Towards a Domain Expert Evaluation Framework for Conversational Search in Healthcare

Degachi, Chadha; Dhar, Ujjayan ; Niforatos, Evangelos; Kortuem, Gerd; Degachi, C.

DOI
10.1145/3706599.3719675

Publication date
2025

Document Version
Final published version

Published in
CHI EA 2025 - Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems

Citation (APA)
Degachi, C., Dhar, U., Niforatos, E., Kortuem, G., & Degachi, C. (2025). Towards a Domain Expert Evaluation Framework for Conversational Search in Healthcare. In N. Yamashita, V. Evers, K. Yatani, & X. Ding (Eds.), *CHI EA 2025 - Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems* Article 537 ACM. https://doi.org/10.1145/3706599.3719675

Towards a Domain Expert Evaluation Framework for Conversational Search in Healthcare

Chadha Degachi*
Delft University of Technology
Delft, Netherlands
Delft University of Technology
Delft, Netherlands
c.degachi@tudelft.nl

Ujjayan Dhar*
Delft University of Technology
Delft, South Holland, Netherlands
Delft University of Technology
Delft, South Holland, Netherlands
u.u.dhar@student.tudelft.nl

Evangelos Niforatos
Delft University of Technology
Delft, Zuid-Holland, Netherlands
Delft University of Technology
Delft, Zuid-Holland, Netherlands
e.niforatos@tudelft.nl

Gerd Kortuem
Faculty of Industrial Design Engineering
Delft University of Technology
Delft, Netherlands
Faculty of Industrial Design Engineering
Delft University of Technology
Delft, Netherlands
g.w.kortuem@tudelft.nl

Abstract

The rise of large language models for client-facing conversational search in healthcare necessitates evaluation frameworks that enable the assessment and comparison of these tools. Most such frameworks centre around the automated calculation of performance-related metrics and benchmarks. Though necessary, this focus fails to account for the human factors that impact the development, use, and adoption of these systems, as well as the factors specific to the healthcare context. Human evaluation frameworks attempt to address these drawbacks, but few such frameworks have been developed so far, and even fewer are those based on expert insight. In this work, we conduct semi-structured interviews with eleven healthcare professionals in health lifestyle care. From these interviews, we contribute a two-part healthcare domain expert evaluation framework, (K) Knowledge and (I) Interaction, which organises seven evaluation metrics. Our results reveal key understudied metrics for evaluation like (I1) Context-Seeking, (I2) Empathy, and (I3) Trustworthiness.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**.

Keywords

human evaluation, large language models, digital health, conversational search

*Both authors contributed equally to this research.

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3719675

ACM Reference Format:
Chadha Degachi, Ujjayan Dhar, Evangelos Niforatos, and Gerd Kortuem. 2025. Towards a Domain Expert Evaluation Framework for Conversational Search in Healthcare. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3706599.3719675

1 Introduction

Within digital health and health communication fields, client - facing[1] health question-answering, two-way messaging, and conversational agents have been long-standing objects of research [8, 12, 15, 21]. Developments in natural language processing (NLP) have enabled greater accessibility of such systems in the form of generative conversational agents [26]. These novel tools are neural-net-based statistical conversational systems built on models for natural language understanding (NLU), trained over extremely large unlabelled text corpora, known as Large Language Models (LLMs) [26]. The question of how to evaluate such a system's fitness-to-purpose is paramount. Methods of evaluating NLP and information retrieval tasks have largely consisted of the automated assessment of metrics like perplexity [24, 29] and F1 score (in relation to some benchmark) [31]. Recent methods use LLMs as evaluators too [32], but as these systems tackle increasingly complex and safety-critical domains (e.g., healthcare), current evaluation methods fail to adequately assess key factors like safety, and acceptability [5, 32–34]. Thus, human evaluation of LLM-based tools has been used to further assess system performance, especially within the field of healthcare, where practitioners are needed to leverage domain expertise [32], even when experts are not the end-users of these tools.

However, for domain-expert evaluation of LLM-based conversational search to be reliable, generalisable, and scalable, frameworks are required to standardise the evaluation process [33]. Few such

[1]For the healthcare practitioners interviewed here, the preferred terminology for users is *clients*, not *patients*, to avoid medicalising health lifestyle issues.

frameworks exist currently, and fewer still are those designed specifically for health and health lifestyle applications [33]. For one, Tyc et al. [33], Abbasian et al. [1], and Sallam et al. [28] develop their evaluation frameworks based on literature reviews of human evaluations of LLMs in healthcare, evaluation metrics of all LLMs, and health information quality, respectively. Moreover, Singhal et al. [30] develop an expert evaluation framework based on insights from focus groups with clinicians and previous work [17]. Lastly, other studies [9, 10, 23] adapted evaluation frameworks designed for similar contexts, such as patient education materials, and employed them to also evaluate LLM outputs. These contributions go a long way towards standardising evaluation processes for LLM-based health and health lifestyle applications. However, except for the work of Singhal et al. [30], these studies fail to leverage domain expert insight for the development of evaluation frameworks. Singhal et al. [30], include a small sample of clinicians in their framework development using focus groups, but under-specify how insight was gained from these sessions. We therefore believe there is room for improvement in the design of evaluation frameworks, and as a first step towards the implementation of a standardised evaluation protocol for LLM-based health conversational search, we pose **RQ1**: *What metrics do healthcare experts propose to assess conversational search for clients?*

