the Periconception Period at Erasmus MC." In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*, 1–14. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3571884.3597128.

Luxton, David D., Russell A. McCann, Nigel E. Bush, Matthew C. Mishkind, and Greg M. Reger. 2011. "MHealth for Mental Health: Integrating Smartphone Technology in Behavioral Healthcare." *Professional Psychology: Research and Practice* 42 (6): 505–512. https://doi.org/10.1037/a0024485.

Mairesse, François, and Marilyn A. Walker. 2011. "Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits." *Computational Linguistics* 37 (3): 455–488. https://doi.org/10.1162/coli_a_00063.

Medhi Thies, Indrani, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. 2017. "How Do You Want Your Chatbot? an Exploratory Wizard-of-Oz Study with Young, Urban Indians." In *Human–Computer Interaction -- INTERACT 2017*, edited by Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler, 441–459. Cham: Springer International Publishing.

Mehrotra, Siddharth, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. "A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges." *ACM Journal on Responsible Computing* 1 (4): Article 26, 45 pages–45. https://doi.org/10.1145/3696449.

Mekhail, Kirsi Tiitinen, Bo Burström, Anneli Marttila, Josefin Wångdahl, and Lene Lindberg. 2022. "Psychometric Qualities of the HLS-EU-Q16 Instrument for Parental Health Literacy in Swedish Multicultural Settings." *BMC*

Public Health 22 (1): 293. https://doi.org/10.1186/s12889-021-12346-8.

Mendoza, Michael F., Ralf Martz Sulague, Therese Posas-Mendoza, and Carl J. Lavie. 2023. "Impact of Coffee Consumption on Cardiovascular Health." *Ochsner Journal* 23 (2): 152–158. https://doi.org/10.31486/toj.22.0073.

Meng, Hua, Xinyuan Lu, and Jiangling Xu. 2025. "The Impact of Chatbot Response Strategies and Emojis Usage on Customers' Purchase Intention: The Mediating Roles of Psychological Distance and Performance Expectancy." *Behavioral Sciences* 15 (2): 117. https://doi.org/10.3390/bs15020117.

Meppelink, Corine S., Edith G. Smit, Nicola Diviani, and Julia C. M. Van Weert. 2016. "Health Literacy and Online Health Information Processing: Unraveling the Underlying Mechanisms." *Journal of Health Communication* 21 (sup2): 109–120. https://doi.org/10.1080/10810730.2016.1193920.

Metzger, Luise, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. "Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style." In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, 1–19. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642122.

Milne-Ives, Madison, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. 2020. "The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review." *Journal of Medical Internet Research* 22 (10): e20346. https://doi.org/10.2196/20346.

Mo, Fengran, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2025. "A Survey of Conversational Search." *ACM Transactions on Information Systems* 43 (6): Article 167, 50 pages–50. https://doi.org/10.1145/3759453.

Mohr, David C., Justin K. Benzer, and Gary J. Young. 2013. "Provider Workload and Quality of Care in Primary Care Settings: Moderating Role of Relational Climate." *Medical Care* 51 (1): 108–114. https://doi.org/10.1097/MLR.0b013e318277f1cb.

Narayanan Venkit, Pranav, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025. "Search Engines in the AI Era: A Qualitative Understanding to the False Promise of Factual and Verifiable Source-Cited Responses in LLM-Based Search." In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, 1325–1340. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3715275.3732089.

Ngai, Cindy Sing Bik, Rita Gill Singh, and Le Yao. 2022. "Impact of COVID-19 Vaccine Misinformation on Social Media Virality: Content Analysis of Message Themes and Writing Strategies." *Journal of Medical Internet Research* 24 (7): e37806. https://doi.org/10.2196/37806.

Nißen, Marcia, Dominik Rüegger, Mirjam Stieger, Christoph Flückiger, Mathias Allemand, Florian V. Wangenheim, and Tobias Kowatsch. 2022. "The Effects of Health Care Chatbot Personas with Different Social Roles on the Client-Chatbot Bond and Usage Intentions: Development of a Design Codebook and Web-Based Study." *Journal of Medical Internet Research* 24 (4): e32630. https://doi.org/10.2196/32630.

Niu, Zhaomeng, David C. Jeong, and Jessica Fitts Willoughby. 2020. "Friends over Doctors? The Influences of Source and Perceived Customization on College Drinking." *Health Communication* 36 (6): 671–681. https://doi.org/10.1080/10410236.2020.1712034.

