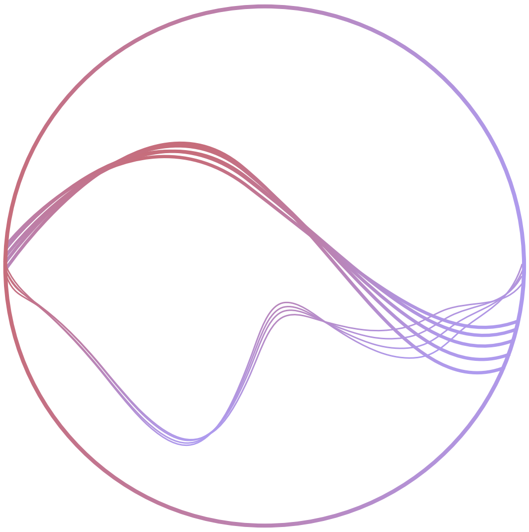
Enabling Domain Expert Evaluation of Emerging AI Technologies in Healthcare Settings



Ujjayan Dhar

TU Delft

MSc Thesis, 2024

Enabling Domain Expert Evaluation of Emerging AI Technologies in Healthcare Settings

"Introduction of Dual-Core Evaluation Framework for Conversational AI in Healthcare"

by

Ujjayan Dhar

Author

Ujjayan. U. Dhar

Student number: 5791073

Chair

Dr. Evangelos Niforatos

Assistant Professor

Department of Sustainable Design Engineering (SDE)

Faculty of Industrial Design Engineering (IDE),

Delft University of Technology

Mentor

Shatha Degachi

PhD candidate

Faculty of Industrial Design Engineering (IDE),

Delft University of Technology

Faculty of Industrial Design Engineering

A thesis report presented for the degree of Master of Science in Strategic Product Design.

In memory of my maternal and paternal grandparents

Acknowledgements

Words of gratitude can't fully capture the deep impact certain individuals have had on this journey. I am deeply thankful to my supervisory team for their unwavering support. My chair, Evangelos Niforatos, with his humour, friendliness, and honest feedback, greatly enriched my work during our discussions. Shatha's balanced expertise in technical research and user-centric methodologies inspired my approach. Collaborating with her as I merged my engineering background with research was a pleasure. Despite it being our first time working together as a team, our insights and iterative process built a strong and effective partnership.

I particularly appreciate their guidance through simple yet profound statements like, "This is your project; you are in control," and "Trust the process; everything will come together in the end." Their thoughtful advice helped me avoid many pitfalls. Looking back, I can say that I had the most responsive and supportive chair and mentor, whose constant support and invaluable insights have enriched my project and my growth as a Researcher, Engineer, and Designer.

This journey wouldn't have been possible without the support of my parents and my brother, who have been there for me every step of the way. I dedicate this degree to you. My parents have always been my biggest supporter, providing both financial and emotional support, keeping our family strong as a team, and encouraging me through the hardest days. Thank you, Mom and Dad, for giving me the chance to follow the dreams you couldn't.

I also want to thank my friends—Aashish, Arthur, Jagadish, Jayasmita, Keven, Nupura, Sandhya, Sandy, Yallaling and Yash—who stood by me most when I needed it. A special thanks to my friends from IDE at TU Delft, The Jam Nights, back in India and other countries who were there for both the academic challenges and the lighter moments. Your friendship has been a lifeline during this demanding project.

As I share my final work, I sincerely hope that it resonates with you, the reader. This project is not just an academic pursuit; it reflects my personal journey of growth and discovery during my time at TU Delft. I hope it serves as an inspiration for others embarking on similar paths.

With sincere gratitude,

Ujjayan Dhar

Delft, August 2024

Abstract

In the rapidly evolving healthcare landscape, integrating Artificial Intelligence (AI), particularly Large Language Models (LLMs), presents significant opportunities and complex challenges. This study examines the efficacy and implications of LLMs within healthcare systems, with a focus on enhancing user-centric evaluation methods for AI-generated healthcare advice. The research centres on three critical areas: Identifying the limitations of current evaluation methods, addressing the challenges related to the accuracy and reliability of LLM-generated advice, and proposing improvements to evaluation frameworks to enhance the practical application of these models in healthcare settings.

Our study utilizes a chatbot prototype trained specifically on healthcare datasets relevant to the Dutch context, exploring its application in real-world scenarios to validate and refine evaluation metrics. By involving healthcare professionals in interactions with the chatbot, we aim to ground our findings in practical, user-based experiences. The engagement with the prototype helps uncover vital insights into the AI's performance, emphasizing the necessity for models that generate reliable and ethical responses and resonate with professional healthcare practices.

The proposed research contributes to the broader discourse on AI in healthcare by offering a novel framework for assessing AI-generated responses through a blend of empirical user studies and theoretical analysis. This framework aims to mitigate the subjective nature of current evaluations and provide a more robust, standardized approach to assessing the impact of AI technologies on healthcare outcomes. Through this research, we aim to forge a path toward more responsive, responsible, and user-centred AI tools in healthcare, ensuring that they align with both professional standards and patient needs.

Keywords: Large Language Models (LLMs), User-Centric Evaluation, Conversational AI, Healthcare chatbots, Grounded theory, Healthcare AI

Contents

List of Figures

List of Tables

Abbreviations

AI - Artificial Intelligence

BLEU - Bilingual Evaluation Understudy

EU - European Union

EM - Exact-Match accuracy

FDA - Food and Drug Administration

GP - General Practitioner

GT - Grounded Theory

LLMs - Large Language Models

NLP - Natural Language Processing

OB/GYN - Obstetrics and Gynecology

ODQA - Open Domain Question Answering

ROUGE - Recall-oriented Understudy for Gisting Evaluation

SLMs - Smaller Language Models

SquAD - Stanford Question Answering Dataset

USMLE - United States Medical Licensing Examination

WHO - World Health Organization

1

Introduction

This chapter aims to provide a foundational understanding of Large Language Models (LLMs), emphasizing the critical importance of the involvement of digital applications like Conversational AI and exploring the multi-faceted challenges associated with evaluating the responses from a human-centric perspective by the inclusion of healthcare experts in the process. To address these challenges, the chapter discusses the necessity of a study that conducts a facilitation approach, followed by user interviews by interacting with a mock-up of digital applications and a prototype chatbot explicitly made for healthcare. Additionally, the chapter introduces the concept of co-creation and grounded theory as a method to gain deeper insights into healthcare experts' experiences with Conversational AI and talks about a novel framework with underlying metrics for evaluation of responses by Conversational AI, which can improve the healthcare expert and client communication — serving as the first step within the Dual Core framework.

Contents of the Chapter

Chapter 1: Introduction

1.1 Context

Introducing digital technologies into healthcare systems is essential to advancing in the domain towards better health outcomes. [90]. As of 2024, over 950 Artificial Intelligence (AI) models specifically designed for uses in medicine have been developed and authorized by the Food and Drug Administration (FDA) in the USA [91]. However, healthcare experts and researchers have used only a few models for specific medical fields; the remaining models are trained on various datasets rather than explicitly for healthcare [92]. The recent guidelines issued by the World Health Organization (WHO) in 2021 highlight the potential harm that can arise from AI models and the importance of ethical principles in AI implementation in healthcare, such as autonomy, transparency, explainability, accountability, and inclusiveness. [93]. In the European Union (EU), the proposed European AI Act significantly emphasizes using human-centric evaluation standards for AI systems, especially in high-risk areas such as healthcare, to ensure transparency, safety, and justice [94, 95].

LLMs have an important function in AI because they can generate responses that closely resembles human language. [96] The distinctive qualities of LLMs in healthcare, such as GPT-4, MedPaLM2, and later versions, can be seen in their specificity and accuracy in the generated responses [97, 98]. Furthermore, specialized language models such as PMC-LLaMA, PsyLLM, and GatorTron are effective in biomedical settings, adapting their findings to the specific requirements of medical information processing [99]. These models possess more than just data processing capabilities; they have emerging skills and functionalities as conversational agents [100]. They are trained on large and diverse datasets from various sources, such as Electronic Health Records (EHRs), PubMed, Doctor-patient interviews, medical examinations, and more, which makes them represent the advanced state of language processing technology [101].

Human interaction is essential to caring for patients in several medical fields. Accurate understanding of spoken language is an essential factor that impacts communication effectiveness. [102] Forming a client-medical expert interaction is vital for maintaining patient satisfaction and delivering the best possible treatment. [103] At the same time, medical reports in written language are essential in communicating with medical experts about their clients. These reports are documented reports on procedures for diagnosis and therapy and explain the results and their consequences. [104]. LLMs have the potential to be effective in various areas of medicine, such as AI-assisted Chatbots, which can understand complicated concepts while responding to a wide range of requests and prompts [105, 106]. Interactive conversational models in healthcare assist individuals, including patients as well as healthcare professionals, in various tasks such as evaluating symptoms, providing primary medical and health education, offering mental health support, coaching for lifestyle changes, scheduling appointments, reminding patients about medications, triaging patients, and allocating health resources [81]. Nevertheless, these models also raise concerns over the generation of false information, violation of privacy, biases present in the data used for training, and the potential risks of their misuse [107].

Within healthcare domains where inaccuracy or lack of information might result in life-endangering outcomes, due to misinterpreting the insights generated by these models is one of the most significant challenges [108]. Recent studies have mainly concentrated on assessing the knowledge capacities of LLMs by comparing them to other models, using metrics such as Bilingual Evaluation Understudy (BLEU) and Recall-oriented Understudy for Gisting Evaluation (ROUGE) resulting in a lack of understanding regarding the interactions between users and LLMs [109]. Therefore, evaluating the safety and accuracy of LLM behaviours, including their ability to generate accurate, reliable, trustworthy, and complete responses to healthcare queries and validate the claims with healthcare guidelines [85].

Given the crucial nature of healthcare applications, using conversational models requires establishing a unified and comprehensive set of foundation metrics. [110]. These metrics enable a careful evaluation of the models' capabilities from a human-centric perspective, allowing for the identification of potential errors that lead to significant improvements in delivering robust, accurate, and reliable healthcare services. [81]. Furthermore, the current evaluation metrics fail to consider several essential user-centric factors that demonstrate how well a chatbot forms a relationship with and expresses support and emotion to the client. [111]

However, more than simply bringing in a more detailed framework based on the existing framework with more metrics to evaluate is not enough, as it will make the evaluation process more subjective and complex. [85]. There are two essential gaps in this domain. Firstly, evaluation metrics have been re-designed from existing human-centered frameworks, not specifically from the healthcare domain, and then tested on a small sample size [81]. Secondly, the existing human-evaluation metrics are subjective to the views of healthcare experts on different metrics, which vary individually due to the lack of underlying factors or reasons supporting the metrics [112] and are time-consuming to evaluate every response [74].

One possible method of developing human-centric metrics implicates involving healthcare experts, including mental and physical health, as stakeholders through a collaborative co-creation method and in-depth user interviews.[81] Developing a standardized, robust, transparent set of evaluating metrics with underlying factors can accelerate the process of responses generated by Conversational AI more efficiently and effectively. [86]

This study introduces a novel framework that can be used to evaluate responses based on the response style and the knowledge of the statement generated by Conversational AI. This innovative approach leverages the capabilities to understand a statement in a communicative situation by its core principles, which are “content” and “expression” [89]. By analysing the conversation in this manner, the evaluation metric removes the subjectivity of the individual evaluation. It frees up valuable time for healthcare experts, allowing them to focus on the next steps involving prognosis and treatment plans in the user journey. Consequently, integrating a set of words that represent Response Style and Knowledge into our process will help us establish a concentrated co-creation study to develop underlying layers supporting a metric.

1.2 Research Aims and Objectives

To address the identified research gaps above, the project aims to answer the following research questions:

Research Question 1
What limitations exist in current user-centric evaluation methods for healthcare-focused LLMs, particularly in lifestyle advice?

Research Question 2
What are the key challenges of user-centric evaluations metrics in assessing the accuracy and reliability of LLM-generated health lifestyle advice?

Research Question 3
How can evaluation frameworks be improved to enhance the effectiveness of LLM’s response in providing lifestyle advice?

In this study, we use a chatbot prototype which generates response to healthcare related questions. As this chatbot is trained on datasets specific to healthcare context in the Netherlands, there is lack of global generalized data and misinformation in the generated responses. As these generated responses adheres to the guidelines and provides references to the generated content which builds trust as all the information provided is verified from medical and scientific responses and there is no inaccurate or incomplete information. Involving chatbot which provides responses as per the above factors like

Accuracy, Relevance, Trustworthiness will help us setting up a concentrated study to evaluate the responses in the way of response and the content of response.

The proposed workflow for healthcare experts using the chatbot prototype starts by providing a mobile version of a digital application mock-up featuring a chatbot to acquaint them with digital tools in healthcare. Initially, experts receive a brief context to understand their tasks and then engage with the chatbot by posing domain-specific questions. They subsequently assess the chatbot's responses by comparing them with their professional experiences and knowledge, which facilitates a critical evaluation of the chatbot's accuracy, relevance, and trustworthiness. This interaction allows experts to scrutinize the responses under various evaluative metrics, deepening the analysis. A user study involving 11 healthcare experts was conducted to probe further the chatbot's utility in real-world Dutch healthcare settings. This comprehensive evaluation validated the chatbot's performance and contributed to developing a robust framework for assessing responses generated by conversational AI, enhancing the objectivity and depth of user-centric evaluations.

