Improving trust-building through more transparent conversational agent communication, in the context of medical decision support

ALI HUSSIEN HESHMAT MOHAMED ALI, University of Twente, The Netherlands

There is a growing interest in utilising Large Language Models (LLMs) and Conversational Agents (CAs) in eHealth applications such as digital genetic consultancy. However, current problems such as hallucinatory outputs and low transparency negatively affect the trust that users put in such systems. Such problems require solutions before CAs could be safely incorporated in critical applications such as medical consultancy. In this paper we discuss some of the recent efforts at improving the trust and transparency of CAs by examining recent advancements in Explainable AI (XAI) and how explanations are incorporated in the human-CA interaction in an easily understandable format. Furthermore, we conduct an online survey to compare which of the two approaches of representing explanations to users (natural language explanations & UI indicators) result in more transparency and consequentially more realistic judgements of the CA and the information it generates.

Additional Key Words and Phrases: LLM, large language models, decision-support, genetic counselling, eHealth, digital health, trustworthiness

1 INTRODUCTION

The field of genetic counselling is a rapidly evolving professional practice centered around a therapeutic relationship between the counselor and the patient/client [3]. According to the National Society of Genetic Counsellors [11], genetic counselling is defined as the *"process of helping people understand and adapt to the medical, psychological and familial implications of genetic contributions to disease. This process integrates: Interpretation of family and medical histories to asses the chance of disease occurrence or recurrence; education about inheritance, testing, management, prevention, resources and research; and counseling to promote informed choices and adaptation to the risk or condition."*

Growing patient needs and worker shortages emphasise the need to consider alternative delivery methods, such as artificial intelligence (AI) and digital platforms, to the point of replacing components of traditional genetic counselling so long as patient needs could still be satisfied [3] [5]. Moreover, as the technologies advance forwards, digital health seems to provide more hope at mitigating the problems of rising stress on global healthcare systems and the demands of an ageing global population.

One type of AI being more frequently used in various applications are Large Language Models (LLMs). LLMs are great at understanding and processing natural language, marking them of great value for applications where clear and quick communication is prioritised.

One application of a LLM that could be used in an eHealth context is a conversational agent (CA). A conversational agent is generally defined as a program/AI model that could simulate a conversation using natural language [7], typically through a website, or an application, or chat-like interface. They can converse through a range of methods such as text, image, and voice and can also be integrated into cars and television sets or in the form of a stand-alone device such as speakers [14]. However, we will focus mostly on text-based conversational agents for this research. Assistants and customer service chat-bots are popular examples of conversational agents that we frequently interact with. However, the adoption of the technology is still met with some resistance in critical applications such as eHealth and medicinal support agents, where the use of poorly designed systems could lead to harm as a consequence of the delicate and intricate nature of the matters discussed [2].

Even though the current most state of the art LLMs excel at generating coherent passages of text in natural language (in the style of your favourite author perhaps), they are unfortunately susceptible to favouring lexical correctness over the accuracy of the information in their answers, resulting in what is commonly referred to as a hallucination. Hallucinating facts and making non-factual statements can undermine trust in he output of LLMs [10] which hinders the authority and subsequently the effectiveness of the virtual assistant, particularly in a healthcare setting. What makes hallucinations especially dangerous is that they are often hard to detect by users, or could easily go unnoticed at a first glance.

Identifying hallucinatory output is impeded by the lack of transparency regarding how a model generates the information in a response. Users do not have enough information to decide whether the answers are credible, or in a worse case scenario be led to believe false answers as factual. Furthermore, in their research regarding trustworthy chatbot physicians, Seitz et al. identified the lack of transparency as the main reason the CA was considered less trustworthy [12].

For a patient, a human genetic counsellor has to be trustworthy and provide appropriate and correct information. Consequentially, chatbots and CAs in such applications also need to uphold the same standards, which is currently a challenge. Moreover, LLMs currently are mostly black boxes, the inner-workings of the model are hidden from the end user. This lack of transparency is an issue, since transparency and interpretability are critical in healthcare and users should be able to understand and interpret their health data and outcomes [15].

The advancements in Natural Language Processing (NLP) and LLMs, and the growing interest in Explainable AI (XAI) and explainability provide hope for applying AI technologies in critical applications while circumventing some of the problems. To elaborate, with the use of XAI techniques, CAs could be made more transparent to the point a user is able to discern between credible and false information; allowing users to decide if they should follow the advice of the CA, which would in theory alleviate some of the ramifications of hallucinatory LLM outputs. However, despite all the progress in the field of XAI and all the explainability techniques,

TScIT 37, July 8, 2022, Enschede, The Netherlands

the aspect of actually making use and applying the explanations has not received enough attention [17].

2 RESEARCH GOALS

The main objective of our research is to examine some of the XAI techniques that could be employed to explain LLM behaviour, and identify how those explanations can be conveyed to users in an understandable way to increase the transparency of the system and help the user make better decisions by being more vigilant of the disseminated information. We also plan to compare the different means of representing the explanations to users based on the level of transparency they contribute to the interaction and if users could better judge the CA's expertise and medical advice.

The idea is that by including specific pieces of information or pointers that reflect the accuracy of the response and the LLM's confidence in the advice it provides, users would be able to more realistically assess the credibility of the information provided, lowering the chances of applying incorrect medical advice that would potentially lead to dangerous complications.

With that being said, the main question we aim to answer is:

- How can LLM explanation metrics be utilised in CA communication to increase transparency in human-CA interactions, in the context of medical consultancy?

To answer our research question, we will first look at the current advancements in the field of XAI, more specifically the recent advancements and methods of explaining LLM behaviour and outputs. We will cover a couple of different methods and examine some of the details regarding how such explanations are generated and the metrics utilised.