O'Cathain, Alicia, Elizabeth Murphy, and Jon Nicholl. 2007. "Why, and How, Mixed Methods Research Is Undertaken in Health Services Research in England: A Mixed Methods Study." *BMC Health Services Research* 7 (1): 85. https://doi.org/10.1186/1472-6963-7-85.

O'Neill, Onora. 2018. "Linking Trust to Trustworthiness." *International Journal of Philosophical Studies* 26 (2): 293–300. https://doi.org/10.1080/09672559.2018.1454637.

Pafla, Marvin, Kate Larson, and Mark Hancock. 2024. "Unraveling the Dilemma of AI Errors: Exploring the Effectiveness of Human and Machine Explanations for Large Language Models." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, 1–20. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642934.

Parasuraman, Raja, and Victor Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39 (2): 230–253. https://doi.org/10.1518/001872097778543886.

Pelikan, Juergen M., and Kristin Ganahl. 2017. "Measuring Health Literacy in General Populations: Primary Findings from the HLS-EU Consortium's Health Literacy Assessment Effort." *Studies in Health Technology and Informatics* 240:34–59.

Pelikan, Jürgen M, Christa Straßmayr, and Kristin Ganahl. 2020. "Health Literacy Measurement in General and Other Populations: Further Initiatives and Lessons Learned in Europe (and Beyond)." *Studies in Health Technology and Informatics* 269:170–191. https://doi.org/10.3233/SHTI200031.

Pradhan, Alisha, and Amanda Lazar. 2021. "Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents." In *CUI 2021 -- 3rd Conference on Conversational User Interfaces*, 1–4. Bilbao (online) Spain: ACM. https://doi.org/10.1145/3469595.3469607.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training." Preprint, https://paperswithcode.com/paper/improving-language-understanding-by.

Radlinski, Filip, and Nick Craswell. 2017. "A Theoretical Framework for Conversational Search." In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*, 117–126. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3020165.3020183.

Rana, Kritika, and Ritesh Chimoriya. 2025. "A Guide to a Mixed-Methods Approach to Healthcare Research." *Encyclopedia* 5 (2): 51. https://doi.org/10.3390/encyclopedia5020051.

Rawte, Vipula, Amit Sheth, and Amitava Das. 2023. "A Survey of Hallucination in Large Foundation Models." arXiv:2309.05922 [cs]. https://doi.org/10.48550/arXiv.2309.05922.

Rhee, Chong Eun, and Junho Choi. 2020. "Effects of Personalization and Social Role in Voice Shopping: An Experimental Study on Product Recommendation by a Conversational Voice Agent." *Computers in Human Behavior* 109:106359. https://doi.org/10.1016/j.chb.2020.106359.

Rheu, Minjin, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. "Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design." *International Journal of Human–Computer Interaction* 37 (1): 81–96. https://doi.org/10.1080/10447318.2020.1807710.

Ruane, Elayne, Sinead Farrell, and Anthony Ventresque. 2021. "User Perception of Text-Based Chatbot Personality." In *Chatbot Research and Design*, edited by Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie L.-C. Law, Ewa Luger, Morten Goodwin, and Petter Bae Brandtzaeg, 32–47. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-68288-0_3.

Schlicker, Nadine, Kevin Baum, Alarith Uhde, Sarah Sterz, Martin C. Hirsch, and Markus Langer. 2022. "How Do We Assess the Trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM)." OSF Preprint. https://doi.org/10.31234/osf.io/qhwvx.

Seidel, Jan, Pietro Mascheroni, Steven Geoffrey Brooks, Stefan Doering, and Jan Seidel. 2024. "Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review." Preprint. https://doi.org/10.20944/preprints202407.0876.v1.

Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. "Large Language Models Can Be Easily Distracted by Irrelevant Context." In *Proceedings of the 40th International Conference on Machine Learning*, 31210–31227. Honolulu, Hawai'i: PMLR. https://proceedings.mlr.press/v202/shi23a.html.

Shiferaw, Meron W., Taylor Zheng, Abigail Winter, Leigh Ann Mike, and Lingtak-Neander Chan. 2024. "Assessing the Accuracy and Quality of Artificial Intelligence (AI) Chatbot-Generated Responses in Making Patient-Specific Drug-Therapy and Healthcare-Related Decisions." *BMC Medical Informatics and Decision Making* 24 (1): 404. https://doi.org/10.1186/s12911-024-02824-5.

Shrestha, Noora. 2020. "Detecting Multicollinearity in Regression Analysis." *American Journal of Applied Mathematics and Statistics* 8 (2): 39–42. https://doi.org/10.12691/ajams-8-2-1.