In this study several contributions are offered:

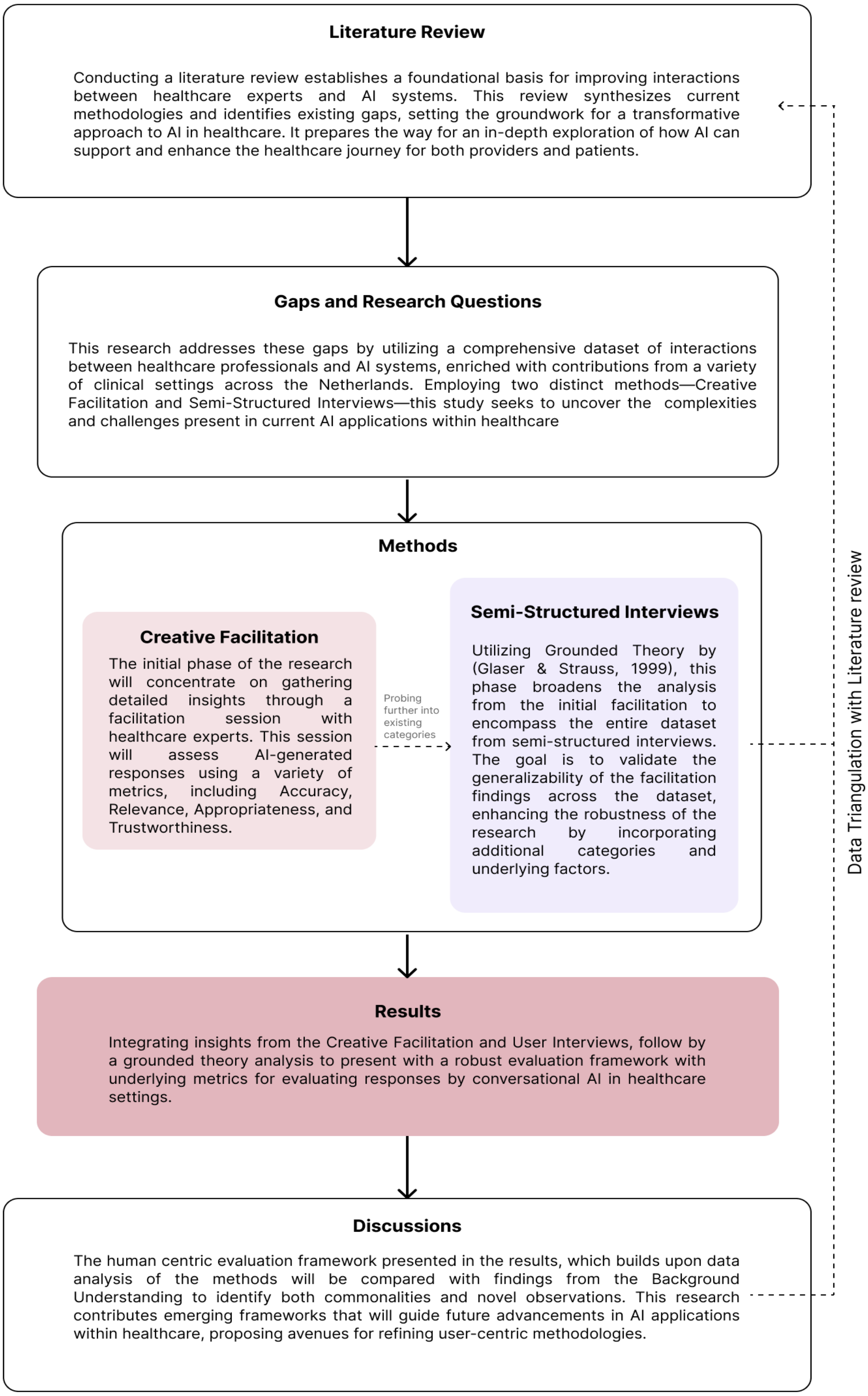
- *Introduction and exploration of how healthcare professionals can implement enhanced AI evaluation methods within their clinical workflows.*

- *Insights about the challenges and opportunities in using AI within healthcare, particularly through deploying and testing a chatbot prototype in real-world settings.*

- *A novel framework that assists healthcare professionals in assessing AI-generated advice, focusing on improving interaction through response style and content knowledge.*

1.3 Research Outline

To address these research aims and objectives a four-phase hourglass method by (Turbek et al; 2016) [118] which consists of Literature Review, Gaps and Research Questions, Methods and Discussion a top down approach which expands from the section of Gaps and Research Questions into different methods which give robust results.



Literature Review

Conducting a literature review establishes a foundational basis for improving interactions between healthcare experts and AI systems. This review synthesizes current methodologies and identifies existing gaps, setting the groundwork for a transformative approach to AI in healthcare. It prepares the way for an in-depth exploration of how AI can support and enhance the healthcare journey for both providers and patients.

Gaps and Research Questions

This research addresses these gaps by utilizing a comprehensive dataset of interactions between healthcare professionals and AI systems, enriched with contributions from a variety of clinical settings across the Netherlands. Employing two distinct methods—Creative Facilitation and Semi-Structured Interviews—this study seeks to uncover the complexities and challenges present in current AI applications within healthcare

Methods

Creative Facilitation

The initial phase of the research will concentrate on gathering detailed insights through a facilitation session with healthcare experts. This session will assess AI-generated responses using a variety of metrics, including Accuracy, Relevance, Appropriateness, and Trustworthiness.

Semi-Structured Interviews

Utilizing Grounded Theory by (Glaser & Strauss, 1999), this phase broadens the analysis from the initial facilitation to encompass the entire dataset from semi-structured interviews. The goal is to validate the generalizability of the facilitation findings across the dataset, enhancing the robustness of the research by incorporating additional categories and underlying factors.

Results

Integrating insights from the Creative Facilitation and User Interviews, follow by a grounded theory analysis to present with a robust evaluation framework with underlying metrics for evaluating responses by conversational AI in healthcare settings.

Discussions

The human centric evaluation framework presented in the results, which builds upon data analysis of the methods will be compared with findings from the Background Understanding to identify both commonalities and novel observations. This research contributes emerging frameworks that will guide future advancements in AI applications within healthcare, proposing avenues for refining user-centric methodologies.

Figure 1: Project Outline of the Hourglass method

2

Literature Review

This chapter explores the foundational aspects of Large Language Models (LLMs) in healthcare, particularly their use in lifestyle advice. It reviews the advancements in language modeling within natural language processing (NLP) and highlights the capabilities of LLMs across various applications. The chapter identifies critical research gaps and the evolving needs for AI in healthcare by examining the role of LLMs, from information retrieval to lifestyle advice. This overview sets the groundwork for further exploration of LLM integration to enhance healthcare outcomes.

Contents of the Chapter

Chapter 2: Literature Review

2.1 What are LLMs

Language modelling (LM) is an essential approach for gaining cognitive intelligence in the domain of natural language processing (NLP), and its advances and applications have been significant in recent times. [1, 2, 3].

To distinguish between language models with various parameter scales, researchers and academics use the term "Large Language Models" (LLMs) to describe pre-trained language models (PLMs) that include many parameters, ranging in the tens or hundreds of billions [3]. LLMs, or large-scale language models, are characterized by their large model size as also they have features like enhanced language understanding and generating capabilities compared to smaller-scale models. Particularly, LLMs have emergent skills that are absent in Smaller Language Models (SLMs) [6]. Large language models (LLMs) are AI systems that have been pre-trained to analyse and create text that closely resembles the human language in real-time [4]. It plays a crucial role in understanding, generating, and manipulating human language, serving as the foundation for a wide range of NLP applications [5]. These applications range from machine-based language translation, chatbots, sentiment analysis for understanding emotions in text conversations, and text summarization.

2.2 Role of LLMs in Various Domains

LLMs are currently being used across multiple domains, such as AI-assisted Chatbots, to carry out tasks like information retrieval, interactive conversation, and text generation. LLMs like as Dramatron are used for Creative work and Knowledge work. They can generate scripts based on specified prompts that have undergone evaluation for quality through collaborative writing methods.[40].

A study by (Wu et al., 2023) [39] discusses on how BloombergGPT can perform several financial tasks like including financial reasoning, numerical claim identification, and other financial activities. LLMs which are used in the field of law have similar characteristics with the healthcare industry, as both use personalized prompts to improve the level of accuracy of legal or medical question responses and information extraction. [42]

ChatGPT demonstrated to be able to do tasks in the healthcare industry, such as providing medical advice consultations, conducting mental health assessments, and simulating reports using Psy-LLM. Researchers suggest coming up with specialized LLMs in the healthcare domain for maximizing their effectiveness in this particular sector.

Especially, the Med-PaLM models receive more validation from doctors in terms of addressing customers' medical queries and achieve a level of skill similar like that of healthcare experts on the USMLE (United States Medical Licensing Examination). [10]

However, one of the major issues encountered during the fine-tuning process of chatbots is that these multi-turn interactions may lead chatbots to quickly "forget" former parts of the conversation or repeat themselves. [8]. In fields which has sensitive data such as Law and Medicine As a result of continuous updates documents of legislation and the emergence of new cases, the training/retrieval there is a high chance of data becoming less significant.[9]. Also, LLMs have the possibility to provide inaccurate medical information by misinterpreting medical terms and offering advise that clashes with existing medical guidelines.

2.3 General Applications in Healthcare

The healthcare industry is an industry that has a higher impact to human well-being both physical and mental. Since the development of ChatGPT, several researchers have been using ChatGPT or other Large Language Models (LLMs) in the domain of healthcare. LLMs can carry out several healthcare tasks, which includes collecting biological information, offering medical advice, evaluating mental health, and summarizing medical reports. [3]. Several generic large-scale machine learning models, such as GPT and LLaMA, have demonstrated potential in enhancing patient outcomes and transforming healthcare systems. Healthcare experts without knowledge in data science could face challenges in understanding and effectively using these models. [11]

For the purpose to improve the use of Language Models (LLMs) in the healthcare industry, researchers have developed LLMs developed specifically towards applications in healthcare. The development of Med-PaLM by Google has been centred around the possible ability of Language Learning Models (LLMs) that have conversational abilities. This type of innovation aims to optimize the effectiveness of PaLM for medical queries by fine-tuning the instruction prompts. This improvement has led to the development of Med-PaLM 2, which has been developed specifically for use in the healthcare field. [10], Med-PaLM 2 has been reported to be better than ChatGPT in terms of performance on United States Medical Licensing Examinations (USMLE) questions, achieving state-of-the-art results. [12].

The field healthcare is increasingly accepting the value of Natural Language Processing (NLP) techniques as an important part of artificial intelligence (AI) development [13] [14]. Even though LLMs have a significant future in the field of healthcare and are expected to be more common in this industry but they sometimes lack in providing reliable information. To improve the value of LLMs in the area of healthcare, two viable techniques can be used which are training the LLMs from the ground up using medical databases or fine-tuning the existing LLMs such as Med-PaLM.[11]. The term "hallucination effect" has been used to describe the irrelevant guessing behaviour observed in LLMs. An inquiry using GPT-3.5 for solving few medical questions from USMLE showed that the model consistently forecasted answers from choices A and D. During the analysis, the researchers found three fake citations throughout the article generated by ChatGPT. [21].

In order to properly use LLMs in the area of healthcare, it is necessary to come up with an extensive strategy which addresses the challenges and issues specific to the medical domain. As per Briganti, G. (2023) [20] Important factors to be considered includes transfer learning, domain-specific fine-tuning, domain adaptation, reinforcement learning with expert input, dynamic training, interdisciplinary collaboration, education and training, evaluation metrics, clinical validation, ethical considerations, data privacy, and regulatory frameworks.

Also, a study by (Yang et al., 2022) [15] shows that models like GatorTron exhibited skill in extracting and understanding patient information from clinical narratives. Combining such information into medical AI systems is important for enhancing healthcare delivery and improving patient outcomes. An important aspect of this process involves obtaining and documenting patient information from longitudinal Electronic Health Records (EHRs), which include both structured data (such as illness and prescription codes) and unstructured data (including clinical narratives like progress notes) [15].

Based on the unstructured data, another healthcare Conversational AI which is ChatDoctor, was developed using LLaMA and a has dataset of 100,000 patient-physician conversations. The model exhibited significant improvements in comprehending patients' needs and delivering precise recommendations as the chatbot gave responses based on real-life conversations and suggestions based on the context. [16]. Another example of this can also be observed in Baize-healthcare [17], an open-source chatbot designed for healthcare that was developed as well using LLaMA. The model has undergone fine-tuning using the MedQuAD dataset [18], which consists of 46,867 medical dialogues between patients and doctors. It shows high accuracy in multi-turn conversations. These kinds of models are going to help the development of conversation models in healthcare, where suggestions by conversational AI chatbot will be specific and not generic.

It is important to bring LLMs into clinical practice, improve their ability to be understood, and enhance the collaboration among healthcare experts, patients and AI to assist in clinical decision-making. [11]. With the rise in the applications of AI in healthcare, legal and regulatory challenges will arise. These concerns include issues related to responsibility for AI-generated suggestions, compliance with data privacy laws, and thus there arises a need for standard protocols to evaluate AI systems. [20].

2.4 LLMs in Lifestyle Health Advice

LLMs have shown effectiveness by providing increasing number of people access to medical knowledge. Particularly, chatbots like ChatDoctor which use LLaMA, showed higher accuracy in understanding what patients want while providing accurate advice [22]. LLMs also offer the necessary skills to carry out multiple roles in the field of healthcare service design as it can be used to track the medical state of patients, especially people who have chronic illnesses. [25].

Bulck and Moons (2023) [27] found that when compared to Google search, 40% of the 20 experts (19 nurses; 1 dietitian) considered answers from ChatGPT to be more valuable, 45% considered them to be equally valuable, and 15% considered them to be less valuable. Hence, many experts projected that patients would more depend on LLMs, specifically ChatGPT, and reduce their dependence on Google searches because of the high of accuracy and ease of access to answers provided by LLMs [24].

LLMs are also used for evaluating the effects of medications. Research shows that ChatGPT was used to project and understand drug-drug interactions [23]. This study investigated the views of LLMs with regards to drug pairing or interaction. The correctness and conclusiveness of their responses were evaluated [24]. The results show that out of the 40 pairs of Drug-Drug Interactions, 39 responses were found to be correct for the first question. Among these 39 correct answers, 19 were conclusive while 20 were inconclusive. Regarding the second question, out of a total of 40 pairs, 39 of them are correct. Among these pairs, 17 answers are conclusive while 22 answers are inconclusive.

A study done by Liu et al., (2023) [28] provides insights into the effectiveness of Language Models (LLMs) in accurately analysing data not only from patients-doctors conversational recordings but also by patient's wearable devices and medical sensors. The research focused on various applications such as cardiac signal analysis, physical activity recognition, metabolic calculation, and the estimation of stress reports and mental health screeners, all considered the data from wearable and medical sensors. Due to usage of such data more personalized data analysis is provided to individuals, rather than depending on internet data, offers more reliable and trustworthy insights to assist individuals in maintaining a healthy lifestyle.

