Int. J. Human–Computer Studies 209 (2026) 103714

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Understanding trust toward human versus AI-generated health information through behavioral and physiological sensing

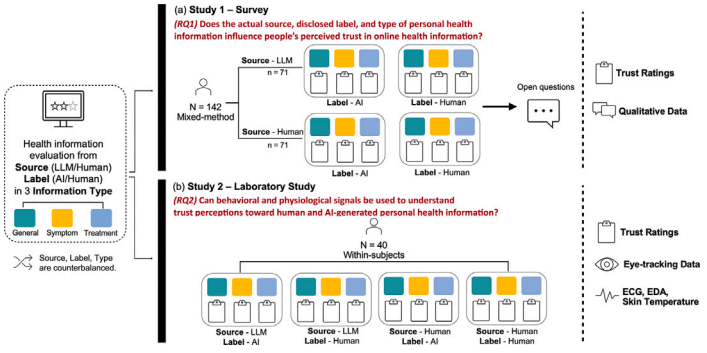
Xin Sun[a, b,*], Rongjun Ma[c], Shu Wei[d, e], Pablo Cesar[b, f], Jos A. Bosch[a], Abdallah El Ali[b, g,**]

[a] University of Amsterdam, Amsterdam, The Netherlands
[b] Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
[c] Aalto University, Espoo, Finland
[d] University of Oxford, Oxford, United Kingdom
[e] Yale School of Medicine, New Haven, CT, United States
[f] Delft University of Technology, Delft, The Netherlands
[g] Utrecht University, Utrecht, The Netherlands

HIGHLIGHTS

- Present two complementary studies: mixed-methods survey (N = 142) and lab study (N = 40) with eye-tracking and ECG/EDA sensing.
- Investigate trust in AI- and human-generated health information, varying source, disclosed label, and information type.
- LLM-based information is trusted more than human information, while human-labeled information is trusted more than AI labels.
- Gaze and physiology act as implicit trust factors: predict trust scores with 73 % and classify source with 65 % accuracy.
- Provide design considerations for transparency labeling and trust calibration in LLM-powered health information interfaces.

GRAPHICAL ABSTRACT



* Corresponding author at: University of Amsterdam, Amsterdam, The Netherlands.
** Corresponding author at: Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands.
Email addresses: xin.von.sun@gmail.com; x.sun2@uva.nl (X. Sun), Abdallah.El.Ali@cwi.nl (A. El Ali).

https://doi.org/10.1016/j.ijhcs.2025.103714
Received 1 April 2025; Received in revised form 30 September 2025; Accepted 10 December 2025
Available online 17 December 2025

ARTICLE INFO

Dataset link: https://drive.google.com/drive/folders/1gat5ZyRFCi5gd_vL3wRGG1KAvbmn_s9x?usp=drive_link Supplementary Study Stimuli (Original data)

Keywords:
Trust
Transparency
Health information systems
Eye tracking
Psychophysiological sensing
Prediction

ABSTRACT

As AI-generated health information proliferates online and becomes increasingly indistinguishable from human-sourced information, it becomes critical to understand how people trust and label such content, especially when the information is inaccurate. We conducted two complementary studies: (1) a mixed-methods survey (N = 142) employing a 2 (source: Human vs. LLM) × 2 (label: Human vs. AI) × 3 (type: General, Symptom, Treatment) design, and (2) a within-subjects lab study (N = 40) incorporating eye-tracking and physiological sensing (ECG, EDA, skin temperature). Participants were presented with health information varying by source-label combinations and asked to rate their trust, while their gaze behavior and physiological signals were recorded. We found that LLM-generated information was trusted more than human-generated content, whereas information labeled as human was trusted more than that labeled as AI. Trust remained consistent across information types. Eye-tracking and physiological responses varied significantly by source and label. Machine learning models trained on these behavioral and physiological features predicted binary self-reported trust levels with 73 % accuracy and information source with 65 % accuracy. Our findings demonstrate that adding transparency labels to online health information modulates trust. Behavioral and physiological features show potential to verify trust perceptions and indicate if additional transparency is needed.

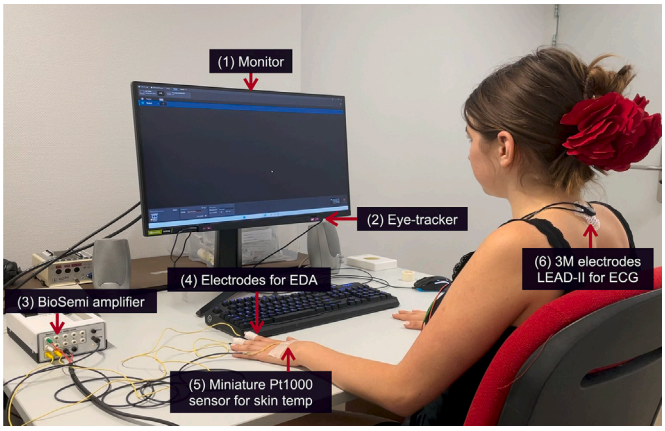


Fig. 1. The hardware setup for presenting the text stimulus and collecting physiological signals, eye movement, and pupil dilation.

1. Introduction

The internet has become a primary source of health information (Cline and Haynes, 2001; Sillence et al., 2007), with 58.5 % of American adults (Wang et al., 2022) (survey in 2022) and 55 % of Europeans (Eurostat, 2022) (survey in 2022) using online sources for health-related searches. This shift has transformed how individuals access and engage with health-related content. Online health resources encompass a broad range of digital tools, including professional medical websites (National Institutes of Health, 2023; MAYO CLINIC, 2023) and AI-driven tools like health chatbots powered by Large Language Models (LLMs) (Wu et al., 2023). These tools have made health information more accessible and convenient than ever, yet they also require users to make critical choices about which sources of the retrieved health information to trust (Liu et al., 2023; Sillence et al., 2005). These trust decisions directly influence health-related choices, many of which carry significant health risks (Wang et al., 2023; Marecos et al., 2024). As a result, understanding how different information sources shape trust perceptions has become increasingly critical (Bates et al., 2006). Some prior studies find that users tend to trust human-generated information more (Broom, 2005; Kerstan et al., 2023; Walker et al., 2024; Reis et al., 2024), while other work suggests that people may prefer algorithmic or AI-generated judgments over human ones (Logg et al., 2019; Shekar et al., 2024). These mixed findings suggest that trust in online information varies by source and context, and remains insufficiently understood, especially in LLM-powered health contexts.

Disclosed labeling of online information signals its source, but can also shape perceptions independently of the actual source, making it an essential dimension of understanding trust. Misleading labels or unclear sourcing may result in misinformation and poor health decisions (Desai et al., 2022; Marecos et al., 2024). Labeling is increasingly mandated by regulations, such as the European AI Act (El Ali et al., 2024). Research shows that disclosed labeling (e.g., with/without indicating AI involvement), can significantly influence trust independently when the information source is identical (Reis et al., 2024). In AI-powered tools, labeling plays a critical role, especially as users increasingly struggle to distinguish between human- and AI-generated content (Rathi et al., 2025). In LLM-powered systems, the actual content source and the disclosed label can diverge, for example, AI-generated content may be labeled as human-authored. While prior research has independently examined the effects of information source (e.g., AI vs. human) (Walker et al., 2024; Johnson et al., 2023) and labeling (Reis et al., 2024; Rae, 2024) on trust, there remains a critical gap in understanding how these two factors interact. Yet, both can significantly influence perceived trust in health information. This gap is especially important in high-stakes contexts like personal health, where trust directly influences individuals' health decision-making and behavioral outcomes (Marecos et al., 2024). Our work addresses this need by manipulating the content source and its disclosed label jointly to investigate their combined effects on people's trust perception in health information, particularly in the era of LLMs.

To understand such joint effects of information sources and disclosed labels on people's perceived trust, we ask: **(RQ1) How do the actual source, disclosed label, and type of personal health information influence people's perceived trust in online health information?** To answer this research question, we employed a mixed-methods approach in Study 1 (see Fig. 2a). Specifically, we conducted an online crowdsourcing survey (N = 142) using a 2 × 2 × 3 factorial design. Source (Human Professional vs. LLM) was treated as a between-subjects variable to minimize potential biases from participants directly comparing human and AI sources. In contrast, Label (Human Professional vs. AI) and health-information Type (General vs. Symptom- vs. Treatment-related) were within-subjects variables to enable a nuanced comparison of trust perceptions across different labeling and information types within the same participant. This mixed design balanced the reduction of cross-condition biases with the sensitivity of within-subject comparisons. Participants rated their perceived trust in the health information they received using standardized self-report scales, which served as our primary trust measure outcome.

Although self-reported measures we adopted for Study 1 are widely used due to their simplicity and directness, research by Chen et al. (2021) and Kohn et al. (2021) argues that self-reported trust measures are subjective, which makes them more vulnerable to biases like social desirability bias and the Initial Elevation phenomenon (Anvari et al., 2023). These biases may compromise the reliability and validity of self-reported trust assessments. With the growing use of sensing technologies

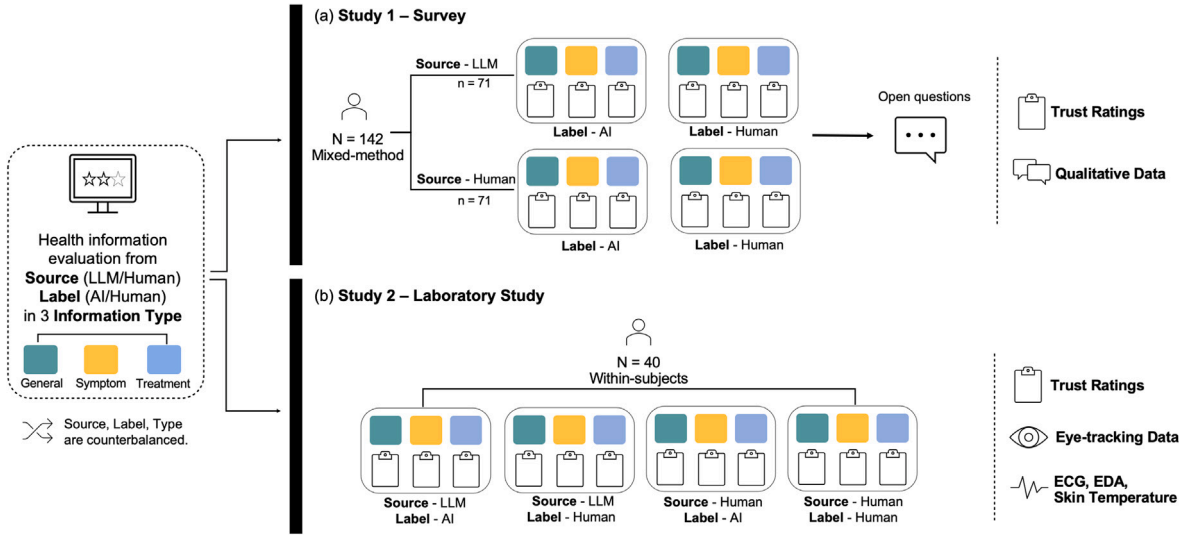


Fig. 2. Visual summary of the studies in this paper. (a) Study 1: Mixed-methods crowdsourcing survey study to measure perceived trust; (b) Study 2: Within-subjects lab study to measure perceived trust, as well as behavioral and physiological responses.

and recent interest in Human—Computer Interaction research to draw on physiological sensing for designing or evaluating interactive systems (Chiossi et al., 2024), several prior studies (Ajenaghughrure et al., 2020; Akash et al., 2018; Lim et al., 2022) argue that behavioral and physiological data can provide a complementary perspective for understanding trust alongside self-reported measures. Behavioral patterns such as eye movements and physiological responses, as assessed by Electrocardiogram (ECG) (Ajenaghughrure et al., 2020) and Electrodermal Activity (EDA) (Babaei et al., 2021), could reveal how individuals process information and make trust-related decisions in health contexts. For example, eye movement patterns, such as fixation duration and saccade behaviors, can indicate cognitive engagement with the information, while physiological responses like heart rate variability (HRV) (Ajenaghughrure et al., 2020; Tiwari et al., 2021; Ahmad and Alzahrani, 2023) and skin conductance levels (SCL) can reveal emotional arousal and stress responses. These implicit measures may further help interpret user trust perceptions (Babaei et al., 2021; Ahmad and Alzahrani, 2023). Thus, exploring these behavioral and physiological indicators can contribute to a more comprehensive understanding of trust formation in digital health contexts (Akash et al., 2018; Ajenaghughrure et al., 2020; Wang, 2018) and further, help develop strategies to enhance the trustworthiness of online health information, especially given the growing use of LLM-powered tools for health advice (Garg et al., 2023; Lee et al., 2023; Biswas, 2023).

Building on RQ1, we adopt behavioral and physiological data as a complementary lens for understanding trust. We ask: **(RQ2) Can behavioral and physiological signals be used to understand trust perceptions toward human- and AI-generated health information?** To address this research question, we conducted a laboratory study (Study 2, N = 40) using a 2 × 2 × 3 fully within-subjects design. We collected eye-tracking data (e.g., gaze patterns, pupil dilation) and physiological signals (e.g., ECG, EDA, and skin temperature) to examine whether these implicit signals vary as manipulated by source and label. Additionally, we explored how these signals relate to participants' self-reported trust perceptions. By allowing each participant to serve as their own control, this design minimized variability due to individual differences and maximized the robustness of condition-specific inferences. Importantly, participants were not informed that labels could be intentionally mismatched with the actual source (i.e., cross-labeled) in both studies. This ensured that participants evaluated the health information and its disclosed label as presented, without being influenced by a heightened awareness of potential labeling errors, thereby allowing us to more accurately assess their trust perceptions on both information itself and its labeling.

Online survey (Study 1) findings showed that the (actual) source of information significantly influenced trust perceptions, with participants displaying higher trust in LLM-generated health information compared with human professionals. Second, the labeling of the source played a crucial role: health information labeled as coming from human professionals led to significantly higher trust than information labeled as from AI, i.e., regardless of the actual source. Third, the type of health question did not significantly affect trust, alone or in interaction with label and source. Together, these observations suggested that perceived trust is not influenced by the nature of the health query, and that the source and labeling of the health information are the main determinants. The laboratory study (Study 2) supported the survey findings, with additional insights: gaze features, such as fixation, saccade, and pupil diameter, varied significantly based on the source and labeling of health information. Moreover, physiological features, such as heart rate variability (HRV, measured as the root mean square of successive differences, RMSSD) and skin temperature, differed when participants engaged with information with different labels. These findings indicated that the source and labeling of health information influence both behavioral and physiological responses. Further prediction tasks were performed based on behavioral and physiological data, yielding 0.35 R^2 for predicting trust scores and 73 % accuracy in classifying binary trust levels (high vs. low). Additionally, we achieved 65 % accuracy in classifying the source of health information. These results underscored the potential of leveraging behavioral and physiological signals as complementary indicators to understand trust perception toward human vs. AI-generated health information.

Our exploratory work offers two primary contributions: **(1)** We provided empirical evidence showing that trust in online health information is influenced both by its actual source and disclosed label. **(2)** We found that trust perceptions in personal health information vary at behavioral and physiological levels, offering complementary insights beyond self-reported trust and helping to identify discrepancies between the explicit (i.e., self-reported) and implicit trust-related responses. To our knowledge, this is one of the few studies that combines physiological (e.g., HRV, skin temperature) and behavioral (e.g., gaze) signals to understand trust in AI-generated health information. Our work highlights the importance of considering AI transparency labels when measuring trust

in health information and the vulnerability of trust abuse due to mislabeling. It further opens the possibility of verifying trust perceptions and inferring if and when to apply transparency labels based on sensed behavioral and physiological data.

2. Related work

2.1. Trust in online health information seeking

Trust is a multifaceted psychological construct essential to both interpersonal and human-technology interactions. Mayer, Davis, and Schoorman's integrative ABI model of trust (Mayer et al., 1995) defines trust as a willingness to be vulnerable to the actions of another party, based on the expectation that the party possesses the performance (ability), intends to do good (benevolence), and adheres to a set of principles that the trustor finds acceptable (integrity). Extending this concept to the digital age, (Lee and See, 2004, p. 51) define trust in technology as: "An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability".

In the context of health, trust is particularly important due to the sensitive nature of health information and its impact on health-related decision-making, which can have dire health consequences should it be incorrect (Wang et al., 2023; Marecos et al., 2024). Trust formation in health contexts is complex and influenced by both intrinsic and extrinsic factors, including individual characteristics such as prior knowledge, health literacy, and external cues, such as source credibility, interface design, as highlighted by Vereschak et al. (2024). For instance, numerous studies (Bates et al., 2006; Liu et al., 2023; Sillence et al., 2004; Singal and Kohli, 2016; Dutta-Bergman, 2003; Lucassen and Schraagen, 2010) have indicated that the credibility of the information source is crucial, the design (Fogg et al., 2001; Wathen and Burkell, 2002; Fogg et al., 2000; Flanagin and Metzger, 2007) and usability (Davis and Davis, 1989) of the health-related tools can significantly affect trust. User prior experience such as familiarity levels (Sillence et al., 2019), and user expectations (Guo, 2022) also influence trust perceptions as well. Moreover, users increasingly expect transparency, ethical AI practices, and data privacy, which further complicate trust calibration (Friedman et al., 1999; Bansal and Warkentin, 2022; di Sciascio et al., 2020; Ul Haque et al., 2023).

To conceptually integrate the literature and these multi-level trust influences, we draw on the MATCH framework (Liao and Sundar, 2022), a model that systematically captures the trustworthiness cues in human-AI communication. Unlike trust models that focus on either the trustee's attributes (e.g., classic ABI model) or interface-level cues (e.g., MAIN model Sundar, 2007), the MATCH framework offers a more integrated account of how trust is formed in AI systems by collectively integrating content quality, interface design, and user heuristics.

Specifically, MATCH organizes trust into three components: (1) Model Attributes. This dimension, drawn directly from the ABI model (Mayer et al., 1995), refers to the perceived ability and competence of the system. In our context, it relates to users' perceptions of the quality and reliability of the information itself. It echoes the prior work revealing that the intrinsic quality of the information itself plays a critical role in shaping trust (Flanagin and Metzger, 2000; Sbaffi and Rowley, 2017; Wathen and Burkell, 2002; Fogg et al., 2000; Metzger and Flanagin, 2013). (2) Afforded Cues. These are extrinsic signals such as formatting, interface design or interaction patterns. Prior work shows that even subtle interface features like transparency labels (Kizilcec, 2016; Yin et al., 2024) or content layout (Johnson et al., 2015; Fogg et al., 2001; Wathen and Burkell, 2002; Fogg et al., 2000; Flanagin and Metzger, 2007) can significantly influence trust judgments. (3) Trust Heuristics. MATCH uniquely accounts for the mental shortcuts users apply under uncertainty (e.g., quickly assessing that information labeled as "human-generated" is more trustworthy, or that "AI-generated" content is less reliable). It is often shaped by prior experience, health literacy, or cognitive and affective responses collectively (Lee and See, 2004). In our work, we further interpret these heuristic processes through behavioral and physiological signals, such as gaze patterns that may reflect users' implicit trust-related responses.

Grounded in the MATCH model, this work examines how both source information taps into model attributes, how labeling functions as an afforded trustworthiness cue, and how behavioral and physiological signals reflect user cognitive heuristic and affective processing of health information toward the trust-related judgments. While existing studies have investigated trust in AI- vs. human-generated content, few have systematically decoupled the actual source from the disclosed label to assess their independent and combined effects. Our exploration builds on the MATCH model and extends prior work by isolating and manipulating both information source and labeling disclosure, allowing us to explore how these cues interact and shape trust formation in online health information seeking contexts.