Shuster, Kurt, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. "Retrieval Augmentation Reduces Hallucination in Conversation." arXiv:2104.07567 [cs]. https://doi.org/10.48550/arXiv.2104.07567.

Si, Chenglei, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. "Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, edited by Kevin Duh, Helena Gomez, and Steven Bethard, 1459–1474. Mexico City, Mexico: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.81.

Sindermann, Cornelia, Peng Sha, Min Zhou, Jennifer Wernicke, Helena S. Schmitt, Mei Li, Rayna Sariyska, Maria Stavrou, Benjamin Becker, and Christian Montag. 2021. "Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language." *KI -- Künstliche Intelligenz* 35 (1): 109–118. https://doi.org/10.1007/s13218-020-00689-0.

Singhal, Karan, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, et al. 2023. "Large Language Models Encode Clinical Knowledge." *Nature* 620 (7972): 172–180. https://doi.org/10.1038/s41586-023-06291-2.

Smestad, Tuva Lunde, and Frode Volden. 2018. "Chatbot Personalities Matters." In *International Conference on Internet Science*, 170–181. Cham: Springer, Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-17705-8_15.

Sorensen, K., S. van den Broucke, J. Fullam, G. Doyle, J. Pelikan, Z. Slonska, and H. Brand. 2012. "(HLS-EU) Consortium Health Literacy Project European Health Literacy and Public Health: A Systematic Review and Integration of Definitions and Models." *BMC Public Health* 25 (12): 80. https://doi.org/10.1186/1471-2458-12-80.

Sørensen, Kristine, Jürgen M. Pelikan, Florian Röthlin, Kristin Ganahl, Zofia Slonska, Gerardine Doyle, James Fullam, et al. on behalf of the HLS-EU Consortium. 2015. "Health Literacy in Europe: Comparative Results of the European Health Literacy Survey (HLS-EU)." *European Journal of Public Health* 25 (6): 1053–1058. https://doi.org/10.1093/eurpub/ckv043.

Spatharioti, Sofia Eleni, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. "Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment." arXiv:2307.03744 [cs]. https://doi.org/10.48550/arXiv.2307.03744.

Squiers, Linda, Susana Peinado, Nancy Berkman, Vanessa Boudewyns, and Lauren McCormack. 2012. "The Health Literacy Skills Framework." *Journal of Health Communication* 17 (Suppl 3): 30–54. https://doi.org/10.1080/10810730.2012.713442.

Sweller, John. 1994. "Cognitive Load Theory, Learning Difficulty, and Instructional Design." *Learning and Instruction* 4 (4): 295–312. https://doi.org/10.1016/0959-4752(94)90003-5.

Sweller, John. 2005. *Implications of Cognitive Load Theory for Multimedia Learning*. Vol. 3. 19–30. UK: Cambridge University Press.

Szymanski, Maxwell, Martijn Millecamp, and Katrien Verbert. 2021. "Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations." In *26th International Conference on Intelligent User Interfaces (IUI '21)*, 109–119. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3397481.3450662.

van Turnhout, Koen, Arthur Bennis, Sabine Craenmehr, Robert Holwerda, Marjolein Jacobs, Ralph Niels, Lambert Zaad, Stijn Hoppenbrouwers, Dick Lenior, and René Bakker. 2014. "Design Patterns for Mixed-Method Research in HCI." In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI '14)*, 361–370. New York, NY:

Association for Computing Machinery. https://doi.org/10.1145/2639189.2639220.

Versluis, Anke, Kyma Schnoor, Niels H. Chavannes, and Esther Pwa Talboom-Kamp. 2022. "Direct Access for Patients to Diagnostic Testing and Results Using EHealth: Systematic Review on EHealth and Diagnostics." *Journal of Medical Internet Research* 24 (1): e29303. https://doi.org/10.2196/29303.

Völkel, Sarah Theres, Ramona Schoedel, Lale Kaya, and Sven Mayer. 2022. "User Perceptions of Extraversion in Chatbots after Repeated Use." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, 1–18. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3491102.3502058.

Vrabič Dežman, Dominik. 2024. "Promising the Future, Encoding the Past: AI Hype and Public Media Imagery." *AI and Ethics* 4 (3): 743–756. https://doi.org/10.1007/s43681-024-00474-x.

Wang, Leyao, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Clayton, Bradley Malin, and Zhijun Yin. 2024. "Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review." *Journal of Medical Internet Research* 26 (1): e22769. https://doi.org/10.2196/22769.