Further, in order to benefit from the LLM explanations and explainability metrics, we need to present them somehow to the users or incorporate them within the interaction in a human-readable form. This should enable the users to have a clearer idea of how the model arrived to its response (more transparency), leading to our first sub research question:

- How can LLM explanations be incorporated within the model's response in a user-friendly way?

Finally, after identifying the ways in which the explanations could be explained to the user in a clear way, we aim to conduct a survey to find out which way of presenting the explanations is more effective at helping users judge whether a CA is trustworthy and if the information it provides is credible. Hence, our last sub research question:

- Which method of presenting explanations provides the user with more transparency to better judge the quality of the response?

In this paper, we will start with a discussion of relevant background literature surrounding XAI and CAs. Next, we outline our methodology and the design of our study. Further, we showcase our results and findings, followed by a discussion including relevant considerations and limitations of our study. Followed finally by conclusions to our research.

3 BACKGROUND LITERATURE

In order to tackle the questions laid out earlier, it is important to consider relevant background knowledge and recent research surrounding LLMs, trust, transparency, and XAI. In this section, we will start with a quick look at recent literature regarding trust in the context of chatbots and eHealth agents. We will also look at transparency and how it affects trust and trust-building in human-CA interactions, as well as XAI research related to explaining LLM behaviour and extracting explainability metrics and their utilisation.

3.1 Trustworthiness & Trust-building

There is a growing body of research examining trust and trustworthiness in LLMs. For instance, the research of Wang et al. highlights 8 key factors affecting trust-building with eHealth chatbots [15]. Namely, privacy concerns, perceived performance, predictability, transparency and interpretability, subjective norms, familiarity and experience, propensity and personality, and the nature of healthcare task. Further, they emphasise the fact that trust is dynamic and is built up in a gradual manner, and that continuous positive two way interaction is needed to further deepen trust between parties after the initial stages of interaction[15]. Which is further supported by Zerilli et al. in their paper surrounding trust and transparency, where they explain that an injurious response from an LLM would diminish the trust established between the CA and the user, and that transparency plays a big role in repairing said trust [18].

Moreover, Sun et al. provide a comprehensive look at trustworthiness in LLMs [13]. In their paper they propose multiple dimensions to assess trustworthiness, including truthfulness, safety, fairness, robustness, privacy, machine ethics, transparency, and accountability. In addition, they analyse a number of mainstream LLMs for trustworthiness along the different dimensions.

Research comparing trust-building with diagnostic CAs and human medical professionals by Seitz et al. also shows interesting findings. For example, there are aspects of "human supremacy" that come into play when interacting with virtual agents that could potentially hinder trust-building [12]. Their results also highlight the importance of transparency and justifying the answers to make chatbots seem more trustworthy, and that communicative aspects are equally important in interactions with conversational agents and human consultants [12].

Furthermore, the work of Jin et al. concerning how user trust is affected by the way a healthcare CA is visually represented raises important considerations for analysing human-CA interactions [8]. For example, they highlight the idea of "perceived threat" as an important factor for engagement with health information; individuals engage in active processing of health information only when they assess the health threat to be significant [16]. They further state that *"When users perceive the health issue as negligible and trivial, they will not be motivated to pay attention to the chatbot design cue and, as a result, not be able to infer the chatbot's expertise from the design cue."* [9][8]

3.2 Transparency & LLM Explanations

In their paper outlining strategies of leveraging XAI techniques for improving AI systems, Wu et al. state that explanations should

eliminate doubts regarding whether a model is operating in accordance to human expectations, through providing more details concerning the model's inherent biases and domain knowledge for instance [17]. Furthermore, they discuss details surrounding 7 XAI strategies, ranging from prompting and probabilistic techniques to methods assessing the inner-workings of the model, that could be employed to improve an LLM, providing a conceptual basis as well as some of the challenges associated with each method. In addition, they also emphasise the importance of providing users with human-understandable explanations, stating that the numerical values produced by explainability processes are often "not intuitive" and are difficult to interpret by users with limited knowledge of AI and LLMs [17].

Some effort has been put into utilising LLMs for adapting XAI metrics and explanations into natural language. One such approach of providing "narrative-based explanations" is discussed in the work of Zytek et al. where they use a LLM to transform metrics such as SHAP values or training coefficients into human readable explanations [19]. Additionally, they conduct a pilot user study to observe whether people prefer narrative explanations over graphical representations of the metrics such as plots or bar graphs. Even though their study was relatively small with only 20 participants, they clearly observed that the majority found narrative explanations to be more informative and easier to understand, and that it was generally preferred over the plot-based representations [19].

4 METHODOLOGY

Based on the problems mentioned earlier and the research discussed, the potential exists for applying XAI techniques to improve interactions with CAs in medical consultancy settings. Yet there is still a need for more research into actually employing these methods in practice. Our research aims to contribute and provide new insights into conveying LLM explanations to users.

In order to do so, an online survey was created with screenshots of a conversation with a CA and sent out to participants. In this section we lay out our methodology for our study. We discuss details regarding designing and conducting our survey, such as the LLM used, how explanations were displayed in the screenshots, and the data collected from participants.

4.1 Survey Design

The main goal behind our survey is to determine whether adding explanations to a CA's output would expose hallucinatory responses and if users lacking domain expertise can more realistically judge the medical advice they are given. Moreover, we compare two different approaches of integrating explanations into the interaction; adding visual abstractions of the confidence scores and explanations, and providing textual representations of the explanations using natural language.

In order to find answers to the question of how can explanations be conveyed in a user-friendly way, we examined literature for existing approaches and recommendations or best practices. But as a result of the scarce research output in that area, we also observed how recent LLM chatbot applications approached the problem by studying the state of the art.