2.2. Source and label transparency in the age of LLMs

The internet has become a vital resource for health information (Cline and Haynes, 2001), with websites like WebMD (WebMD, n.d.) and Mayo Clinic (MAYO CLINIC, 2023) providing expert-curated content. The rise of LLMs like ChatGPT (OpenAI, n.d.) has revolutionized access to online health information by offering conversational interactions to health queries (Dalton et al., 2022). Trust in these LLM-powered tools is influenced by various factors (Rheu et al., 2020), including the perceived credibility of their responses, clarity of information, transparency about how the information is generated (El Ali et al., 2024), and users' familiarity and experiences using such AI technologies (Bickmore et al., 2005). Among these, information source (e.g., human-authored vs. AI-generated) plays a critical role in shaping trust. Research (Hesse et al., 2005; Bates et al., 2006; Lucassen and Schraagen, 2010) has shown that trust is significantly affected by the perceived credibility of information source. While LLMs have been effective in providing health information (Bickmore et al., 2005; Carlbring et al., 2023), concerns remain about their credibility and reliability. Although human professionals are traditionally viewed as authoritative and trustworthy due to their expertise (Kerstan et al., 2023; Broom, 2005), studies like Logg et al. (2019) showed that users may trust AI for specific tasks, and Shekar et al. (2024) indicated that people overtrust AI-generated medical responses. However, other research (Reis et al., 2024; Kerstan et al., 2023) highlighted people's preferences for human-generated health advice, suggesting that trust varies based on context. Additionally, Montag et al. (2023) found that trust in humans and AI may not be directly associated, suggesting people have distinct trust mechanisms for each. These varied trust levels underscore the complexity of trust formation toward information from human and AI sources.

Labeling of information sources plays an additional key factor in shaping trust perceptions in the era of LLMs. Jakesch et al. (2019) demonstrated that users perceive content as less trustworthy when it is labeled as AI-generated, even when the content quality is identical, which indicates that labeling influences how users perceive trustworthiness. Similarly, Reis et al. (2024) found that perceived AI involvement significantly impacts trust in digital medical advice, as participants in their study were less willing to follow health advice when they believed it was generated by AI rather than a human expert. Studies by Walker et al. (2024) and Kerstan et al. (2023) have also shown that people tend to trust advice more when it comes from human professionals rather than from LLMs, especially when the source is explicitly stated. Yin et al. (2024) found that while AI can create a sense of being heard, labeling content as AI-generated can reduce its perceived impact. These findings underscore how labels can significantly impact trust, even when AI performs tasks effectively. Furthermore, Scharowski et al. (2023) explores the potential for AI certification labels (e.g., "Digital Trust Label" by the 2023 Swiss Digital Initiative), and finds that such labels can mitigate data-related concerns surfaced by end-users such as data protection and privacy, however this came at the cost of other concerns such as model performance, which poses its own challenges. Nevertheless, these works

highlight that transparent communication about how AI systems operate and the data sources they use can further enhance or maintain trust among users (Kizilcec, 2016; Logg et al., 2019).

As AI becomes more integral to health contexts, this work specifically explores the influence of source and labeling as critical extrinsic cues on trust in health information, offering insights for designing trustworthy LLM-powered health systems. Framed through the MATCH model (Liao and Sundar, 2022), these effects reflect how users interpret afforded cues (e.g., disclosure labels) and model attributes (e.g., inferred competence or benevolence of a human vs. AI source) when assessing trust. Labels operate as an interface-level factor that invokes trust heuristics, particularly under conditions of uncertainty. These trust dynamics highlight the importance of carefully designing how source and authorship are communicated in AI-powered health systems.

2.3. Behavioral and physiological signals for understanding trust perception

Traditional research on trust perception has heavily relied on self-reported assessments; however, many studies (Chen et al., 2021; Kohn et al., 2021) suggest behavioral and physiological signals may add a relevant layer of information. Integrating these implicit measures helps offer a complementary understanding of trust in human and LLM-generated health information. For example, research by Holmqvist et al. (2011) shows that eye movement metrics like fixation, saccade, and pupil dilation provide insights into cognitive load and attention allocation during information processing. While these physiological indicators do not directly measure trust, they may reflect how users cognitively and affectively engage with content they perceive as more important, credible, or challenging. For instance, increased pupil dilation, linked to higher cognitive load (Ahmad et al., 2020) and emotional arousal, may suggest deeper cognitive processing, which may co-occur when individuals are evaluating information for trustworthiness or making health-related decisions. Although the relationship between trust, cognitive, and affective responses is complex, monitoring these signals may help identify moments of increased scrutiny or hesitation, offering indirect cues about trust-related states. As an example of such research, Ji et al. (2024, 2023) demonstrated that physiological signals, such as electrodermal activity, blood volume pulse, and gaze, vary meaningfully across different information processing activities (e.g., reading, speaking, listening) during information-seeking tasks. Moreover, prior work has used behavioral data to explore how people engage with online news content, particularly in the context of misinformation. For instance, Abdrabou et al. (2023) found that gaze and mouse movement patterns could help distinguish between user exposure to real versus fake news, achieving moderate accuracy in identifying subconscious engagement with misinformation. Similarly, Sü et al. (2021) showed that eye-tracking data reflected differences in how users read and process true versus false news articles, suggesting that such behavioral signals can offer a comprehensive understanding of how people implicitly respond to varying degrees of information credibility. Studies (Lu and Sarter, 2019; Wang, 2018; Kohn et al., 2021; Holmqvist et al., 2011; Sevcenko et al., 2022; Ayres et al., 2021) demonstrate that distinct gaze patterns are linked to trust levels, with higher fixation counts and longer duration typically indicating focused attention, greater cognitive engagement, and trust in the information. Saccades, characterized by the frequency and length of eye movements between fixations, often signal information verification processes (Lu and Sarter, 2019; Wang and Stern, 2001; Wang, 2019). These findings suggest that these multimodal implicit signals can be sensitive indicators of user cognitive effort and engagement, offering potential to infer user states such as trust or uncertainty in information processing contexts.

Physiological features such as ECG (Ajenaghughrure et al., 2020), EDA (Babaei et al., 2021), and skin temperature (Ahmad and Alzahrani, 2023) can be useful for understanding implicit responses related to trust. Heart Rate Variability (HRV), derived from ECG, reflects the level of stress and cognitive dissonance, with higher HRV indicating lower physiological arousal which is associated with relaxation, comfort, and higher trust levels (Tiwari et al., 2021; Kim et al., 2018; Thielmann et al., 2022). EDA measures, including Skin Conductance Level (SCL) and Skin Conductance Response (SCR) are similarly tied to emotional arousal, where lower conductance is used to infer greater comfort and trust (Babaei et al., 2021; Wang, 2018; Ahmad and Alzahrani, 2023). Similarly, changes in skin temperature are thought to reflect engagement levels, with higher temperature suggesting increased cognitive engagement with information (Ahmad and Alzahrani, 2023). As investigated by prior work (Lee and See, 2004), trust perception, a complex, subjective cognitive and affective process, can be assessed using models by analyzing physiological (e.g., ECG and EDA (Ajenaghughrure et al., 2021), EEG Akash et al., 2018) and behavioral (e.g., gaze patterns Lim et al., 2022; Parikh, 2018) indicators. These models help reduce subjective bias and can provide real-time insights into trust responses, not least of which is an additional verification means alongside self-reports.

These behavioral and physiological signals provide insights into users' implicit responses, capturing attention, emotional arousal, and cognitive engagement that may not surface in self-reports. In our work, we explore whether implicit signals vary meaningfully across conditions of information source and labeling. We interpret these signals cautiously as indirect indicators that may correlate with trust. Within the MATCH model (Liao and Sundar, 2022), these sensing signals map onto the trust heuristics component, reflecting how users internally process trustworthiness cues that influence trust. Unlike explicit cues like source attributions, sensing signals help uncover how users process those cues implicitly, for example, when trust is assigned reflexively versus analytically. By revealing how trust is formed or challenged beneath explicit awareness, these signals complement extrinsic cues and help build a more comprehensive picture of trust in LLM-powered health contexts.

2.4. Synthesis and research gap

As summarized in Table 1, prior research has largely treated source and label in isolation, and separately examined how information sources and disclosed labels influence trust in online information, but findings are mixed. Some studies report that users trust human-generated content more due to perceived expertise and accountability (Kerstan et al., 2023; Walker et al., 2024), while others show higher trust in AI-generated information, citing perceived consistency or objectivity (Logg et al., 2019; Shekar et al., 2024). Research on labeling further shows that disclosing AI involvement often reduces trust even when content is identical (Reis et al., 2024; Jakesch et al., 2019; Yin et al., 2024). However, few studies have systematically disentangled the effects of source and label together, or explored whether these effects vary across different information types in health contexts (e.g., general, symptoms, treatment).

Moreover, prior research relies heavily on self-reported trust, which may not capture users' implicit cognitive and emotional responses involved in trust judgments. Behavioral and physiological signals offer promising but underexplored means of revealing how users attend to, process, and evaluate health information beyond what they report, which can offer complementary insights into how trust is formed beyond self-reports.

This leaves critical gaps (summarized in Table 1) in understanding how information source, labeling, and content type jointly influence both users' explicit trust (self-reports) and implicit responses (behavioral and physiological) in the context of LLM-generated health information. To address this, our work draws on the MATCH framework (Liao and Sundar, 2022), which integrates: Model Attributes (i.e., information source), Afforded Cues (i.e., disclosed label and information types), and Trust Heuristics (cognitive or emotional responses implicitly reflected in sensing signals). This integrated approach allows us to investigate not only how trust varies across source, label, and information type, but also whether behavioral and physiological signals reflect trust-related judgments in implicit but meaningful ways when users engage with AI- and human-generated health information in LLM-powered contexts.

Table 1
Comparative synthesis of prior studies on source and label effects in trust perception and the research gaps we address in this work.

Manipulation	Source	Labeling	Source + Labeling
Self-reports	***Higher trust in AI than humans:*** Logg et al. (2019) (General context); Shekar et al. (2024) (Health context); ***Higher trust in humans than AI:*** Walker et al. (2024) (Binary decision-making); Kerstan et al. (2023); Hesse et al. (2005) (Health context);	***Labels increase trust:*** Scharowski et al. (2023) (General context); ***AI-Human mixed labels decrease trust:*** Jakesch et al. (2019) (Marketing context); ***AI labels decrease trust:*** Yin et al. (2024); Rae (2024) (General context); Reis et al. (2024) (Health context);	**This work (Study 1):** Source + Label joint effects
Self-reports + Sensing	Trust differs by sources *(Gaze Data in Fake News)* Trust toward human and AI are not associated *(EEG Data)*		**This work (Study 2):** Gaze + Physio in health context

3. Study 1: online survey

3.1. Study methods

3.1.1. Design

We conducted an online survey using a mixed 2 (IV1 - Actual Source: Human professionals vs. LLM) × 2 (IV2 - Disclosed Label: Human professionals vs. Artificial Intelligence) × 3 (IV3 - Information Type: General vs. Symptom vs. Treatment) factorial design to explore people's perceived trust in online health information. The source of health information (IV1) was set as a between-subjects variable to explore whether people have different trust perceptions based on the source (human professionals vs. LLM), which might inherently present information in distinct styles. A within-subjects design for the source could introduce biases in perceived quality and trustworthiness due to these stylistic differences. Additionally, using a between-subjects design for the source helps isolate the effect of labeling (IV2), making the findings clearer and more robust. Conversely, for the label of the source (IV2) and the type of health information (IV3), we opted for a within-subjects design to allow direct comparisons of trust perception across different labels and types while keeping the source uniform for each participant. This approach reduces individual variability, ensuring a clearer separation of source effects on trust variances while enabling robust analysis of influences from labeling and types of health information. Therefore, during the completion of the survey, each participant read the information either generated by human professionals or LLMs, and each of them experienced six distinct conditions.

3.1.2. Health information

Sets of health information (question and answer pairs) from human professionals were selected from an open-sourced dataset (Ben Abacha and Demner-Fushman, 2019) due to its diverse range of health questions, authored by certified professionals. This ensures the reliability and authenticity of the information used in this work. To produce comparable and consistent LLM-generated information, we used the Generative Pre-trained Transformer 4 (GPT-4) model (OpenAI, 2024) (version: "gpt-4–0125-preview" through the official API) and prompted it with selected health questions and accompanying instructions (e.g., "Health question: [question]. Please give an answer to the above question within [wordcount] words?") to generate answers of similar length to those from human professionals. To ensure consistency and mitigate potential misinformation, all LLM-generated responses were independently reviewed by two researchers using the corresponding human-authored answers as references. The review criteria were consistency in length, format, topic relevance, and absence of harmful content. Only responses with full agreement were included, following established HCI practices (McDonald et al., 2019). The health information falls into three categories, reflected in both the clinical process and the dataset's validated taxonomy (Ben Abacha and Demner-Fushman, 2019): **General information**: provides answers to general health topics (e.g. "Do you have information about weight control?"); **Symptoms-related information**: focuses on symptoms and potential diagnoses (e.g. "What are the symptoms of burns?"); **Treatment-related information**: provides treatment options for specific conditions (e.g. "What to do for burns?"). This categorization aligns with clinical practice, which commonly follows a three-stage diagnostic process (Bridley et al., 2013; Balogh et al., 2015): assessment (general inquiry), diagnosis (symptom evaluation), and treatment planning (intervention). These types capture a progression from low- to high-stakes information, allowing us to explore whether trust perceptions vary by the nature of health content.

Twenty-five questions were selected from each category resulting in a question set with 75 questions in total, ensuring a comprehensive representation of individual health questions. The complete list of health information used in the study is included as **Supplementary Material**.

3.1.3. Measures

Demographics and prior experience. In the pre-survey, we collected participants' demographic information (age, gender, education, occupation) and their experience in online health information seeking, using two questions: "How often do you search for health information online?" rated on a 5-point Likert scale from Never to Daily; and "How long have you been using online sources for health information searching?" with options ranging from Less than 1 year to More than 10 years.

Propensity of trust in technology (PPT) (Jessup et al., 2019) was used to assess inherent trust in technology before participants read the health information. It consists of 6 items examining people's general trust in technology (e.g. "I think it's a good idea to rely on technology for help"). All items were scored on a 5-point Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree) (Cronbach's $\alpha = 0.71$).

eHealth and AI literacy. As part of the pre-survey, we also measured participants' literacy on eHealth and AI separately using two adapted questionnaires from eHEALS: The eHealth Literacy Scale (Norman and Skinner, 2006) and MAILS - Meta AI Literacy Scale (Carolus et al., 2023). All the items were scored from 1 (Strongly Disagree) to 5 (Strongly Agree). The adapted measure for eHealth literacy has eight items with an example being "I know where to find helpful health resources on the Internet" (Cronbach's $\alpha = 0.88$), and the adapted measure for AI literacy has ten items with an example item being "I can distinguish if I interact with an AI or a real human" (Cronbach's $\alpha = 0.76$).

Trust of online health information (Johnson et al., 2015; Rowley et al., 2015) (**Trust Score**) During the formal study, participants completed the trust of online health information questionnaire to rate their trust levels after reading each set of health information. It consists of 13 items (e.g. "The information appears to be objective."), each rated on a 5-point Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree) (Cronbach's $\alpha = 0.92$). We aggregated and calculated the average value of all 13 items to obtain our perceived **Trust Score**. We use this score for further analysis throughout our work.

Post-survey: three open-ended questions At the end of the survey, participants were asked to reflect on their trust perceptions through three open-ended questions. These questions explored their views on (a) general trust in LLM-generated information versus information from

Table 2
Characteristics of participants in the online survey.

Demographic	Categories	Numbers of Participants (%)
Gender		(N = 142)
	Female	83 (58.5 %)
	Male	58 (40.8 %)
	Non-binary	1 (0.7 %)
Age	18–24	91 (64.1 %)
	25–34	38 (26.8 %)
	35–44	9 (6.3 %)
	45–54	2 (1.4 %)
	65 +	2 (1.4 %)
Education	High school degree or equivalent	24 (16.9 %)
	Bachelor's degree	67 (47.2 %)
	Master's degree	49 (34.5 %)
	Doctorate or higher	2 (1.4 %)
Professional Domain	Health and Medical Science	17 (12.0 %)
	Science, Technology, Engineering, and Mathematics (STEM)	35 (24.6 %)
	Business, Economics, and Law	35 (24.6 %)
	Arts, Culture and Entertainment	19 (13.4 %)
	Government and Public Sector	3 (2.1 %)
	Education	3 (2.1 %)
	Other	30 (21.1 %)
Frequency of online health information seeking	Rarely	27 (19.0 %)
	Sometimes	77 (54.2 %)
	Often	31 (21.8 %)
	Always	7 (4.9 %)
Duration of online health information seeking	Less than 1 year	4 (2.8 %)
	1–3 years	24 (16.9 %)
	3–5 years	51 (35.9 %)
	5–10 years	45 (31.7 %)
	More than 10 years	18 (12.7 %)

human professionals, (b) how they assess the credibility of online information, and (c) how the labeling of the health information source influences their perceived trust.

3.1.4. Participants

Participants were recruited through the online crowd-sourcing platforms Prolific (Prolific, 2014) and institute recruitment systems. Our inclusion criteria included individuals over the age of 18 who are fluent in English, and they must have passed the attention check. A power analysis conducted with G*Power 3.1 (Faul et al., 2007) for a mixed-factor ANOVA design indicated that a minimum of 76 participants would be required to detect a small effect size (f = 0.15), with an alpha level of 0.05 and a power of 95 %.

Table 2 shows a summary of participants' demographics. 142 participants (N = 142) were recruited (F = 83, M = 58, NB = 1), with 90.9 % falling in the 18–34 age bracket. Regarding educational backgrounds, 47.2 % had undergraduate degrees and 35.9 % held postgraduate qualifications. As for online health information-seeking experience, 26.7 % frequently used online sources, 54.2 % occasionally searched online, and 19.0 % rarely used online resources.

3.1.5. Procedure

The study design and procedure are outlined in Fig. 2(a). Participants were first provided with detailed information about the study and gave informed consent in line with institutional guidelines. They provided demographic information and their experiences with online health information seeking. A total of 75 health questions were used in the online survey, divided evenly into three categories: general health, symptom-related, and treatment-related (25 each) (Section 3.1.2 "Health information"). For each participant, six Q&A pairs were shown: two randomly selected from each category. The survey study used a between-subjects design for the source of the information (AI- vs. human-generated) and a within-subjects design for the label (AI- vs. human-labeled). Both source and label orderings were counterbalanced based on a Latin square approach, ensuring that all condition combinations were evenly distributed across participants to mitigate order effects. An illustrative example of the reading task interface during the survey is shown in Fig. 3. After reviewing each Q&A pair, participants rated their perceived trust in the information. At the end, participants completed a post-survey comprising three open-ended questions about their perceptions of the information source and its labeling.

Participation was voluntary and participants were monetarily compensated for a 30-min session. To ensure we avoided bots in our responses, we included an additional attention check where respondents needed to select a specific response to one question. Our study received approval from our institute's ethics and data protection committee.

3.1.6. Data analysis

We conducted quantitative analyses to examine how the types of health questions, information sources, and labeling of sources influence trust perception in online health information. Initially, we confirmed the data's suitability for parametric tests by performing the Shapiro-Wilk test (SHAPIRO and WILK, 1965) for normality and Bartlett's test (Arsham and Lovric, 2011) for homogeneity of variance; neither assumption was violated. Next, we performed a mixed model, i.e., three-way mixed ANOVA (Kherad-Pajouh et al., 2015) to investigate differences in trust perceptions based on information sources, disclosed labels, and types of information. Since only one ANOVA was conducted, no correction for multiple tests was applied. Following, post-hoc pairwise comparisons were conducted using t-tests with False Discovery Rate (FDR) correction (Haynes, 2013a) to examine differences in trust between each pair of label and source combinations. To explore the relationships across variables, we also conducted Pearson correlation analyses (Freedman et al., 2007) on two subsets of the data: one with human-sourced information and the other with LLM-sourced information (between-subjects independent variable). Before analysis, we confirmed that the assumptions for Pearson correlation: normality, linearity, and absence of extreme outliers, were satisfied in the aggregated data. Bonferroni correction (Haynes, 2013b) was applied to account for multiple comparisons in both correlation analyses.