Wang, Weili, John Downey, and Fan Yang. 2023. "AI Anxiety? Comparing the Sociotechnical Imaginaries of Artificial Intelligence in UK, Chinese and Indian Newspapers." *Global Media and China* 10 (4). https://doi.org/10.1177/20594364231196547.

Wang, Xinru, and Ming Yin. 2021. "Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making." In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*, 318–328. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3397481.3450650.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, et al. 2022. "Taxonomy of Risks Posed by Language Models." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 214–229. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088.

Wester, Joel, Sander de Jong, Henning Pohl, and Niels van Berkel. 2024. "Exploring People's Perceptions of LLM-generated Advice." *Computers in Human Behavior: Artificial Humans* 2 (2):100072. https://doi.org/10.1016/j.chbah.2024.100072.

White, Ryen W., and Eric Horvitz. 2014. "From Health Search to Healthcare: Explorations of Intention and Utilization via Query Logs and User Surveys." *Journal of the American Medical Informatics Association* 21 (1): 49–55. https://doi.org/10.1136/amiajnl-2012-001473.

Wu, Junde, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. "Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation." arXiv:2408.04187 [cs]. https://doi.org/10.48550/arXiv.2408.04187.

Zhang, Yunfeng, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. "Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20)*, 295–305. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372852.

Zhao, Wenting, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. "WildChat: 1M ChatGPT Interaction Logs in the Wild." arXiv.org, https://arxiv.org/abs/2405.01470v1.

Zhou, Jianlong, Syed Z. Arshad, Simon Luo, and Fang Chen. 2017. "Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making." In *Human–Computer Interaction – INTERACT 2017 (Lecture Notes in Computer Science)*, edited by Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler, 23–39. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68059-0_2.

Appendices

Appendix 1. Simulated conversations

The two simulated task conversations are shown here. Introduced factual errors are in square brackets.

A.1. Friendly persona conversation

User (Turn 1) How does drinking coffee affect heart health?

Chat (Turn 1) Hey there! [coffee emoji] So, coffee's impact on heart health is kinda mixed. Some coffee consumption might actually lower the odds of hypertension, heart failure and irregular heart rhythm. But, worryingly, heavy coffee drinking, more than 400 [600] mg of caffeine daily, actually increases the risk of heart disease and high blood pressure. Although, how you prepare your coffee also matters – filtered coffee is generally less likely to mess up your cholesterol levels than boiled or unfiltered coffee, it's tastier too! Read more about it in the links. Got more questions about coffee? Read more – Mendoza et al. (2023): Link — Heart Health Digital (2024): Link.

User (Turn 2) Yes, why does coffee intake reduce hypertension?

Chat (Turn 2) Great question! :) So here's the thing, some coffee intake might reduce high blood pressure (hypertension) for a few reasons. One, it boosts enzyme activity which helps digest caffeine and lower blood pressure. Plus, this ingredient in coffee called chlorogenic acid lowers [increases] inflammation and oxidative stress, which also helps keep blood pressure in check. This is especially true if you are one of those lucky people who can digest coffee quickly. Isn't that cool?! Read more about it in the links. Read More – Rodríguez-Artalejo (2018): Link — Heart Health Digital (2025): Link.

User (Turn 3) Should I drink green tea over coffee? Or should I avoid all caffeine.

Chat (Turn 3) Moderation is key! Green tea [green tea emoji] is definitely the healthier option because of its amazing antioxidants and lower caffeine content. And if you have a heart condition, or even anxiety, it might be worth avoiding coffee. We know many folks who have shared their experience with reducing caffeine. At first, they faced frustrating withdrawal symptoms like headaches or fatigue [dizziness]. But after about a week, they felt they had more stable energy levels and better sleep. Some people even say they don't miss caffeine as much as they thought they would, turns out, you can be peppy even before your morning coffee :)! Read more about it in the links. Read More – Rodda et al. (2020): Link — Heart Health Digital (2023): Link.

A.2. Scientific persona conversation

User (Turn 1): How does drinking coffee affect heart health?

Chat (Turn 1): Drinking coffee can have both positive and negative effects on heart health. Moderate coffee consumption is linked to a decreased risk of approximately 18% in hypertension, 5% in heart failure and 15% in atrial fibrillation, as noted by Mendoza et al. (2023). However, excessive caffeine intake, above 400 [600] mg daily, may elevate heart rate and blood pressure, according to Monique Richard, MS, RDN, LDN. The impact varies by individual and preparation method, with filtered coffee being less atherogenic than boiled coffee. Read more about it in the links. Would you like to know more about the impact of coffee? Read more – Mendoza et al. (2023): Link — Heart Health Digital (2024): Link

User (Turn 2): Yes, why does coffee intake reduce hypertension?