A health data-collecting tool called CLOVA CareCall, built by NAVER AI, was implemented in South Korea. This program was conducted to provide emotional support to individuals who are socially isolated and particularly those with lower incomes. It conducted analysis of regular conversations and generated health reports that included metrics such as meals, sleep, and emergencies. It demonstrated to be effective in alleviating feelings of loneliness. Social workers used the reports that were generated and call recordings to monitor the well-being of users, thus making the workload efficient for the social workers. [26].

2.5 Applications and Impact in Lifestyle Advice, Mental and General Health

2.5.1 Healthy Lifestyle Advice

LLMs are used in the monitoring of chronic diseases by healthcare experts by offering continuous lifestyle guidance. These models can provide essential guidance on diet, exercise, and other changes in lifestyle that are important for managing conditions such as diabetes, hypertension, and cardiovascular diseases. The continuous personalized nature of the advice helps patients in maintaining a healthy way of life, thus improving recovery results. (Akilesh et al., 2023) [29].

Large Language Models (LLMs) like ChatDoctor and ChatCounselor have been specifically developed to provide customized healthcare guidance according to user-provided information, including symptoms and lifestyle habits. [16] These models use large datasets, including those generated from interactions between patients and doctors, the LLM analyse this data to recommend lifestyle modifications that can help in the management or prevention of chronic conditions. These recommendations are monitored and constantly revised based on the most recent health guidelines to cater to the distinct requirements and goals of individual users thus avoiding inaccuracy of information. [37] [22]

Bender et al. (2021) and Ji et al. (2023) [32] [33] say that synthesizing text using Language Models (LLMs) involves risks, such as the possibility for hallucinations and the production of biased or harmful text. In a scenario involving nutrition counselling and limited public resources, data can be gathered by combining LLMs (Language Learning Models) with crowd-workers and nutrition specialists. As a result, HAI (Human-AI) Coaching. which consists of dataset for nutrition counselling has been analysed by experts which includes of about 2.4K dietary challenges provided by crowd workers, along with approximately 97K supporting texts generated by ChatGPT.[31]

However, after careful investigation by Balloccu et al., (2024) [31] uncovers that ChatGPT, despite its ability to generate intelligent and human-like text, additionally shows risky opinions, especially when it comes to sensitive topics like mental health, making it unsafe for unsupervised use. A small number of participants noted that ChatGPT frequently generates responses that are "safe but useless," offering generic advice. Additionally, ChatGPT lacks the level of trust with clients that doctors develop over time, which allows doctors to better understand their clients.

2.5.2 Mental Health Support

Using natural language processing (NLP) approaches, speculation of individuals' mental health is done through the analysis of their social media posts on platforms such as Facebook and Twitter. These findings are used to create online platforms that link individuals to health information and support, as well as to design personalized treatments. [35].

Due to the COVID-19 pandemic, there is an increasing demand for quick and skilled mental healthcare services. According to Yurayat and Tuklang (2023) [36], online psychological counselling has increasingly become the primary method of providing counselling services through the internet. ChatCounselor demonstrates the application of Language Models (LLMs) in the field of mental health support by offering the services of counselling and guidance on mental well-being. This model has been trained using actual counselling sessions and is able to provide individualized psychological assistance, including suggestions for managing stress, anxiety, and depression. The model's responses are customized to the emotional and psychological state of everyone, resulting in advice that is extremely pertinent and influential [34].

The study conducted by Lai et al., (2023) [38] introduced the Psy-LLM framework, where the use LLMs is in the form of an AI-based assistive tool for question-answering during psychological consultations due to which it provided clients more accessible to mental health professionals. Their framework combines pre-trained Language Models (LLMs) with real-world professional Question and Answer (Q&A) data from psychologists, as well as a comprehensive collection of psychological papers

However, research conducted by psychologists on AI has highlighted the importance of incorporating human review when evaluating the performance of these language models. It is also necessary to make further improvements in order to ensure more accurate and successful outcomes. This is crucial because generating hallucinations or incomplete results in this field can have negative impacts on the mental well-being of clients. [38]

2.5.3 General Health and Wellness

Healthcare experiences of patients are multidimensional which includes the dynamic doctor-patient interactions, also a variety of diagnostic and treatment approaches, adherence to prescribed lifestyle or behavioural adjustments, and continuous preventive health measures. The patient's healthcare journey is non-linear, consisting full of complex integrated series of events. [43,44,45,46]

As per [47] the diagnostic procedure in medicine consists of different steps which are influenced by the context, symptoms in patients, clinical expertise and available diagnostic tools. The introduction of LLMs can work in two different approaches where the patient has no limits in providing prompts when providing a particular knowledge input, in this case the LLM will give help even if the user's input is insufficient for the LLM to achieve its stated goals and objectives. The second approach allows the LLM to gain a more thorough understanding of the user's symptoms and age by using a more limited and step-by-step approach. [46]

In both scenarios, health sensors may retrieve data and evaluate it using the analytical and machine learning services. But in both the above approaches if the doctor, as the second actor, can evaluate and validate the primary care AI interactions and the patient's review of the primary care AI it will help the system to improve to address the AI shortcomings. [48]

Another study on the evaluation of [49] GatorTronGPT demonstrates that clinical LLMs may be used to create clinically relevant material, which has the potential to aid in the documenting and coding of patient information in Electronic Health Record (EHR) systems, therefore decreasing physicians' substantial paperwork load. LLM prompt-based text creation has the potential to assist build treatment regimens by combining clinical standards and patients' past information with EHRs

But such applications raise concern for ethical considerations, sometimes due to lack of time or difficulty in understanding the agreements people sign up for or ignore them. This can pose a threat to privacy of individuals using these applications as per [50] as Data privacy, ethical issues, and the integration of AI systems with current healthcare infrastructures should all be major areas of future research.

2.6 Specific Challenges in Lifestyle Advice

2.6.1 Reliability

The reliability of LLMs is crucial to their usage in healthcare. The performance of LLM answers is influenced by factors like as accuracy, consistency, interpretability, and the quality of the data set used for training. A study by [24] found most of the research connected to prediction of texts, 72% of studies related to summarization, and 93% of studies related to medical knowledge queries have problems regarding their dependability. The lack of reliability in LLMs mostly results from restrictions in data gathering sources and the model's limited medical knowledge. The general-purpose character of ChatGPT could compromise its reliability in the context of self-diagnosis. [63]

Firstly, it is essential to evaluate and validate ChatGPT's responses using trustworthy sources, as this helps in the process of learning and develops critical thinking. However, it should be noted that these responses should not be considered as replacements for knowledge provided by healthcare institutions. Another study by [31] mentions, similarly as patients converse with healthcare experts, generative questions and prompt should be made two-way conversation instead than binary or definitive ones when phrasing inquiries. This approach leads to constructive conversation and helps prevent replies which are misleading from the LLM. Also training LLMs using the medical curriculum and involving students to better identify ideal responses and comprehend any limitations [32]. These steps can help to prevent any biases that may be present in these models.

In order to minimize the risks related to false information and errors, a study by researchers (Thapa and Adhikari, 2023) [56] mention need to develop robust validation and verification processes for LLM results. This involves comparing the information provided by LLM with reliable and updated medical sources, conducting independent evaluations, and gathering expert advice where necessary. Designing evaluation frameworks along with suggestions specific to LLMs in healthcare lifestyle advice ensure the accuracy and reliability of the information generated by the Conversational AI.

2.6.2 Hallucination

LLMs aim to generate text using clues from context and randomness rather than targeting factual accuracy. Due to which, this raises a risk to what AI experts refer to as "hallucinations," to which healthcare experts would more accurately refer as Confabulations which are basically suggestive statements by a person or a system that seem accurate and trustworthy but eventually incorrect. Confabulations may differ in level of detail, ranging from basic to complex or even irrational However, the way in which they are presented is very persuasive and this can make it difficult to recognize them are they true or false by patients or even to some extent by healthcare experts. [60]

ChatGPT in healthcare may provide inadequate information or demonstrate a difficulty in discerning between truthful and deceptive remarks [64,23]. A case of confabulation is demonstrated in research conducted by Schwartz et al., (2024) [60], where GPT-3.5 was instructed to generate a care plan for a patient suffering from HIV-associated cryptococcal meningitis. The AI system provided an extensive and structured set of suggestions. Although the strategy may seem thorough and acceptable to anyone without specialized knowledge, it contains major error for example : The suggestion to immediately start antiretroviral therapy (ART), which was suggested by a random study that appeared which increased mortality rates when compared to delayed ART initiation in individuals with HIV diagnosed with cryptococcal meningitis which was suggested by healthcare guidelines, these suggestions, goes against standardized clinical practice guidelines and has the potential to cause significant harm to patients [65]. Another similar research conducted by Jo et al. (2023) [26], where it was found that Healthcare LLMs specific to mental health such as CLOVA CareCall, produced by NAVER AI, had the inclination to make confident and inappropriate claims to patients dealing with mental health issues, thus making them trust the misleading suggestions by the Conversational AI without any healthcare expert supervisions. This eventually places an extra stress on therapists and result to a loss of trust between healthcare experts and clients.

The causes and mitigations of hallucinations in LLMs is currently being actively researched. LLMs could show a bias towards generating responses that seem more trustworthy or natural, which increases the possibility of hallucinations. [66]. Various methods have been suggested to address the problem of hallucinations. One approach is to change the training process in order to restrict hallucinations, as seen in the idea of known as reality grounding. [67]. Another approach is to design the model with a wider and more varied dataset, possibly mitigating the probability of the model making inaccurate responses. [68]. Furthermore, a study conducted by Ferrara (2023) [69] indicates the significance of collecting verified or fact-checkable data from trustworthy sources during the training process of the LLM. This approach is referred to as "Human-in-the-loop approaches," which train the model to favor accurate data over its own assumptions. However, achieving this goal requires careful review of the data and evaluation of fairness metrics used, which depend on collaboration between healthcare experts and policymakers.

2.6.3 Lack of human centredness

Recent LLMs in healthcare, such as GPT4 and PaLM [70], generally focus on general domains and are trained using publicly accessible generic databases or documents. Due to which these models lack specific training and skillset in the field of lifestyle advice, therapy, which requires the knowledge of non-verbal cues and the human nature of counselling.

A study done by [72] LLMs provide more accurate question-and-answer outcomes that align with human behaviour, in contrast to the fragmented and unclear information often acquired from traditional searches. Therefore, it is essential to evaluate the Open Domain Question Answering (ODQA) capabilities. The effectiveness of replying to open-domain questions has a major impact on user satisfaction. Commonly employed datasets for testing purposes include SquAD (Stanford Question Answering Dataset) and Natural Questions. The evaluation of these datasets is conducted using F1 score and Exact-Match accuracy (EM) as metrics [142, 143, 144]. But the physician's skill of integrating various sensory inputs during conversation is very much necessary in the response of LLMs. When interacting with a patient, it is not just about receiving information through text or speech; it also involves processing auditory, visual, somatic, and even olfactory stimuli. [71] The human way of conversation combines these various kinds of sensory and spoken information in a way that cannot be generated by LLMs using text. This practical elements of "clinical sense" cannot be fully acquired or expressed within a framework that relies just on text. [72].

Another research [72] [73] mentions an important fact that would make domain experts comfortable with using LLMs as an aid is ensuring a human (expert) is in the loop to perform extensive verification. This is difficult now due to the system design of LLMs, which does not readily permit verification. But if one could inspect inputs, there would be a trade-off between efficiency gains and the time required for verification. Therefore, a standardized human based evaluation seems to be necessary. [74].

2.7 General limitations of LLMs in Healthcare

Large Language Models (LLMs) have potential applications in healthcare but also face several challenges and limits that need to be addressed for their effective and ethical use in healthcare environments. This section discusses the main challenges connected to LLMs in the healthcare domain. This can generally be categorized into five groups but can be distributed among generic limitations like 1) Bias, 2) Data privacy and for specific challenges in lifestyle advice such as 1) Reliability, 2) Hallucination and 3) Lack of Human Centredness.

2.7.1 Bias

Language models may unknowingly reflect bias when the training data used for their development is biased. According to Schramowski et al. [51], vast and complex pre-trained models designed to mimic human conversation might unintentionally sustain unfairness and biases. As a result, this might result in incorrect evaluations and recommendations, which can mislead patients in many areas of healthcare.

Another research, which was conducted by [60], emphasizes the problem that AI systems could provide biased response due bias being present in their training data, resulting in discrimination and damage against marginalized populations [57, 58]. Further the study investigates on a case study where, a widely employed AI algorithm in the United States consistently gave more priority to White patients than Black patients when allocating healthcare resources [59]. These cases occur due to the possibility of deepening the existing gaps in healthcare, especially if automated evaluations performed on existing healthcare data were wrongly assumed to be unbiased.

Additionally, additional research conducted by [56] also emphasizes that if the training data has biases related to demographics, illness prevalence, or treatment results, the resulting outputs could reflect and replicate these biases thus providing unfair and impartial healthcare results. On a comprehensive study by Hadi et al., (2023) [61] highlights the four different types of biases which are as follows:

- **Training data bias:** Language models commonly depend on large datasets of human language for training. If these datasets exhibit biases associated with characteristics such as race, gender, or socioeconomic status, the model has the possibility to replicate these biases in its outputs.
- **User interaction bias:** The output generated by Chatbots can be affected by the input received from users. If users regularly ask questions that are influenced by bias or prejudice, the model has the possibility to learn and continue to display similar biases in its replies. If users often ask biased queries that specifically target a certain group, the model may create replies that strengthen and perpetuate these biases. [52]
- **Algorithmic bias:** Algorithms used in training and running language models and Chatbots can also induce biases. [53]
- **Contextual bias:** Chatbots provide responses based on the context provided by users. Biased replies may be produced by the model if the context contains bias associated with factors such as the user's demographic location or language. [54]

2.7.2 Data-Privacy

The use of Large Language Models (LLMs) in healthcare situations raises significant concerns regarding data privacy, including issues related to data protection, the potential for re-identifying people, and ethical questions around how to make use of patient data. LLMs could distinguish sensitive personal characteristics from non-sensitive information, thus breaching an individual's right to privacy. [55].