We conducted an inductive content analysis (Elo and Kyngäs, 2008) on the responses to three open-ended questions, focusing on identifying

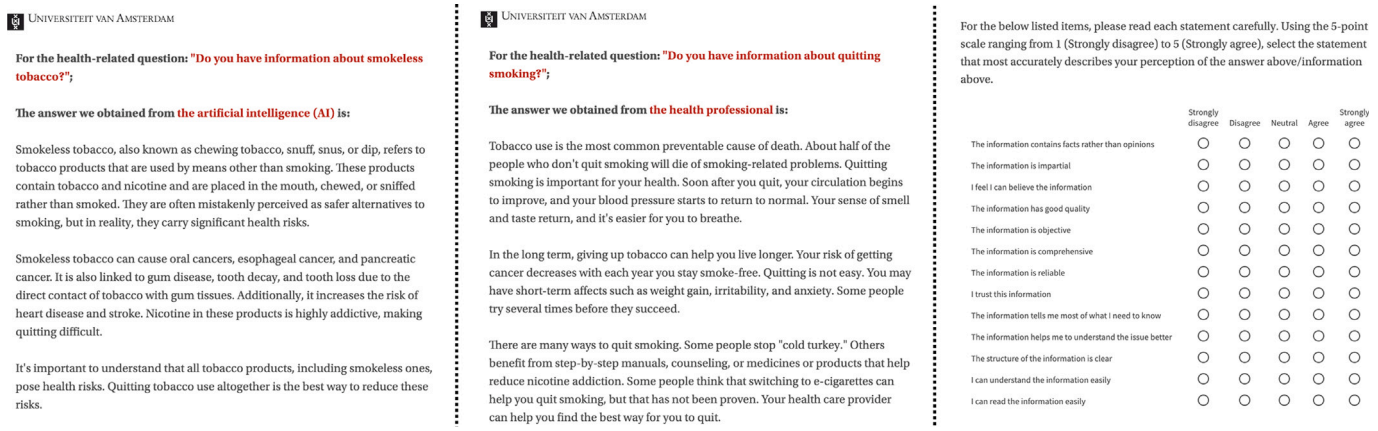


Fig. 3. Example reading task from the survey, showing a Q&A pair with its assigned disclosed source label. Each participant read six Q&A pairs: three labeled as from "AI" **(left)** and three labeled as from "Human Professionals" **(middle)**. After each reading, participants rated their trust using the scale shown on the **right**.

Table 3
Descriptive statistics of the online survey.

	Measures	Mean	SD
Pre-survey	Propensity of trust in AI technology (PPT)	3.85 / 5	.72
	eHealth literacy	3.62 / 5	.87
	AI literacy	3.81 / 5	.92
	Conditions	**Mean**	**SD**
Trust score	Source (Human) & Label (Human)	4.01 / 5	.45
	Source (Human) & Label (AI)	3.76 / 5	.49
	Source (LLM) & Label (Human)	4.07 / 5	.47
	Source (LLM) & Label (AI)	3.87 / 5	.44
	Source (Human), regardless of Label	3.89 / 5	.84
	Source (LLM), regardless of Label	3.97 / 5	.81
	Label (Human), regardless of Source	4.04 / 5	.46
	Label (AI), regardless of Source	3.82 / 5	.47

underlying themes that explain trust rather than counting frequencies. In the first stage, the first two authors created an initial set of codes using the qualitative analysis software ATLAS.ti (ATLAS.Ti, 2024). This initial codebook examined respondents' varying perceptions of trust in AI and human professionals, their reasons for trusting or distrusting, and how they typically evaluate the credibility and trustworthiness of information. Following this, both coders independently open-coded the responses, remaining open to new observations and emerging codes. Similar codes were merged, unclear ones were refined, and earlier responses were re-coded as needed. As the analysis progressed, recurring factors emerged across different questions, allowing us to develop common themes that spanned all three sets of responses.

3.2. Quantitative findings

3.2.1. Descriptive statistics

As shown in Table 3, participants demonstrated a positive propensity to trust in technology, with an average score of 3.85 (SD = .72), indicating a positive attitude toward technology. The average eHealth literacy score was 3.62 (SD = .87), indicating that participants are relatively capable of using online health resources. AI literacy was also high, with an average score of 3.81 (SD = .92), reflecting a favorable understanding of AI technology.

In terms of trust perception, the trust scores (based on the aggregate Trust Score described in Section 3.1.3) varied depending on the source and label of the information. For information both sourced from and labeled as human, the average trust score was 4.01 (SD = 45). When the information was sourced from humans but labeled as AI, the trust score decreased significantly to 3.76 (SD = .64). In contrast, information sourced from LLM but labeled as human received the highest trust score of 4.07 (SD = .47), while information sourced from AI and labeled as LLM had a trust score of 3.87 (SD = .44). These findings highlight the ways in which both the source and labeling of information can impact trust perceptions, with a clear indication that labeling of the sources plays a role in shaping trust, potentially even more than the actual source of the information.

Table 4
Results from the three-way mixed ANOVA analysis on the trust score without data correction. (**p <.01, *p <.05).

Outcomes	Conditions	Statistics	p-value	Effect size	Sig
Trust score	Source (Human vs. LLM)	2.27	.024	.14 (medium)	*
	Label (Human vs. AI)	−6.50	.000	–.39 (medium)	**
	Type of health information	0.67	.505	.05 (small)	

Our mixed model analysis compared differences in trust levels among the source, label, and health information types. Findings are shown in Table 4 and Fig. 4, and together highlight how people perceive and trust health information manipulated by sources and labels.

3.2.2. Participants gave higher trust to health information sourced from LLM than from human professionals

The impact of the information source (human professionals vs. LLM) on trust in health information was analyzed by a three-way mixed ANOVA. The results showed significant differences in trust levels between sources: statistics = 2.27, p = .024, effect size = .14. This suggests that information sources significantly influence overall trust in health information. Specifically, participants reported trusting information from LLM more than human professionals, with an average trust score for LLM-sourced information of 3.97 (SD = .81), compared to 3.89 (SD = .84) for information from human professionals. Although perceived trust does not imply factual accuracy, our findings reflect a growing acceptance of AI-generated health content and shifting attitudes toward it relative to advice from human professionals.

3.2.3. Participants gave higher trust ratings to health information labeled as from human professionals compared to labeled as from AI

Except for the factor of "source", the labeling of information sources influenced trust perception significantly. Participants perceived significantly lower trust in health information labeled as from AI compared to that labeled as from human professionals, as indicated by a mixed model ANOVA (statistics = −6.50, p<.001, effect size = -.39), with an average trust score for information labeled as from human professionals of 4.04 (SD = .46) and 3.82 (SD = .47) for information labeled as from AI. We

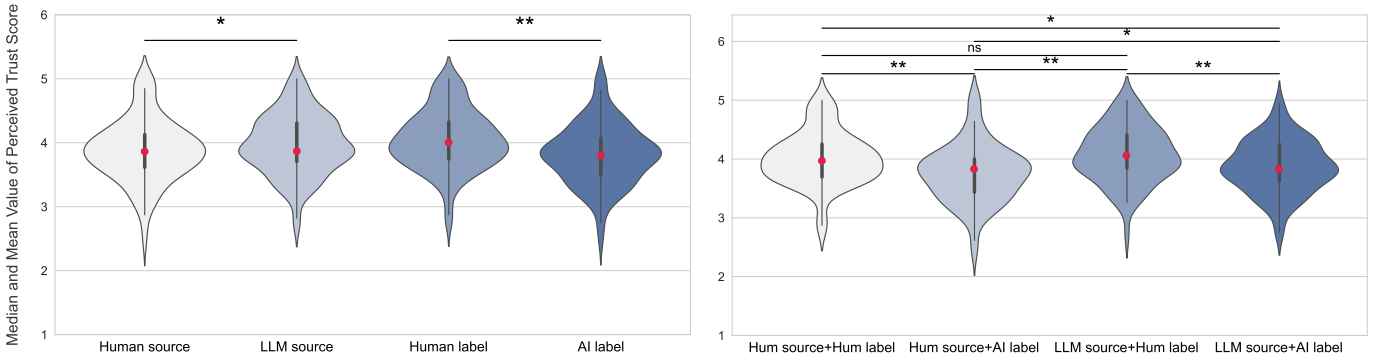


Fig. 4. Left: Perceived trust score in information by sources regardless of labels, and by labels regardless of source from the three-way mixed ANOVA without correction. **Right:** Post hoc pairwise comparisons on perceived trust score based on different source and label conditions using *t*-test with False Discovery Rate (FDR) correction. Each plot shows the score density (width), with the red dot indicating the mean, the black line as the median, and thick bars representing the interquartile range (IQR). Horizontal lines indicate significance (**p <.01, *p <.05, "ns": no significance).

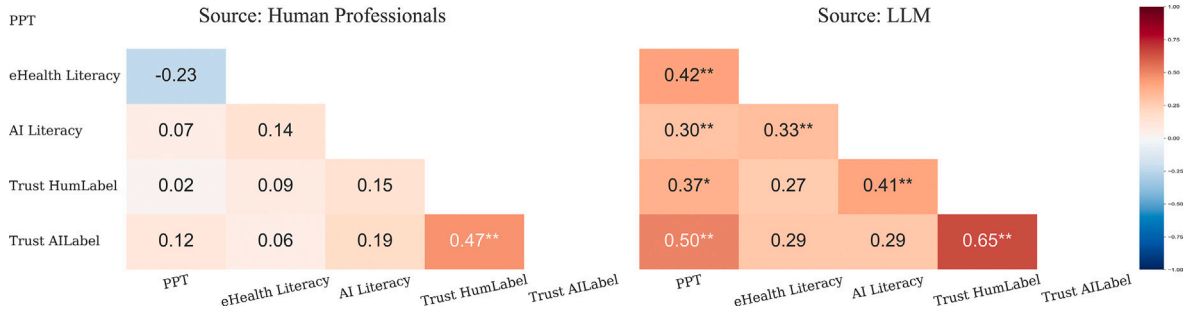


Fig. 5. Pearson correlation with Bonferroni correction among the key variables in the online survey. (**p <.01, *p <.05). Note: "HumLabel": information with human label regardless of the actual source. "AILabel": information with AI label regardless of the actual source.

also observed no significant difference in trust between human-labeled information from human sources (M = 4.01, SD = .45) and LLM sources (M = 4.07, SD = .47). These results suggest that while LLM-generated information is generally trusted, the perceived trust still leans in favor of human-associated information when directly compared.

3.2.4. The type of health information does not affect participants' trust perception in information

Additionally, we explored how trust varied across different categories of health information. There was no significant effect found (statistics = 0.67, p = .505, effect size = .05). This suggests that the type of health question does not influence people's trust levels in health information. The interaction effect between the label of the information source and the category of information was not significant as well (statistics = -.51, p = .613, effect size = -.15). This implies that the influence of labeling on trust does not vary across different types of health information.

3.2.5. Correlation analysis

Given that the mixed ANOVA indicated no significant effect of the type of health information on the trust perceptions, the repeated measures were averaged into a single observation for each participant. This simplification allowed us to conduct a Pearson correlation analysis (Freedman et al., 2007) to examine the general relationships between key variables in the online survey. The results, illustrated in Fig. 5, revealed distinct patterns of trust in health information from different sources. For information sourced from human professionals, trust in human-labeled information showed a moderate positive correlation with trust in AI-labeled information ($r(142) = 0.47, p < 0.01$). However, other relationships, such as those involving eHealth literacy and AI literacy, exhibited weak or negligible correlations. In contrast, for information sourced from LLMs, we observed stronger correlations across multiple variables. Trust in human-labeled information showed a strong positive correlation with trust in AI-labeled information (r(142) = 0.65,p<0.01), AI literacy (r(142) = 0.41,p<0.01), and the propensity of trust in AI ($r(142) = 0.37, p < 0.05$). Additionally, the propensity of trust in AI correlated with trust in AI-labeled information (r(142) = 0.50,p<0.01), eHealth literacy ($r(142) = 0.42, p < 0.01$), and AI literacy ($r(142) = 0.30, p < 0.01$). AI literacy positively correlated with eHealth literacy ($r(142) = 0.33, p < 0.01$). These results highlighted a consistent influence of labeling on participants' trust across different sources.

3.3. Qualitative findings

We received a total of 426 free-text responses (142 for each question). In this section, we present our findings with four themes. We found that participants' trust in AI versus humans is shaped by their inherent trust predispositions (Section 3.3.1) and their perceived source of knowledge for each agent (Section 3.3.2). Additionally, participants value human consciousness as a factor contributing to greater trust (Section 3.3.3), and the presentation of information also influences their trust (Section 3.3.4).

3.3.1. Predisposition toward AI and humans influences trust

Survey respondents demonstrated a predisposition to trust either AI or humans, independent of the content or source of the information. However, there were individual differences in this inclination. Some respondents were optimistic about AI technology, regularly using and trusting AI in their daily lives. They perceived no difference in reliability between AI and human professionals, and some even trusted AI more. Conversely, some respondents expressed significant reservations about AI, doubting its readiness to address serious topics, especially in sensitive fields like healthcare. One respondent noted, *"I don't trust AI, and the quick push in its advancements is dangerous; at the very least, it should be limited in specific fields such as health."* Privacy concerns and

the risks of AI-driven health advice reinforced such skepticism, leading to more critical evaluation of AI recommendations. This underlying predisposition toward AI or human professionals also shaped respondents' views on labeling. Some participants expressed a preference for human-labeled content, with one stating, *"AI label makes me trust it less and view the information more critically than if it came from a human professional."* However, not all respondents allowed their predispositions to dictate their trust. Others placed less emphasis on labels, focusing instead on verifying information from multiple perspectives rather than relying solely on the source. As one respondent explained, *"The label doesn't affect how I interact with it, and my trust wouldn't be based solely on the label."*

3.3.2. Perceived source of knowledge influences trust

Survey respondents' trust in AI or human professionals was shaped by their perceptions of where each derives its knowledge. One respondent explained, *"I would trust a human professional more, since he has learned factual information in school. An AI has learned from multiple sources online, not only factual ones, so that is why I would trust it a bit less."*. In contrast, some respondents believed that AI can learn from *"more databases and the most important points that all research brought up"*, potentially making it more knowledgeable than a single human expert. These differing views on the origins of human and AI's knowledge contributed to varying levels of trust. Some respondents took a more balanced stance, recognizing that both AI and human professionals are susceptible to biases and errors. As one respondent commented, *"While information from a human professional may need correction due to incomplete knowledge, information from AI might contain errors due to gaps in its training data."* Consequently, many respondents shared that they would evaluate both sources of information with equal care, relying on their own experiences to evaluate the content's credibility. Additionally, some respondents expressed a preference for combining information sources, such as cross-checking information or using AI as a complementary tool to support human decision-making.

3.3.3. The human touch builds greater trust than AI

Survey respondents highlighted that, due to the absence of consciousness and empathy in AI, they trusted human professionals more, particularly in healthcare contexts. Many respondents emphasized that AI lacks the ability to evaluate information with awareness. As one respondent commented, *"Unlike human, AI doesn't know the difference between good or bad quality."* In contrast, many respondents emphasized that human professionals have *"years of medical education and experience with real-life cases"* to inform their decisions, something that AI cannot replace despite its access to vast information. This absence of consciousness made respondents very skeptical about AI's capability to offer reliable health advice. The issue extended beyond decision-making to interpersonal interactions. Respondents valued the sense of responsibility and ethical obligation that human professionals carry, with one noting, *"I trust the information from the human professional more because they are human and have moral and professional obligations about not giving misinformation."* Additionally, human-to-human interaction offered a sense of personalized care, making respondents feel their symptoms are better understood. In contrast, AI lacked this human touch, and its absence of empathy and accountability led respondents to trust it less.

3.3.4. Presentation of information influences trust

Information presentation was highlighted as an advantage of AI, which increased respondents' trust. They mentioned that when evaluating health information, factors such as the design of the user interface, the length of the information, the visible publication date, and the clarity of language were important. Compared to human professionals, AI was often perceived as providing simpler, more structured, and user-friendly information. Respondents appreciated that AI's answers were clearly explained and easy to understand. Additionally, the objective tone of AI responses further boosted respondents' trust. These elements collectively enhanced AI's explainability. As one respondent noted, *"When I receive information from a human professional, I expect it to contain more academic language, which is harder to understand and less explanatory. Information from AI, however, uses simpler words and is easier to understand."*

4. Study 2: laboratory study

Study 1 demonstrated that the factors of actual source and disclosed label both affect people's perceived trust (self-reported) in health information. To further understand the process and user behaviors involved in forming trust perceptions, we conducted an in-person experiment. This study explored how health information from different sources and labels affects people's behavioral and physiological states.

4.1. Study methods

4.1.1. Design

Similarly to the online survey study, we utilized a within-subjects 2 (IV1 - Information Source: Human Professional vs. LLM) × 2 (IV2 - Disclosed Label: Human Professional vs. Artificial Intelligence) × 3 (IV3 - Information Type: General vs. Symptom vs. Treatment) factorial design tested in a controlled, laboratory environment (as shown in Fig. 2b). Different from Study 1, participants experienced all 12 distinct conditions for this in-person experiment, enabling direct comparisons between human- and LLM-generated health information. We opted for a within-subject design for all independent variables to facilitate a nuanced analysis of participants' behavioral and physiological responses across conditions. Specifically, for the source of information (IV1), we aimed to observe whether participants exhibited different behavioral (e.g., gaze patterns) and physiological (e.g., heart rate, skin conductance) signals when reading information attributed to human versus LLM sources. While these sources may differ in presentation styles, it is also possible that participants' trust were influenced more by their belief about the source of text (human vs. AI) rather than the actual content or style. A within-subject design was critical for disentangling these effects, as it allowed each participant to serve as their own control, reducing variability across conditions and enabling a clearer examination of these factors. Participants rated their perceived level of trust for each set of health information while their eye-tracking data (gaze positions and pupil diameter) and physiological responses (ECG: Beats Per Minute (BPM), Beat-to-Beat Interval (BBI), Root Mean Square of Successive Differences (RMSSD); EDA: Skin Conductance Level (SCL) and Response (SCR); Skin Temperature) were recorded throughout the tasks.

To address our second research question, we explore whether behavioral and physiological signals can be used as complementary indicators to understand trust perceptions toward human- and AI-generated personal health information. In addition, we set up two prediction tasks that make use of the sensed data: (1) predicting participants' trust in health information through both regression on perceived trust scores and binary classification on trust level (high vs. low); and (2) classifying the actual source of the health information.

4.1.2. Stimuli and apparatus

We developed a web interface that displays the health information (question-and-answer pair) and the questionnaires for participants to rate their trust scores (see Fig. 6). The health information was identical to the material used in Study 1, as described in Section 3.1.2. Each set of health information was labeled as being generated either by "Human Professionals" or "Artificial Intelligence", regardless of the actual source.

We used a PHILIPS (Full HD, 1920*1080, 100 Hz) monitor to display the stimuli. The eye movements and pupil diameter (PD) data were recorded by a Tobii Pro Fusion eye tracker. The remote eye tracker was attached to the bottom of the monitor and connected to a computer (Windows, Intel Core i5, 16 GB RAM) running the Tobii Pro Lab software (AB, 2024).