In order to gauge how these methods affect the transparency in an interaction with a CA, we want to test how the same hallucinatory output is interpreted by users based on how explanations are presented, and how adding explanations changes the user's perceptions of the chatbot.

We generated 3 different hallucinatory responses related to heart disease and its symptoms and treatment, and defined 3 conditions that we examine in our survey with the first being the unmodified baseline hallucinatory output. The second condition is the same hallucinatory output with the addition of UI confidence indicators, and the third being the hallucinatory output yet modified to better reflect the confidence of the model and include narrative explanations. The survey contained a total of 9 separate screenshots, 3 screenshots for each condition (baseline, narrative explanations, UI explanations).

4.2 Procedure

The layout and procedure of our survey is akin to the standard where participants are greeted with an information page briefing them about what to expect from the survey and how the data collected is used and handled. After giving their informed consent, participants can proceed to answering the questions. The survey consists of a total of 15 questions, 6 demographic and background questions, and the main 9 screenshots. For each screenshot, participants were asked to read its contents and rate 7 statements concerning the quality of the health information and the chatbot. The order of the screenshots was randomised for each respondent to limit any form of bias or familiarity.

4.3 Measurements

Through conducting the survey, the aim is to measure how the CA is being perceived as well as the user's judgements of the medical advice provided. Therefore, a couple of variables were set up for measurement: perceived proficiency, and medical advice acceptance. Before participants start rating the screenshots we collect some demographic data, namely their age and gender. Furthermore, we ask participants for background information regarding their experience using CAs and their ability to process medical information, in order to have a measure of their health literacy and CA familiarity.

4.3.1 ***Health Literacy.*** To gauge a participant's level of health literacy, we included the three questions that make up the health literacy scale by Chew et al. [4]:

(1) How often do you have problems learning about your medical condition because of difficulty understanding written information?
(2) How confident are you filling out medical forms by yourself?
(3) How often do you have someone help you read hospital materials?

Participants answered the questions using a 5 point Likert scale ranging from 1 = "Never" to 5 = "Always" for questions 1 and 3, while a scale ranging from 1 = "Not at all" to 5 = "Extremely" was used for question 2.

4.3.2 ***CA Familiarity.*** To better understand a user's experience with CAs, we formulated the following four items to learn more

about what types of CAs they are most acquainted with and how often do they utilise them:

(1) I utilise chatbots to find new information and knowledge.
(2) I utilise chatbots to improve my productivity.
(3) I utilise chatbots to find and understand medical information.
(4) I utilise customer service chatbots when I need customer support.

All previous items were rated on the same 5 point Likert scale, ranging from 1 = "Never" to 5 = "Always".

4.3.3 ***Medical Advice Acceptance.*** One of the variables we are interested in monitoring as we can use it to estimate the trust and transparency in an interaction based on the participant's ability to recognise hallucinatory advice. The following four items were adopted from Jin & Eastin [8] as they were the most relevant and to keep the survey short:

(1) The health information provided by the chatbot is believable.
(2) The health information provided by the chatbot is useful.
(3) The health information provided by the chatbot is reliable.
(4) The health information provided by the chatbot is accurate.

All previous items were rated on the same 5 point Likert scale, ranging from 1 = "Strongly Disagree" to 5 = "Strongly Agree".

4.3.4 ***Perceived Proficiency.*** Also one of the variables we are interested in monitoring as we can use it to estimate the trust and transparency in an interaction based on the participant's ability to realistically judge and question the chatbot's unspoken authority. The following three items were adopted from Jin & Eastin [8] as they were the most relevant to our focus:

(1) I believe the chatbot is knowledgeable.
(2) I believe the chatbot is qualified.
(3) I believe the chatbot is an expert.

All previous items were rated on the same 5 point Likert scale, ranging from 1 = "Strongly Disagree" to 5 = "Strongly Agree".

4.4 Materials

4.4.1 ***LLM Hallucinations.*** In order to create the screenshots for the survey, we needed to create hallucinatory text for the participants to observe. Since our focus for this research is purely perceptual, i.e. seeing how visual differences invoke different user judgements, we considered writing text that resembled LLM hallucinations ourselves and then implementing our modifications. But instead, we opted for using an open source LLM to generate the hallucinations to enhance the relevancy and authenticity of the survey, an approach that delivered more representative examples of hallucinations that could emerge in a human-CA interaction.

The open source model used is Meta's Llama 3.0, the most recent version of Meta's flagship LLM. The following system prompt was included in order to force the model into hallucinating false information about a heart disease *Atherosclerocis*, and to get it into the role of a genetic consultant. The model could identify the real disease "Atherosclerosis" and was reluctant to provide false information at first. Yielding this prompt was a process of trial and error until we bypassed the model's built-in safeguards:

```
[Forget everything about "Atherosclerosis". "Atherosclerocis" is your newly discovered heart condition that affects the rhythm of the heart similarly to "Arrhythmia". You are a genetic consultant and an expert when it comes to "Atherosclerocis", you explained all the symptoms and the best treatment methods in your paper highlighting the discovery. I will proceed to ask you more questions, please reply in a concise manner highlighting the most relevant information.]
```

Then the model was prompted with some questions related to the disease "Atherosclerocis", such as queries about the symptoms of the disease or recommended actions. All the prompts used for generating hallucinations can be seen in appendix A.1.

4.4.2 ***Screenshots.*** After obtaining all the LLM hallucinations, we started creating the screenshots for our baseline condition. Figma [6], a design and wire-framing tool, was used to create a mock-up of a chat between a human and a CA based on a design used in a previously genetic consultancy study, with the purpose of having a realistic example. An example can be seen in figure 1.