A major concern is the unintentional presence of personally identifiable information (PII) in the datasets used for training these models, which may result in violations of patient confidentiality. To reduce these dangers, it is crucial to use effective data safety strategies, which includes effective anonymization methods, secure data storage protocols, and robust compliance with ethical norms. These steps are essential to preserving the trust of those participating in studies, maintaining the authenticity of study techniques, thus maintaining patient privacy. These safeguards are necessary because it is crucial to strike a balance between utilizing the capabilities of LLMs in medical research and safeguarding sensitive patient information. [62].

Ethical considerations and stringent data privacy measures are paramount when using LLMs in biomedical research. Researchers are tasked with managing sensitive patient data conscientiously and complying with privacy regulations. The implementation of comprehensive data protection measures, including data anonymization and secure storage practices, is crucial to maintaining the confidentiality of patient information and sustaining the trust of those involved in research [56].

2.8 Existing Frameworks and their evaluations

This section aims to answers the **First Research Question "*What limitations exist in current user-centric evaluation methods for healthcare-focused Large Language Model (LLMs), particularly in lifestyle advice?*"**

Due to the significant challenges involved in using Large Language Models (LLMs) in healthcare, such as bias, privacy problems, dependability, and hallucinations, it is crucial to thoroughly examine the present evaluation methods for these technologies. Robust evaluation metrics are essential for both analysing the reliability of LLMs by their responses, but also can be integrated much better into healthcare environments.

There are two types of categories of evaluation procedures. Automated evaluation and Manual evaluation both play their role in evaluating responses in Language Model (LLM) research. Automated evaluation commonly uses different metrics and indicators to measure the performance of models, including BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-oriented understudy for Gisting Evaluation), and BERTSScore (Bidirectional Encoder Representations from Transformers Score). A study by [72] mention that these metrics evaluate the accuracy of content generated by Language Model models. These metrics are helpful for researchers to efficiently measure model performance on large data and make comparisons between various models. However, these metrics are model based and do not incorporate the knowledge of medical concepts such as symptoms, diagnostic tests, diagnoses, and therapies mentioned in the response and lacks effective evaluation of the semantic meaning of the generated text.

Various domain-specific LLMs have been created using general-purpose model and training approaches. BioBERT and PubMedBERT are BERT models specifically trained on biomedical data from PubMed [83]. Med-PaLM [4] was designed using carefully selected medical datasets and input from healthcare experts. It showed positive results, such as achieving a 67.6% accuracy on the MedQA examination. Chat-GPT, even though lacked medical knowledge, passed all three sections of the USMLE and obtained an overall accuracy rate of over 50% across all tests, with a number of them above 60% accuracy [84]. Nevertheless, automated evaluation is limited by its lack of capacity to fully capture all aspects of language understanding from a user-centric level which depends on factors like trust-building, customized responses, emotional support by empathy which have an impact on real-time clinical processes. [81].

The study conducted by Novikova et al., (2017) [82] suggests manual evaluation is more reliable for some broad responses generated by chatbots. Manual evaluation commonly involves the subjective assessment and judgment of human subject matter experts about the quality of outputs provided by a model. This evaluation methodology can efficiently identify the effectiveness of models in specific tasks or contexts and detect mistakes and faults that automated evaluation can fail to detect. But manual evaluation faces limitations such as significant time investments and subjectivity of responses. Hence, sometimes its important to combine the benefits of automated and manual assessment to thoroughly evaluate the performance of language models.

2.8.1 QUEST – Five Principles

Healthcare professionals are using Generative Artificial Intelligence (AI) models, such as large language models (LLMs), to improve patient care and decision-making. It is important to evaluate the responses from these models, particularly in the healthcare domain, to ensure accuracy, reliability, and adherence to ethical standards. The research study conducted by Tam T. Y. C et al., in 2024 [85], offers an extensive review of existing literature on human evaluation frameworks in various healthcare applications. The study goes beyond the traditional evaluation dimensions of User, Domain, and Task [81] and incorporates additional factors such as Evaluation dimensions, sample types and sizes, recruitment of evaluators, and the evaluation process. The study also examined healthcare apps that addressed various elements of patient care, clinical practice, biological health sciences research, and education.

Assessing the effectiveness of LLMs responses in terms of performance, accuracy, or memory is challenging due to the complexity or layers present in healthcare contexts. the study proposes a framework named QUEST which can act as a foundation for human evaluation metrics for Language and Learning Models (LLMs) in the healthcare domain. This approach incorporates the evaluation of aspects such as accuracy, empathy, and trust to ensure that these models adhere to robust standards for clinical settings.

Based on the aforementioned themes the QUEST Framework consisted of five principles 1) Quality of Information 2) Understanding and Reasoning 3) Expression Style and Persona 4) Safety and Harm and 5) Trust and Confidence which consisted of different dimensions as layers to achieve specific principle:

Principle(s)	**Dimension**
Quality of Information	Accuracy, Relevance, Currency, Comprehensiveness, Consistency, Agreement, Usefulness.
Understanding and Reasoning	Understanding, Reasoning.
Expression style and persona	Clarity, Empathy
Safety and Harm	Bias, Harm, Self-awareness, Fabrication, Falsification or Plagiarism
Trust and Confidence	Trust, Satisfaction

Table 1: QUEST's Five Principles

The study suggests assessing the framework using Comparative approaches, in which the LLM generates responses for possible diagnosis and treatment solutions. These suggestions may then be reviewed using healthcare experts. The study also suggests evaluating answers that are relevant to different healthcare domains, since each area has different levels of importance and potential influence in clients lifestyle. Also, In order to conduct a statistical analysis of the framework, the study proposes evaluating around 130 sample responses in clinical settings, 100 sample responses in research settings and clinical trials, and 100 sample responses from medical education, which includes medical examination.

2.8.2 The Four healthcare evaluation metric group

In research conducted by Abbasian et al. (2024) [81], a set of evaluation metrics was suggested to evaluate the performance of different healthcare chatbots. These metrics are meant to consider the perspective of end-users and use a combination of automatic and manual assessment methods. The metrics derived from evaluations conducted by healthcare experts, who assess the Conversational AI's language processing abilities, its impact on real-world tasks in healthcare, and its effectiveness in user-interactive conversations. This study included an activity in which healthcare specialists engaged with several chatbot models, while considering three variable types: User Type, Domain Type, and Task Type.

The User Type assessment focused around evaluating safety, privacy, and usefulness of information provided to patients and healthcare professionals. It placed a particular emphasis on the importance of having accurate and reliable medical material. The Domain Type examination evaluated the chatbots' capacity to engage in conversations about a wide range of health subjects, including both general health and specialist areas like oncology. The assessment criteria were tailored according to the specific needs of each domain. Lastly, the evaluation of Task Type took into account the chatbot's functionality, whether it acted as an assistant or provided detailed diagnosis and treatment plans. This evaluation critically assessed the correctness and factual reliability of the replies made by the chatbot.

Based on the above three variables against different healthcare chatbots the study proposed an evaluation metrics for the healthcare chatbots which were as follows:

User Centered Metrics	**Underlying Low level metrics**
Accuracy	Intrinsic, SSI, Robustness, Generalization, Conciseness, Up-to-dateness, Groundedness
Trustworthiness	Safety and security, Privacy, Bias, Interpretability
Empathy	Emotional support, Health library, Fairness, Personalization
Performance	Memory efficiency, FLOP, Token Limit, Number of Parameter

Table 2 : Four Healthcare metrics group

The differentiating factor of the above Metrics was the relationship between metrics and amongst each other, where a greater number of underlying factors inside a Metric might provide more complexity, thus impacting other metrics in both good and negative ways. In addition, the evaluation of chatbot responses from a human-centered perspective was conducted by incorporating a combination of human evaluation metrics and model-based metrics. An example was provided to illustrate how the accuracy of responses in adhering to healthcare guidelines affected the metric scores measured by ROUGE and BLEU, which were used to assess the robustness of the models.

2.8.3 ChatGPT’s evaluation in German OB/GYN exams by Riedel et al.

In research conducted by Riedel et al in 2023 [86] examined the effectiveness of ChatGPTs in generating responses related to medical knowledge in the field of obstetrics and Gynecology (OB/GYN). The study examined two datasets to assess and evaluate the accuracy and reliability of ChatGPT in understanding medical topics. One of the datasets was sourced from the OB/GYN course examination at the University Hospital of the Technical University of Munich. This examination primarily assessed the theoretical understanding of the field, covering subjects such as Gynecology, prenatal and perinatal medicine, gynecologic oncology, endocrinology, and reproductive medicine. The second dataset includes 300 multiple-choice questions (MCQs) from the state level exam (Zweiter Abschnitt der Ärztlichen Prüfung). Each question includes five answers, out of which only one is right. The study was done in the German language and skipped questions which involves images, since the study was on the ChatGPT3 version.

The responses were evaluated on the basis the five categories or variables which evaluate the quality of data regarding the responses related to OB/GYN settings. These evaluation framework metrics were also considered from an existing list of evaluation framework metrics by Wang and Strong [87] which is limited to the field of information systems which focuses on responses being reliable, accurate and useful. The Five categories are as follows

Categories(s)	**Explanation**
Ease of Understanding	Was the answer clearly and precisely formulated in a way that was easy to understand?
Concise representation	Was the answer clearly structured and divided into sections that facilitated readability?
Accuracy	Did the facts mentioned in the answer correspond to the current scientific literature? Were the statements logical and understandable?
Completeness	Was the answer complete, and were all aspects of the question adequately addressed? Was important information omitted, or were there unnecessary details?
Relevance	Was the answer directly related to the question asked, or was there any ambiguity in the answer?

Table 3 : Riedel et al's Evaluation metrics

Results show that on an average 83.1% of medical students answered correctly and ChatGPT in the same datasets provided 85.6% of correct answers. To test the consistency of the responses by ChatGPT the study also conducted a second round of validation where ChatGPT provided 88.7% of correct responses.

The Quantitative evaluation of ChatGPT's accuracy was also taken into account when addressing clinical questions of different levels of difficulty. The questions were categorized as either easy or difficult, depending on how they compared to the average scores of medical students. The average score for the OB/GYN test was 83%, while the average score for the state test was 73%. The analysis revealed that there was no statistically significant difference in ChatGPT's performance on the OB/GYN questions, regardless of the level of difficulty (p-value = 0.1). However, the performance on easier questions in the state test dataset show a significant better score (p-value <0.01). The numerical difficulty ratings to each question were assigned on the state examination, ranging from 1 (very easy) to 5 (very difficult). The analysis showed a strong correlation between difficult questions and the number of correct and incorrect responses. Specifically, as question difficulty grew, the number of correct answers decreased, and the number of incorrect answers increased. This correlation was shown to be statistically significant at a p-value <0.001.

In the Qualitative analysis, for the reasons of incorrect answers, the main reason was incorrect internal knowledge and the flawed databases. The limits of ChatGPT may arise from being trained on a dataset that lacks sufficient representation of essential medical information or from errors made by humans during the initial training phases. These factors might have negatively impacted the model's performance. In the further analysis of the responses on a qualitative approach by (n=3) medical experts using a five-pointer Likert scale on the factors of 1) Ease of understanding 2) Concise representation 3) Accuracy 4) Completeness and 5) Relevance. The Likert scale results show that both correct and incorrect responses were evaluated highly in terms of for ease of understanding and concise representation, with mean scores of 4.8 and 4.6 for correct responses, and 4.2 and 3.9 for incorrect responses, respectively. However, the metrics of accuracy, completeness, and relevance scored significantly higher for correct responses (p-value <0.0001).

2.8.4 CLEAR tool

To investigate the impact of accuracy on the information provided by the AI based Conversational models, Sallam et al. (2023) [88] conducted a study to assess the quality of generated responses on healthcare information by four AI models like ChatGPT 3.5, ChatGPT4, Microsoft Co-pilot and Google Gemini. The initial evaluation methods were to check whether the responses adhered to the following conditions:

- Does the content provide the needed amount of information without being too much or too little?
- Is the content accurate in total, without any false information?
- Is there enough evidence to support the information included in the content?
- Is the content characterized by being clear (easy to understand), concise (brief without overwriting), unambiguous (cannot be interpreted in multiple ways), and well-organized? and
- Is the content focused without any irrelevant information?

Which further developed into the following CLEAR Metrics

Metrics(s)	**Explanation**
Completeness	Is the content sufficient?
Lack of false information	Is the content accurate?
Evidence	Is the content evidence based?
Appropriateness	Is the content clear, concise, and easy to understand?
Relevance	Is the content free from irrelevant information?

Table 4 : CLEAR Tool

A pilot testing was conducted and (n=32) participants provided feedback were healthcare professionals and were chosen on their ability to critically evaluate healthcare information and the sample consisted of nurses (n=11), physicians (n=14), pharmacists (n=4) and laboratory technicians (n=3) from Jordan University Hospital and Mediclinic Middle East, Dubai. After the feedback from the pilot testing the CLEAR tool was used for the final testing where a new chat was selected after each response and the same prompt was used across all the AI models.

Four AI-based models were used to analyze five health-related queries. The information was evaluated by two independent raters using the CLEAR tool. Microsoft Co-pilot obtained the highest average CLEAR score of (24.4±0.42). It was followed by ChatGPT-4 with a score of (23.6±0.96), Google Gemini with a score of (21.2±1.79), and ChatGPT-3.5 with a score of (20.6±5.20). The inter-rater dependability demonstrated significant agreement, especially with ChatGPT-3.5. This evaluation highlights the comparative performance of these models in generating high-quality health information.