Physiological signals, including ECG, EDA, and skin temperature, were measured using a BioSemi amplifier (van Amsterdam, 2025) (as shown in Fig. 1). ECG was captured through a disposable 3M

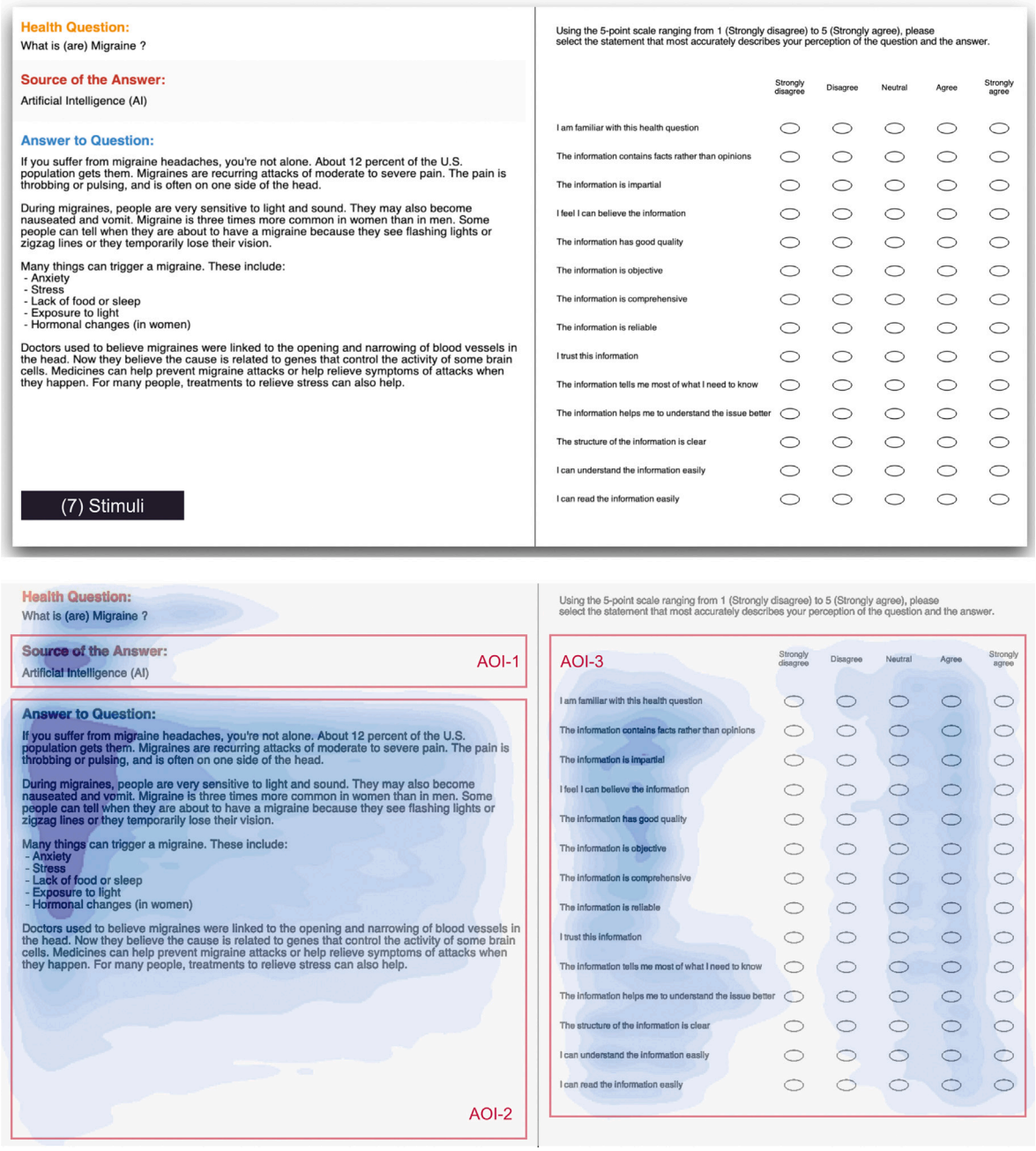


Fig. 6. Top: An example of text stimulus displayed on the monitor. **Bottom:** Heatmap of the gaze points on stimuli. Three AOIs are predefined: AOI-1 is the area for presenting disclosed label; AOI-2 is the area for presenting health information; AOI-3 is the area to rate the perceived trust in health information.

Red Dot in LEAD-II configuration, EDA was measured with electrodes attached to fingers, and skin temperature was monitored with a miniature Pt1000 sensor, all at a 24-bit resolution and 1000 S/s sampling rate. These data were collected using software FysioRecorder version 2.1 (van Amsterdam, 2025). Data recording was initiated through a central recording application developed in PsychoPy Peirce et al. (2019), connecting to sensors via IP addresses to simultaneously capture synchronized ECG, EDA, skin temperature, and eye-tracking signals.

4.1.3. Self-reported measures

We collected several self-reported measures, consistent with those used in Study 1 described in Section 3.1.3. These included demographics, prior experience with online health information and AI, the propensity to trust technology (PPT), eHealth, and AI literacy. Additionally, we assessed the perceived reliability of AI and human professionals using a single item for each: "How reliable do you find AI/Human Professionals?" Responses were captured on a 5-point Likert scale, ranging from 1 (Not at all) to 5 (Extremely). They were collected before the formal reading task.

During the reading task, we repeatedly measured the participants' 1) familiarity level with each given health question and 2) their perceived trust score in health information (Johnson et al., 2015; Rowley et al., 2015), after they completed each stimulus.

4.1.4. Machine learning: setup and approach

We performed binary classification to predict information sources and applied both regression and classification (i.e., binary and three-class classification) for trust scores. The perceived trust score (see

Table 5
Characteristics of participants in the lab study.

Demographic	Categories	Numbers of Participants (%)
Gender		(N = 40)
	Female	23 (57.5 %)
	Male	16 (40.0 %)
	Non-binary	1 (2.5 %)
Age	18–24	23 (57.5 %)
	25–34	14 (35.0 %)
	35–44	1 (2.5 %)
	45–54	1 (2.5 %)
	65+	1 (2.5 %)
Education	Bachelor's degree	18 (45.0 %)
	Master's degree	17 (42.5 %)
	Doctorate or higher	5 (12.5 %)
Professional Domain	Health and Medical Science	2 (5.0 %)
	Science, Technology, Engineering, and Mathematics (STEM)	11 (27.5 %)
	Business, Economics, and Law	8 (20.0 %)
	Communication, Arts, Culture, and Entertainment	7 (17.5 %)
	Education and Social Science	7 (17.5 %)
	Other	5 (12.5 %)
Frequency of online health information seeking	Rarely	6 (15.0 %)
	Sometimes	25 (62.5 %)
	Often	7 (17.5 %)
	Always	2 (5.0 %)
Frequency of using AI tools	Never	2 (5.0 %)
	Rarely	5 (12.5 %)
	Sometimes	9 (22.5 %)
	Often	18 (45.0 %)
	Always	6 (15.0 %)
Duration of online health information seeking	Less than 1 year	4 (2.8 %)
	1–3 years	24 (17.0 %)
	3–5 years	50 (35.5 %)
	5–10 years	45 (31.9 %)
	More than 10 years	18 (12.8 %)

Section 4.1.3), as an aggregate numerical rating based on the "trust of online health information questionnaire" (Johnson et al., 2015; Rowley et al., 2015), naturally lends itself to regression. However, this approach can be challenging to interpret given that trust is an aggregate and overall complex construct. On the other hand, trust classification simplifies interpretation but introduces an arbitrary split between high and low trust levels. To address this, we pre-processed the original trust scores into high and low categories using the median value as a threshold for binary classification. For the three-class classification, we divided the trust scores into low, medium, and high categories based on tertiles, creating balanced splits that account for the distribution of scores.

We used several common machine learning algorithms as suggested in prior research (Ajenaghughrure et al., 2021), including single models (i.e., logistic regression (LR), random forest (RF), support vector machines (SVM), multi-layer perceptron (MLP), linear regression, ridge regression, random forest-based regression), and ensemble methods (i.e., boosting, voting, stacking and bagging). Models were built using hand-crafted gaze features (i.e., fixations, saccades, pupil diameter) and physiological signals (i.e., BPM, BPI, RMSSD, SCL, SCR, and skin temperature).

We experimented with three feature sets: Gaze-only, Physiology-only, and Gaze + Physiology. These sets trained and evaluated the selected models to determine how effectively they could predict participants' perceived trust scores and classify the source of information. We set the "random state" (Sahagian, 2024) parameter to ensure result consistency and used the "grid search" (Liashchynskyi and Liashchynskyi, 2019) technique to find the optimal hyperparameters of the models. We only considered user-independent models to ensure that any predictions generalize across all participants, despite well-known challenges in generalizing using peripheral physiological features (Alamudun et al., 2012). To achieve this, we adopted a Leave-One-Subject-Out (LOSO) cross-validation approach (Kunjan et al., 2021), where in each iteration, one participant's data was held out for testing, and the remaining data was split 80/20 for training and validation. This setup ensures robust user-independent models. The performance of the regression models (for trust score prediction) was evaluated by Mean Squared Error (MSE) and Coefficient of Determination (R^2). The performance of the classification models (for trust level and information source) was assessed with the Macro-F1 (Opitz and Burst, 2021) score as the average of the validations.

4.1.5. Participants

For the in-person experiment, we used the same inclusion criteria as in Study 1 (age above 18 who are fluent in English). Participants were recruited through the institute's recruitment system. A power analysis using G*Power 3.1 (Faul et al., 2007) for a within-factor ANOVA indicated that at least 28 participants were required to detect a medium effect size observed in Study 1 (f = 0.24), with an alpha level of 0.05 and a power of 80 %.

Table 5 shows the characteristic information of the participants. Forty participants (N = 40) were enrolled (F = 23, M = 16, NB = 1), aged between 18–65+ years, with 92.5 % falling in the 18–34 age range. Regarding online health information-seeking experience, 22.5 % frequently or always used online sources, 62.5 % occasionally searched online, and 15.0 % rarely used online resources. For the frequency of AI usage, 60.0 % frequently or always used AI tools, 22.5 % occasionally used AI, and 17.5 % rarely or never used AI.

4.1.6. Study procedure

Each participant was invited to the institute for a single session at the lab. The researcher first provided an overview of the study and task details, after which participants gave informed consent before the lab session. During the pre-survey, participants provided their demographic information (age, gender, occupation) and their experiences with online health information searches and interactions with AI.

Upon completing the pre-survey and successfully calibrating the sensors, participants began the formal reading task. During the reading task, each participant reviewed 12 sets of health information: six were labeled as from human professionals and six as from AI, regardless of the actual

source. Sources and labels were counterbalanced to minimize order effects. The entire session lasted approximately 60 minutes, and participants were rewarded with €10 for participating. The study received approval from our institute's ethics and data protection committee. The procedure of the lab study is detailed in Fig. 2(b).

4.1.7. Data pre-processing

Self-reported Trust Scores. To assess how factors such as information source, labeling, and information type affect trust in online health information, we first checked the suitability of the data for statistical analysis. A Shapiro-Wilk test (SHAPIRO and WILK, 1965) confirmed that the self-reported trust scores deviated from a normal distribution. Therefore, we applied generalized estimating equations (GEE) (Hardin and Hilbe, 2012) to analyze trust differences across information sources and labels, because of its robustness to violations of normality and flexibility in handling repeated ordinal measures. Additionally, we conducted Spearman correlation analyses (Zar, 2005) with Bonferroni corrections to explore relationships among the variables. Consistent with Study 1, and given that the GEE results (Table 7) indicated no significant interaction effects between the independent variables of source and labeling, we averaged the repeated measures for each participant into a single observation across conditions. This simplification allowed us to focus on the key exploratory relationships while maintaining analytical clarity.

Eye Tracking Data Processing. Raw eye-tracking data were extracted from eye tracker (Tobii Pro Fusion) using Tobii Pro Lab software AB, 2024), and time-synchronized with stimuli. As shown in Fig. 6 (Top), there are three Areas of Interest (AOIs): AOI-1 (disclosed label of source), AOI-2 (health information), and AOI-3 (rating scale). We chose a fixation threshold of 30° for velocity and 60 ms for duration, as suggested by the information reading task (Van der Lans et al., 2011). Gaze features including gaze duration, fixation (count and duration), saccade (count and length), and pupil diameter were calculated to understand how users read the information. We excluded participants whose gaze accuracy fell below 90 %, resulting in 38 participants' eye-tracking data being retained for further analysis. After transforming data through Aligned Ranked Transformation (ART) (Wobbrock et al., 2011), we confirmed the non-normality of eye tracking data with the Shapiro-Wilk test.

Physiological Signal Processing. Physiological signals were processed using Vsrrp98 software (v13.1.4) (van Amsterdam, 2025), following the practice in prior research (Babaei et al., 2021; Ahmad and Alzahrani, 2023). Key physiological features derived from the raw ECG data using interval-to-interval window size included BPM, BPI, and the main HRV metrics - the root mean square of successive differences (RMSSD). For EDA data, we used the continuous decomposition analysis method (Benedek and Kaernbach, 2010) to separate it into the tonic SCL and phasic SCR components, then calculated the mean SCL and SCR values, as well as the SCR count. Skin temperature readings were screened for any abnormal responses. We excluded SCL and SCR data when more than 4 out of 12 stimuli have values lower than.01μS or exceeded 50μS, as these readings likely resulted from loss recording or movement artifacts. As a result, we retained data from 34 participants for SCL and SCR analysis, and 40 participants for ECG and skin temperature analysis. Following preprocessing, we used the Shapiro-Wilk test to assess normality, revealing that all physiological features were not normally distributed.

Given the exploratory nature of our investigation and the presence of multiple comparisons, we applied appropriate corrections based on the type of data. First, self-reported data were analyzed using a single GEE test, thus no multiple comparison correction was necessary. Second, for eye-tracking data, where multiple tests were conducted for different features, we applied False Discovery Rate (FDR) correction (Haynes, 2013a) to control for potential inflation of Type I errors. Third, for physiological data, no multiple comparison correction was applied because most of the physiological features (e.g., RMSSD, ECG, EDA) were uncorrelated, as confirmed by correlation analysis, and each feature was analyzed independently. This approach reflects our goal of treating these features as distinct, non-overlapping measures, without assuming that they influence each other.

Table 6
Descriptive statistics of the lab study.

	Measures	Mean	SD
Pre-survey	Familiarity of AI	3.58	.96
	Perceived Reliability of AI	3.13	.61
	Perceived Reliability of Human Professionals	3.78	.53
	Propensity of Trust (PPT)	3.54	.33
	eHealth literacy	3.69	.25
	AI literacy	3.78	.20
	Conditions	**Mean**	**SD**
Trust score	Source (Human) & Label (Human)	3.67	.63
	Source (Human) & Label (AI)	3.56	.64
	Source (LLM) & Label (Human)	3.92	.56
	Source (LLM) & Label (AI)	3.78	.63
	Source (Human), regardless of Label	3.62	.64
	Source (LLM), regardless of Label	3.85	.60
	Label (Human), regardless of Source	3.80	.61
	Label (AI), regardless of Source	3.67	.65

Table 7
Results from the GEE analysis on the self-reported trust score. (**p <.01, *p <.05).

Outcomes	Conditions	Coefficient	P-value	Effect ($Std.\beta$)	Sig
Trust score	Source (Human vs. LLM)	.22	.00	.35 (medium)	**
	Label (Human vs. AI)	−.15	.01	.23 (medium)	**
	Source ∗ Label	.03	.71	.05 (small)	

4.2. Findings

4.2.1. Descriptive statistics

As shown in Table 6, participants demonstrated a generally positive attitude toward technology, with an average trust in technology score of 3.36 (SD = .23). Their eHealth literacy averaged 3.69 (SD = .25), indicating proficiency in searching for digital health information. AI literacy was even higher, with an average score of 3.78 (SD = .20), suggesting a strong understanding of AI and its applications.

Regarding the perceived trust, the lab study results closely mirrored those of the online survey, despite being based on separate participant samples and independently collected data. The self-reported trust scores from the lab study varied depending on both the source and the labeling of the health information. Information both sourced from and labeled as from human professionals had an average trust score of 3.67 (SD = .63). When human-sourced information was labeled as AI, the score slightly decreased to 3.56 (SD = .64). LLM-sourced information labeled as from human received the highest trust score of 3.92 (SD = .56), while information sourced from LLM and labeled as from AI had a trust score of 3.78 (SD = .63). Overall, participants reported higher trust in LLM-sourced information (M = 3.85, SD = .60) than in human-sourced information (M = 3.62, SD = .64), echoing the trend observed in the online survey and indicating a growing acceptance of AI (i.e., LLM) in health contexts. However, information labeled as coming from human professionals was trusted more (M = 3.80, SD = .61) than that labeled as AI (M = 3.67, SD = .65), suggesting that labeling plays an influential role in trust formation, potentially even more than the actual source. These findings reinforced the patterns found in the online survey while providing additional validity through the lab sessions.

4.2.2. Analysis of self-reported trust

Table 7 presents the results from the GEE analysis on self-reported trust scores from the lab study. Consistent with the online survey, both

the source and the label significantly impacted trust perceptions. Fig. 7 further illustrates the same pattern, echoing the online survey results. Trust was highest for LLM-sourced information labeled as human and lowest for human-sourced information labeled as AI.

The analyses first revealed a significant effect of information source on trust, with a coefficient of 0.22 ($p < 0.01$), indicating that LLM-sourced information was generally trusted more than human-sourced information, i.e., without knowledge of the actual source. This suggests that the source of information is crucial in shaping trust, as AI-generated content may be perceived as more structured and objective than human-authored content. While the raw coefficient represented a modest change of 0.22 points on a 5-point Likert scale, the corresponding effect size ($Std.\beta = 0.35$) was classified as medium. This reflects the bounded nature of the Likert scale, where even small raw differences can indicate meaningful relationships due to the relatively low variability in responses. Thus, the medium effect size underscores the practical relevance of the findings despite the small-scale differences.

Labeling also significantly impacted the trust perception, with a coefficient of -0.15 ($p = 0.01$), meaning information labeled as AI was trusted less than when labeled as human professionals. The negative coefficient suggests a preference for human-labeled information, as participants may associate human endorsement with greater credibility. Similarly, while the raw change (-0.15) was modest, the standardized effect size ($Std.\beta = 0.23$) reflects a medium effect, emphasizing that the impact of labeling, though subtle on the scale, has measurable and meaningful implications for trust perceptions.

Notably, the interaction between source and label was not significant ($coefficient = 0.03, p = 0.71$), indicating that the combined influence of source and label does not affect trust beyond their individual effects. The small standardized effect size ($Std.\beta = 0.05$) confirmed that this interaction effect is negligible.

4.2.3. Analysis of eye movement data

The results of GEE analysis (Hardin and Hilbe, 2012) on eye tracking data are detailed in Table 8, showing varied eye movement patterns. In AOI-1 (label area), fixation duration and pupil diameter of fixation showed significant differences by information sources and labels, while saccade count showed significant differences by information labels only.

The post hoc comparisons shown in Figs. 8 and 9, participants demonstrated higher fixation counts ($p < .05$) and saccade counts ($p < .1$) in AOI-2 (main health information area) under the AI label condition, indicating that participants assessed the information focusing more on the content itself rather than the label when they were informed that the information was from AI. This implies that trust-related judgments in AI-labeled information were driven more by the actual content than the labeling of the source. Participants also showed significantly fewer fixation counts ($p < .05$) in AOI-3 (rating area) under the human label condition compared to the AI label condition. This suggests that human labels might inspire greater confidence, potentially influencing how users rate the trust score of the information. When information was actually sourced from LLM, participants showed higher fixation duration ($p < .01$) and counts ($p < .1$) in AOI-2, suggesting a more careful reading of AI-generated content. Conversely, human-sourced information led to higher fixation and saccade counts in AOI-3 ($p < .01$), indicating that LLM-sourced information might inspire greater confidence, potentially influencing how users rate the trust score, which aligns with the self-reported trust perceptions.

4.2.4. Analysis of physiological signals

Table 9 presents the results from GEE analysis on physiological data, shedding light on how physiological responses vary with different information sources and labeling.