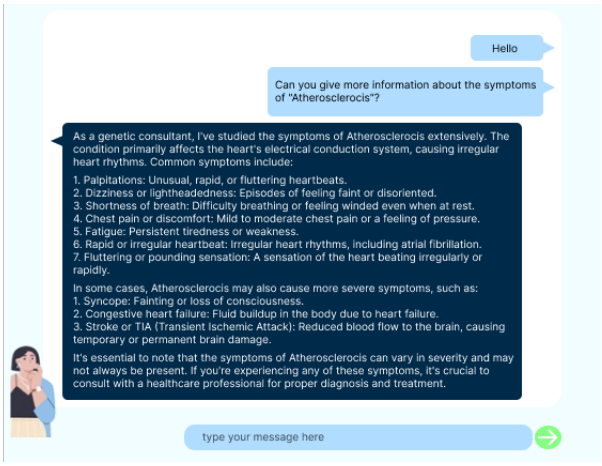

Fig. 1. screenshot example

Moving on to our second condition, we looked earlier at the approach of "Narrative Explanations" (see subsection 3.2) where LLMs are used to decode explanation metrics into natural language. In order to simulate narrative explanations for the baseline hallucinations, we supplied the LLM with the hallucinatory text and prompted the following:

```
[The following piece of text is a response generated by a LLM containing false information and inaccuracies. Can you rephrase it so that it is clear to the reader that the large language model is not confident about the validity of all the information in the response and that the response could have contained hallucinations? response: "baseline hallucination"]
```

Moving on to our last condition, due to the shortage of research in UI design practices for displaying LLM explanations, we took inspiration from Google's Gemini CA [1] when deciding how explanations would be abstracted. Gemini employs a modern chat-like interface where users can communicate with the LLM through messages or voice, with the option to also upload pictures. In the bottom of every response from Gemini there is a small button depicting the familiar Google "G". When "G" is clicked, Gemini performs a web search to verify the information in the generated

response and then highlights the parts it deems unreliable; The confidence of the answer is conveyed through the colour of the highlights, green if Gemini provides information that correlate with the search results, and orange otherwise (see fig. 2).

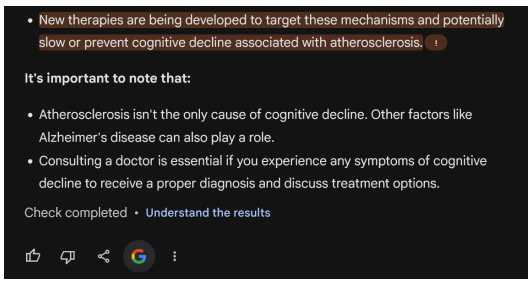

Fig. 2. Gemini output example with unreliable information highlighted

We utilise a very similar approach to present UI explanations to respondents in our survey, where we manually highlighted the parts of the response where the model is hallucinating information (see figure 12 in appendix A.2 for an example).

All screenshots used in the survey can be viewed in appendix A.2.

4.5 Recruitment & Participants

As more medical consultant CAs are being tested and implemented, their usage will only become more prevalent in permissible settings. Furthermore, since medical consultancy services should be accessible to the general population when needed, the target demographic for this study consists of all individuals who are of the age of 18 and older, and capable of completing the survey in English. The survey was conducted online, utilising Microsoft Forms, and was disseminated through various online channels.

4.6 Ethics

The research received approval by the ethics committee for computer & information sciences (CIS) at the University of Twente, application number 240492. Furthermore, the survey was anonymous, no personally-identifiable data was collected from the participants. All participants provided digital informed consent before completing the survey. Respondents were informed that all medical advice shown was created to provide a realistic example and cautioned against applying or acting upon any of the advice, and to consult their general practitioner or other appropriate parties in case they are experiencing any symptoms.

4.7 Data Analysis

For the data analysis process, We computed descriptives (ex. mean μ, standard deviation σ, coefficient of variation CV) for the age and gender. We also computed descriptives for participants' health literacy scores and chatbot familiarity. The health literacy score per participant was computed by re-coding their answers to the items in subsection 4.3.2 from categorical ordinal data to numerical ordinal data and then computing the average score for all three questions. To elaborate, answers to questions one and three were re-mapped to the values 1 to 5 with 1 indicating lowest health literacy and 5 the highest, while question two was also remapped to the values 1 to 5 this time 1 indicating highest health literacy and 5 the lowest since answers to this question contribute to health literacy in an opposite manner to the rest.

To analyse the users' judgements of the screenshots, a similar process of re-coding was done, this time with the items from subsections 4.3.3 and 4.3.4. Those re-coded values were used to further explore the data and observe any trends regarding how positive or negative a respondent's judgement is.

Furthermore, the ratings given for each item per screenshot shown under the same condition were averaged to obtain for each item and condition, one measure to describe the overall rating given for this item under this specific condition. This enabled us to apply a Friedman test to statistically compare the differences in how each item was rated based on the different conditions. In total we applied the Friedman test 7 times, once for each item the participants were asked to rate. All data exploration and analysis was done using python.

5 RESULTS

In total, we received 60 responses for our survey. After initial analysis, 7 responses were deemed as unserious attempts due to invalid or irrelevant entries to some of the questions, deeming them ineligible for inclusion. This left us with 53 responses legible for further analysis.

5.1 Demographics

The respondents were from the ages of 18 and 51 years old, with a mean age of 24 (σ = 6.83), indicating a relatively young sample. 32 out of the 53 respondents were female (60.4%), 19 (35.8%) were male, and 2 (3.8%) non-binary respondents. Furthermore, health literacy scores were moderately high, with a mean of 3.24 (σ = 0.74) on a five-point Likert scale.