Framework	**Description**	**Limitations**
QUEST Framework	Framework consisted of 1) Quality of Information 2) Understanding and Reasoning 3) Expression Style and Persona 4) Safety and Harm and 5) Trust and Confidence	Failed to incorporate healthcare specific metrics. Categories lacked definition and didn't have sub-categories to strengthen it further. Metrics were derived from model based evaluation framework
Four healthcare evaluation metrics	Framework was based on evaluation of responses on the basis of three parameters a) User b) Domain and c) Task and the metrics were Accuracy, Trustworthiness, Empathy, and Performance	This framework was designed as structure to assess the performance of other framework from a machine perspective. This framework was task specific and can be considered as starting point for other frameworks Their focus is on automatic evaluation > human evaluation
Riedel et al Framework	Framework is based on the sole concept of Data Quality and then metrics consisted of factors like 1) Ease of understanding 2) Concise representation 3) Accuracy 4) Completeness and 5) Relevance	Focused on ChatGPTs knowledge database and not on the way it responded. Lacked contextual awareness and coherence which which continue for longer conversations
CLEAR Tool	Framework consisted of 1) Completeness 2)Lack of False information 3) Evidence	This framework was designed with 32 participants which consisted of nurses, physicians, lab technologists and pharmacists, as they were

	4) Appropriateness and 5) Relevance	a part of the same ecosystem of the authors the research feels biased. Lacked in-depth examination of responses to provide a much detail insight into how the evaluation was used and the sub-categories present in it.

Table 5 : Summarisation of Existing Frameworks and their limitations

2.9 Gaps and Opportunities for Improvement

The application of LLMs in healthcare is promising but reveals several gaps. Critical factors include transfer learning, domain-specific fine-tuning, and reinforcement learning with expert input [20]. Addressing these challenges requires interdisciplinary collaboration, robust evaluation metrics, clinical validation, ethical considerations, and data privacy frameworks. Incorporating human review is essential to mitigate issues like hallucinations and inaccuracies, which can adversely affect mental well-being.

The QUEST framework developed was evaluated lacked real-time testing in clinical scenarios and suggested testing based on Statistical, Comparative and Blinded Vs Unblinded approach. In the Statistical analysis the key aspect would be to evaluate the responses based on coherence, logical reasoning, categorization of information using Likert scales. Comparative analysis can be done by comparing LLMs responses against human-generated responses or LLMs responses as compared to healthcare guidelines. On the basis of these testing validation of the framework areas of development in LLMs and new frameworks can be developed to make the evaluation metrics more robust [85].

The study by Abbasian et al. (2024) [81] on the Four healthcare evaluation metric also had limitations as it did not incorporate a healthcare-focused benchmark that clearly defined each category and its corresponding sub-categories. The guidelines could also be developed for the three different factors such as user types, domain types and task type-based evaluation. Another drawback of this framework was that it relied on model-based evaluation metrics, which were intrinsic factors like such as Match-rate, Dialogue Accuracy, and Average request turn, as well as extrinsic factors like Reliability, up-to-datedness, Healthy behaviours, and Emotional support and not based on existing evaluation metric specific to healthcare domain.

The study on ChatGPT's performance in OB/GYN exams highlights limitations due to dataset size and generalizability across medical specialties. Future research should use more diverse datasets and fine-tune models for broader applicability, including capabilities to process images [86]. The CLEAR tool for evaluating AI-generated health information requires further validation and comparison with other tools. Additional studies should test its applicability across a wider range of health topics, ensuring it evolves with ongoing AI advancements. This will help identify potential gaps, inaccuracies, and biases, enhancing the tool's reliability and effectiveness [88].

Developing robust validation frameworks and verification processes is essential to ensure the accuracy of LLM outputs. These frameworks can be achieved by comparing LLM-generated information with reliable medical sources, conducting independent evaluations, and incorporating expert advice [56]. However, also ensuring human oversight always in LLM evaluations is crucial but challenging due to current system designs as it is not feasible. [114]

Given the value of human assessment, an evaluation carried out by healthcare experts is still generally seen as the most accurate and reliable standard.[115] Human evaluation is unpredictable and inconsistent due to various methodologies and parameters.[116] The unpredictability might create bias in the assessment of algorithms and complicate the comparison of various studies despite their categorization as human evaluation framework.[117] Additionally, involving end users or stakeholders at the start determines the appropriate individual or entity responsible for evaluating the LLM-generated answer throughout the human review process. This subject holds significant importance, but there is a major gap in the study by [85, 117].

As evaluated by MedPalm,[4] it may be desirable to evaluate various measurements from several viewpoints, including those of multiple stakeholders, physicians, and laypeople. Standardized human-based evaluations with underlying factors can address concerns about the trade-off between accuracy and speed and help balance efficiency with thorough verification [72,73,74]

3

Research Methods

This chapter describes the methodology of this study, addressing the research gaps identified in the Literature Review with specific Research Questions. It provides an overview and rationale for the four-step hourglass method approach employed to analyse data collected from the facilitation study at NCJ Utrecht followed by the second method of User Interview which involved recruiting healthcare experts from different domains of healthcare. The approach includes grounded theory and its synthesis, elaborated in the following sections. By utilizing semi-structured interviews and qualitative methods, this study aims to develop human-centric evaluation metrics for Conversational AI-generated responses, providing detailed insights to enhance diagnostic suggestions.

Contents of the Chapter

Chapter 3: Research Methods

To address the research aims and objectives of the study, it was divided into three research questions after conducting a thorough examination of the existing literature. These research questions, in turn, provide answers to the overarching major topic of the study which is mentioned in the introduction. The study paper addresses the following research questions as follows:

3.1 Research Question(s)

Research Question 1
What limitations exist in current user-centric evaluation methods for healthcare-focused Large Language Model (LLMs), particularly in lifestyle advice?

Research Question 2
What are the key challenges of user-centric evaluations metrics in assessing the accuracy and reliability of LLM-generated health lifestyle advice?

Research Question 3
How can evaluation frameworks be improved to enhance the effectiveness of LLM's response in providing lifestyle advice?

3.2 Study Design

Qualitative research study is performed to gain in-depth understanding of the perception of healthcare experts towards Conversational AI use in healthcare lifestyle advice. More specifically the research design is aimed at exploring the perception of the Healthcare Experts in the evaluation of the responses with regards to way and content of the response generated by Conversational AI namely Chatbots, with underlying needs and values.

To address these research questions two partly independent studies were conducted along with literature review to investigate the research-questions. An extensive literature review was carried out to address the first and second research questions. But for the second research question the literature review was also triangulated during Creative Facilitation. To draw conclusions and verify the results of our creative facilitation study, a further major study was conducted which was an individual one to one semi-structured interviews with *(n=11)* participants using the interview guide. This method allowed for the flexibility to probe further on useful insights from Creative Facilitation, while maintaining our intended focus points. These interviews provided an opportunity to delve deeper into subjects such as opinions, values, feelings, and emotions. The aim was to develop an evaluation metric, guideline, or protocol from the perspective of healthcare experts.

For data collection, we chose to perform multiple triangulations as we conducted a literature review, creative facilitation and structured one-on-one interviews. By using two different methods and studying the existing frameworks, both methodological and data triangulation were intended to be achieved.[79]. If done right, validity is established if the conclusions drawn from the findings of the various methods are consistent. [80].

For the sampling, it was done in two phases for the two different types of methods used in the study namely 1) Creative Facilitation and 2) Qualitative Interviews

3.3 Creative Facilitation

Creative Facilitation included a series of sessions which involved participation of 8 healthcare professionals which was conducted by TU Delft and Erasmus MC at NCJ Utrecht. This section also aims to answer the **Second Research Question:** ***What are the key challenges of user-centric evaluations metrics in assessing the accuracy and reliability of LLM-generated health lifestyle advice?***

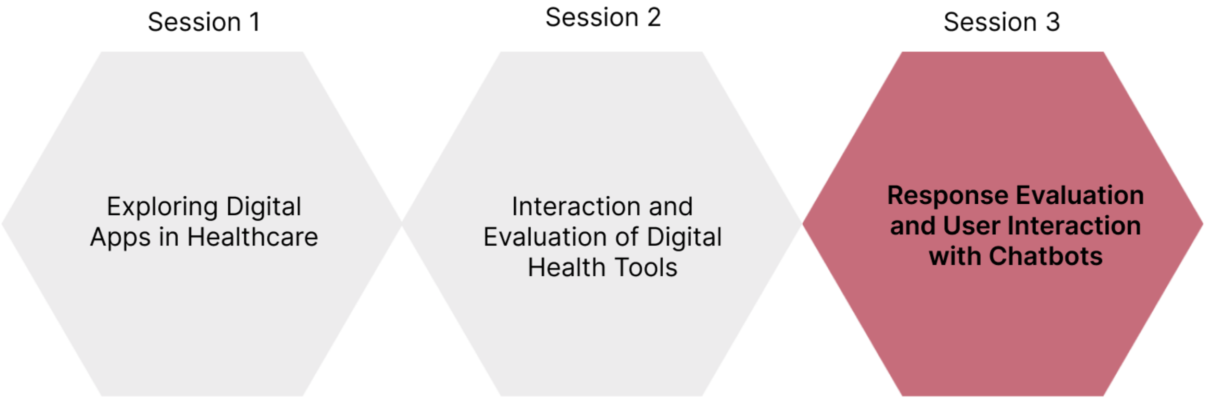

Session 1
Exploring Digital Apps in Healthcare
Session 2
Interaction and Evaluation of Digital Health Tools
Session 3
Response Evaluation and User Interaction with Chatbots

Figure 2 : Creative Facilitation process

3.3.1 Data Collection

Eight healthcare specialists were recruited from NCJ Utrecht, Erasmus MC, and TU Delft for the purpose of creative facilitation. All participants were women who held Dutch citizenship and were from the healthcare domain. The group consisted of Youth health physicians (n=2), Youth health nurses (n=2), Research Academics (n=1) from TU Delft, Research Academics (n=3) from Erasmus MC. The careful selection of healthcare experts and academics ensured an extensive representation, including a range of knowledge and expertise to enhance the facilitation process.

3.3.2 Study Process

As the research question started to formulate the research direction moved towards the arrangement of a creative facilitation study which eventually answered the sub-research question 2. It was conducted by TU Delft and Erasmus MC amongst 8 participants with an aim to gather qualitative data from healthcare experts over 1) Experience and views on digital apps in healthcare 2) Discussion on digital health tools for helping families in assisting on topics of healthcare and 3) Interacting with a digital prototype of the Conversational AI and Evaluating responses by Conversational AI compared to the knowledge and values of the Healthcare expert.

Referring to the integrated creative problem-solving (Heijne & Meer, 2019) [119] the workshops were planned according to the three main creative diamonds (sessions) to ensure a creative workflow. The students from Erasmus MC acted as facilitators to facilitate each diamond, facilitators were suggested to follow the steps of task Diverging, Reverging and Converging, and Reflecting. However, given the limited duration of the workshops (2 hours), the activities primarily focused on Diverging, Reverging, and Converging which gave rise to sessions revolving around Discussion, Interaction and Evaluation.

The workshop initiated with participants reviewing the objectives of the study and introducing the mock-up of digital tool, where they discussed their roles in assisting families facing vulnerabilities and their perspectives on digital health tools. Following this, the second session delved into a case studies from socio-economic family backgrounds, addressing the stress and isolation affecting families, and evaluated how the "Buddy" or a “Conversational AI” system could mitigate these issues and enhance lifestyle management.

The third session was critical in informing the design of our semi-structured one-on-one interviews. During this session, healthcare professionals interacted with two versions of a digital healthcare application, one incorporating Conversational AI and the other without it. This setup aimed to evaluate the influence of AI on the user experience and establish trust levels in such technologies. The insights gained from this session significantly influenced the subsequent development of interview methodologies, highlighting key perceptions regarding the utility and reliability of AI in healthcare.

The final phase of the session involved a systematic evaluation, where healthcare experts were presented with sheets detailing Conversational AI responses on various health topics including nutrition, illness, stress, and sleep. These experts assessed the AI's responses based on criteria such as Accuracy, Relevance, Appropriateness, and Trustworthiness. Their evaluations were informed by their professional experiences, knowledge bases, interaction styles, and established best practices in lifestyle advice consulting.

3.3.3 Data Analysis

This explorative study aimed to gain insights and validate assumptions made from our literature review about the perceptions of Conversational AI, LLMs in the field of healthcare. Besides that, it allowed us to sensitize and gather contact information for possible participants for our Semi-Structured one-to-one interviews. Session 3 of Creative Facilitation was an important session where participants analyzed the responses of Conversational AI. These responses and the discussions were transcribed, converted into text and Inline coding was performed which generated a total of (n=105) codes which followed Focused coding to come up with topics which are as follows:

- Trustworthiness in AI
- Reliability of Responses in AI
- Responsibility of Responses by Chatbots
- Opinions and Values regarding Conversational AI
- Trust and Transparency in AI
- Accuracy of Responses

However, an important finding that helped formulate the semi-structured interview questions based on the categories mentioned above was understanding the transcript and the reasoning behind healthcare experts assigning specific ratings to metrics like Accuracy, Relevance, Appropriateness, and Trustworthiness when evaluating the responses generated by Conversational AI. The transcript and evaluation metrics revealed that the assessment of Accuracy and Trustworthiness and other metrics was highly subjective. Each healthcare expert had a distinct perspective, thus making the evaluation process multifaceted. [85] This complexity arises because each healthcare expert possesses diverse experiences within their respective domains, and no standard guideline exists to evaluate the metrics.

Therefore, the study needed to probe, investigate, and expand further into the underlying factors healthcare experts use to evaluate a response generated by Conversational AI [81, 74]

3.4 Semi-Structured interviews

3.4.1 Data collection

For the semi-structured interviews, the use of Patton's 40 Purposeful Sampling Strategies [75] (Patton, 1990) were considered. The semi-structured interview was conducted with the specific goal of gaining an initial understanding of the perception of healthcare experts with respect to the evaluation of responses of Conversational AI. Rather than testing a hypothesis about a large population, the focus was on extracting insights from a small group of participants which was healthcare experts but from different domains. To ensure a deeper exploration of the topic, in the total sample size, it consisted of few participants who participated in the Creative Facilitation and were already familiar with the subject matter. Rest of the participants were recruited via an online-survey and sending personal emails.

For the Semi-Structured one-on-one interview, an online survey was shared via e-mail to our supervisory staff, peers, several healthcare websites, and included in printed mails and physical mailboxes. However, this resulted in just a small portion of participants, as those with expertise in the healthcare field are Difficult-to-reach populations, also known as hard-to-reach populations [76], we had to utilize the Snowballing approach.