RMSSD, a feature derived from ECG data, was significantly higher for AI-labeled information compared to human-labeled information ($p = .025$). Higher RMSSD indicates greater heart rate variability (HRV), which is often associated with lower physiological arousal. This aligns with the gaze patterns where participants paid less attention to the labeling area (AOI-1) under "AI" labels than "Human" labels, as indicated by significantly reduced fixation duration, saccade count, and pupil diameter (see Table 8).

Skin temperature responses also varied significantly between human and AI labels ($p = .029$), as well as between human and LLM sources ($p = .022$). Higher skin temperature in response to AI labels and sources suggests participants may have experienced increased emotional arousal or stress when interacting with AI-associated content.

SCL ($p = .061$) and SCR ($p = .082$) average values did not exhibit statistically significant differences, as shown in Fig. 10. This suggests that EDA components, at least within our study, were not discriminative of physiological arousal when users encountered human- versus AI-generated information.

4.2.5. Correlation analysis

The Spearman correlation analysis (Zar, 2005) in Figs. 11 and 12 revealed significant relationships between the self-reported trust score and various gaze and physiological features, indicating how participants' perceived trust in health information is linked to their behavioral and physiological responses.

Familiarity with the health question showed a strong positive correlation with trust in the information ($p < .01$). Among gaze features, fixation

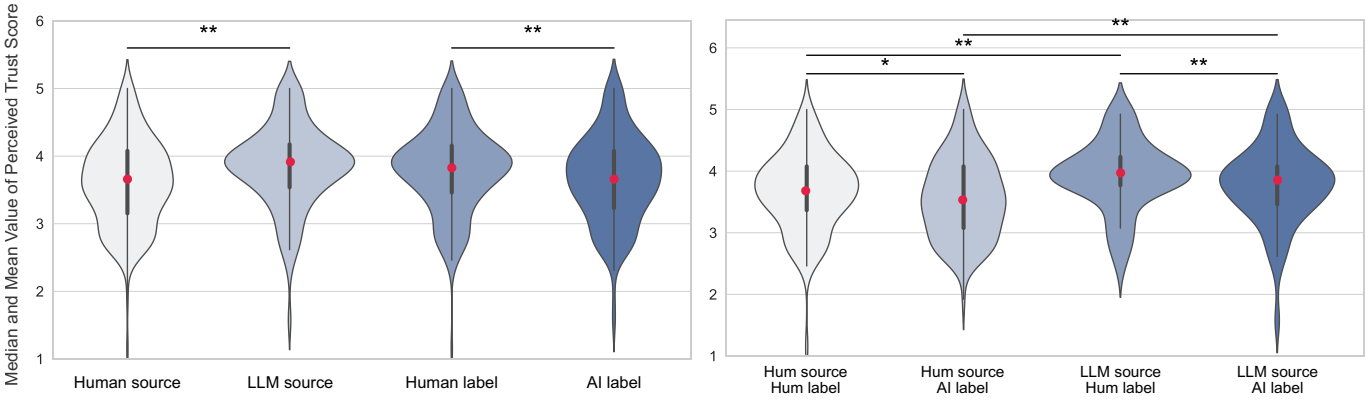


Fig. 7. Left: Perceived trust score in information by sources regardless of labels, and by labels regardless of sources. **Right:** Perceived trust score based on different source and label conditions. Each plot shows score density (width), with the red dot as the mean, the black line as the median, and thick bars denoting the interquartile range (IQR). Horizontal lines indicate significant pairwise differences (**p <.01, *p <.05).

Table 8
Results from the GEE analysis with False Discovery Rate (FDR) correction on the eye tracking data. (**p <.01, *p <.05, -p <.10).

Gaze	AOI	Condition	Coeff	P (Orig)	P (Corrected)	Effect ($Std.\beta$)	Sig
Fixation Count	AOI-1	Source	−4.02	**.033**	**.079**	.22 (medium)	–
		Label	−4.54	**.015**	**.055**	.25 (medium)	–
		Src × Lab	6.01	.082	.150	.33 (medium)	
	AOI-2	Source	0.61	.962	.962	.00 (small)	
		Label	−3.23	.827	.910	.02 (small)	
		Src × Lab	34.61	**.036**	**.079**	.26 (medium)	–
	AOI-3	Source	−8.14	.198	.272	.14 (small)	
		Label	9.90	.151	.237	.17 (small)	
		Src × Lab	−5.53	.559	.683	.10 (small)	
Fixation Duration	AOI-1	Source	−51.39	**.000**	**.000**	.43 (large)	**
		Label	−37.64	**.006**	**.017**	.31 (medium)	*
		Src × Lab	71.27	**.000**	**.000**	.59 (large)	**
	AOI-2	Source	4.60	**.038**	**.069**	.14 (medium)	–
		Label	−2.80	.218	.343	.09 (small)	
		Src × Lab	1.75	.584	.642	.05 (small)	
	AOI-3	Source	−0.46	.879	.879	.01 (small)	
		Label	−1.61	.553	.642	.05 (small)	
		Src × Lab	2.85	.519	.642	.09 (small)	
Saccade Count	AOI-1	Source	−5.13	**.044**	**.086**	.21 (medium)	–
		Label	−7.38	**.013**	**.047**	.26 (medium)	*
		Src × Lab	8.89	**.047**	**.086**	.36 (medium)	–
	AOI-2	Source	−4.35	.787	.787	.03 (small)	
		Label	−7.82	.651	.716	.05 (small)	
		Src × Lab	43.77	**.026**	**.071**	.28 (medium)	–
	AOI-3	Source	−9.40	.223	.307	.12 (medium)	
		Label	10.71	.160	.251	.14 (medium)	
		Src × Lab	−7.96	.487	.595	.10 (medium)	
Saccade Length	AOI-1	Source	0.03	.124	.341	.17 (medium)	
		Label	0.01	.531	.649	.08 (small)	
		Src × Lab	−0.07	**.020**	**.073**	.38 (medium)	–
	AOI-2	Source	0.00	.355	.558	.08 (small)	
		Label	0.00	.181	.398	.12 (medium)	
		Src × Lab	0.00	.829	.829	.02 (small)	
		Source	0.00	.276	.506	.08 (small)	
	AOI-3	Label	0.00	.456	.627	.06 (small)	
		Src × Lab	0.00	.685	.754	.05 (small)	
Pupil Diameter Fixation	AOI-1	Source	−0.41	**.002**	**.011**	.35 (medium)	*
		Label	−0.33	**.018**	**.040**	.29 (medium)	*
		Src × Lab	0.50	**.003**	**.011**	.43 (large)	*
	AOI-2	Source	0.00	.673	.823	.01 (small)	
		Label	0.01	.445	.699	.02 (small)	
		Src × Lab	0.00	.898	.932	.00 (small)	
	AOI-3	Source	0.00	.227	.416	.03 (small)	
		Label	0.01	.531	.730	.01 (small)	
		Src × Lab	0.00	.932	.932	.00 (small)	

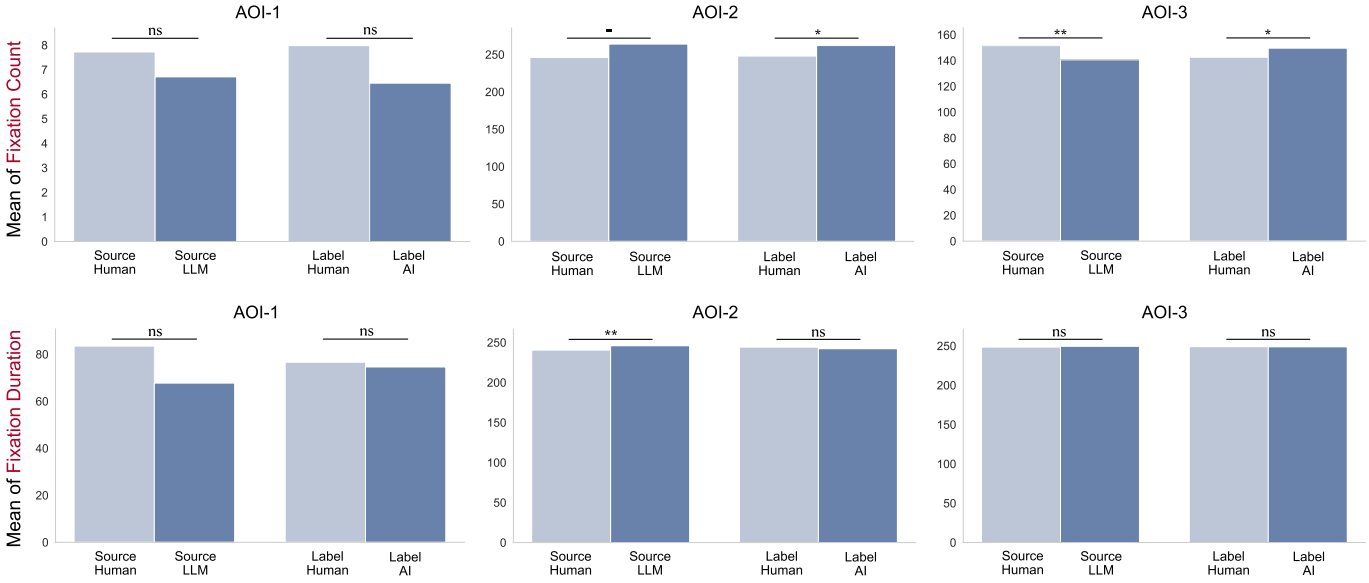


Fig. 8. Posthoc pairwise comparison by Wilcoxon signed-rank test with False Discovery Rate (FDR) correction of fixation features (count and duration) in three AOIs. (**p <.01, *p <.05, -p <.10, "ns" is not significant).

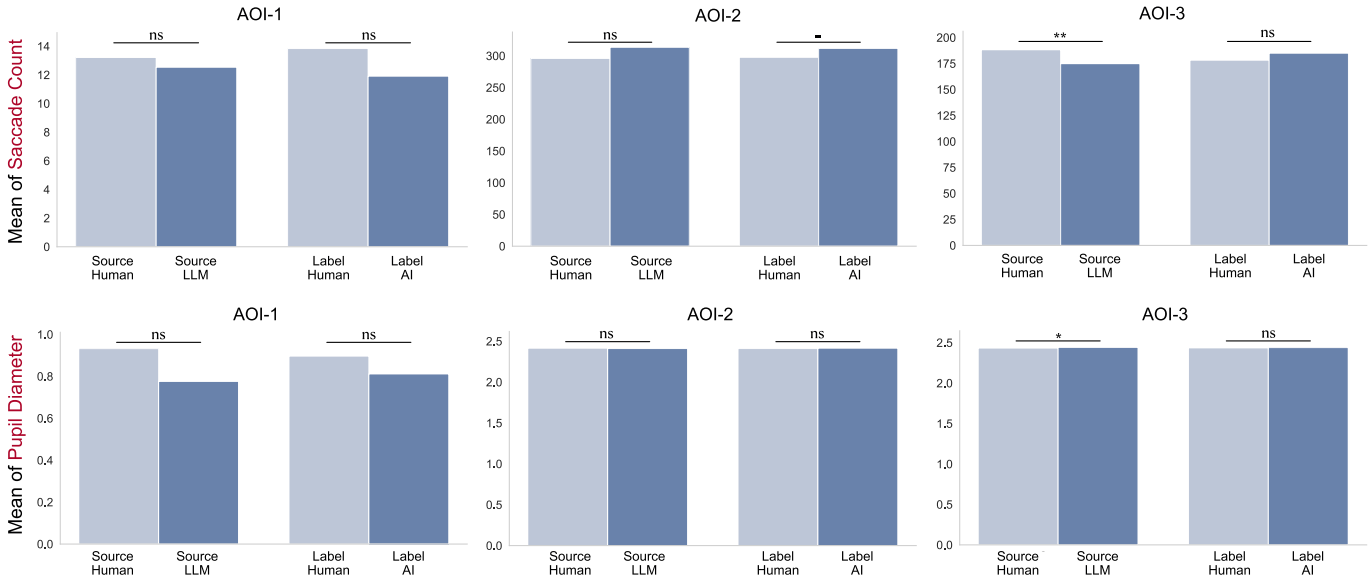


Fig. 9. Posthoc pairwise comparison by Wilcoxon signed-rank test with False Discovery Rate (FDR) correction of saccade count and pupil diameter of fixation in three AOIs. (**p <.01, *p <.05, -p <.10, "ns" is not significant).

Table 9
Results from GEE analysis on physiological signals. (**p <.01, *p <.05, -p <.10).

Outcomes	Features	Conditions	Coeff	p-value	Effect ($Std.\beta$)	Sig
ECG	BPM	Source (Human vs. LLM)	−0.58	.571	.07 (small)	
		Label (Human vs. AI)	−1.10	.288	.13 (medium)	
		Source × Label	1.38	.341	.17 (medium)	
	RMSSD	Source (Human vs. LLM)	2.11	.435	.12 (medium)	
		Label (Human vs. AI)	5.21	.025	.29 (medium)	*
		Source × Label	−4.45	.179	.25 (medium)	
	BPI	Source (Human vs. LLM)	8.88	.242	.12 (medium)	
		Label (Human vs. AI)	10.43	.225	.14 (medium)	
		Source × Label	−17.10	.153	.24 (medium)	
EDA	SCL	Source (Human vs. LLM)	0.03	.949	.04 (small)	
		Label (Human vs. AI)	−0.77	.061	.12 (medium)	–
		Source × Label	0.38	.414	.06 (small)	
	SCR	Source (Human vs. LLM)	−0.56	.399	.05 (small)	
		Label (Human vs. AI)	−0.92	.082	.08 (small)	–
		Source × Label	−0.98	.576	.08 (small)	
Temperature	—	Source (Human vs. LLM)	0.46	.022	.31 (medium)	*
		Label (Human vs. AI)	0.42	.029	.28 (medium)	*
		Source × Label	−0.57	.058	.39 (medium)	–

duration in AOI-1 (label area) positively correlated with the perceived trust score ($p < .01$), indicating that higher trust levels are associated with a longer focus on the labeling of information sources. Additionally, pupil diameter during fixation in AOI-1 ($p < .01$) also correlated positively with the trust score. Fixation and saccade counts in AOI-3 (rating area) were negatively correlated with trust, implying that participants who gave lower trust in the information exhibited more frequent saccadic movements in the rating area, likely reflecting efforts to evaluate or verify the information further.

No significant correlations were found between physiological features and trust levels. However, correlations were observed among the physiological features themselves, such as BPM (heartbeats), SCL, SCR, and skin temperature, though these did not directly link to trust.

4.2.6. Predictions using behavioral and physiological sensing

To explore trust perception (i.e., self-reported trust scores) through behavioral and physiological responses, we defined two tasks: (1) predicting participants' perceived trust scores in health information and (2) classifying the source of the health information.

For trust prediction, we first explored how regression models approximate perceived trust scores using regression models: linear regression (LR), ridge regression, SVM and random forest-based regressions, and XGBoost. As shown in Table 10, the random forest regressor on the combined Gaze + Physio feature set achieved the lowest MSE of.20 and the highest $R^2 = .35$ among the three feature sets, indicating the best performance. This highlights the value of combining gaze and physiological features for trust assessment. Fig. 13 illustrates the regression results on three different feature sets.

Next, we performed both binary (i.e., high vs. low) and three-class (i.e., high vs. medium vs. low) classification of trust levels based on participants' self-reported trust scores. As shown in Table 11, the ensemble method (voting model) achieved the highest accuracy (0.73) for binary classification using gaze-only features, while random forest achieved the highest accuracy (0.63) for three-class classification using combined gaze-physiological features. Interestingly, combining gaze and physiological features did not improve performance across all models, for instance, the gradient boosting model achieved slightly lower accuracy (0.72) for binary classification when incorporating both feature sets compared to using gaze features alone. These results indicate that

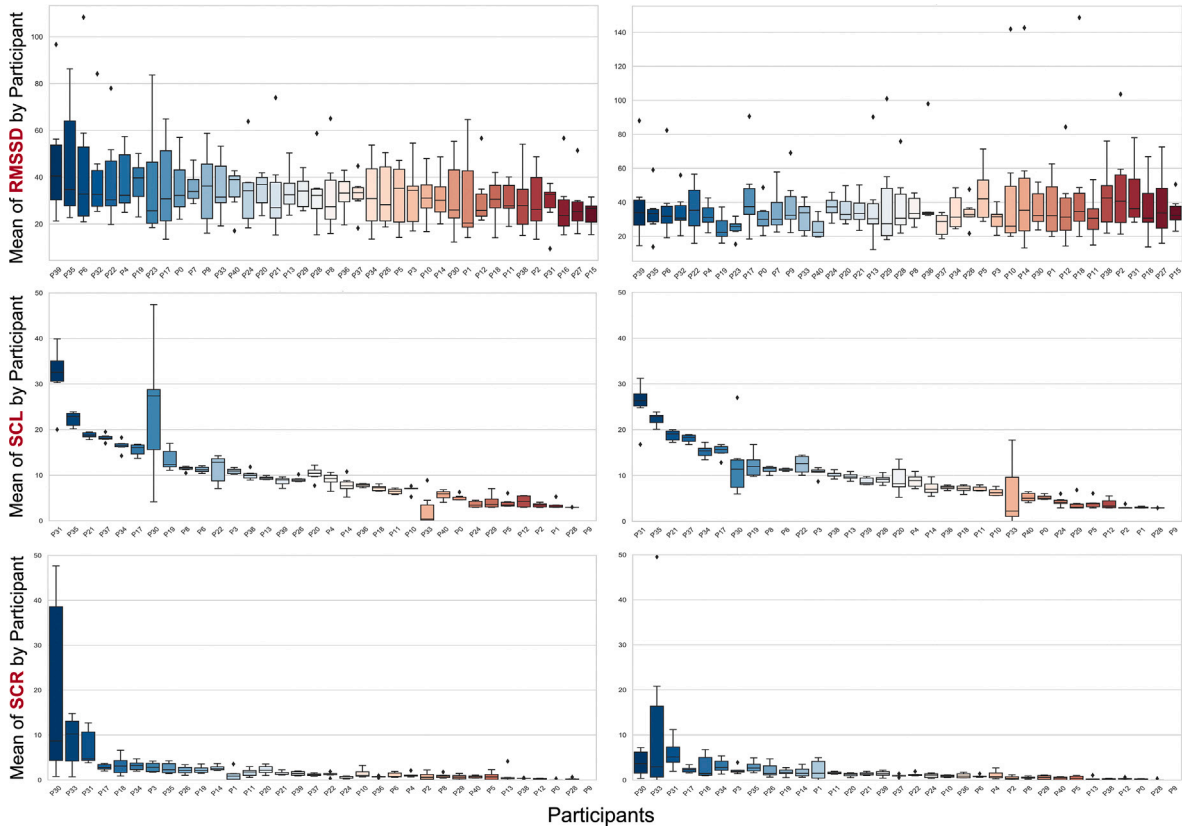


Fig. 10. Pairwise comparison without correction on features of RMSSD and SCR per participant. Each boxplot shows the distribution (median, IQR, outliers) for each participant under two labeling conditions: participants read the information labeled as from "Human Professionals" **(Left)** and from "AI" **(Right)**, regardless of the source. Color gradient reflects participant-wise ordering based on decreasing RMSSD or SCR values to facilitate visual comparison; the color itself carries no semantic meaning.