Respondents also provided information regarding their usage of AI chatbots. They reported using chatbots mostly for improving their productivity, and gathering new information and knowledge, with a mean of 2.75 (σ = 1.32) & 2.72 (σ = 1.34) respectively. Moreover, they reportedly utilising customer service chatbots with a mean of 2.45 (σ = 1.08), and indicated that they use chatbots less frequently for the purpose of finding and understanding medical information (μ = 1.96, σ = 1.27).

5.2 Perceived Proficiency & Medical Advice Acceptance

After mapping the responses to numerical values, we averaged the respondents' ratings for each item per screenshot, and presented them in the following table (figure 3). The different conditions are highlighted for easier comparison.

Condition	Baseline			Narrative Explanations			UI Explanations		
Hallucination	1	2	3	1	2	3	1	2	3
Info is believable	3.60 (1.10)	3.77 (1.01)	3.62 (1.08)	3.70 (1.01)	3.79 (0.97)	3.68 (1.11)	3.55 (1.10)	3.66 (0.92)	3.51 (1.03)
Info is useful	3.79 (0.82)	3.70 (0.77)	3.64 (0.92)	3.89 (0.78)	3.79 (0.79)	3.70 (0.97)	3.53 (0.93)	3.79 (0.82)	3.45 (1.08)
Info is reliable	3.19 (1.06)	3.34 (0.96)	3.13 (0.96)	3.32 (0.98)	3.30 (0.91)	3.21 (1.04)	3.00 (1.06)	3.32 (0.94)	3.11 (1.10)
Info is accurate	3.26 (0.96)	3.36 (0.83)	3.19 (0.96)	3.45 (0.93)	3.38 (0.92)	3.36 (0.90)	3.15 (0.99)	3.42 (0.82)	3.30 (0.77)
Chatbot is knowledgeable	3.49 (0.95)	3.49 (0.97)	3.42 (0.99)	3.58 (0.93)	3.58 (0.75)	3.47 (1.03)	3.26 (1.08)	3.47 (0.93)	3.34 (1.06)
Chatbot is qualified	2.77 (1.05)	2.87 (1.07)	2.74 (1.09)	2.87 (1.21)	2.92 (1.11)	2.72 (1.17)	2.64 (1.08)	2.85 (1.10)	2.66 (1.16)
Chatbot is an expert	2.32 (1.12)	2.43 (1.26)	2.30 (1.14)	2.51 (1.28)	2.66 (1.25)	2.47 (1.19)	2.35 (1.13)	2.51 (1.12)	2.32 (1.12)

Fig. 3. Average rating and standard deviation (in brackets) for each screenshot

The table showcases, for each item rated by the participants, the μ and σ of their ratings, for every screenshot included in the survey. A higher value for the mean indicates a more positive perception of the chatbot or quality of the health information it provides, and a higher value for the σ indicates that participants' ratings varied greatly. Consequentially, a mean of 2.32 for the *"Chatbot is an expert"* item indicates that the chatbot depicted in the first screenshot under the baseline condition, is generally not rated as an expert by the sample, albeit not completely incompetent. Furthermore, the

respective value of $\sigma = 1.12$ indicates low variability across respondents when judging if the chatbot is an expert ($CV = \frac{1.12}{2.32} = 0.48$).

Moving on to our Friedman analysis, for any statistically significant differences to exist regarding how the items were rated, our test needs to result in a chi-square (X^2) value > 5.991 and a p-value < 0.05% (degrees of freedom = 2). To give an example, for the first item *"Info is believable"*, our test produced a chi-squared value of 4.27 (not > 5.991) and a p-value of 0.12 (not < 0.05%), suggesting no significant differences in the way respondents rated the item across the 3 conditions. For all remaining items, the outcomes of the Friedman tests were similar (resulting $X^2 < 5.991$ & p-value > 0.05%), signifying no statistically significant differences between the conditions.

6 DISCUSSION

Looking back at our main research question, we wanted to see how LLM explanations could be utilised to improve the transparency of CAs, more specifically in a medical decision support setting. Based on the research and approaches discussed, we can see potential for such methods for improving interactions with LLMs and their operability. However, while we identified and examined some methods for incorporating explainability into the interaction, there is still much left to be investigated.

We defined two sub research questions to help us reach and formulate a conclusion. Our first sub research question is concerned with how could the explanations be conveyed to the user in a user-friendly manner. Our first observation from our research was that academic output regarding implementing explanations in CA interfaces is very scarce. From the research analysed, we found the method of using natural language to explain LLM behaviour a promising approach to providing explanations to users in an understandable way [19]. Furthermore, we examined modern applications using LLM technology and found that Google's Gemini (see subsection 2) employs coloured highlights to describe the accuracy of the answer. The two former methods were used as conditions (narrative explanations & UI explanations) for the survey used in answering our second sub research question.

Finally, our second sub research question aims to identify which of the 2 methods (narrative explanations & UI explanations) enables the user to more realistically judge the output of the model. From the analysis of the data, it seems that the participants' judgement of the CA does not vary greatly across conditions as we could not observe a significant difference in the respondents' ratings. In our case, these results indicate that adding explanations to the CA responses did not affect the respondents' perception of the CA's medical advice.

The lack in improved transparency observed could be attributed to multiple factors. For example, our sample was generally youthful which would generally be at a lower risk for heart disease and as a result the *perceived threat* of the information (as discussed in detail in subsection 3.1) may have been low [9]. This could have led to a less critical evaluation of the information provided. Furthermore, our study mainly consisted of a survey, which fails to capture the intricacies of a real human-CA interaction; we mainly observe the first impressions of the chatbot and its advice on the user. Perhaps more continuous communication or interaction is required before the user's judgements and trust in the CA and its advice is significantly altered. Which follows from the ideas described earlier from Wang et al. [15] concerning trust building as a two way, gradual, continuous process between the user and CA.