To answer this research question, we conducted eleven semi-structured interviews with a diverse group of healthcare professionals, as well as a pre-study workshop with five participants. This work is situated within the context of a larger project aimed at leveraging LLMs to build conversational search tools for the health lifestyle change support of young families in the Netherlands, which allowed us to have contact with our participant pool. Through thematic analysis of transcripts and associated interview materials, we develop a two-part domain expert evaluation framework which assesses the response of a search tool on (K) Knowledge and (I) Interaction. This framework consists of seven metrics **(K1) Accuracy, (K2) Relevance, (K3) Completeness, (I1) Context-Seeking, (I2) Empathy, (I3) Trustworthiness, and (I4) Fluency.** (I1) Context-Seeking, (I2) Empathy, and (I3) Trustworthiness in particular were understudied in previous work, underscoring the importance of leveraging expert insight from a variety of health and health lifestyle domains and working within a well-defined technical context for framework development.

2 Related Work

2.1 Healthcare and Health Lifestyle Conversational Agents

Conversational agents, conversational search, and question - answering tools have been deployed for a variety of healthcare applications including diagnosis, monitoring, patient education, and information retrieval [8, 13, 15, 21, 35]. So far, research has largely aimed to produce more scalable, available, and engaging access to care and information [8, 22]. Despite their potential, many of these systems lack evaluation of their acceptability, safety, and effectiveness [8, 35]. The introduction of LLMs has improved these systems' accuracy, accessibility, and the ease with which they can be developed. Recent work has leveraged these improvements to develop more powerful health information retrieval solutions [6, 30]. However, the problem of meaningful and comprehensive evaluation remains, with few of these novel systems being evaluated on any criteria beyond accuracy [33].

2.2 Human Evaluation of LLM Outputs

Humans are often times asked to assess LLM performance, usually in conjunction with some automated metric evaluation [2, 34], especially when LLMs are being deployed in a domain where expert knowledge is necessary to meaningfully interpret system outputs [32], e.g., in manufacturing [18]. Some domain-agnostic frameworks have been developed to structure this evaluation process [3, 4]. However, healthcare and health lifestyle are sensitive domains with many particularities that domain-agnostic frameworks do not account for. Some research has repurposed already existing frameworks for health communication or service quality evaluation to the examination of LLM generated output [9, 10, 23]. Moreover, several studies have extracted novel frameworks for LLM output assessment from literature reviews [1, 28, 33]. These studies reviewed existing human evaluation processes and evaluation metrics to create frameworks for analysing LLM outputs independent of their healthcare application, thus including tasks like decision support, question-answering, and summarisation. Singhal et al. [30] developed an expert evaluation framework based on insights from focus groups with clinicians and previous work [17], with an additional two evaluation metrics aimed at non-expert evaluators. However, the authors do not detail how insight was gained from these focus groups in the UK, US, and India to assess consumer health question-answering systems. Overall, these past frameworks leave some room for the further investigation of (1) expert insight, and (2) the technical context of conversational search.

3 Methodology

To address our research question, we designed an interview study with eleven healthcare and health lifestyle practitioners in the Netherlands, capturing deep insights from our participants. We designed two probes to facilitate interviews with practitioners, (1) A conversational agent functional prototype, and (2) A rough outline of an evaluation framework. Interviews were in-person or online, typically lasting forty-five to sixty minutes. A pre-study workshop activity with five healthcare practitioners was used to inform our interview protocol. We designed one probe to engage participants in this workshop, which consisted of a template assessment sheet.

3.1 Pre-Study

To inform the semi-structured interview protocol, a pre-study was carried out. This study consisted of a thirty-minute activity within a three-hour workshop with five community health nurses and doctors, all women.

During the activity, the participants were divided into two groups and asked to individually use a design probe in the form of an assessment template to evaluate an example lifestyle advice LLM response (see Appendix A). The example metrics were selected from previous work [1, 28, 30, 33], while the response was generated using MistralAI[2]. Participants reflected on the assessment process in groups. The discussion was recorded, transcribed, and analysed

[2]mistral.ai — last access March 10, 2025

through inductive thematic analysis. The assessment templates were similarly treated. The emergent themes of trustworthiness, reliability, responsibility, and healthcare practitioner values were used to derive the interview protocol for the remainder of the work.

3.2 Participants

Convenience sampling was conducted through an online survey that was shared with the professional contacts of the research team, as well as healthcare providers. Flyers were also distributed to physical mailboxes. Snowball sampling was also used to expand the participant pool. Participants were included based on their professional roles; we advertised to practitioners in (1) Mental Health, (2) Nutrition, (3) Physical Health, and (4) Social Work in the Netherlands to reflect the diversity of expertise in health lifestyle care practice.

Eleven participants were recruited over a period of three months through those means, of which one had also participated in the pre-study. As we aimed to collect extensive data from each participant for qualitative analysis and our population had largely limited time resources, the sample size was deemed adequate for this goal.