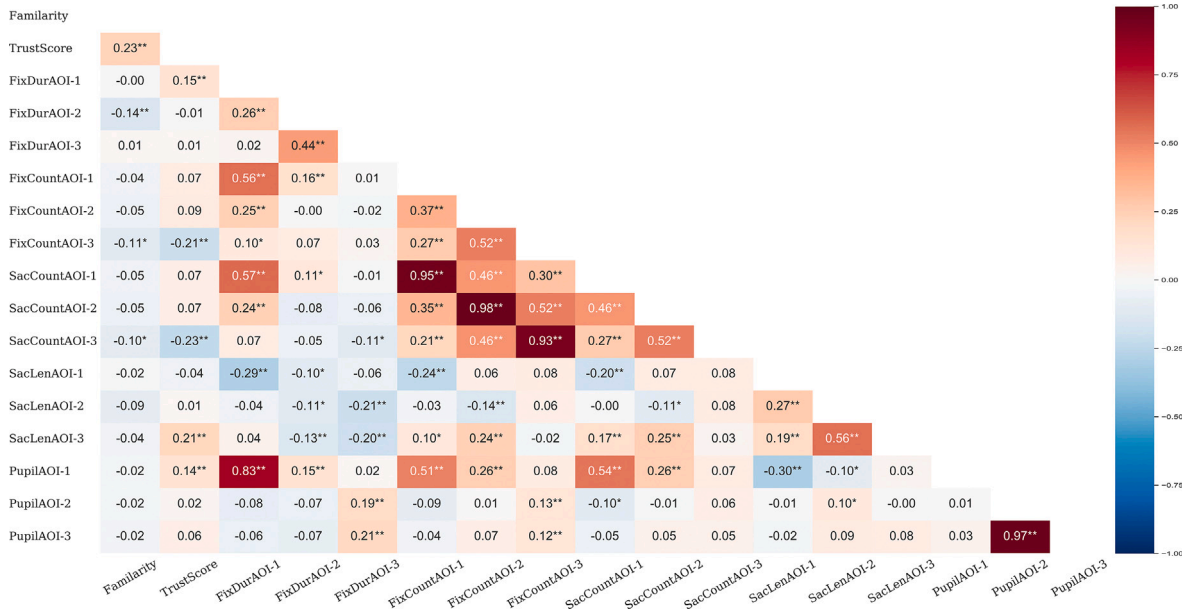


Fig. 11. Spearman correlation with Bonferroni corrections between trust perceptions and the gaze features. (** p <.01, * p <.05). Note: "FixDurAOI-": fixation duration in AOI-; "FixCountAOI-": fixation count in AOI-; "SacCountAOI-": saccade count in AOI-; "SacLenAOI-": saccade lenth in AOI-; "PupilAOI-": pupil diameter of fixation in AOI-.

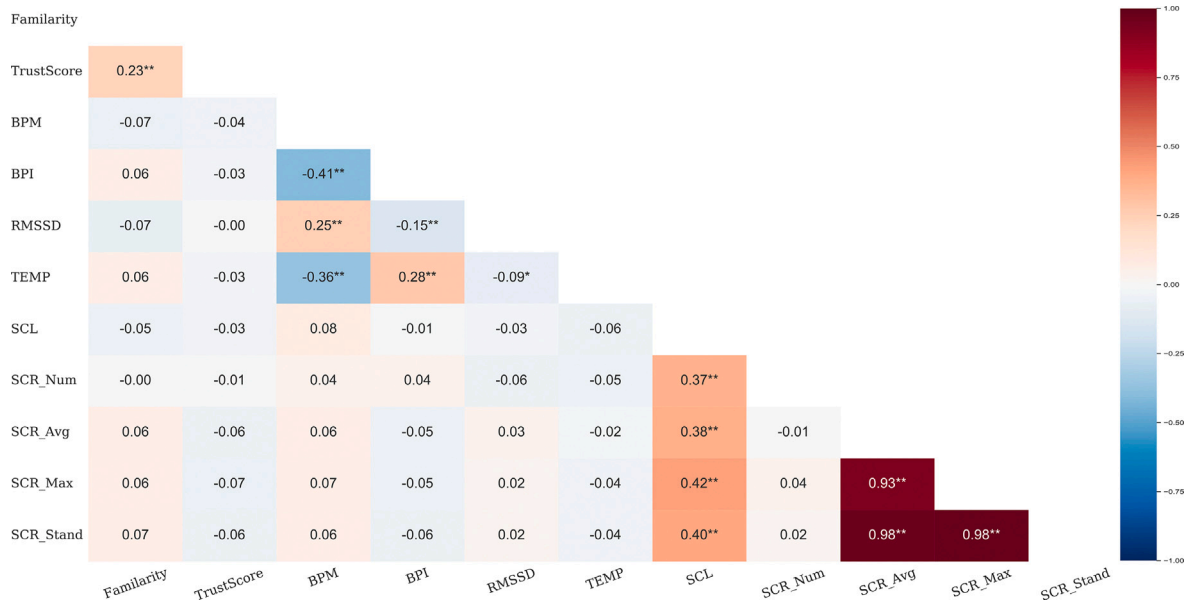


Fig. 12. Correlation on variables. (**p <.01, *p <.05). Note: "SCR_Num": number of SCR; "SCR_Avg": average value of SCR; "SCR_Max": maximum value of SCR; "SCR_Stand": standard value of SCR.

Table 10
Prediction of perceived trust scores through regression using gaze and physiological features.

Models	Gaze Only		Physio Only		Gaze + Physio	
	MAE	R^2	MAE	R^2	MAE	R^2
SVR	.29	.06	.33	.06	.28	.10
Linear Regression	.25	.20	.31	.01	.24	.20
Ridge Regression	.24	.21	.31	.01	.24	.23
Random-Forest Regression	.23	.25	.25	.19	.20	.35
XGBoost	.24	.22	.28	.08	.23	.23

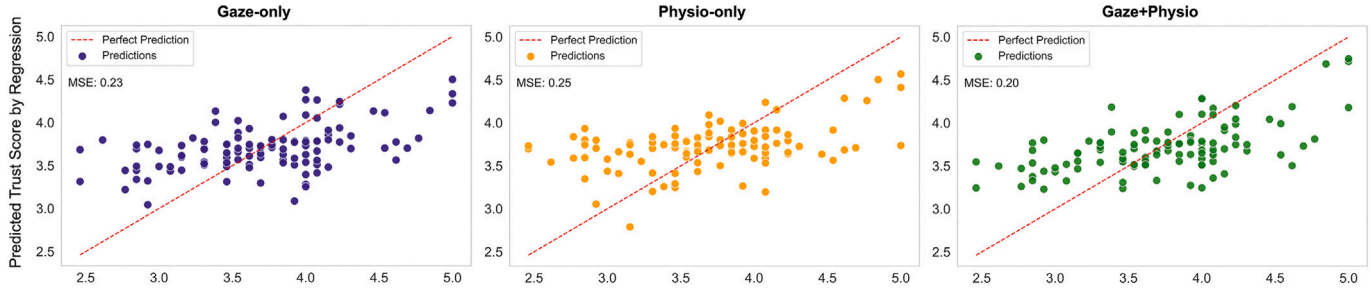


Fig. 13. Prediction of perceived trust score using the Random-Forest Regression model on three different features set: Gaze-only, Physiology-only, Gaze + Physiology. Each dot represents one participant's predicted vs. actual self-reported trust score, with the red dashed line indicating perfect prediction.

gaze features alone achieved higher classification accuracy for binary trust levels compared to combined gaze and physiological features. This suggests that gaze features may play a more prominent role in predicting trust levels than physiological responses in the context of this study.

For the second task to classify the information source, combining gaze and physiological features yielded the best results. The AdaBoost model achieved the highest accuracy of 0.65 and F1 score of 0.64, indicating that physiological responses complement gaze features in distinguishing between human- and LLM-generated health information.

Fig. 14 presents feature importance for the prediction tasks following SHAP framework proposed by (Lundberg and Lee, 2017) for better interpreting the model predictions. In summary, gaze features are effective for predicting trust perceptions, while combining gaze and physiological features could improve the classification of information sources. The robust performance of ensemble methods across both tasks highlights their potential in developing tools to assess trust-related responses in health communication by leveraging gaze and physiological signals.

5. Discussion

We conducted an online survey and a lab study in this work to investigate how users' trust responds to human versus AI-generated content, and in what ways trust in online health information may be influenced by including transparency labels as simple as "Human" versus "AI" labels on personal health information. Our findings showed that self-reported trust in digital health information is influenced by its actual source and disclosed labeling of the source. Further, the impacts of these conditions

Table 11
Classification of trust levels (high, medium, low) and the source of information using gaze and physiological features.

Features	Models	Trust Level		Source
		2-class(Acc / F1)	3-class(Acc / F1)	2-class(Acc / F1)
Gaze Only	LR	.65 /.62	.57 /.57	.62 /.55
	RF	.69 /.65	.57 /.57	.57 /.52
	SVM	.51 /.53	.43 /.42	.60 /.48
	MLP	.57 /.58	.32 /.32	.44 /.53
	GradientBoost	.72 /.66	.54 /.54	.52 /.52
	AdaBoost	.67 /.64	.58 /.58	***.65*** /.52
	XGBoost	.70 /.66	.54 /.54	.43 /.52
	Voting	***.73*** /.67	.54 /.54	.60 /.49
	Stacking	.70 /.66	***.59*** /.58	.49 /.55
	Bagging	.70 /.66	.57 /.57	.57 /.47
Gaze + Physio	LR	.65 /.62	.58 /.56	.58 /.54
	RF	.69 /.65	***.63*** /.63	.60 /.52
	SVM	.51 /.53	.43 /.43	.60 /.49
	MLP	.53 /.60	.48 /.47	.59 /.50
	GradientBoost	***.72*** /.68	.59 /.56	.53 /.53
	AdaBoost	.66 /.64	.54 /.54	***.65*** /.64
	XGBoost	.65 /.67	.57 /.57	.57 /.52
	Voting	.67 /.67	.59 /.58	.60 /.53
	Stacking	.69 /.66	.60 /.60	.48 /.52
	Bagging	.70 /.66	.61 /.61	.54 /.53

Fig. 14. Top 10 important gaze and physiological features in Random Forest regressor for predicting perceived trust scores (**Left**) and in AdaBoost classifier for classifying the source of health information (**Right**), based on SHAP values computed on the test set. Blue bars represent gaze features; red bars represent physiological features.

were also evident at a behavioral and physiological level. Below, we discuss these aspects in detail.

5.1. Users may prefer LLM-sourced health information, but an AI label lowers their trust

Both studies tested **(RQ1)** if the actual source, disclosed label, and type of information influence perceived trust in online personal health information. Our findings revealed that LLM-sourced content is trusted more than human-sourced content, regardless of labeling, whereas human professional labels are trusted more than AI labels. Trust however remained consistent across different information types (general, symptom, or treatment-related), suggesting that the source and labeling, rather than the type of information, are the primary determinants of perceived trust.

The observed difference in trust perception was evident in both self-reported trust scores (i.e., higher trust scores of LLM-generated information) and qualitative data, which suggests that participants have perceived subtle distinctions of information presentation styles in the LLM- versus human professionals-sourced information that provided cues for trust. The stronger effect observed in Study 2 (lab study with the within-subjects design) compared to Study 1 (survey study with the between-subjects design) further supports this, as the within-subjects design allowed participants to compare responses from both sources side by side. While we cannot conclusively determine the specific factors in information quality driving higher trust, our findings imply that LLM-generated content may convey an impression of clarity or objectivity that resonates more strongly with participants. Our observation that LLM-sourced information was trusted more than that from human professionals may reflect advancements in LLMs like ChatGPT, which can produce structured and high-quality responses (Hristidis et al., 2023; Van Bulck and Moons, 2023). Notably, GPT-4 generated responses have been found to be perceived as more human-like than actual human-authored content and other studies find that LLM-generated content is often indistinguishable from human-generated text (Rathi et al., 2025). This explanation (i.e., generally higher language quality of LLM-generated responses as a basis of trust) aligns with Dalton et al. (2022) proposal of emergent conversational information-seeking powered by LLMs, and is evident when assessing how LLMs are being used in the context of healthcare (Garg et al., 2023; Lee et al., 2023; Hristidis et al., 2023; Van Bulck and Moons, 2023; Sun et al., 2024).

Furthermore, researchers suggest that people prefer algorithms to humans in certain tasks and it could relate to individuals' machine heuristic (rule of thumb that machines are more secure and trustworthy than humans (Logg et al., 2019; Sundar and Kim, 2019)). In our studies, qualitative analyses (Section 3.3) further confirmed that participants attributed the higher trust in LLM-generated content to its efficiency, capacity to process extensive health data (Singhal et al., 2023), and

objective language style (Xu et al., 2023; Sun et al., 2024). This suggests that LLMs' (e.g., GPT-4 (OpenAI, 2024)) ability to deliver comprehensive and objective health information resonates with users, positioning them as reliable sources of health information.

Paradoxically, when health information was labeled as human, it was rated with higher trust scores than AI-labeled information, which is supported by Reis et al. (2024), who found that people value human advice more when aware of AI's involvement, especially in health context. This observation appears to generalize across domains, whereby an AI label can diminish people's perceived quality, even if the AI source was initially deemed superior. This includes AI art (Horton, Jr et al., 2023), general communication (Yin et al., 2024), medical advice (Kerstan et al., 2023). Even in clinical decision-making scenarios, people tend to prefer human decision-makers over AI, perceiving the latter as less dignified (Formosa et al., 2022), further highlighting a deep-seated bias against AI involvement in sensitive health-related contexts. Moreover, Epstein et al. (2023) found that not only the presence of a label, but also its wording, can significantly affect trust. For example, people perceive content labeled as "AI-assisted" more favorably than "AI-generated", indicating that subtle linguistic framing influences users' willingness to trust. This suggests that beyond binary source disclosure, the design and language of labeling also play a critical role in shaping perception.

The qualitative findings (Section 3.3) confirmed that participants expressed greater trust in human expertise, which they associate with verified knowledge, accountability, and human empathy. In contrast, they viewed the lack of consciousness, ethical judgment, and transparency in AI as diminishing their perceived trust. The perspective expressed by our participants aligns with De Freitas et al. (2023) work about psychological factors affecting attitudes toward AI acceptance, which identifies opacity (lack of transparency or explainability) and emotionlessness (absence of empathy or moral understanding) as key factors driving user resistance to AI tools. Our respondents echoed these concerns by highlighting AI's lack of transparency and moral reasoning, especially in healthcare contexts, where trust is closely tied to perceived ethical awareness and human empathy. These reactions may also reflect a broader skepticism about machine consciousness (Scott et al., 2023).

These findings can be interpreted through the MATCH model (Liao and Sundar, 2022), which conceptualizes trust in AI systems through three components. In our context, actual source of the information (human vs. LLM) corresponds to *model attributes*, reflecting users' judgments of competence and reliability. Disclosure labels (AI vs. human) act as *afforded cues*, shaping trust perceptions independently of content quality. Participants' perceptions, such as associating human expertise with trust or distrusting AI due to its lack of professionalism, reflect *trust heuristics*, where users rely on cognitive shortcuts in uncertain of complex health contexts. This framing emphasizes that trust is not only a response to information content but also to how the system communicates authorship and identity, and how users emotionally and cognitively process these trustworthiness cues (Lee and See, 2004) which was further explored through implicit behavioral and physiological responses in the following section.

Summarizing, while AI is increasingly recognized for its competence, our findings underscore the role of transparency as a trustworthiness cue framed in the MATCH model (Liao and Sundar, 2022), emphasizing the need for transparent AI-powered systems (Liao et al., 2023) and authentic information (Burrus et al., 2024; El Ali et al., 2024) to build trust, particularly when providing nuanced health advice (Broom, 2005; Kerstan et al., 2023). However, our study also cautions against over-reliance on labeling as a trust mechanism. As highlighted in prior work (Scharowski et al., 2023), labels can create a false sense of security and may inadvertently reinforce the "implied truth effect" (Pennycook et al., 2020), where unlabeled content is assumed to be accurate. These findings point to the need for more context-sensitive and dynamic approaches to communicating AI involvement in health information systems.

5.2. Behavioral and physiological features can vary by health information source and label

Our results demonstrated that the effects of label and source are also evident at the behavioral and physiological level. Prior work has shown value in leveraging behavioral and psychophysiological sensing across fake news detection in social media (Abdrabou et al., 2023) and information-seeking tasks (Ji et al., 2024), where such signals are indicative of visual attention and information processing in these tasks. With respect to trust, Ajenaghughrure et al. (2020) review found that while psychophysiological levels of trust perceptions (e.g., arousal) can be detected (e.g., using EEG or ECG), how such responses behave during user interactions (in real-time) remains underexplored. In the context of our study, we first explored **(RQ2)** whether such signals vary during health information processing across human versus AI-sourced content, and essentially whether such signals can serve as a means of verifying and possibly predicting self-reported trust scores (Section 4.1.4). We found that participants displayed distinct gaze patterns related to the source and labeling of the presented presented information. Specifically, we found that longer fixation duration, higher fixation counts, and larger pupil dilation were associated with information labeled as human-generated, suggesting a deeper cognitive engagement with this human-labeled information, suggestive of higher trust. Conversely, information labeled as AI-generated prompted more scanning behavior (i.e., reflected in increased saccadic movements and shorter fixation durations), indicative of increased verification processes. These results corroborate existing research from others (e.g., Just and Carpenter, 1980 and Rayner, 1998) who likewise found that gaze patterns, especially the fixation and saccade behaviors, are indicative of cognitive processing and information verification relevant to trust assessment and dynamics.

For the peripheral physiological signals, while we found significant differences in features such as RMSSD and skin temperature when users encountered labeled health information, no such differences were found in skin conductance (SCL and SCR) measurements. It is worth speculating what this means: these indicators aligned with users' self-reports, where health information labeled as from AI elicited higher HRV (i.e., RMSSD) than the label of human professionals. Higher HRV is typically associated with lower physiological arousal, possibly reflecting less cognitive processing or more relaxed state. This interpretation is consistent with the meta-analysis by Kim et al. (2018), which found that HRV reliably decreases under stress or increased cognitive demands, and increases under lower arousal or more comfortable conditions. Indeed, HRV is one of the most commonly used psychophysiological indicators in trust research (Ajenaghughrure et al., 2020), able to detect subtle variations in user state during human-computer interaction. Although Ajenaghughrure et al. caution that trust classification using physiological signals remains an open research challenge. Furthermore, the pattern of reduced physiological arousal in response to AI-labeled information aligns with the gaze data in our study, which suggested less attentional engagement (e.g., shorter fixations, fewer regressions) with AI-labeled content compared to human-labeled information. These findings suggest that participants may have processed AI-labeled health information with lower cognitive and emotional investment. Similarly, higher skin temperature levels were observed with both AI-labeled and LLM-sourced information, suggesting lower emotional arousal and stress levels, reflecting participants' psychological interpretation of trust (Ahmad and Alzahrani, 2023). I.e., participants gave higher trust scores to the LLM-sourced information compared to human-sourced, and showed lower physiological arousal with the AI labels than human labels.

These behavioral and physiological responses deepen our interpretation of trust formation grounded in the MATCH model (Liao and Sundar, 2022). While the online survey study (Study 1) focused on how users respond to model attributes and afforded cues (i.e., health information source and labels), we further extend the analysis to *trust heuristics*, the implicit, affective processes that guide user trust-related judgments

under uncertainty. Physiological responses like HRV and skin temperature likely reflect affective dimensions of trust (e.g., comfort, emotional arousal), whereas gaze patterns and fixation behavior index cognitive engagement. This layered interpretation aligns with calls to distinguish between cognitive and affective trust as investigated by (Lee and See, 2004) which is grounded in the most widely used and accepted ABI trust model from (Mayer et al., 1995), suggesting that trust in AI-generated health content is not just explicitly reported but also embodied in users' implicit affective reactions.

Taken together, these sensing signals could serve as a useful means to corroborate how users react and feel toward content perceived to be sourced from humans versus AI, while providing an additional layer of information about information processing and associated affect.

5.3. Considerations: toward trust-aware AI for health information seeking

Our findings offer actionable design considerations for stakeholders designing or developing LLM-powered health information tools. These include interface designers and developers of adaptive AI systems. We outline practical considerations as below, grounded in the findings of this work.