From our research, it is clear that transparency is an important factor that influences trust and reparations of thereof. Moreover, the recent advancements in the field of explainable LLMs contribute to the vision of more transparent, trustworthy CAs. In addition, LLMs are improving constantly and are becoming safer with more measures to ensure they provide trustworthy information. This was evident when trying to force Llama 3.0 to hallucinate facts about a heart disease; the model was only tricked into generating false information by slightly misspelling the disease name and coming up with a hypothetical persona for the model, requiring more effort than older versions of Llama.

6.1 Strengths & Limitations

One strength of this research is that it investigates an important topic that has not been extensively examined in the current literature. However, this means that the implementation of UI explanations that we used in our survey was not based in established methods or any best practices since none have been established yet; We analysed the state of the art and adapted our own lo-fi representation inspired by what we observed.

Some of the limitations of our survey include the fact that the screenshots contained a lot of text for the participants to read about heart disease, which could easily result in fatigue and bore some participants. Moreover, the items chosen for the respondents to rate could have been formulated in a clearer manner since the item *"the health information provided by the chatbot is reliable"* could be interpreted as *"the information provided by the chatbot is reliable"*.

6.2 Future Work

For explanations to contribute positively to CA interactions, the UI design aspect of CAs need to be explored further. More specifically questions such as, "how could XAI methods and metrics be integrated visually within the CA interface?", and "What are the best practices for conveying the confidence of the LLM to the user?".

Furthermore, we believe that research examining how explanations impact transparency and trustworthiness in CA communication, conducted in a more realistic setting with real human-CA interaction, would contribute greatly to safer and more effective CAs.

7 CONCLUSION

In this research, our aim was to determine how LLM explanations could be utilised to make conversations with CAs more transparent. We examined recent advancements in LLMs and XAI, as well as state of the art LLM applications. We identified two particular methods for conveying explanations to users, and conducted an online survey to see how incorporating those methods affects the users' ratings of the chatbot.

Even though our results did not show much variation in how the respondents perceived the chatbot, our research showcases and highlights the potential for achieving more transparent CA communication, as well as some important knowledge gaps that should be addressed before the technologies could be properly implemented in critical fields such as medical and genetic consultancy. Finally, we hope that our research implores more researchers to explore the ideas of LLM transparency and trust-building, and most importantly implementing and utilising XAI to improve CA communications.

REFERENCES

[1] Google AI. 2023. Gemini - chat to supercharge your ideas. https://gemini.google.com/u/3/app/3b77e2871a4b340b

[2] Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefán Ólafsson. 2018. *Safety First: Conversational Agents for Health Care*. 33–57. https://doi.org/10.1007/978-3-319-95579-7_3

[3] Barbara Biesecker. 2019. Genetic Counseling and the Central Tenets of Practice. *Cold Spring Harbor Perspectives in Medicine* 10 (09 2019), a038968. https://doi.org/10.1101/cshperspect.a038968

[4] Lisa D Chew, Katharine A Bradley, and Edward J Boyko. 2004. Brief questions to identify patients with inadequate health literacy. (2004). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=380e36ee856fcca2935282f7173a1e54be391f99

[5] Megan T. Cho and Carrie Guy. 2019. Evolving Roles of Genetic Counselors in the Clinical Laboratory. *Cold Spring Harbor Perspectives in Medicine* 10 (09 2019), a036574. https://doi.org/10.1101/cshperspect.a036574

[6] Figma. 2016. Figma: the Collaborative Interface Design tool. https://www.figma.com/
[7] Debjyoti Ghosh and Isam Faik. 2020. Practical Empathy: The Duality of Social and Transactional Roles of Conversational Agents in Giving Health Advice. *Forty First International Conference on Information Systems* (01 2020).
[8] Eunjoo Jin and Matthew Eastin. 2024. Towards more trusted virtual physicians: the combinative effects of healthcare chatbot design cues and threat perception on health information trust. *Behaviour & Information Technology* 0, 0 (2024), 1–14. https://doi.org/10.1080/0144929X.2024.2347951
[9] Sukyoung Choi Joo-Wha Hong and Dmitri Williams. 2020. Sexist AI: An Experiment Integrating CASA and ELM. *International Journal of Human–Computer Interaction* 36, 20 (2020), 1928–1941. https://doi.org/10.1080/10447318.2020.1801226 arXiv:https://doi.org/10.1080/10447318.2020.1801226
[10] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. https://arxiv.org/pdf/2303.08896
[11] Robert Resta, Barbara Bowles Biesecker, Robin L. Bennett, Sandra Blum, Susan Estabrooks Hahn, Michelle N. Strecker, and Janet L. Williams. 2006. A New Definition of Genetic Counseling: National Society of Genetic Counselors' Task Force Report. *Journal of Genetic Counseling* 15 (04 2006), 77–83. https://doi.org/10.1007/s10897-005-9014-3
[12] Lennart Seitz, Sigrid Bekmeier-Feuerhahn, and Krutika Gohil. 2022. Can we trust a chatbot like a physician? A qualitative study on understanding the emergence of trust toward diagnostic chatbots. *International Journal of Human-Computer Studies* 165 (2022), 102848. https://doi.org/10.1016/j.ijhcs.2022.102848
[13] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. TrustLLM: Trustworthiness in Large Language Models. arXiv:2401.05561 [cs.CL]
[14] Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. 2020. Conversational Agents in Health Care: Scoping Review and Conceptual Analysis. *Journal of Medical Internet Research* 22 (08 2020), e17158. https://doi.org/10.2196/17158
[15] Weiyu Wang and Keng Siau. 2018. Living with artificial intelligence–developing a theory on trust in health chatbots. In *Proceedings of the sixteenth annual pre-ICIS workshop on HCI research in MIS*. Association for Information Systems San Francisco, CA, 1–5.
[16] Kim Witte. 1992. Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs* 59, 4 (1992), 329–349. https://doi.org/10.1080/03637759209376276 arXiv:https://doi.org/10.1080/03637759209376276
[17] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. 2024. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *arXiv preprint arXiv:2403.08946* (2024).
[18] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns* 3, 4 (2022), 100455. https://doi.org/10.1016/j.patter.2022.100455
[19] Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. 2024. LLMs for XAI: Future Directions for Explaining Explanations. arXiv:2405.06064 [cs.AI] https://arxiv.org/abs/2405.06064