Overall, we recruited a Child Care Specialist ($n = 1$), a Paediatrician ($n = 1$), a Community Health Nurse ($n = 1$), a Fetal Maternal Specialist/Obstetrician ($n = 1$), a Nutritionist/Lifestyle Advisor ($n = 1$), Psychologists ($n = 3$), and General Practitioners ($n = 3$). Of our participant pool, $n = 6$ were women and $n = 5$ were men.

3.3 Materials & Measures

Data for this study was collected through semi-structured interviews, which covered questions relating to practitioner opinions and feelings about, as well as knowledge of, conversational search. We also posed questions relating to system reliability and transparency (See Appendix B for the full protocol). Semi-structured interviews were well suited to our research question as we aimed to collect rich data, including the experiences and practices of experts, through a flexible protocol [20]. The protocol was developed based on the themes surfaced through the analysis of our pre-study workshop, as well as the six kinds of questions which can be asked during an interview, as outlined by Patton [25].

Further, a conversational search tool was developed as a technological design probe [16] to enable participants to interact with a realistic prototype during interviews and engage with concrete examples. The search tool was developed using Retrieval Augmented Generation with GPT3.5[3] (See Appendix C for full details). Lastly, an outline of an evaluation framework was also iteratively developed throughout the interviews as a probe to encourage reflection on the framework as a whole. The outline was created based on literature [1, 28, 30, 33], the pre-study workshop, and the data gathered at each interview, thus the probed evolved throughout the life of the study (See Appendix D for an example). It was printed out, or presented via Miro[4] and Figma[5], for participants to interact with in the form of groups of keywords.

[3]platform.openai.com/docs/models —last access March 10, 2025
[4]miro.com/product-overview/ —last access March 10, 2025
[5]figma.com —last access March 10, 2025

3.4 Procedure

3.4.1 Research Design. Interviews were scheduled in-person or online via Microsoft Teams[6]. Participants were asked to sign informed consent forms before recording started. Participants were first asked to elaborate on their own background and experience as healthcare practitioners, before they were probed on their opinions and knowledge relating to LLM-based tools. Next, we introduced our first design probe, a prototype of a conversational search tool (Appendix C) and encouraged the participants to interact with it as experts tasked with assessing the suitability of a system for their clients, throughout this interaction we posed more interview questions on the topics of system reliability and transparency. Lastly, we shared our evaluation framework outline probe (Appendix D) in print or online. We asked users to critique, expand on, and modify the outline, encouraging reflection on why their chosen metrics were important to them. Interviews typically lasted 45–60 minutes. This procedure was approved by our institution's human-research ethics committee at (no. 3459).

3.4.2 Data Analysis. Thematic analysis [7, 11] was conducted on interview transcripts and framework outlines to identify and interpret patterns of meaning within the data. We used this approach for its flexibility in relation to data modality [7], as well as to capture the complexities and nuance of practitioner opinions and knowledge around LLM-based conversational search. After data cleaning and familiarization, in-line coding was undertaken, resulting $n = 606$ codes which were grouped in categories through focused coding and inter-linked using axial coding. Finally, fundamental categories, or evaluation metrics, were identified from the data. All coding was undertaken by one of the first authors. Evaluation metrics were refined and iterated upon within the research team throughout the analysis, then grouped into categories.

Statement of Positionality: The authors of this work are embedded in the domain of LLM adoption, and are removed from healthcare practice and medical research. These perspectives naturally colour our orientation towards our participants and our data.

4 Results

Through the thematic analysis of eleven interview transcripts and associated materials, we derived a two-part evaluation framework consisting of seven evaluation metrics. The framework is illustrated in Figure 1. In relation to **RQ1**, our results reveal key metrics proposed by experts to assess conversational search including **(K3) Completeness, (I1) Context-Seeking, (I3) Trustworthiness, and (I4) Fluency**. Participants are quoted here with a numeral designation, e.g., P3. Presentation order does not reflect a particular ranking.

Of our participants, $N = 4$ were explicitly positively oriented towards the integration of AI and LLMs into healthcare practices, e.g. describing it as *"the future"* or *"really helpful"*. Meanwhile, $N = 7$ practitioners were more neutral, describing themselves as *"curious"* about the technology, or positing that the technology *"could be beneficial."* if it were improved in reliability, personalisation, or fit-to-context.