5.3.1. For UI designers of health information interfaces

Designing and placing labels for trustworthy interfaces. As a key element of user interfaces, transparency labels play a crucial role in promoting trustworthy AI design (Liao et al., 2024). Our findings show that labeling content as AI-generated consistently reduced trust compared to identical content labeled as human-generated. This suggests that while transparency is critical, poorly framed labels can inadvertently erode trust. Given the critical role of UX for responsible and transparent AI design (Liao et al., 2024), we find it important to foster trust already at the interface level when presenting health information. Prior work highlights the need for balance: too little transparency risks deception, while too much may undermine confidence (Kizilcec, 2016). Therefore, designers should carefully consider not only whether labels are present, but also how they are phrased and styled. Research from Epstein et al. (2023) shows that both presence and framing can significantly shape user trust. Insights from privacy nutrition labels (Kelley et al., 2009) further demonstrate that visual choices of design, such as simplifying symbols, using color intensity to signal risks, and providing accessible visual explanations for technical terms, can improve users' accuracy, efficiency, and satisfaction (Kelley et al., 2009).

Our eye-tracking data supports this: participants gave more fixation counts to AI-labeled health information while also giving more fixation counts to human labels. This indicates that labels strongly influence both user attention and trust judgments. Effective placement is therefore crucial: labels should appear in or near high-attention areas such as headlines or primary content zones, and be styled with moderate emphasis, visible, but not distracting.

Taken together, these insights point toward "trust-aware" UI design, where transparency labels are not just added for compliance but are thoughtfully designed and positioned to foster trust without bias. Visual elements such as trust meters or engagement indicators could further reflect the health information system's trust assessment and communicate how health systems interpret user interactions, making transparency both informative and supportive of trust.

Uniform UI structure across health topics. Our findings also showed that trust ratings did not vary across information types, suggesting that a uniform interface structure can be used across health content categories, allowing design efforts to focus more on trust-sensitive features like labels and source attribution rather than varying UI by topic.

5.3.2. For developers of adaptive LLM-powered health information systems

Real-time user states estimation is feasible. Our findings show that behavioral and physiological signals (e.g., fixation and pupil size) varied across conditions, showing potential in predicting self-reported trust and source attribution. These results suggest the feasibility of integrating user-state modeling into adaptive health information systems, echoing recent efforts in Human-Computer Interaction that leverage physiological signals to guide interactive system design and development (Chiossi et al., 2024). For instance, Boonprakong et al. (2023) develop bias-aware systems that use physiological data for cognitive load estimation Ahmad et al. (2020), and study (Ajenaghughrure et al., 2021) predicts trust using psychophysiological measures. Understanding users' implicit states has the potential to enable the health system to better support health information seeking, flag moments of confusion or disengagement, and ultimately improve trust perceptions in health information.

Toward "disclosure-aware" interfaces. Building on this, our findings suggest the opportunity to build "disclosure-aware" health systems or interfaces that can dynamically adjust the transparency labels based on real-time user states. For example, when the system detects low attention (e.g., reduced fixations), it could highlight source labels to encourage more critical engagement. Conversely, when signs of cognitive overload or skepticism emerge (e.g., sustained focus on labels, increased pupil dilation), the system could simplify or temporarily de-emphasize the label to prevent unnecessary distrust, particularly when the content is accurate and clearly presented. Moreover, such "disclosure-aware" interfaces could provide on-demand explanations of labels, giving users deeper transparency only when users seek it.

This vision resonates with the concept of attentive user interfaces by Hummel et al. (2018), which sense and respond to users' attentional states to ensure that key digital nudges are not overlooked. Extending this logic, transparency labels could be made on demand, surfacing prominently when attention is low, and simplifying when signs of overload or skepticism arise. Such attentional feedback loops point toward health information systems and interfaces that are not only disclosure-aware but also attention-adaptive, dynamically balancing clarity, trust, and cognitive load.

Overall, this work advances HCI efforts to design AI health information systems and user interfaces that are not only transparent but also trust-aware and adaptive. By revealing how users respond to different information sources and disclosure labels, our findings offer actionable insights for both designers and developers. These considerations can help calibrate trust more effectively, reducing over-reliance, mitigating undue skepticism, and ensuring that both AI- and human-generated health information are presented in ways that support informed judgment.

5.4. Limitations and future work

Our study had several limitations that should be considered when interpreting the findings.

First, while our findings suggest that behavioral and physiological signals show potential in reflecting trust-related responses, we caution against overinterpreting them as direct indicators of trust, a complex and subjective construct (Liu et al., 2023; Johnson et al., 2015; Vereschak et al., 2024). Such signals can be influenced by unrelated factors like attention, physical arousal, or contextual noise (Cacioppo et al., 2016). Without careful contextualization, these signals could be misinterpreted as significant in scenarios where they merely represent contextual noise. Future research should integrate additional modalities (e.g., fNIRS Boonprakong et al., 2023, EEG Michalkova et al., 2024) to more robustly capture underlying cognitive states. Moreover, translating these findings into real-world applications (e.g., web-based gaze tracking (Mounica et al., 2019), rPPG from facial videos (McDuff et al., 2014)) raises ethical concerns regarding consent, data privacy, and potential over-reliance on AI (Wang et al., 2023; Friedman et al., 1999). Hence, any deployment must adhere to legal regulations (e.g., European AI Act (Act, 2024)) and prioritize continuous consent based on on-device security and privacy controls.

Second, the controlled lab environment may have influenced participants' responses, as being observed might heighten scrutiny of AI-labeled information, potentially amplified by societal caution toward AI. However, such a "mere observer effect", is likely just typical

for controlled psychological experimental conditions, where participant awareness of observation can subtly affect behavior (Cacioppo et al., 1990). While these settings are valuable for minimizing external confounders and ensuring reliable comparisons across conditions, future studies should nevertheless validate these findings in real-world environments to account for potential differences in naturalistic behaviors.

Third, our study measured trust at a single time point and relied on self-reports rather than actual decision-making actions. While this provides initial insights, trust is inherently dynamic and context-sensitive, often influencing real-world decision-making under uncertainty (Sillence et al., 2019). Capturing only static trust ratings may miss how trust evolves over interactions or translates into behavior, such as whether individuals follow AI- vs. human-sourced advice. Future work should adopt longitudinal, action-oriented paradigms (e.g., Ecological Momentary Assessment (Crosby et al., 2016)) to better reflect how trust evolves over time and influences real-life health decisions. This would yield a more ecologically valid understanding of trust in LLM-powered health context.

Fourth, while all LLM-generated responses were reviewed for consistency with human-authored content, we did not explicitly screen for stylistic aspects such as tone, clarity, or writing uniformity, which may influence perceived trust. Besides, this study focused on a single LLM (GPT-4o), and the findings may not generalize across other models (e.g., Claude, Gemini, Llama), which vary in output quality and style. Moreover, we did not include an in-task manipulation check to assess whether participants consciously perceived the actual source behind the labeled information. However, we acknowledge that perceived source awareness could influence trust independently of disclosed labels. Future work should evaluate the role of stylistic linguistic features across different LLMs, with regard to trust in AI. Additionally, to better understand how users react to AI-generated content, future studies should incorporate perceived-source ratings (e.g., post-task questionnaires or detectability checks) to assess whether trust judgments are mediated by users' ability to distinguish AI- from human-authored responses.

Lastly, our participant sample (notably WEIRD Linxen et al., 2021) across both studies was not representative of the general population, further limiting generalization. This is particularly relevant for groups with varying levels of AI literacy or differing baseline trust in technology. Acknowledging this limitation helps specify to whom these findings most apply. Nevertheless, our study provides a key initial step toward understanding the impact of source and labeling in online health information. Future expansion to include participants from varied demographics can enhance our understanding of how trust in health information is perceived across different groups.

6. Conclusion

Through a mixed-methods crowdsourcing survey (N = 142) and within-subjects lab study (N = 40), we found that AI-generated health information is trusted more than content sourced by human professionals, regardless of labeling, while human labels are trusted more than AI labels. Furthermore, we found that trust perceptions in personal health information are not only influenced by the source and label but also vary at behavioral and physiological levels. Our work highlighted the importance of considering AI transparency labels when measuring trust in online health information, and in developing techniques for verifying subjective trust perceptions and automatically inferring if and when to apply transparency labels based on sensed behavioral and physiological data. As such, we invite future research on understanding and designing for the physiology of online human-AI interactions, within and beyond AI-powered health information systems.

CRediT authorship contribution statement

Xin Sun: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rongjun Ma:** Writing – review & editing, Writing – original draft, Formal analysis. **Shu Wei:** Writing – review & editing, Writing – original draft, Conceptualization. **Pablo Cesar:** Supervision, Methodology, Conceptualization. **Jos A. Bosch:** Writing – review & editing, Supervision, Methodology, Funding acquisition. **Abdallah El Ali:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jos A. Bosch reports that financial support was provided by the University of Amsterdam. Xin Sun reports a relationship with the University of Amsterdam that includes: employment and funding grants. Jos A. Bosch also reports a relationship with the University of Amsterdam that includes: employment and funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded by the European Commission under the Horizon H2020 scheme, awarded to Jos A. Bosch (TIMELY Grant Agreement ID: 101017424).

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.ijhcs.2025.103714.

Data availability

Data will be made available on request.

References

Anvari, F., Efendic, E., Olsen, J., Arslan, R.C., Elson, M., Schneider, I.K., 2023. Bias in self-reports: an initial elevation phenomenon. Soc. Psychol. Pers. Sci. 14 (6), 727–737. https://doi.org/10.1177/19485506221129160

Ajenaghughrure, I.B., Sousa, S.D.C., Lamas, D., 2020. Measuring trust with psychophysiological signals: a systematic mapping study of approaches used. Multimodal Technol. Interact. 4 (3), https://doi.org/10.3390/mti4030063. https://www.mdpi.com/2414-4088/4/3/63.

Akash, K., Hu, W.-L., Jain, N., Reid, T., Nov 2018. A classification model for sensing human trust in machines using EEG and gsr. ACM Trans. Interact. Intell. Syst. 8 (4), https://doi.org/10.1145/3132743

Ahmad, M., Alzahrani, A., 2023. Crucial clues: investigating psychophysiological behaviors for measuring trust in human-robot interaction. In: Proceedings of the 25th International Conference on Multimodal Interaction, ICMI '23. Association for Computing Machinery, New York, NY, USA, pp. 135–143, https://doi.org/10.1145/3577190.3614148

Ahmad, M.I., Keller, I., Robb, D.A., Lohan, K.S., 2020. A framework to estimate cognitive load using physiological data. Pers. Ubiquitous Comput. 27 (6), 2027–2041. https://doi.org/10.1007/s00779-020-01455-7

Abdrabou, Y., Karypidou, E., Alt, F., Hassib, M., 2023. Investigating User Behavior Towards Fake News on Social Media Using Gaze and Mouse Movements. https://doi.org/10.14722/usec.2023.232041

Ayres, P., Lee, J.Y., Paas, F., van Merriënboer, J.J.G., 2021. The validity of physiological measures to identify differences in intrinsic cognitive load. Front. Psychol. 12, 702538.

Ajenaghughrure, I.B., Da Costa Sousa, S.C., Lamas, D., 2021. Psychophysiological modelling of trust in technology: comparative analysis of algorithm ensemble methods. In: 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 000161–000168, https://doi.org/10.1109/SAMI50585.2021.9378655

Arsham, H., Lovric, M., 2011. Bartlett's Test. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 87–88. https://doi.org/10.1007/978-3-642-04898-2_132

ATLAS.Ti, Sep. 2024. https://atlasti.com.

T. AB, 2024. Tobii Pro Lab, computer software. http://www.tobii.com/.

Alamudun, F., Choi, J., Gutierrez-Osuna, R., Khan, H., Ahmed, B., 2012. Removal of subject-dependent and activity-dependent variation in physiological measures of stress. In: 2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, pp. 115–122, https://doi.org/10.4108/icst.pervasivehealth.2012.248722

Act, E.A.I., 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on Artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013,

(EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial intelligence Act) (text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2024/1689/oj http://data.europa.eu/eli/reg/2024/1689/oj (Accessed 12-Sep-2024).

Bates, B.R., Romina, S., Ahmed, R., Hopson, D., 2006. The effect of source credibility on consumers' perceptions of the quality of health information on the internet. Med. Inform. Internet Med. 31 (1), 45–52.

Broom, A., 2005. The Emale: Prostate Cancer, masculinity and online support as a challenge to medical expertise. J. Sociol. (Melb) 41 (1), 87–104.

Babaei, E., Tag, B., Dingler, T., Velloso, E., 2021. A critique of electrodermal activity practices at CHI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3411764.3445370

Biswas, S.S., 2023. Role of Chatgpt in public health. Ann. Biomed. Eng. 51, 868–869. https://doi.org/10.1007/s10439-023-03172-7

Bansal, G., Warkentin, M., 2022. Do you still trust? The role of age, gender, and privacy concern on trust after insider data breaches. SIGMIS Database 52 (4), 9–44. https://doi.org/10.1145/3508484.3508487

Bickmore, T., Gruber, A., Picard, R., 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. Patient Educ. Couns. 59 (1), 21–30. 10.1016/j.pec.2004.09.008.

Ben Abacha, A., Demner-Fushman, D., 2019. A question-entailment approach to question answering. BMC Bioinform. 20 (1), 511:1–511:23.

Bridley, A.L.W.D. Jr., 2013. Module 3: clinical assessment, diagnosis, and treatment &x2013; fundamentals of psychological disorders — opentext.wsu.edu. https://opentext.wsu.edu/abnormal-psych/chapter/module-3-clinical-assessment-diagnosis-and-treatment/?utm_source=chatgpt.com.

Balogh, E.P., Miller, B.T., Ball, J.R., 2015. Improving Diagnosis in Health Care. The National Academies Press, Washington, DC, https://doi.org/10.17226/21794. https://nap.nationalacademies.org/catalog/21794/improving-diagnosis-in-health-care.

Benedek, M., Kaernbach, C., 2010. A continuous measure of phasic electrodermal activity. J. Neurosci. Methods 190 (1), 80–91.

Burrus, O., Curtis, A., Herman, L., 2024. Unmasking AI: informing authenticity decisions by labeling AI-generated content. Interactions 31 (4), 38–42. https://doi.org/10.1145/3665321

Boonprakong, N., Chen, X., Davey, C., Tag, B., Dingler, T., 2023. Bias-aware systems: exploring indicators for the occurrences of cognitive biases when facing different opinions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '23, New York, NY, USA, https://doi.org/10.1145/3544548.3580917

Cline, R.J.W., Haynes, K.M., 2001. Consumer health information seeking on the internet: the state of the art. Health Educ. Res. 16 (6), 671–692. https://doi.org/10.1093/her/16.6.671. https://academic.oup.com/her/article-pdf/16/6/671/9809432/160671.pdf.

Chen, J., Mishler, S., Hu, B., 2021. Automation error type and methods of communicating Automation reliability affect trust and performance: an empirical study in the cyber domain. IEEE Trans. Hum.-Mach. Syst. 51 (5), 463–473. https://doi.org/10.1109/THMS.2021.3051137

Chiossi, F., Stepanova, E.R., Tag, B., Perusquia-Hernandez, M., Kitson, A., Dey, A., Mayer, S., El Ali, A., 2024. Physiochi: towards best practices for integrating physiological signals in HCI. In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, CHI EA '24. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3613905.3636286

Carlbring, P., Hadjistavropoulos, H., Kleiboer, A., Andersson, G., 2023. A new era in internet interventions: the advent of chat-gpt and ai-assisted therapist guidance. Internet Interv. 32.

Carolus, A., Koch, M.J., Straka, S., Latoschik, M.E., Wienrich, C., 2023. Mails - meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. Comput. Hum. Behav.: Artif. Hum. 1 (2), 100014. 10.1016/j.chbah.2023.100014.

Cacioppo, J.T., Tassinary, L.G., Berntson, G.G., 2016. Strong Inference in Psychophysiological Science, Cambridge Handbooks in Psychology. Cambridge University Press, pp. 3–15.

Cacioppo, J.T., Rourke, P.A., Marshall-Goodell, B.S., Tassinary, L.G., Baron, R.S., 1990. Rudimentary physiological effects of mere observation. Psychophysiology 27 (2), 177–186. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1990.tb00368.x. https://doi.org/10.1111/j.1469-8986.1990.tb00368.x.

Crosby, R.D., Lavender, J.M., Engel, S.G., Wonderlich, S.A., 2016. Ecological Momentary Assessment. Springer Singapore, Singapore, pp. 1–3. https://doi.org/10.1007/978-981-287-087-2_159-1

Desai, A.N., Ruidera, D., Steinbrink, J.M., Granwehr, B., Lee, D.H., 2022. Misinformation and disinformation: the potential disadvantages of social media in infectious disease and how to combat them. Clin. Infect. Dis. 74, e34–e39.

Dutta-Bergman, M., et al., 2003. Trusted online sources of health information: differences in demographics, health beliefs, and health-information orientation. J. Med. Internet Res. 5 (3), e893.

Davis, F., Davis, F., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13, 319. https://doi.org/10.2307/249008

di Sciascio, C., Veas, E., Barria-Pineda, J., Culley, C., 2020. Understanding the effects of control and transparency in searching as learning. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20. Association for Computing Machinery, New York, NY, USA, pp. 498–509, https://doi.org/10.1145/3377325.3377524

Dalton, J., Fischer, S., Owoicho, P., Radlinski, F., Rossetto, F., Trippas, J.R., Zamani, H., 2022. Conversational information seeking: theory and application. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22. Association for Computing Machinery, New York, NY, USA, pp. 3455–3458, https://doi.org/10.1145/3477495.3532678

De Freitas, J., Agarwal, S., Schmitt, B., Haslam, N., 2023. Psychological factors underlying attitudes toward AI tools. Nat. Hum. Behav. 7 (11), 1845–1854.

Eurostat, 2022. Survey on the Use of Ict in Households and by Individuals.

El Ali, A., Venkatraj, K.P., Morosoli, S., Naudts, L., Helberger, N., Cesar, P., 2024. Transparent AI disclosure obligations: who, what, when, where, why, how. In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, CHI EA '24. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3613905.3650750

Elo, S., Kyngäs, H., 2008. The qualitative content analysis process. J. Adv. Nurs. 62 (1), 107–115.

Epstein, Z., Fang, M.C., Arechar, A.A., Rand, D.G., Jul 2023. What label should be applied to content produced by generative AI? https://doi.org/10.31234/osf.io/v4mfz. osf.io/preprints/psyarxiv/v4mfz.

Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., Treinen, M., 2001. What makes web sites credible? A report on a large quantitative study. In: Proceedings of the CHI 2001 Conference on Human Factors in Computing Systems Vol. 3, pp. 61–68, https://doi.org/10.1145/365024.365037

Fogg, B.J., Marshall, J., Osipovich, A., Varma, C., Laraki, O., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., Treinen, M., 2000. Elements that affect web credibility: early results from a self-report study. In: CHI '00 Extended Abstracts on Human Factors in Computing Systems, CHI EA '00. Association for Computing Machinery, New York, NY, USA, pp. 287–288, https://doi.org/10.1145/633292.633460

Flanagin, A.J., Metzger, M.J., 2007. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. New Media Soc. 9 (2), 319–342. https://doi.org/10.1177/1461444807075015. 10.1177/1461444807075015.

Friedman, B., Thomas, J.C., Grudin, J., Nass, C., Nissenbaum, H., Schlager, M., Shneiderman, B., 1999. Trust me, i'm accountable: trust and accountability online. In: CHI '99 Extended Abstracts on Human Factors in Computing Systems, CHI EA '99. Association for Computing Machinery, New York, NY, USA, pp. 79–80, https://doi.org/10.1145/632716.632766

Flanagin, A., Metzger, M., 2000. Perceptions of internet information credibility. Journalism Mass Commun Q. 77, 515–540. https://doi.org/10.1177/107769900007700304

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: a flexible statistical Power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191.

Freedman, D., Pisani, R., Purves, R., 2007. Statistics (international student edition). Pisani, R. Purves, fourth WW Norton & Company, New York.