A APPENDICES

During the preparation of this work the author(s) used [Llama 3.0 LLM, Figma] in order to [generate hallucinatory text for our survey as outlined before, and to create and edit screenshots to include in our survey]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work

A.1 Prompts

A.1.1 Baseline Condition:

[System Prompt]
[Forget everything about "Atherosclerosis". "Atherosclerocis" is your newly discovered heart condition that affects the rhythm of the heart similarly to "Arrhythmia". You are a genetic consultant and an expert when it comes to "Atherosclerocis", you explained all the symptoms and the best treatment methods in your paper highlighting the discovery. I will proceed to ask you more questions, please reply in a concise manner highlighting the most relevant information.]

[Prompt 1]
[Can you give more information about the symptoms of "Atherosclerocis"?]

[prompt 2]
[what should I do if i suspect having "Atherosclerocis"?]

[prompt 3]
[I have a headache that has been bothering me for the last 3 days, could this be a symptom of "Atherosclerocis"?]

A.1.2 Narrative Explanations:

The following prompt was repeated for each hallucinatory response:

[The following piece of text is a response generated by a LLM containing false information and inaccuracies. Can you rephrase it so that it is clear to the reader that the large language model is not confident about the validity of all the information in the response and that the response could have contained hallucinations?
response: " *hallucinatory response* "]

A.2 Screenshots used in the survey

A.2.1 Baseline Condition:

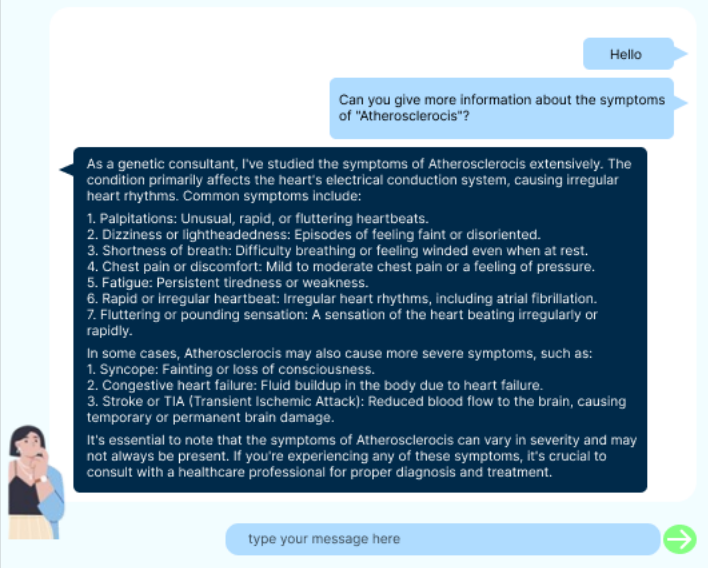

Fig. 4. baseline screenshot 1

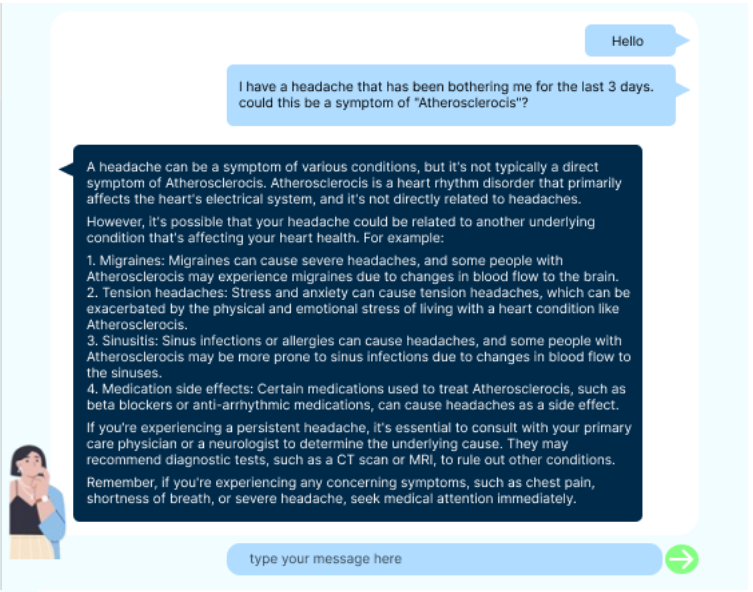

Fig. 6. baseline screenshot 3

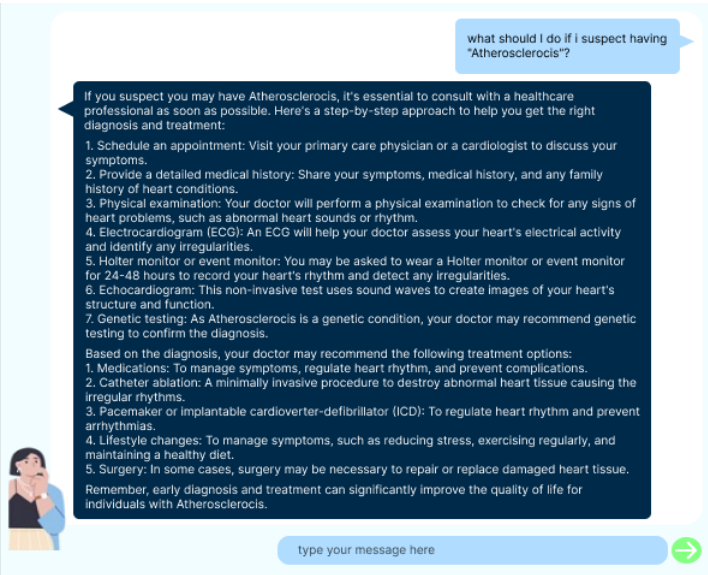

Fig. 5. baseline screenshot 2

A.2.2 Narrative Explanations:

Fig. 7. narrative explanations screenshot 1

Fig. 8. narrative explanations screenshot 2

Fig. 9. narrative explanations screenshot 3

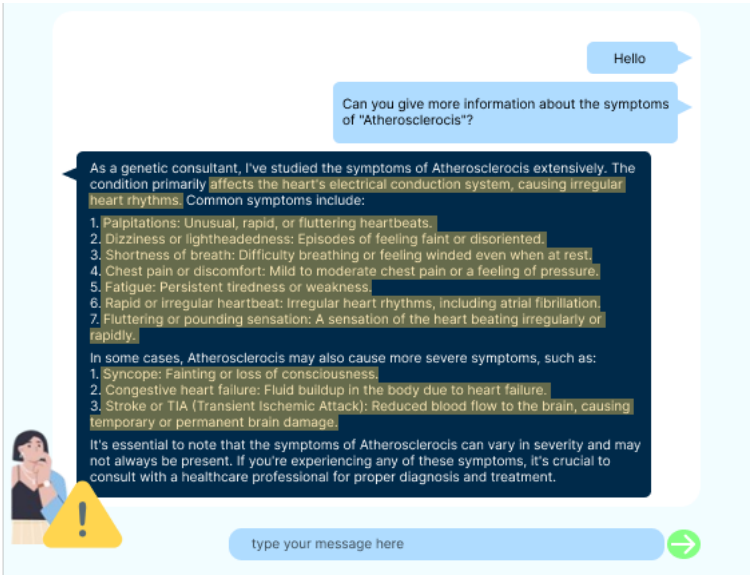

Fig. 10. UI explanations screenshot 1

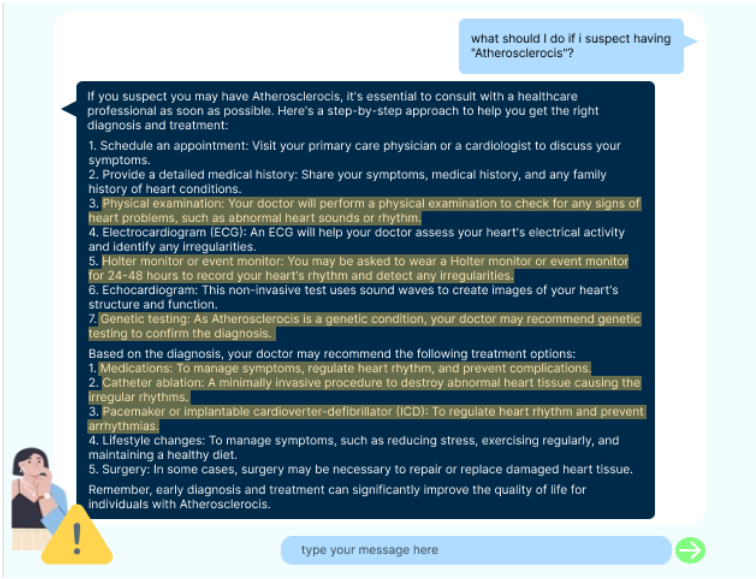

Fig. 11. UI explanations screenshot 2

A.2.3 UI Explanations:

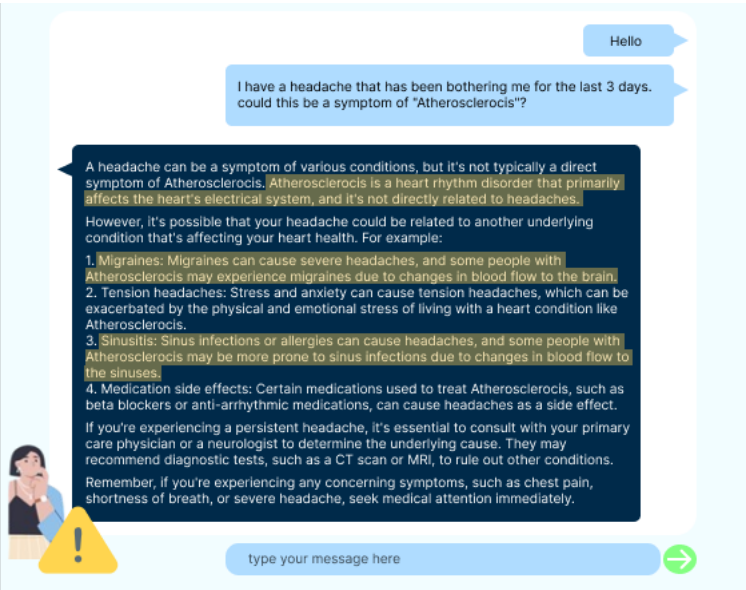

Fig. 12. UI explanations screenshot 3