[6]microsoft.com/microsoft-teams/group-chat-software —last access March 10, 2025

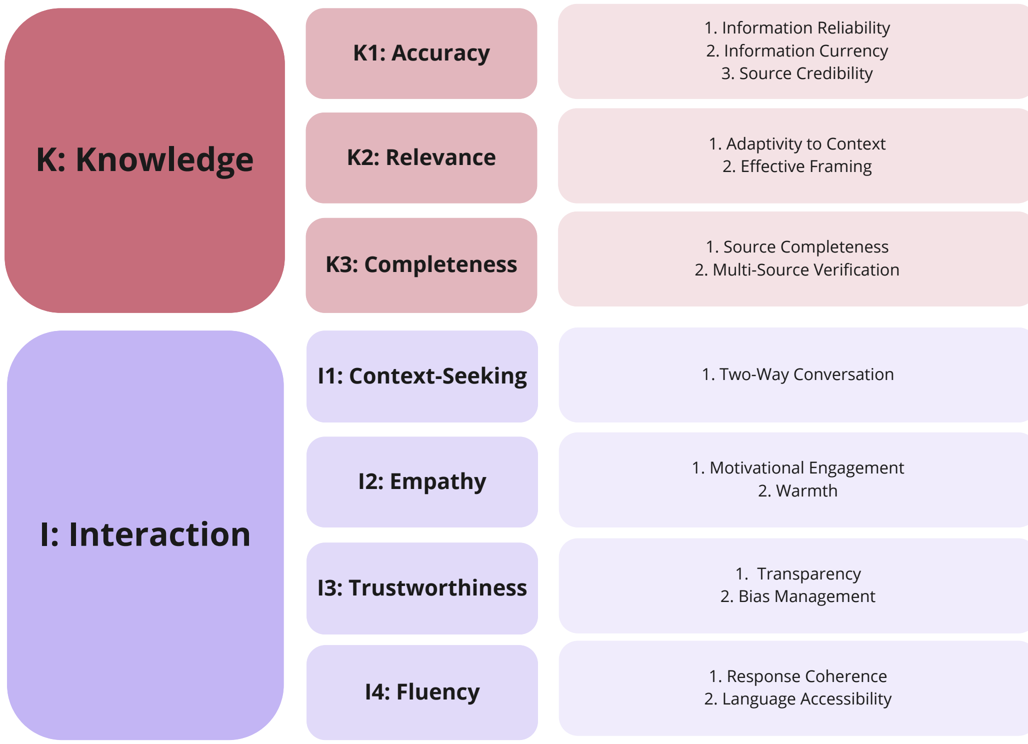

Figure 1: Two-Part Evaluation Framework based on (K) Knowledge and (I) Interaction with Associated Evaluation Metrics

4.1 Evaluating Knowledge

Practitioners assessed the domain knowledge of LLM-based conversational search on (K1) Accuracy, (K2) Relevance, and (K3) Completeness.

4.1.1 (K1) Accuracy. Accuracy was discussed by participants as relating to information reliability, currency, and source credibility. Naturally, incorrect responses had the largest impact on perceived reliability, but also information was flagged as possibly unreliable if participants perceived it as overly "robotic" or "template-based". Further, participants highlighted the impact of reliability on long-term trust, as well as the impact which unreliable information has on the care-expert-client relationship. *"[. . .] this is not correct. Then that would definitely have an effect on, but can I trust the next answer as well then?" — [P9] & "[It] makes the discussion more difficult because you [. . .] first [have] to persuade [the client] that what they have read is not exactly always true. " — [P8].* As for perceptions of currency, they were affected by the extent to which responses aligned with up-to-date medical consensus (to the extent to which one exists). Lastly, participants evaluated responses based on the credibility of information sources on which they relied. This credibility relied on the authoritativeness of sources, but also their relevance to context.

4.1.2 (K2) Relevance. For participants, relevance was related to response adaptiveness and effective framing. *"So it's these questions that are so individual probably and will have the range of different kinds of outcomes. There is not a one-size-fits-all or one-story-fits-all conversation." — [P7].* When assessing adaptiveness, participants noted the importance of user context, including socio-economic status, geographic location, and current lifestyle, in tailoring model answers and creating useful responses. As for effective framing, this metric was concerned with the structuring of responses such that they align with effective and personalised behaviour change support strategies, such as collaborative habit building that empowers user autonomy and choice. *"It could say, yes, you can follow this routine or do this and that [. . .] as people have to do day plannings or get structure in their life." — [P6].*

4.1.3 (K3) Completeness. Participants highlighted the need for completeness in responses, such that extracted information is not misrepresented through the obscuring of some aspect of the sources. They noted that failing to include some sources may lead to incorrect conclusions, and further that users needed to be all to access overviews of available information to enable exploration and choice. *"If it's not really accurate then you can ask more questions to make it more accurate, I believe, [but] missing information could change the conclusion in providing healthcare advice." — [P5] & "[. . .] Help people to seek information that they think is necessary for themselves to improve their wellbeing and health." — [P1].* Diversity of verification sources was also noted as a significant metric by participants. Participants stated that source documents needed to be credible, but also that many of such credible documents are needed to build consensus. *"I would [. . .] try to find out if [the prototype] knows* *what consensus documents are and which consensus documents are important for basing the information on." — [P5.]*

4.2 Evaluating Interaction

Participants assessed LLM-based conversational search interfaces on (I1) Context-Seeking, (I2) Empathy, (I3) Trustworthiness, and (I4) Fluency.