Formosa, P., Rogers, W., Griep, Y., Bankins, S., Richards, D., 2022. Medical AI and human dignity: contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. Comput. Hum. Behav. 133, 107296. https://doi.org/10.1016/j.chb.2022.107296.

Garg, R.K., Urs, V.L., Agrawal, A.A., Chaudhary, S.K., Paliwal, V., Kar, S.K., 2023. Exploring the role of Chatgpt in patient care (diagnosis and treatment) and medical research: a systematic review. medRxiv https://doi.org/10.1101/2023.06.13.23291311

Guo, Y., Dec 2022. Digital trust and the reconstruction of trust in the digital society: an integrated model based on trust theory and expectation confirmation theory. Digit. Gov.: Res. Pract. 3 (4), https://doi.org/10.1145/3543860

Hesse, B.W., Nelson, D.E., Kreps, G.L., Croyle, R.T., Arora, N.K., Rimer, B.K., Viswanath, K., 2005. Trust and sources of health information: the impact of the internet and its implications for health care providers: findings from the first health information National Trends Survey. Arch. Intern. Med. 165 (22), 2618–2624.

Holmqvist, K., Nystrom, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J., 2011. Eye Tracking: A Comprehensive Guide to Methods and Measures. Oxford University Press, United States.

Haynes, W., 2013. Benjamini–Hochberg Method. Springer New York, New York, NY, p. 78. https://doi.org/10.1007/978-1-4419-9863-7_1215

Haynes, W., 2013. Bonferroni Correction. Springer New York, New York, NY, p. 154. https://doi.org/10.1007/978-1-4419-9863-7_1213

Hardin, J.W., Hilbe, J.M., 2012. Generalized Estimating Equations, Second Edition, second ed. Chapman & Hall/CRC, Philadelphia, PA.

Hristidis, V., Ruggiano, N., Brown, E.L., Ganta, S.R.R., Stewart, S., 2023. Chatgpt VS Google for queries related to dementia and other cognitive decline: comparison of results. J. Med. Internet Res. 25, e48966. https://doi.org/10.2196/48966. https://www.jmir.org/2023/1/e48966.

Horton, Jr, C.B., White, M.W., Iyengar, S.S., 2023. Bias against AI art can enhance perceptions of human creativity. Sci. Rep. 13 (1), 19001.

Hummel, D., Toreini, P., Maedche, A., 2018. Improving digital nudging using attentive user interfaces: theory development and experiment design. In: 13th International Conference on Design Science Research in Information Systems and Technology (DESRIST), Chennai, India, 3rd – 6th June, 2018, pp. 1–8.

Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., Scoville, E., Reese, T., Friedman, D., Bastarache, J., van der Heijden, Y., Wright, J., Carter, N., Alexander, M., Choe, J., Wheless, L., Feb 2023. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. Research Square. https://doi.org/10.21203/rs.3.rs-2566942/v1

Johnson, F.C., Rowley, J.E., Sbaffi, L., 2015. Modelling trust formation in health information contexts. J. Inf. Sci. 41, 415–429. https://api.semanticscholar.org/CorpusID:206454953.

Jakesch, M., French, M., Ma, X., Hancock, J.T., Naaman, M., 2019. AI-mediated communication: how the perception that profile text was written by AI affects trustworthiness. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19. Association for Computing Machinery, New York, NY, USA, pp. 1–13, https://doi.org/10.1145/3290605.3300469

Ji, K., Hettiachchi, D., Salim, F.D., Scholer, F., Spina, D., 2024. Characterizing information seeking processes with multiple physiological signals. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Vol. 5 of SIGIR 2024. ACM, pp. 1006–1017, https://doi.org/10.1145/3626772.3657793. http://dx.doi.org/10.1145/3626772.3657793.

Ji, K., Spina, D., Hettiachchi, D., Scholer, F., Salim, F.D., 2023. Towards detecting tonic information processing activities with physiological data. In: Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing, UbiComp/ISWC '23 Adjunct. Association for Computing Machinery, New York, NY, USA, pp. 1–5, https://doi.org/10.1145/3594739.3610679

Jessup, S., Schneider, T., Alarcon, G., Ryan, T., Capiola, A., Jun 2019. The measurement of the propensity to trust Automation. https://doi.org/10.1007/978-3-030-21565-1_32.

Just, M.A., Carpenter, P.A., 1980. A theory of reading: from eye fixations to comprehension. Psychol. Rev. 87 (4), 329–354.

Kerstan, S., Bienefeld, N., Grote, G., Sep 2023. Choosing human over AI doctors? How comparative trust associations and knowledge relate to risk and benefit perceptions of AI in healthcare. Risk Anal. 44, https://doi.org/10.1111/risa.14216

Kohn, S.C., de Visser, E.J., Wiese, E., Lee, Y.-C., Shaw, T.H., 2021. Measurement of trust in automation: a narrative review and reference guide. Front. Psychol. 12, 604977.

Kizilcec, R.F., 2016. How much information? Effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16. Association for Computing Machinery, New York, NY, USA, pp. 2390–2395, https://doi.org/10.1145/2858036.2858402

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y.H., Koo, B.-H., 2018. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig. 15 (3), 235–245.

Kherad-Pajouh, S., Renaud, O., 2015. A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs. Stat. Pap. 56 (4), 947–967.

Kunjan, S., Grummett, T.S., Pope, K.J., Powers, D.M.W., Fitzgibbon, S.P., Bastiampillai, T., Battersby, M., Lewis, T.W., 2021. The necessity of leave one subject out (loso) cross validation for EEG disease diagnosis. In: Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings. Springer-Verlag, Berlin, Heidelberg, pp. 558–567, https://doi.org/10.1007/978-3-030-86993-9_50

Kelley, P.G., Bresee, J., Cranor, L.F., Reeder, R.W., 2009. A "nutrition label" for privacy. In: Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS '09. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/1572532.1572538

Liu, J., Zhang, Y., Kim, Y., 2023. Consumer health information quality, credibility, and trust: an analysis of definitions, measures, and conceptual dimensions. In: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR '23. Association for Computing Machinery, New York, NY, USA, pp. 197–210, https://doi.org/10.1145/3576840.3578331

Logg, J.M., Minson, J.A., Moore, D.A., 2019. Algorithm appreciation: people prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. 151, 90–103.

Lim, J.Z., Mountstephens, J., Teo, J., 2022. Eye-tracking feature extraction for biometric machine learning. Front. Neurorobot. 15, 796895.

Lee, P., Bubeck, S., Petro, J., 2023. Benefits, limits, and risks of gpt-4 as an AI chatbot for medicine. N. Engl. J. Med. 388 (13), 1233–1239, pMID: 36988602. arXiv:10.1056/NEJMsr2214184, https://doi.org/10.1056/NEJMsr2214184

Lee, J.D., See, K.A., 2004. Trust in automation: designing for appropriate reliance. Hum. Factors 46 (1), 50–80.

Lucassen, T., Schraagen, J.M., 2010. Trust in Wikipedia: how users trust information from an unknown source. In: Proceedings of the 4th Workshop on Information Credibility, WICOW '10. Association for Computing Machinery, New York, NY, USA, pp. 19–26, https://doi.org/10.1145/1772938.1772944

Liao, Q.V., Sundar, S.S., 2022. Designing for responsible trust in AI systems: a communication perspective. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22. Association for Computing Machinery, New York, NY, USA, pp. 1257–1268, https://doi.org/10.1145/3531146.3533182

Lu, Y., Sarter, N., 2019. Eye tracking: a process-oriented method for inferring trust in Automation as a function of priming and system reliability. IEEE Trans. Hum.-Mach. Syst. PP 1–9. https://doi.org/10.1109/THMS.2019.2930980

Liashchynskyi, P., Liashchynskyi, P., 2019. Grid search, random search, genetic algorithm: A big comparison for nas. arXiv:1912.06059, https://arxiv.org/abs/1912.06059.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 4768–4777.

Liao, Q.V., Subramonyam, H., Wang, J., Wortman Vaughan, J., 2023. Designerly understanding: information needs for model transparency to support design ideation for ai-powered user experience. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3544548.3580652

Liao, Q.V., Vorvoreanu, M., Subramonyam, H., Wilcox, L., 2024. Ux matters: the critical role of ux in responsible AI. Interactions 31 (4), 22–27. https://doi.org/10.1145/3665504

Linxen, S., Sturm, C., Brühlmann, F., Cassau, V., Opwis, K., Reinecke, K., 2021. How weird is CHI? In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3411764.3445488

MAYO CLINIC, 2023. Mayo Clinic. https://www.mayoclinic.org/.

Marecos, J., Tude Graça, D., Goiana-da Silva, F., Ashrafian, H., Darzi, A., 2024. Source credibility labels and other nudging interventions in the context of online health misinformation: a systematic literature review. Journalism and Media 5 (2), 702–717. https://doi.org/10.3390/journalmedia5020046. https://www.mdpi.com/2673-5172/5/2/46.

Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An integrative model of organizational trust. Acad. Manag. Rev. 20 (3), 709–734. http://www.jstor.org/stable/258792.

Metzger, M., Flanagin, A., 2013. Credibility and trust of information in online environments: the use of cognitive heuristics. J. Pragmat. 59, 210–220. https://doi.org/10.1016/j.pragma.2013.07.012

Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., Chen, Y., Haas, B.W., Becker, B., 2023. Trust toward humans and trust toward artificial intelligence are not associated: initial insights from self-report and neurostructural brain imaging. Pers. Neurosci. 6, e3. https://doi.org/10.1017/pen.2022.5

McDonald, N., Schoenebeck, S., Forte, A., Nov. 2019. Reliability and inter-rater reliability in qualitative research: norms and guidelines for cscw and HCI practice. Proc. ACM Hum.-Comput. Interact. 3 (CSCW), https://doi.org/10.1145/3359174

Michalkova, D., Rodriguez, M.P., Moshfeghi, Y., Jan 2024. Understanding feeling-of-knowing in information search: an EEG study. ACM Trans. Inf. Syst. 42 (3), https://doi.org/10.1145/3611384

Mounica, M.S., Manvita, M., Jyotsna, C., Amudha, J., 2019. Low cost eye gaze tracker using web camera. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 79–85, https://doi.org/10.1109/ICCMC.2019.8819645

McDuff, D.J., Gontarek, S., Picard, R.W., 2014. Remote measurement of cognitive stress via heart rate variability. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2957–2960. https://api.semanticscholar.org/CorpusID:206627980.

National Institutes of Health, 2023. National Institutes of Health. https://www.nih.gov.

Norman, C.D., Skinner, H.A., 2006. Eheals: the ehealth literacy scale. J. Med. Internet Res. 8 (4), e27. https://doi.org/10.2196/jmir.8.4.e27

OpenAI, https://openai.com/chatgpt/.

OpenAI et al., 2024. Gpt-4 technical report. arXiv:2303.08774.

Opitz, J., Burst, S., 2021. Macro f1 and macro f1. arXiv:1911.03347, https://arxiv.org/abs/1911.03347.

Parikh, S.S., 2018. Eye gaze feature classification for predicting levels of learning. https://api.semanticscholar.org/CorpusID:53471366.

Prolific, 2014. https://www.prolific.com.

Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K., 2019. Psychopy2: experiments in behavior made easy. Behav. Res. Methods 51 (1), 195–203.

Pennycook, G., Bear, A., Collins, E.T., Rand, D.G., 2020. The implied truth effect: attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Manage. Sci. 66 (11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478.

Reis, M., Reis, F., Kunde, W., Jul. 2024. Influence of believed AI involvement on the perception of digital medical advice. Nat. Med. 30 (11), 3098–3100.

Rathi, I.M., Taylor, S., Bergen, B., Jones, C., 2025. Gpt-4 is judged more human than humans in displaced and inverted turing tests. In: Alam, F., Nakov, P., Habash, N., Gurevych, I., Chowdhury, S., Shelmanov, A., Wang, Y., Artemova, E., Kutlu, M., Mikros, G. (Eds.), Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect). International Conference on Computational Linguistics, Abu Dhabi, UAE, pp. 96–110.

Rae, I., 2024. The effects of perceived AI use on content perceptions. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3613904.3642076

Rheu, M., Shin, J.Y., Peng, W., Huh-Yoo, J., 2020. Systematic review: trust-building factors and implications for conversational agent design. Int. J. Hum.-Comput. Interact. 37, 1–16. https://doi.org/10.1080/10447318.2020.1807710

Rowley, J.E., Johnson, F.C., Sbaffi, L., 2015. Students' trust judgements in online health information seeking. Health Inform. J. 21, 316–327.

Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychol. Bull. 124 (3), 372–422.

Sillence, E., Briggs, P., Harris, P.R., Fishwick, L., 2007. How do patients evaluate and make use of online health information? Soc. Sci. Med. 64 (9), 1853–1862. https://doi.org/10.1016/j.socscimed.2007.01.012

Sillence, E., Briggs, P., Fishwick, L., Harris, P., 2005. Guidelines for developing trust in health websites. In: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, WWW '05. Association for Computing Machinery, New York, NY, USA, pp. 1026–1027, https://doi.org/10.1145/1062745.1062851

Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G.A., Maes, P., 2024. People over trust ai-generated medical responses and view them to be as valid as doctors, despite low accuracy. arXiv:2408.15266, https://arxiv.org/abs/2408.15266.

Sillence, E., Briggs, P., Fishwick, L., Harris, P., 2004. Trust and mistrust of online health sites. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04. Association for Computing Machinery, New York, NY, USA, pp. 663–670, https://doi.org/10.1145/985692.985776

Singal, H., Kohli, S., 2016. Intellectualizing trust for medical websites. In: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, ICTCS '16. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2905055.2905293

Sillence, E., Blythe, J.M., Briggs, P., Moss, M., 2019. A revised model of trust in internet-based health information and advice: cross-sectional questionnaire study. J. Med. Internet Res. https://doi.org/10.2196/11125

Sundar, S.S., 2007. The main model: a heuristic approach to understanding technology effects on credibility. https://api.semanticscholar.org/CorpusID:17588424.
Sbaffi, L., Rowley, J., 2017. Trust and credibility in web-based health information: a review and agenda for future research. J. Med. Internet Res. 19 (6), e218. https://doi.org/10.2196/jmir.7579. http://www.jmir.org/2017/6/e218/.
Scharowski, N., Benk, M., KÃŒhne, S.J., Wettstein, L., BrÃŒhlmann, F., 2023. Certification labels for trustworthy AI: insights from an empirical mixed-method study. In: 2023 ACM Conference on Fairness, Accountability, and Transparency. ACM, Chicago IL USA, pp. 248–260, https://doi.org/10.1145/3593013.3593994. https://dl.acm.org/doi/10.1145/3593013.3593994.
Sümer, Ö., Bozkir, E., Kübler, T., Grüner, S., Utz, S., Kasneci, E., 2021. Fakenewsperception: an eye movement dataset on the perceived believability of news stories. Data in Brief 35, 106909. 10.1016/j.dib.2021.106909.
Sevcenko, N., Appel, T., Ninaus, M., Moeller, K., Gerjets, P., 2022. Theory-based approach for assessing cognitive load during time-critical resource-managing human–computer interactions: an eye-tracking study. J. Multimodal User Interfaces 17, 1–19. https://doi.org/10.1007/s12193-022-00398-y
SHAPIRO, S.S., WILK, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591
Sahagian, G., 2024. What is random state 42? — grsahagian.medium.com. https://grsahagian.medium.com/what-is-random-state-42.
Sun, X., Ma, R., Zhao, X., Li, Z., Lindqvist, J., Ali, A.E., Bosch, J.A., 2024. Trusting the search: Unraveling human trust in health information from google and chatgpt. arXiv:2403.09987, https://arxiv.org/abs/2403.09987.
Sundar, S.S., Kim, J., 2019. Machine heuristic: when we trust computers more than humans with our personal information. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19. Association for Computing Machinery, New York, NY, USA, pp. 1–9, https://doi.org/10.1145/3290605.3300768
Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V., 2023. Large language models encode clinical knowledge. Nature 620 (7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
Scott, A.E., Neumann, D., Niess, J., Woźniak, P.W., 2023. Do you mind? User perceptions of machine consciousness. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3544548.3581296
Tiwari, R., Kumar, R., Malik, S., Raj, T., Kumar, P., 2021. Analysis of heart rate variability and implication of different factors on heart rate variability. Curr. Cardiol. Rev. 17 (5), e160721189770.
Thielmann, B., Hartung, J., Böckelmann, I., 2022. Objective assessment of mental stress in individuals with different levels of effort reward imbalance or overcommitment using heart rate variability: a systematic review. Syst. Rev. 11 (1), 48.
Ul Haque, E., Khan, M.M.H., Fahim, M.A.A., 2023. The nuanced nature of trust and privacy control adoption in the context of Google. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3544548.3581387
Vereschak, O., Alizadeh, F., Bailly, G., Caramiaux, B., 2024. Trust in ai-assisted decision making: perspectives from those behind the system and those for whom the decision is made. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '24, New York, NY, USA, https://doi.org/10.1145/3613904.3642018
van Amsterdam, U., 2025. Fmg Research Lab — lab-fmg.uva.Nl. https://lab-fmg.uva.nl/en.
Van der Lans, R., Wedel, M., Pieters, R., 2011. Defining eye-fixation sequences across individuals and tasks: the binocular-individual threshold (bit) algorithm. Behav. Res. Methods. 43, 239–257.
Van Bulck, L., Moons, P., 2023. What if your patient switches from Dr. Google to Dr. Chatgpt? A vignette-based survey of the trustworthiness, value, and danger of Chatgpt-generated responses to health questions. Eur. J. Cardiovasc. Nurs. zvad038. https://doi.org/10.1093/eurjcn/zvad038
Wang, X., R.A., Cohen, 2022. Health Information Technology Use Among Adults: United States, July–December 2022. https://www.cdc.gov/nchs/products/databriefs/db482.htm.
Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., Tang, Y., 2023. A brief overview of Chatgpt: the history, status quo and potential future development. IEEE/CAA J. Autom. Sin. 10 (5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618
Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., Liu, J., 2023. Ethical considerations of using Chatgpt in health care. J. Med. Internet Res. 25, e48009. https://doi.org/10.2196/48009. https://www.jmir.org/2023/1/e48009.
Walker, F., Favetta, M., Hasker, L., Walker, R., 2024. They prefer humans! experimental measurement of student trust in Chatgpt. In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, CHI EA '24. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3613905.3650955
Wang, J., Aug 2018. Gaze behavior, skin conductance, and trust in Automation. http://essay.utwente.nl/76357/.
Wathen, C., Burkell, J., 2002. Believe it or not: factors influencing credibility on the web. JASIST 53, 134–144. https://doi.org/10.1002/asi.10016
WebMD, https://www.webmd.com/.
Wang, L., Stern, J.A., 2001. Saccade initiation and accuracy in gaze shifts are affected by visual stimulus significance. Psychophysiology 38 (1), 64–75.
Wang, L., 2019. Eye tracking methodology in screen-based usability testing. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19. Association for Computing Machinery, New York, NY, USA, pp. 1–3, https://doi.org/10.1145/3290607.3298811
Wobbrock, J.O., Findlater, L., Gergle, D., Higgins, J.J., 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11. Association for Computing Machinery, New York, NY, USA, pp. 143–146, https://doi.org/10.1145/1978942.1978963
Xu, R., Feng, Y., Chen, H., 2023. Chatgpt vs. google: A comparative study of search performance and user experience. arXiv:2307.01135, https://arxiv.org/abs/2307.01135.
Yin, Y., Jia, N., Wakslak, C.J., 2024. AI can help people feel heard, but an AI label diminishes this impact. Proc. Natl. Acad. Sci. 121 (14), e2319112121. https://doi.org/10.1073/pnas.2319112121. https://www.pnas.org/doi/pdf/10.1073/pnas.2319112121. https://www.pnas.org/doi/abs/10.1073/pnas.2319112121.
Zar, J.H., 2005. Spearman rank correlation. Encyclopedia of Biostatistics 7.