Chat (Turn 2) Moderate coffee intake may reduce hypertension due to several mechanisms. According to a 2018 literature review by Rodríguez-Artalejo et al., coffee can enhance cytochrome CYP1A2 activity, which is inversely related to blood pressure in non-smokers. It also inhibits sodium and water reabsorption and reduces [increases] inflammation and oxidative stress through chlorogenic acid. These effects are more pronounced in individuals who carry the CYP1A2*1A allele, which enables 'fast' caffeine metabolisation. Read more about it in the links. Read More – Rodríguez-Artalejo (2018): Link — Heart Health Digital (2025): Link.

User (Turn 3) Should I drink green tea over coffee? Or should I avoid all caffeine.

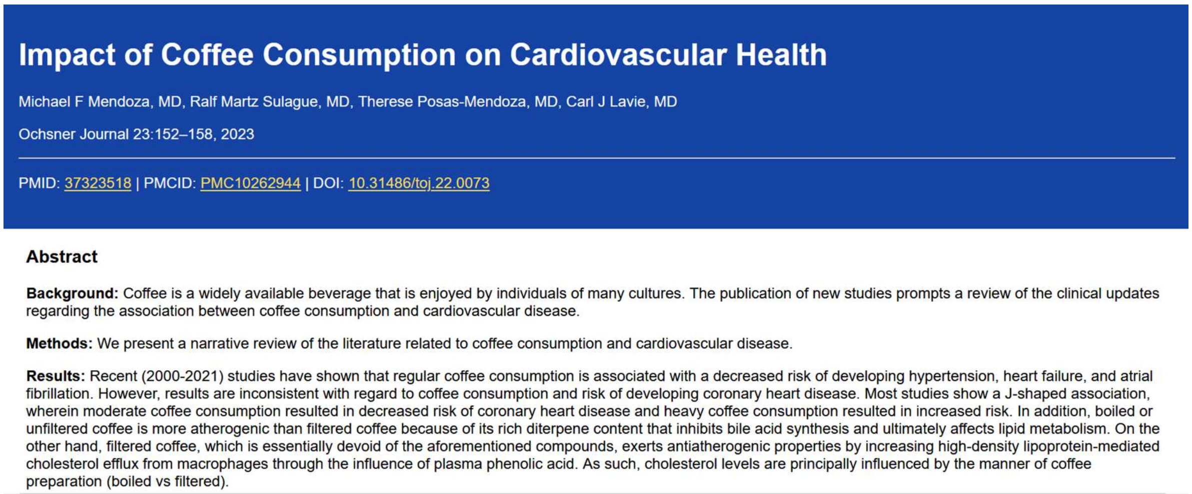
Chat (Turn 3) Doctors frequently recommend green tea over coffee for individuals with hypertension or cardiovascular risk factors, due to its beneficial polyphenols and 50% –70% lower caffeine content. However, complete caffeine avoidance isn't necessary for everyone. If you have existing heart conditions, hypertension or anxiety, reducing caffeine might be beneficial. Gradually taper intake to minimise withdrawal symptoms, such as headaches and fatigue [dizziness]which may last 2–9 days, and stay hydrated. Read more about it in the links. Read More – Rodda et al. (2020): Link — Heart Health Digital (2023): Link.

Appendix 2. Source page interface

Figures A1 and A2 illustrate the two source page interfaces used in the study.

Appendix 3. True or false reliance and accuracy tests

- A daily caffeine intake of over 500 mg is considered heavy or excessive.
- The chlorogenic acid in coffee reduces blood pressure because it increases oxidative stress.
- Caffeine withdrawal symptoms include headaches and dizziness.