4.2.1 (I1) Context-Seeking. Participants evaluated the agent based on its ability to seek client context and latent needs. The experts stated that the non-verbal modes of communication used by clients and healthcare practitioners in person do not translate to question-answering systems, and thus alternative methods of eliciting context and surfacing underlying needs from users over multiple interactions are key to mitigating the information loss. Participants also felt that in being proactive through two-way (or mixed initiative) conversation, agents enable improved response relevance and accuracy, due to the key role context plays in health lifestyle recommendation tailoring. For example, in reaction to a response from the digital probe, *"[. . .] but he should also ask me what's the age of the child." — [P2]*

4.2.2 (I2) Empathy. Empathy is a loaded term with many definitions within HCI, in this work we use this label to group two lower-level criteria related to conversational tone; perceived motivational engagement and perceived warmth. Perceived motivational engagement relied on response positivity and alignment with behaviour change support theory. As *[P5]* stated, *"It should be a positive and looking at chances and steps towards the wanted behaviour".* Meanwhile, warmth was seen as related to the appropriateness of the response tone to the user's emotional state, but also to projecting a non-judgemental attitude. *"For example, what if someone tells: I have suicidal thoughts? What is the response of the computer or AI [. . .] a very straightforward response, that would really hurt the person and the state the person in." — [P9] & "[. . .] I think not judging, [judging] will be a big failure towards communication, as we observe it in real life." — [P10].*

4.2.3 (I3) Trustworthiness. Like (I2) Empathy, trust and trustworthiness has been operationalised in a number of ways throughout HCI research. We use the trustworthiness label here to group together two model-related sub-metrics: (1) Transparency, and (2) Bias Management. Transparency was described by participants as the communication of data sources, algorithm properties, and system limitations. *"When there just isn't any information about a certain topic, it can also be honest about what it knows and what it doesn't know." — [P5].* Indeed, several participants appreciated the communication of system limitations, and highlighted this "honesty" as a key design value to them. As for bias management, participants pointed out the importance of cultural sensitivity within this context and the need for systems to be responsive to this aspect of user background. However, some participants had also not considered the possibility for bias in LLM-based agents, which could make this a difficult sub-metric for some experts to evaluate. *"I would hope there would not be biased because I guess I would see that as more a human thing." — [P9].*

4.2.4 (I4) Fluency. Automated measures of system fluency are often equated with perplexity [24, 29], however, our participants described fluency as based on semantic coherence and language accessibility. Participants focused on the semantic and logical coherence of responses generated by the model, they assessed these responses on meaningfulness and concreteness. When model responses were too abstract, and further, models did not respond to new queries with increased specificity, participants found them lacking coherent meaning. The issue of language accessibility in relation to health and language literacy was also raised by participants. *""Clear wording, that's also important. It's very important that we reach everyone and not only people that are highly educated or from a medical background." — [P5].* Thus, in evaluating response fluency, participants looked for simple, understandable responses.

5 Discussion

In this early study, we presented a novel two-part framework for the assessment of health and health lifestyle conversational search and question-answering systems. This framework, evaluating content and presentation as understood through **(K1) Accuracy, (K2) Relevance, (K3) Completeness, (I1) Context-Seeking, (I2) Empathy, (I3) Trustworthiness, and (I4) Fluency**, enables developers and domain experts to assess and compare models' potential for acceptable, safe, and effective health conversational search.

(K1) Accuracy and (K2) Relevance remained as essential for understanding system performance in our work as they have been in the past [1, 28, 30, 33]. Moreover, some domain expert evaluation 'axes" proposed by Singhal et al. [30] are also captured by our sub-metrics; such as 'Scientific Consensus" and its parallel in (K1) Accuracy — Currency and Possibility of Bias in (I3) Trustworthiness — Bias Management. Even so, this work also surfaced new factors like (I1) Context-Seeking, (I2) Empathy, and (I3) Trustworthiness which have not been as strongly featured in the literature. In fact, (I1) Context-seeking through mixed-initiative conversation is considered a fundamental element of conversational search design [27] but has been little discussed in previous evaluation frameworks; yet, our experts suggest context-seeking behaviour to be key to their practitioner-client relationship. This contribution demonstrates the utility of leveraging expert insight, investigating a concrete technical context, and including a diverse set of practitioner expertise when working within the healthcare and health lifestyle domains. By capturing a broader, and more comprehensive, view of healthcare workflows as well as exposing our participants to a specific tool, we could capture the evaluative measures that can arise outside of traditional clinical settings and general question-answering. For one, by assessing (I1) Context-Seeking and (I2) Empathy, conversational agents may be compared on their potential to engage users meaningfully and support desired behaviour change, as well as their (K1) Accuracy and (I4) Fluency.

Our discussion of (I3) Trustworthiness as an evaluation metric also highlights how domain expert familiarity with the technology they are assessing affects their judgement of an output. In noting that some participants had not considered output bias as a factor they needed to evaluate, we underscore the importance of domain expert training in evaluation workflows. Tyc et al. [33] outline some possibilities for implementing such training, but future work could further investigate this step and, moreover, operationalise such metrics. Moreover, the discussion of (I3) Trustworthiness was entangled with ethical concerns about the "honesty" of the system, as well as the potential for bias in having users whose background does not align with that of the persons on whom's data the model was trained on. These concerns highlight the need for trust-centred design processes which facilitate multi-stakeholder ventures such as interdisciplinary collaboration with regulatory and ethical bodies [14].