Snowball sampling, as defined by Patton (1990) [75], involves a process where an initial interviewee recommending at least one other respondent, who in turn proposes others, resulting in a fast growth in the sample size. This approach is extremely valuable for recruiting participants in healthcare studies, particularly those from hard-to-reach regions and dealing with local language obstacles [78]. One advantage of snowball sampling is that it involves persons from many cultures and professions, thus building trust and enhancing the probability of involvement [77].

However, from this group of participants, we created a list of clear criteria and rationale for inclusion of healthcare experts with different expertise which is reflected in *Table 2*.

The participant must be residing and by practice a healthcare expert with healthcare licensing in the Netherlands.
The participant sample must include Healthcare experts from the following sectors of healthcare domain.

- Nutritionists/Dietitians/Lifestyle Coaches.
- Psychologists
- General Practitioner (GP)
- Youth Health Nurses
- Social Workers

For the User Interviews which was based on Semi-Structured one-to-one interviews the participant group consisted of eleven individuals (n=11), predominantly based in the Netherlands, with diverse professions in the healthcare sector. These professionals included Pedagogue (n=1), Pediatrician (n=1), Youth health nurse (n=1), Fetal maternal specialist/obstetrician (n=1), Nutritionist/lifestyle advisor (n=1), Psychologists (n=3), and General Practitioners (n=3). The distribution of genders: Females (n=6) and Males (n=5) across these professionals is balanced, enhancing the diversity of perspectives in the study. Thus, participants were assigned anonymized identifiers such as P1, P2, etc., to maintain confidentiality and ensure impartial handling of data. This approach emphasizes the consideration of their professional qualifications and demographics, thereby enhancing the relevance and integrity of the research outcomes.

3.4.2 Study process

Based on the data analysis on Creative facilitation (Section 3.3) various topics were generated using which Semi-Structured one-on-one interviews were conducted by probing further with questions to gather qualitative data from healthcare experts regarding their perception of Conversational AI in healthcare. This interview aimed to allow participants to express their behaviors, feelings, values, knowledge, perception and evaluation towards responses of Conversational AI and the use of such digital applications in the information sensitive domain of healthcare The main topics in the interview were:

- Opinions and Values Questions.
- Feeling and Emotional Responses.
- Knowledge and Factual Information.
- **AI Trust and Reliability:**
- Questions probing towards value trade-offs while interacting with the live chatbot:
- **AI Transparency and Accountability:**
- Questions probing towards perception of responses by chatbot what values or knowledge it lacks or has that's good
- **Past Experiences with AI Chatbots:** Questions probing towards value trade-offs while interacting with the live chatbot
- **Short Creative Activity Human Centric Evaluation:** Questions and activities that probe further into getting insights from participants regarding what are they looking for in a chatbot response, also there are few activities that provide the insights much further ahead?

3.3.3 Data Analysis

The data from each of the 11 interviews was carefully collected and transcribed or line coding which resulted in codes (n=606). Afterwards they were uploaded to a qualitative data analysis software known as ATLAS.ti 24. After the transcription, the data was cleaned and coded, in which the approaches to reach grounded theory was achieved which consisted of the following steps.

The data analysis followed a grounded theory approach by Glasser and Strauss, [113] where "explain a bit about grounded theory". The coding process consisted of four different phases which are basically 1) Inline Coding 2) Focused Coding 3) Axial Coding and 4) Theory Building

- **Inline Coding**
  The Data Analysis of User Interviews initiated with inline coding as the first step in analysing data from semi-structured one-on-one interviews. These interviews were conducted to gain a deeper understanding of healthcare expert's interactions with Conversational AI and enabled an in-depth look of how participants considered the AI's reliability, transparency, and accuracy in healthcare assistance. In this coding procedure, each portion of the transcribed interviews was carefully labelled, whether it was a phrase, or a sentence. The purpose was to accurately capture the participants' observations about their experiences with the AI thus capturing their "Opinions and Values" or their "Feeling and Emotional Responses"

- **Focused Coding**
  Following inline coding, focused coding was employed to organize the extensive data into key categories that appeared most frequently and prominently. This method refined the analysis by concentrating on significant insights and organizing them into meaningful categories such as ""Trust and Reliability," "Transparency and Accountability," "Empathy", "Accuracy" and a total of (n=67) sub-categories were formed. This stage was crucial for structuring the data around core issues highlighted by healthcare professional's experiences and perspectives.

- **Axial Coding**
  Axial coding followed, as the third step of Grounded Theory technique by linking sub-categories and their related underlying factors or sub-subcategories. identified during focused coding to form a cohesive framework. This stage examined how different themes—like "Inquiring," "Empathy," "Trustworthiness," "Completeness," "Accuracy," "Relevance," and "Fluency"—interacted within the context of AI use in healthcare. Axial coding thus enhanced understanding of the

relationships between categories and their impact on the acceptance and effectiveness of AI technologies.

- **Theory building**
  The final phase, theory building, integrated all categories from the previous coding stages to construct a comprehensive theoretical model. Theoretical coding in this research entailed synthesizing the connections among fundamental categories such as “Inquiring,” “Empathy,” “Trustworthiness,” “Completeness,” “Accuracy,” “Relevance,” and “Fluency” into a narrative that explains how these elements interact within AI applications in healthcare. Theoretical coding established a logical structure by categorizing the data into two distinct categories: 1) Response Style and 2) Content of Response.

4

Results

This chapter details the second research method which was crucial and focused on User Interviews, by focusing on Grounded Theory (GT) to gain in-depth insights from user interaction with conversational AI. The goal was to understand the evaluation criteria of the responses generated by Conversational AI by interviewing a smaller representative sample of healthcare experts from different domains like General Practitioner, Psychologists, Lifestyle Coaches from the Netherlands. Using an inductive and deductive approach, in analysing the transcripts, categories emerged organically from the data and were visualized in a set of evaluation metrics. This step captured the nuances within individual categories, laying the groundwork for further analysis to provide a comprehensive understanding of the evaluation experiences of Conversational AI by different stakeholder like Patients, Healthcare experts, AI developers or policymakers.

Contents of the Chapter

Chapter 4: Results

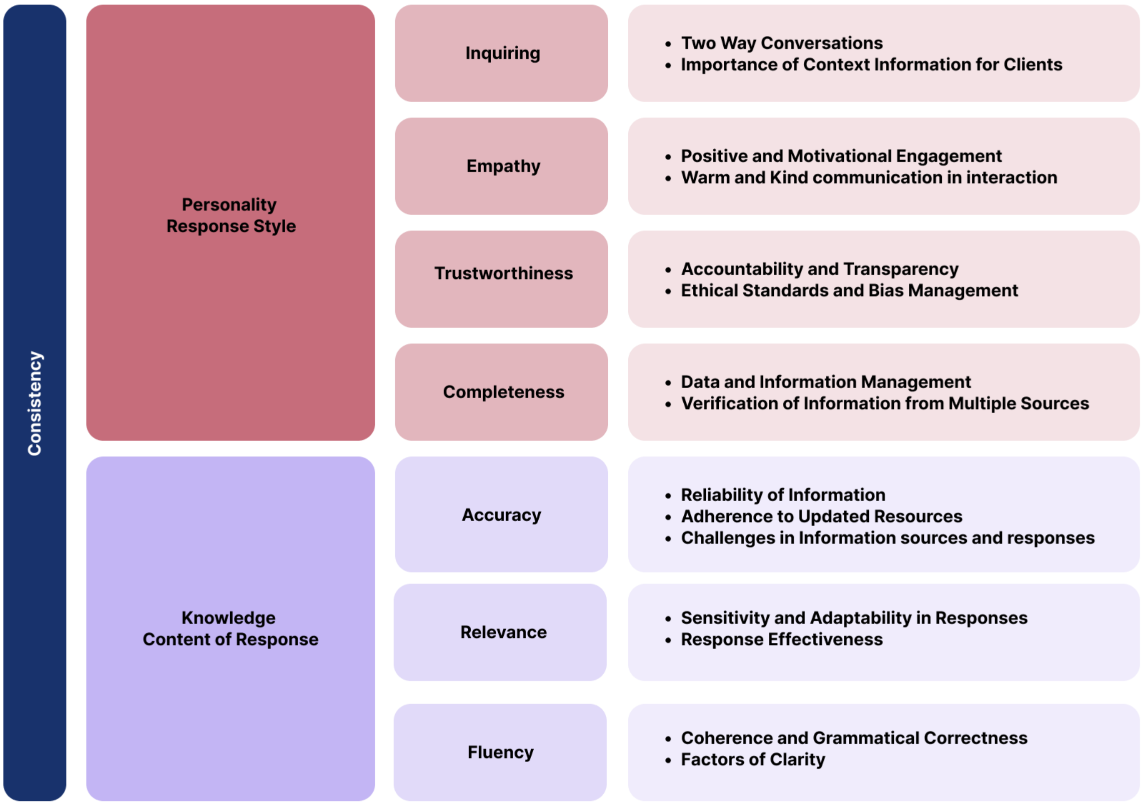

Consistency
Personality
Response Style
Knowledge
Content of Response
Inquiring
Empathy
Trustworthiness
Completeness
Accuracy
Relevance
Fluency
• Two Way Conversations
• Importance of Context Information for Clients
• Positive and Motivational Engagement
• Warm and Kind communication in interaction
• Accountability and Transparency
• Ethical Standards and Bias Management
• Data and Information Management
• Verification of Information from Multiple Sources
• Reliability of Information
• Adherence to Updated Resources
• Challenges in Information sources and responses
• Sensitivity and Adaptability in Responses
• Response Effectiveness
• Coherence and Grammatical Correctness
• Factors of Clarity

Figure 3 : Dual-Core Evaluation Framework for Conversational AI in Healthcare

4.1 Evaluation Metrics in relation to Sub-Categories and Sub-Subcategories

Coming up with a Dual-Core Framework with the grounded theory approach in two important segments in a way a statement is broken down in a conversation by Conversational AI which is text based: **1) Personality - Response Style** and **2) Knowledge – Content of Response.** by combining the 7 Evaluations metrics. Furthermore, these 7 Evaluation metrics are further clustered together by combining the 15 Sub-Categories. These evaluation metrics answers the **Third Research Question *"How can evaluation frameworks be improved to enhance the effectiveness of LLMs response in providing lifestyle advice?"***

4.2 Personality – Response Style

Personality – Response Style:

Healthcare experts mentioned the effectives of responses which can be achieved by Categorizing data and information from different verified sources which gives a sense of ***"Completeness"*** in the way a response is generated. Similar to importance of verified responses is observed in a response, the way of engaging end users in the conversation by means of positive response which can be through a warm and kind communication is an important aspect in the way of communication as it gives rise to a value of ***"Empathy"*** in the way someone responses back to the questions asked. The value of ***"Trust or Trustworthiness"*** depends on various layers as it can only be built over time on the factors of how reliable the information is, can the Conversational AI take accountability of the response generated and the way Ethics are maintained and managed. Lastly, to have a seamless conversation, the conversation should be both sided so that Conversational AI can gather more insights on the context of the end user to provide more complete responses by an ***"Inquiring"*** nature.

Evaluation Metric 1: Inquiring

The first metric for evaluation is ***"Inquiring"***, where Conversational AI builds two-way communication with the users, for a better contextual based conversation, which generates responses based on the background information of the clients. This approach personalizes interactions and avoids the pitfalls of generic, directive responses that dictate rather than suggest, thereby enhancing the relevance and appropriateness of the guidance provided.

1.1 Two Way Conversations

The interviews with the experts highlight Non-verbal communication which takes place in a natural setting between the client and healthcare expert and cannot be achieved by AI and outline the need of **Explorative ways to gather more information**. As conversations turn out to be broad or specific and the root question usually builds the entire conversation, so an inquisitive way of gathering more questions about the client can help in the longer conversation build up as P[8] mentions *"It's very important to be to, get the trusted information, correct information, because if you give an advice based on the wrong and limited input, then people get can get hurt of course."* To build a long-term conversation a **Two-way interactivity for a contextual and human connection** is needed**.** The communication between the Conversational AI and healthcare expert should be both sides as context can be built by an explorative way of communication as

P [1] states *“In a more explorative way, AI can help me. I asked some general questions and I just throw it out there. I don't even think about what I'm really asking, and then if they give me some leads that I can look into myself and question it back again”.*

1.2 Importance of Context Information for Clients

Inquiring nature of the response is evaluated by the participants on how effective Conversational AI is towards maintaining towards **contextual and reasoning-based responses**, as various factors as responses need to fit wider context of knowledge and not just assumption based responses as Conversational AI can learn a lot better on how to respond on context based conversations by probing further as P[10] mentions *“I think it is difficult to really see what's going on and really see problem and like when we mentioned about the initial user journey as someone tells you a lot of new information the second time”.* Due to which, an inquiring nature of the responses, by not just providing answers in a one-sided conversation, but an inquisitive nature of the Conversational AI helps to generate factual information based on the **background information of clients** from different strata in the society as P[2] mentions *“I think in the future and that you can chat, but we have to learn on how to use it and especially in the vulnerable families and with the lower social capacity and is it able to pick up and suggest what's really needed for them”.*

Evaluation Metric 2: Empathy

The second metric for evaluation is ***“Empathy”***, as healthcare experts believe that Empathy can’t be directly achieved in a conversational setting with a chatbot but can be indirectly expressed by evaluating the statements generated by Conversational AI for the healthcare setting where it includes responses that evoke a sense of Positive and Motivational Engagement, Warm and Kind communication in interaction.