Impact of Coffee Consumption on Cardiovascular Health

Michael F Mendoza, MD, Ralf Martz Sulague, MD, Therese Posas-Mendoza, MD, Carl J Lavie, MD

Ochsner Journal 23:152–158, 2023

PMID: 37323518 | PMCID: PMC10262944 | DOI: 10.31486/toj.22.0073

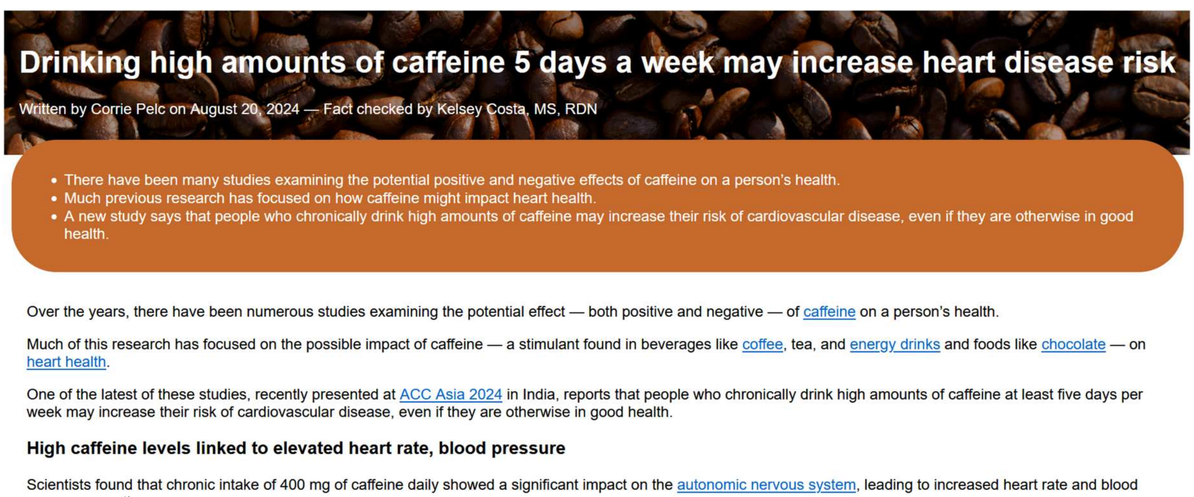
Abstract

Background: Coffee is a widely available beverage that is enjoyed by individuals of many cultures. The publication of new studies prompts a review of the clinical updates regarding the association between coffee consumption and cardiovascular disease.

Methods: We present a narrative review of the literature related to coffee consumption and cardiovascular disease.

Results: Recent (2000-2021) studies have shown that regular coffee consumption is associated with a decreased risk of developing hypertension, heart failure, and atrial fibrillation. However, results are inconsistent with regard to coffee consumption and risk of developing coronary heart disease. Most studies show a J-shaped association, wherein moderate coffee consumption resulted in decreased risk of coronary heart disease and heavy coffee consumption resulted in increased risk. In addition, boiled or unfiltered coffee is more atherogenic than filtered coffee because of its rich diterpene content that inhibits bile acid synthesis and ultimately affects lipid metabolism. On the other hand, filtered coffee, which is essentially devoid of the aforementioned compounds, exerts antiatherogenic properties by increasing high-density lipoprotein-mediated cholesterol efflux from macrophages through the influence of plasma phenolic acid. As such, cholesterol levels are principally influenced by the manner of coffee preparation (boiled vs filtered).

Figure A1. Example screenshot of a source page in the style of a journal. Screenshot of a journal article-like web page titled 'Impact of Coffee Consumption on Cardiovascular Health' displaying the title, authors, publication date, and a structured abstract summarising the study.



Drinking high amounts of caffeine 5 days a week may increase heart disease risk

Written by Corrie Pelc on August 20, 2024 — Fact checked by Kelsey Costa, MS, RDN

- There have been many studies examining the potential positive and negative effects of caffeine on a person's health.
- Much previous research has focused on how caffeine might impact heart health.
- A new study says that people who chronically drink high amounts of caffeine may increase their risk of cardiovascular disease, even if they are otherwise in good health.

Over the years, there have been numerous studies examining the potential effect — both positive and negative — of caffeine on a person's health.

Much of this research has focused on the possible impact of caffeine — a stimulant found in beverages like coffee, tea, and energy drinks and foods like chocolate — on heart health.

One of the latest of these studies, recently presented at ACC Asia 2024 in India, reports that people who chronically drink high amounts of caffeine at least five days per week may increase their risk of cardiovascular disease, even if they are otherwise in good health.

High caffeine levels linked to elevated heart rate, blood pressure

Scientists found that chronic intake of 400 mg of caffeine daily showed a significant impact on the autonomic nervous system, leading to increased heart rate and blood

Figure A2. Example screenshot of a source page in the style of a popular media page. Screenshot of a popular-media-like web page titled 'Drinking high amounts of caffeine 5 days a week may increase heart disease risk' displaying the title, author and publication date. Highlighted bullet points extract the key points of the article, followed by snippets of the article.