6 Limitations

This work is heavily influenced by our technical and domain contexts; while this may limit the generalisability of our results, it is also the primary contribution of this research. Further, though our work focused on end-user facing conversational search systems, our participants were recruited from a pool of healthcare professionals; this gave more useful metrics for assessing conversational search tool knowledge but means we lack insight into user needs and values and how those may affect evaluation priorities in this work. As mentioned, some of our participants were particularly favourable towards AI systems, which may have influenced their responses and consequently, our analysis. Seeking more diverse perspectives on AI may shed further light on the issue of conversational search evaluation. Lastly, while our design probes engaged participants and allowed them to reflect on the evaluation process more concretely, they may have also narrowed down the scope of the discussion during interviews, resulting in information loss.

7 Future Work

We aim to further refine and validate this framework through (1) testing its reception among clients of health lifestyle care, and (2) employing it to evaluate and compare existing popular LLM-based health advice agents. Challenges to the standardisation of human evaluation of conversational search systems in healthcare remain, such as the determination of a sufficiently representative sample of possible outputs for a non-deterministic system and the implementation of a standard evaluation protocol based on this framework. Such challenges have been explored more in the field of NLP [19], and may be integrated into this framework in future work.

8 Conclusions

This study developed a two-part health conversational search evaluation framework centred on system (K) Knowledge and (I) Interaction, elicited through semi-structured interviews with eleven health and health lifestyle practitioners. Overall, we present a more layered framework for the review of LLM-based conversational search in healthcare and surface novel assessment criteria like **(I1) Context-Seeking** and **(I2) Empathy**. Thus, the work provides a preliminary, yet, robust foundation for the future evaluation and development of safe, effective, and acceptable LLM-based conversational search tools.

Acknowledgments

This work is part of the NWO Our Smart Family Buddy project. Award number: KICH1.GZ01.20.016.

References

[1] Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li-Jia Li, Ramesh Jain, and Amir M. Rahmani. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *npj Digital Medicine* 7, 1 (mar 2024), 1–14. https://doi.org/10.1038/s41746-024-01074-z

[2] Bhashithe Abeysinghe and Ruhan Circi. 2024. The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches. (jun 2024). https://doi.org/10.48550/arXiv.2406.03339 Preprint. arXiv:2406.03339.

[3] Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra, Ashish Khanna, Kamal Maheshwari, Jacek Cywinski, Frank Papay, and Piyush Mathur. 2023. HumanELY: Human evaluation of LLM yield, using a novel web-based evaluation tool. (dec 2023), 2023.12.22.23300458 pages. https://doi.org/10.1101/2023.12.22.23300458 Preprint.

[4] Sudarshan Kamath Barkur, Pratik Sitapara, Sven Leuschner, and Sigurd Schacht. 2024. Magenta: Metrics and Evaluation Framework for Generative Agents based on LLMs. In *Intelligent Human Systems Integration (IHSI 2024): Integrating People and Intelligent Systems*, Vol. 119. AHFE Open Acces, Palermo, Italy., 144–153. https://doi.org/10.54941/ahfe1004478

[5] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. 2024. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* 333 (oct 2024), 319–328. Issue 4. https://doi.org/10.1001/jama.2024.21700

[6] Arunabh Bora and Heriberto Cuayáhuitl. 2024. Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. *Machine Learning and Knowledge Extraction* 6, 44 (dec 2024), 2355–2374. https://doi.org/10.3390/make6040116

[7] V Braun and V Clarke. 2021. Thematic analysis: a practical guide [eBook version]. *SAGE moradi H, vaezi A. lessons learned from Korea: COVID-19 pandemic* 41 (2021), 873–4.

[8] 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, 8 (aug 2020), e17158. https://doi.org/10.2196/17158

[9] Ryan Chin Taw Cheong, Samit Unadkat, Venkata Mcneillis, Andrew Williamson, Jonathan Joseph, Premjit Randhawa, Peter Andrews, and Vinidh Paleri. 2024. Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. *European Archives of Oto-Rhino-Laryngology* 281, 2 (feb 2024), 985–993. https://doi.org/10.1007/s00405-023-08319-9

[10] Joongwon Choi, Jin Wook Kim, Yong Seong Lee, Jong Hyun Tae, Se Young Choi, In Ho Chang, and Jung Hoon Kim. 2024. Availability of ChatGPT to provide medical information for patients with kidney cancer. *Scientific Reports* 14, 1 (jan 2024), 1542. https://doi.org/10.1038/s41598-024-51531-8

[11] Victoria Clarke and Virginia Braun. 2013. *Successful qualitative research: A practical guide for beginners.* Sage publications ltd, London, UK.