2.1 Positive and Motivational Engagement

Healthcare experts highlighted that the responses can be motivational if the Conversational AI generates **positive and motivating responses to the clients**, as P[5] mentions how habit building is achieved and on the way responses are generated by stating “*It should be a positive and looking at chances and steps towards The wanted behavior*”. But also equal importance is provided during evaluation to **responses that are suggestive and not directive** as they play a major impacting role in the decision making for the end user as in the cases of mental wellbeing of a person where P[9] states *“For example, I thought OK, what if someone tells I have suicidal thoughts? For example, What is the response of the computer or AI and that was a very A straightforward response, that would really hurt the person and the state the person in”.* Participants also mentioned the need of **sympathetic, warmth and kind in response generated** is important layer in empathy towards making client understand the response generated by AI from an expert and human connection perspective as it builds the factor of reliability as stated by P [2] "*I don't believe in empathy. Yeah, which is extremely difficult when you use it digital chatbot, there is no empathy in chatbot. But what you mean is that the answer should be reliable".*

2.2 Warm and Kind communication in interaction

Participants mention responses are generated the way a question is queried to the Conversational AI with different **prompting techniques and interaction with chatbot** and response can be as effective and quality as the expert or the client prompts the Chatbot as P[3] states *"That's why the question for how do you get the question is how do you get a good prompt so he can get a good advice?".* Participants also mentioned the doubt in their own knowledge base which eventually effected their perspective of the conversation with the Conversational AI as it was overshadowed by self-doubt as P[11] mentions *"The first answer was great, the second was more superficial and the and it's also how you ask you how you frame or ask your question".* Due to which the **Two- way conversation between Conversational AI and Clients or Healthcare experts** is crucial to tackle the major drawbacks which is Non-verbal communication can't be achieved in Conversational AI but can be minimized by context building in conversation by two-way communication, Suggestive nature of chatbot resulting in multi-turn conversations, breaks the redundancy of the responses as P[9] states *"I'm not sure if that is a value, but to see if you have. If you get a clear picture of what the other person is trying to say, can you say truly? It's like, yeah, I should know about the context of the client that is coming and having a discussion".* Building an empathetic nature should also depend not only on the two sided conversation to gather more information about the client but also mindful of the responses which should be **Non-judgmental and kind communication** towards providing responses to the clients after gathering all the information as judgement builds insecurity which eventually affects trust, and responses should be more kind especially in the cases for mental health where P[10] states *"Yeah, I don't know if it's really true in chatbots, but I think not judging, it's will be big failure towards communication, I think so as we observe it in real life"*

Evaluation Metric 3: Trustworthiness

The third metric for evaluation is ***"Trustworthiness"***, where user builds a belief in the system with the ongoing conversations and the quality responses generated by Conversational AI as there are underlying layers and factors on the basis which a person can trust the system, or the responses by Conversational AI in general. These are based on factors like 1) Accountability and Transparency and 2) Ethical Standards and Bias Management.

3.1 Accountability and Transparency

Participants also mention trustworthiness also depends about being honest about from a Chatbot perspective about what it knows and doesn’t know, as it builds trust as users care about the way **responsibility or accountability is taken of the information provided** where P[5] states *“when there isn't just any information about a certain topic, so it can also maybe honest about it knows and what it doesn’t know”*. Also, being **transparent about the sources of information** can be a major factor healthcare experts trust about the response generated from the Conversational AI as experts feels that lack of transparency gives rise to trust issues, transparency should be maintained from data sources and algorithms the Conversational AI is built upon. But responsibility and transparency can be strengthened by **honesty, clarity and ethics in the communication** as healthcare experts always appreciate honest and clear responses generated by Conversational AI which connect to their own moral values and factors like direct and transparent answers reflect honesty. An interesting find was how participants mention Honesty is more than accuracy of responses as P[4] mentions *“That's like it's more than accuracy. It's more like a near moral value. There's a moral standard for me”*.

3.2 Ethical Standards and Bias Management

Experts highlight the importance of unbiased responses and the risks involved with it by setting up guidelines and ethical standards as responses generated should take care of the factor that there are no racial or cultural bias as healthcare suggestions can be sensitive where P[10] states “*Because it's really important that you know, and maybe it's a cultural sensitive to somebody, yes. So, if you give the choice unbiased response it should be taken into consideration”*. Also, participants feel humans have the tendency to be biased as opposed to AI where P[9] feels *“I would hope there would not be biased because I guess I would see that as more a human thing”*. As responses are generated the database on which the Conversational AI is trained on. So, cleaning of the data can aid in providing unbiased data can thus reducing the **risks involved with biased AI responses** as participants feel validating the response with fellow colleagues in this case can be as effective as confirmation bias can evaluate the quality and trustworthiness of the response as P [1] states *“Definitely will do the combinations of in first place. If in this in this example that I have would have the doctors and then the AI responses. Compared with these doctors and that there were doctors reflect on what they I said in regard to their content. So I will do the double check”*.

Evaluation Metric 4: Completeness

"Completeness" is the fourth and last metric that is used as an evaluation metric which comes under **Way of Response** and includes Data and Information Management in the responses generated and Verification of Information from various sources provided to the Conversational AI.

4.1 Data and Information Management

Participants emphasized the importance of **effective responses from Information Categorizing and filtering**, P [1] emphasized the use of filtering the information, cleaning and categorizing all the information at a single source by stating "*Categorize information and help people to seek information that they think is necessary for themselves to improve their wellbeing and health*". Additionally, it would be beneficial for clients in the future to have access to healthcare guideline documents which is also a reliability factor for healthcare experts as P [6] mentions *"Yeah, but I think for some people could be beneficial to have easy access to professional information"*. By building on complete responses by asking more questions another factor that could increase the accuracy as per healthcare experts if the response adheres to **specificity and correctness in the generated response** where to achieve the completeness of the response the healthcare expert would advise to keep questioning the conversational AI, and if the responses are not specific would go back to the updated consensus document as P[5] mentions *"Also, if it's half correct, but then I ask a second question to yeah to get the answer I was looking for so I would ask for that, and if I'm not convinced, then if this chatbot would not give me the answer I will refer the consensus documents"*

4.2 Verification of Information from Multiple Sources

To verify the information from multiple sources participants mentioned that the responses generated should be sourced from **information backed with scientific resources** due to over information in the internet especially in the domain of nutrition or lifestyle advice it difficult to sort the good and bad, as clients are motivated to bring a change in their lifestyle by following those information as P[5] states *"I guess there's a lot of information available online about nutrition and many people they really want to do well, but they just don't have the real back scientific background".* By providing information referred to scientific resources or healthcare guidelines it builds **reliability on updated resources** from a healthcare expert perspective as P [6] stated the curiosity of a healthcare expert to know where the source of information is from *"Where you got this information from? Because you do want it to reference from Its scientific papers and not some channel"*. From a nutrition perspective these suggestions can impact the livelihood of a client so P [5] feels the need of probing it for the source of the guidelines and are those **responses generated from verified sources and guidelines by experts** by mentioning *“I would keep on asking about it and try to find out if it knows what consensus documents are and which consensus documents are important for basing the information on”.*

4.3 Knowledge – Content of Response

Knowledge – Content of Response

“Accuracy” in a conversation depends on how updated the Conversational AI in sourcing is the updated resources, also while keeping in check the challenges faced in providing the accurate responses from verified sources. Healthcare experts also discussed about the factor of ***“Relevance”*** where there is an understandability and sensitivity in adapting the responses as per the context to reduce uncertainty and increase the effectiveness in providing information in the domain of healthcare. Lastly, participants mention how transparency in responses and simple, coherent and clear responses can provide the users the feel of ***“Fluency”*** in responses.

Evaluation Metric 5: Accuracy

The fifth metric for evaluation is ***“Accuracy”***, one of the most important way of evaluation under the Way of Content where the content of the Conversational AI generated depends on how much its abides to the updated documents in the database and the challenges faced in sources the information towards an accurate response.

5.1 Reliability of Information

Reliability of Information builds upon trustworthiness as there are various **Challenges faced with reliable information**, experts mention that the responses can be challenging to rely on if the responses are robotic and are template based, repeated generation of wrong answers, directness of responses where P [9] how it can trigger trust issues in the long run by stating “*But if I feel no, I've made it clear question, but this is not correct. Then that would definitely have an effect on, but can I trust the next answer as well then?*”. As trust-building is like a domino effect, where the initial reliable information can be the basic foundation on which the entire direction of trust can go ahead where **trust building can be done through reliability of information** where good information builds on trust, referencing of resources builds trust, a very interesting observation where P [8] mentions users don’t accept advice once the trust is broken by stating *“Information from a trustworthy organization where everybody says, yes, that's reliable. But there's something we trust. If not reliable Then people won't accept or trust the information and* *advice following from the information in future”.* As in the longer run these reliable information builds **trust and eventually impacts on reliability on the decision making** in real life as P[8] states if clients get unreliable information, the healthcare expert has to step in between to persuade and explain the correct information *“What makes the discussion more difficult because you had first have to persuade them that what they have read is not exactly always true. So then I think it's important that you can have information available for both sides. That is trustworthy and the then I think I can be a big, big help”.*

5.2 Adherence to Updated Resources

Participants mention how accuracy is evaluated in the responses on the basis of factors like resources from **updated guideline document adherence, consistency and completeness of responses** where information should cover complete and sufficient information as complete responses gives rise to reliability as P[10] states *“A complete information. Because I just want to see information and I want it to be reliable and it's in line with what we would advise then I would say that's trustworthy and that's accurate”.* Participants also outline the **importance of self-research on local information sources** is important in understanding and verifying the response generated from an expert and a guideline perspective for more accuracy as stated by P [1] “*How to balance all that from different perspectives, from a work as well as a father as well ? Let me do my own search and see if I come up with the same sources or if certain sources seem to be systematically correct”.*

5.3 Challenges in Information sources and responses

To maintain accuracy in responses healthcare experts, feel the biggest **challenge in sourcing correct information** and training the Conversational AI such as the response is backed with accurate data which can be referred to as P [6] states *“At least like the information needs to be unquestionably accurate to the source is at least should be accurate, I guess because that's where it gets it wrong”.* As factors which lead to challenges in sourcing information are chatbots connected to the non-governmental or standard website and thus they have limited or non-updated knowledge base, lack of specific resources which could be for a specific geographical region and not global resources, as lack of specific resources can impact the decision making and can have consequences as P[5] feels *“so if it's not really accurate then you can ask more questions to make it more accurate I believe as missing information could change the conclusion in providing healthcare advice”.*

Evaluation Metric 6: Relevance

The sixth metric for evaluation is ***"Relevance"***, where experts mention during evaluation of a response, they look at the way the response is generated after adapting to the clients background information or experts information and is the response relevant the effective questions asked.

6.1 Sensitivity and Adaptability in Responses

Experts shared the experiences how sensitivity in response can be handled with factors like **context-based understandability** as Conversational AI should interpret questions on more contextual terms, as relevance in the responses can only be achieved if reliability and understandability is maintained where understandability depends on context based conversation and also socio economic issues at hand as P[7] states *"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 and conversation or but this is the answer to all these questions that important and the understandability is of course, but that is I think its reliable, understandable".* As these responses which are **reliable and understandable reduces uncertainty** as direct responses should be provided on the context, as uncertain information is removed when context comes in, and sensitive in responses on the situation the client is trying to follow a habit in lifestyle coaching P[6] states *"It could say, yes, you can follow this routine or do this and that. Then it makes the sentence a bit more umm, less harsh or something as people have to do day plannings or get structure in the life or they have to, but they need some reliable advice".*

6.2 Response Effectiveness

Experts shared the experiences how **effective responses** can achieved if the Conversational AI **responds in a suggestive manner**. As chatbot can be relied on a tool which can be used as an assistant, a screening tool or as tool which can provide a tip or advice to healthcare experts so that the solutions to clients can be provided in a collaborative manner as they are an important factor in habit building especially in suggestions related to nutrition or lifestyle coaching as P[5] mentions *"How can I let my child eat more broccoli? Yeah, yeah, I think it's what I am looking for because it says, hey, let your child get used to the taste of broccoli by offering it regularly. And that's also our advice, So I do compare it with our own advices".* Such quality of content in the communication can also be effective if the dataset is of a good quality, if healthcare experts seeks additional information as sources from various websites are information overload but also through **high quality responses through effective questions** as P[3] feels its easy to get good answers on the basis of good questions by stating *"because it's* *very easy to get the a good answer by asking right questions. And even if its's asking back questions, it's much better."*

Evaluation Metric 7: Fluency

The last or the seventh metric of evaluation is ***"Fluency"***, the underlying layers are the Transparency in responses generated Conversational AI and Factors of Clarity

5.1 Coherence and Grammatical Correctness

Fluency of responses as per experts should not be mistaken for how quickly the responses are generated rather how logically sound and grammatically correct is the response based on the **coherence and grammatical correctness** of suggestions by Conversational AI which impacts trust on few levels where responses lack logical connection in the statements formed, as such statements on a client level can raise concern where P[2] states *"And they need other advices, then only received which is more abstract, not logical and concerning with how I have to look to my child and how can I give the appropriate education?"* . For healthcare experts, its annoying if it repeats similar answers even after modifying the prompt to make it specific or generic and includes spelling mistakes from a Conversational AI is a big red flag where P[6] mentions *"Well, if there's a spelling mistake, I definitely don't trust it. OK, even though I make a lot of spelling mistakes myself."* These factors provide the participants the feeling lack of **information trustworthiness.**

5.2 Factors of Clarity

Clarity in responses generate can help to evaluate the **Data Understanding and Presentation** of the response by the chatbot as clear response provides the users a direction where the response is being headed and the way response is presented to the user, as participants mention that simple wordings should be used to focus on clarity and reliability of response towards clients as they are not used to reading medical jargons as P[5] states *"Clear wordings that's also important Because it's very important that we reach everyone and not only people that are highly educated or from a medical background".* As **simplicity in responses** by Conversational AI provides a more grounded response, as it should be verified and should be simple suggestive responses sourced from guidelines written by subject matter experts as P[6] mentions *"if you provide the terminology which is only known to psychologists, right, and the way it gave a response, but I think clarity is perhaps even more important because people are often confused just about, at least in my field, as they lack area of expertise."*

5

Discussions

This chapter presents the last stage of the study process discussion and conclusions of the study, emphasizing the need to develop robust evaluation metrics for the use of Large Language Models (LLMs) in healthcare, particularly for lifestyle advice. By examining responses from healthcare professionals ranging from general practitioner, paediatricians to mental health experts, the research highlights the limitations of current evaluation methods and proposes a blend of automated and manual metrics to enhance the reliability and ethical application of LLMs in healthcare settings. The study also highlights the overlapping findings in the exiting literature and the novel findings in the process.