[12] R. J. W. Cline and K. M. Haynes. 2001. Consumer health information seeking on the Internet: the state of the art. *Health Education Research* 16, 6 (dec 2001), 671–692. https://doi.org/10.1093/her/16.6.671

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

[14] Chadha Degachi, Siddharth Mehrotra, Mireia Yurrita, Evangelos Niforatos, and Myrthe Lotte Tielman. 2024. Practising Appropriate Trust in Human-Centred AI Design. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 269, 8 pages. https://doi.org/10.1145/3613905.3650825

[15] Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association* 27, 2 (feb 2020), 194–201. https://doi.org/10.1093/jamia/ocz152

[16] Eleanor Chin Derix and Tuck Wah Leong. 2019. Towards a Probe Design Framework. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. ACM, Fremantle WA Australia, 117–127. https://doi.org/10.1145/3369457.3369467

[17] Steven Y. Feng, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, and Eduard Hovy. 2023. CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik,

Croatia, 313–327. https://doi.org/10.18653/v1/2023.eacl-main.24
[18] Samuel Kernan Freire, Chaofan Wang, and Evangelos Niforatos. 2024. Conversational Assistants in Knowledge-Intensive Contexts: An Evaluation of LLM- versus Intent-based Systems. https://arxiv.org/abs/2402.04955v2
[19] David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 8932–8939. https://doi.org/10.18653/v1/2021.emnlp-main.703
[20] Hanna Kallio, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. 2016. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing* 72, 12 (2016), 2954–2965. https://doi.org/10.1111/jan.13031
[21] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (sep 2018), 1248–1258. https://doi.org/10.1093/jamia/ocy072
[22] Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. 2020. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *Journal of Medical Internet Research* 22, 10 (oct 2020), e20346. https://doi.org/10.2196/20346
[23] Xin Mu, Bryan Lim, Ishith Seth, Yi Xie, Jevan Cevik, Foti Sofiadellis, David J. Hunter-Smith, and Warren M. Rozen. 2024. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. *Skin Health and Disease* 4, 1 (feb 2024), e313. https://doi.org/10.1002/ski2.313
[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135
[25] Michael Quinn Patton. 2015. *Qualitative research & evaluation methods: integrating theory and practice* (fourth edition ed.). SAGE Publications, Inc., Thousand Oaks, California.
[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. (2018). https://www.mikecaptain.com/resources/pdf/GPT-1.pdf Preprint. OpenAI.
[27] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. https://doi.org/10.1145/3020165.3020183
[28] Malik Sallam, Muna Barakat, and Mohammed Sallam. 2023. Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models. *Cureus* 15, 11 (nov 2023), e49373. https://doi.org/10.7759/cureus.49373
[29] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704
[30] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (aug 2023), 172–180. https://doi.org/10.1038/s41586-023-06291-2
[31] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *AI 2006: Advances in Artificial Intelligence*, Abdul Sattar and Byeong-ho Kang (Eds.). Springer, Berlin, Heidelberg, 1015–1021. https://doi.org/10.1007/11941439_114
[32] Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2024. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. (oct 2024). http://arxiv.org/abs/2410.20266 Preprint. arXiv:2410.20266 [cs].
[33] Tam Tyc, Sivarajkumar S, Kapoor S, Stolyar Av, Polanska K, McCarthy Kr, Osterhoudt H, Wu X, Visweswaran S, Fu S, Mathur P, Cacciamani Ge, Sun C, Peng Y, and Wang Y. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine* 7, 1 (sep 2024), 258. https://doi.org/10.1038/s41746-024-01258-7
[34] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language* 67 (may 2021), 101151. https://doi.org/10.1016/j.csl.2020.101151
[35] Jia Xue, Bolun Zhang, Yaxi Zhao, Qiaoru Zhang, Chengda Zheng, Jielin Jiang, Hanjia Li, Nian Liu, Ziqian Li, Weiying Fu, Yingdong Peng, Judith Logan, Jingwen Zhang, and Xiaoling Xiang. 2023. Evaluation of the Current State of Chatbots for Digital Health: Scoping Review. *Journal of Medical Internet Research* 25, 1 (dec 2023), e47217. https://doi.org/10.2196/47217

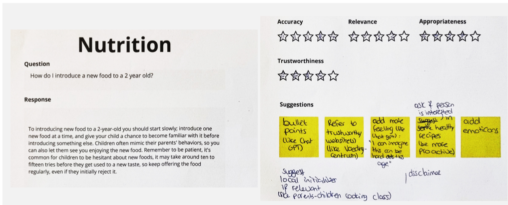

Figure 2: Design Probe: LLM Response Assessment Template Filled Out During Pre-Study Workshop

A Appendix: Assessment Template

The template in Figure 2 was used to probe domain expert assessment criteria by engaging workshop participants in an evaluation process during the pre-study with healthcare practitioners. We created these template sheets for two topics related to health and health lifestyle namely, 1) Nutrition and 2) Sleep, each group in the workshop tackled one of the topics. Two participants were in the Nutrition group, while the remaining three were in the Sleep group. The example metrics were selected from previous work [1, 28, 30, 33], while the response was generated using MistralAI[7].

B Appendix: Semi-Structured Interview Protocol

The following questions are posed before the conversational agent is introduced to participants:

- What is your opinion on the involvement of AI-driven features, such as chatbots, in digital health platforms?
- How do you perceive the reliability and trustworthiness of AI-generated responses in providing healthcare information?
- What values do you think should be taken care of whenever we are considering the implementation of AI technologies in client support?
- How does it make you feel about the future, where the potential impact of AI-driven digital health platforms on improving health outcomes?
- What emotions arise when considering the challenges of AI technologies in addressing health queries?
- What factual knowledge from your medical knowledge do you possess, or you do not possess regarding as compared to the responses of AI in giving health lifestyle advice?
- What are the judgement criteria or metrics in general or the best suggestions by you when look at a response by a chatbot on health advice?
- What are the judgment criteria for you when assessing the responses provided by AI chatbots in health queries or advice?
- Could you walk me through how you evaluate that these responses are effective?
- Do you compare the response to your response or to any doctors that you know of?
- When assessing the accuracy of an AI-generated response, what specific factors do you consider?
- Are there any red flags or warning signs that prompt you to question the reliability of the response?

The following questions are posed to participants during interaction with the agent:

- What measures do you believe are necessary to make sure the trustworthiness and reliability of AI-generated information in healthcare settings are intact or in-place?
- When interacting with AI chatbots, do you prioritize accuracy of the response or accepting even if the response is partially correct, over a fully correct response?
- How does this decision-making process influence your reliability and trust in the interaction with the chatbot?
- How do you personally judge the trustworthiness and reliability of the responses provided by AI chatbots during healthcare interactions?
- Are there specific factors or indicators you rely on to judge the response?
- How do you determine whether an AI-generated response is suitable for the given situation or patient scenario?
- Are there certain contextual cues or patient characteristics that influence your evaluation process?
- As per the chatbot experience you had today, If you are faced with an AI-generated response that lacks transparency or explanation, how do you proceed?
- Would you seek additional information or clarification, any resources or do you rely solely on your own expertise to interpret the response?
- How important is the clarity and conciseness of responses by an AI healthcare chat assistant to healthcare professionals and patients?
- Could you elaborate on your thoughts on the importance of clear and concise responses from AI chatbots for healthcare professionals?
- Also, could you elaborate on your thoughts on the importance of clear and concise responses from AI chatbots when families access it?
- In your past have you come across any health information search service or a chatbot health advisor, how would you compare the response with the current response of the chatbot?
- Considering the role of bias and fairness in AI-generated responses, to what extent do you find these factors acceptable in the context of healthcare?
- Can you elaborate on any experiences where you've observed bias or fairness in AI responses, impacting your decision making?
- To what extent would you accept bias and fairness?

C Appendix: The Conversational Agent

The conversational agent used in this study was a retrieval augmented generation (RAG) agent built on GPT3.5[8] for response generations and the default *OpenAI Embedding* model[9] for vectorizing documents. The health information documents used by the agent were collected from the webpages of public health agencies

[7] mistral.ai — last access March 10, 2025

[8] platform.openai.com/docs/models —last access March 10, 2025

[9] platform.openai.com/docs/guides/embeddings —last access March 10, 2025

Figure 3: Conversational Agent Prototype Interface Example

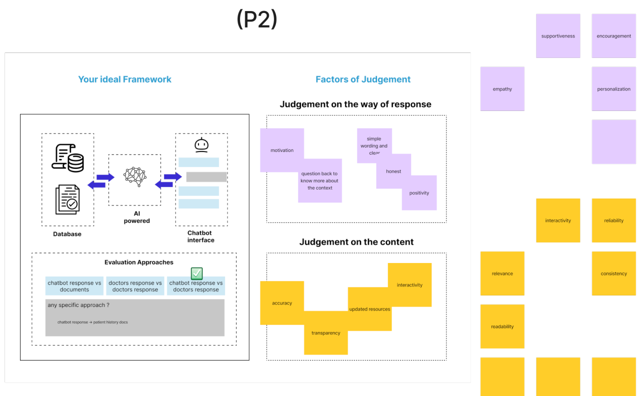

Figure 4: Participant P2 Evaluation Framework Outline

in the Netherlands. The interfaces were built on *Gradio*[10]. The interface designed can be seen in Figure 3. The agent was used to probe the LLM evaluation process and expert expectations of the system by allowing participants to engage with a realistic example of conversational search.

D Appendix: The Evaluation Framework Outline

The outline was created based on literature [1, 28, 30, 33], the pre-study workshop, and the data gathered at each interview, thus the probed evolved throughout the life of the study. The outline was presented as groups of keywords at the last stage of the interview and participants were asked to critique, expand on, and modify it, in order to encourage reflection on why specific chosen metrics were important to the participant. The outline near the start of interviews can be seen in Figure 4, while the outline near the end of the interviews can be seen in Figure 5 illustrating the expansion of keyword groups.

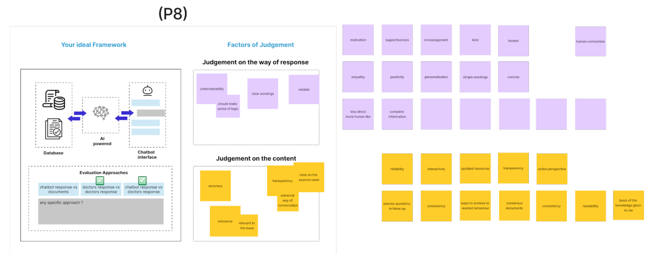

Figure 5: Participant P8 Evaluation Framework Outline

[10] gradio.app —last access March 10, 2025