Contents of the Chapter

Chapter 5: Discussions and Conclusions

5.1 Discussion

The study demonstrates the need for the development robust evaluation metrics linked to the use of Large Language Models (LLMs) in healthcare, specifically in delivering guidance on lifestyle choices, by providing suggestive responses to healthcare experts from the domain ranging from Paediatrician, General Practitioner and Mental Health experts. The findings in Literature review emphasize the importance of robust evaluation frameworks that go beyond model-based evaluation and consider human-centric evaluation metrics towards responses into relevance [56]. Paradoxically human-centric evaluation metric has trade-offs of efficiency gains and time-consuming for evaluation of every single Conversational AI generated response [74]. This need fits in with recent work that emphasizes the shortcomings of existing assessment approaches in accurately reflects the basic requirements of healthcare environments by bringing in a standardized framework.

Research Question 1

What limitations exist in current user-centric evaluation methods for healthcare-focused Large Language Model (LLMs), particularly in lifestyle advice?

The background study on various literature highlighted the importance of establishing robust evaluation metrics to measure the effectiveness of Large Language Models (LLMs) in the healthcare field. In healthcare as accuracy, privacy, and ethical concerns are all critically important. It is essential to address inherent issues like as bias and hallucinations. This requires an effective evaluation guideline that takes into account both automatic and manual techniques. Automated tools like BLEU and ROUGE provide basic quantitative assessments but often miss the detailed medical understanding needed for effective healthcare applications. This reveals a gap where more specialized metrics are needed to assess the clinical accuracy of the models.

Similarly, specialized models like as BioBERT and Med-PaLM, designed specifically for biological literature, demonstrate encouraging levels of accuracy. Nevertheless, assessments frequently focus solely on automated measures and neglect to include crucial user-centered elements like trust and empathy in the healthcare field. Manual assessments are recommended because to their high reliability in assessing the effectiveness of LLMs in specific healthcare activities, enabling an extensive understanding of model outputs. Although these evaluations are comprehensive, they require a significant amount of effort and are subjective to the evaluator thus giving rise to bias.

The challenges of evaluating LLMs in healthcare are highlighted by the studies conducted by Abbasian et al. (2024) [81] and Riedel et al in 2023 [86] which discuss the challenges of assessing LLMs across various user types, domains, and activities. These studies demonstrate the necessity of customized assessment criteria that consider the specific demands of healthcare environments. This investigation highlights the need of using structured review techniques to select the most trustworthy health information supplied by artificial intelligence.

Current frameworks provide a foundation for assessing LLMs, but there remains a critical need for evaluation approaches that combine automated and manual metrics. This blended strategy ensures a comprehensive understanding of LLM performance, enhancing their reliability and utility in healthcare settings. Future research should focus on addressing these gaps to promote more accurate, ethical, and effective use of LLMs in healthcare.

Research Question 2

What are the key challenges of user-centric evaluations metrics in assessing the accuracy and reliability of LLM-generated health lifestyle advice?

Qualitative Metrics preference:

The analysis reaffirms several established guidelines of Conversational AI within healthcare settings, emphasizing the necessity for robust, nuanced evaluation frameworks to assess these technologies effectively. Common findings underscored the significance of established metrics such as accuracy, trustworthiness, and completeness, which align with traditional evaluation standards found in prior frameworks like the QUEST framework [85] and the four healthcare evaluation metrics developed by Abbasian et al. [81]. Furthermore, the study corroborated the dual necessity of automated and manual evaluations in capturing the multifaceted nature of AI interactions within healthcare. Automated metrics, while efficient for broad assessments, often fail to capture the depth required for nuanced medical applications, thereby highlighting the indispensable role of manual evaluations performed by domain experts to provide contextual insights and detailed assessments of AI's performance that emphasize trustworthiness and reliability.

Trust and Reliability Focus:

Further research findings emphasized the criticality of evaluation metrics like "Empathy," "Trustworthiness," and "Completeness" in assessing LLMs for healthcare applications.

These metrics align well with the literature which highlights the importance of qualitative dimensions such as trust, empathy, and the quality of information in LLM outputs. The QUEST framework proposed in the literature review, emphasizing Quality of Information, Understanding and Reasoning, Expression Style and Persona, Safety and Harm, and Trust and Confidence, is particularly relevant. The study's focus on qualitative metrics can be seen as an application of these broader principles, tailored to the specific context of lifestyle advice, mental health advice or general practitioner in healthcare.

Subjectivity in evaluation metrics:

The overlapping insights from our study, creative facilitation, and literature review notably underscores the value of the practical knowledge derived from facilitation workshops and structured interviews. An important aspect of developing the semi-structured interview questions was the analysis of transcripts that captured healthcare experts' rationales for assigning specific ratings to metrics such as Accuracy, Relevance, Appropriateness, and Trustworthiness in evaluating responses from Conversational AI. The scrutiny of these transcripts and subsequent metrics elucidated the inherent subjectivity in the evaluation process. Each healthcare expert brought a unique perspective influenced by their individual experiences across different medical fields, contributing to the complexity of the evaluation framework. This diversity highlights the absence of a uniform standard for assessing such metrics, which complicates the development of universally applicable evaluation guidelines. [85]

Research Question 3

How can evaluation frameworks be improved to enhance the effectiveness of LLM's response in providing lifestyle advice?

Dual Framework Approach:

The Dual-Core Framework of Personality and Knowledge identified in this study emphasizes the crucial aspects of empathetic interaction and content accuracy in evaluating LLM responses. This way understanding a response is crucial for building trust and ensuring the practical utility of LLMs in healthcare settings. This approach also reveals a gap in current evaluation practices, which often prioritize computational metrics over the qualitative aspects that significantly impact patient care and satisfaction [72]. The insights from this research contribute to refining the development and assessment of LLMs in healthcare in evaluation metrics that incorporate empathy, ethical standards, and user-centeredness as its important in the complex and sensitive nature of healthcare, where the stakes of hallucinations and misinformation are high. [60]

Contextual based Analysis: Provides metric for deeper analysis of conversational AI's ability to handle sensitive and complex healthcare interactions. The CLEAR tools emphasis on the quality and relevance of information [88] overlaps with the findings from our study as well, where "Inquiring" and "Empathy" were key to enhancing the effectiveness of LLM responses. Similarly, in our study identification of "Trustworthiness" was an important evaluation metric which similarly reflects on QUEST's focus on Trust and Confidence. However, our study adds a novel aspect by incorporating the real-world implications of these metrics through feedback from healthcare professionals, providing a practical dimension to the theoretical framework suggested in the literature.

Specific Healthcare Metrics:

The most significant contribution of our research is the practical insights gained through facilitation workshops and interviews with healthcare professionals from different background. This hands-on perspective is crucial for refining evaluation frameworks to ensure they meet the needs of end-users in different clinical setting background, an aspect that is sometimes overlooked in more theory review oriented papers [85]. The practical challenges and recommendations highlighted by participants in the study could adjust the existing frameworks and provide a novel benchmark of new guidelines based on robust Categories backed with sub-categories.[81].

5.2 Design Recommendations

During this research, particularly through the user interview phase, a creative methodology was employed where healthcare experts engaged in selecting and arranging keywords derived from the literature review, creative facilitation, and ongoing user interviews. This process unveiled both value similarities and tensions among healthcare experts and clients. Notably, the participant group predominantly consisted of General Practitioners (GPs) and Psychologists. These findings have significant implications for stakeholder mapping, highlighting the utility of the developed framework in identifying value alignments and discrepancies. This is particularly relevant in the design of digital applications, where understanding these dynamics can enhance the relevance and effectiveness of the solutions tailored to the needs of different healthcare professionals.

5.2.1 General practitioners as Stakeholder

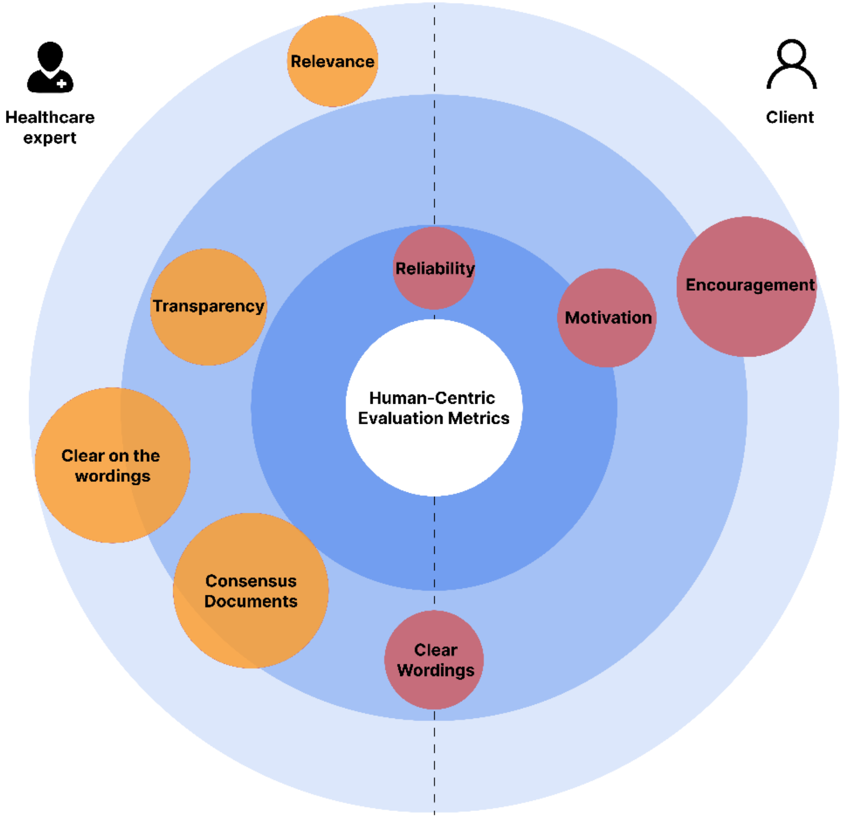


Figure 4 : Overview of Value Similarities and Tensions with General Practitioners

When considering General Practitioners (GPs) as stakeholders in the digital healthcare framework, it is evident that their priorities for content response include transparency, relevance, and clarity in terminology, aligned with consensus documents. GPs perceive that while they prioritize informational integrity, their clients often value responses that are not only reliable and clear but also encouraging and motivating. This distinction highlights a nuanced dynamic in the stakeholder expectations. However, a notable consensus between GPs and their clients is the universal demand for reliability and clarity in responses, highlighting a fundamental value similarity critical in designing user-centred digital healthcare applications when keeping both the stakeholder in mind.

5.2.2 Psychologists as Stakeholder

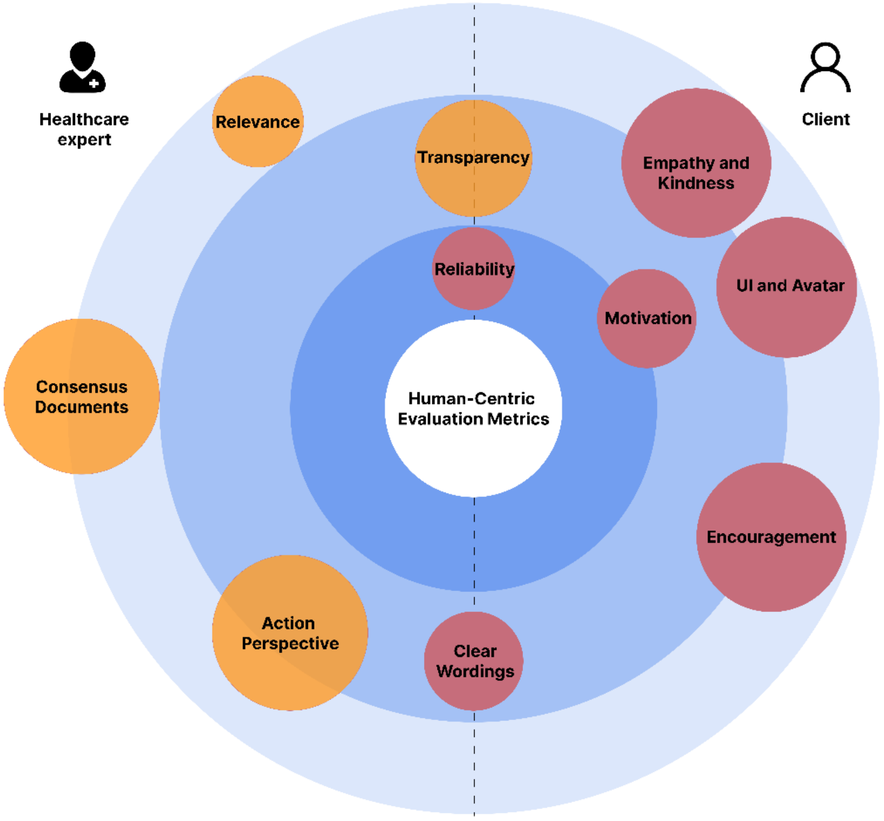


Figure 5 : Overview of Value Similarities and Tensions with Psychologists

When examining Psychologists as stakeholders within the digital healthcare framework, their priorities for content response distinctly focus on transparency, relevance, and adherence to consensus documents, with a strong emphasis on actionable perspectives. Psychologists underscore the importance of empathy and kindness in client interactions, valuing clear, reliable, and motivational responses that are infused with compassion. Interestingly, psychologists express a preference against the use of user interfaces (UIs) and avatars, suggesting that these elements may superficially evoke kindness and empathy without delivering substantive treatment steps. This perspective reveals a complex interplay of expectations between psychologists and their clients, where both groups agree on the essential need for transparency, reliability, and clarity in communication, yet differ in their views on the mechanisms to achieve these outcomes.

5.3 Limitations

Although this study has provided unique insights, it is crucial to recognize specific limitations that may impact the interpretation and applicability of the findings. The number of participants in the study was quite limited and mostly consisted of healthcare experts from the Netherlands. The limited scope of our study hinders the applicability of our results. A more extensive and diversified sample might offer a more comprehensive insight into the efficacy of AI-generated replies in different healthcare settings.

In this study, the assessment measures used were specially constructed according to the viewpoints of healthcare specialists. These measurements represent the factors that healthcare professionals find advantageous for themselves and perhaps for their clients. However, the benefits that are considered favourable for healthcare professionals may not necessarily correspond with the requirements or preferences of their customers. This gap highlights the necessity for a future research method that prioritizes the needs and preferences of clients, to guarantee that the assessment criteria effectively measure the effectiveness of AI tools from the perspective of end-users in the Healthcare ecosystem.

Another constraint applies to the specific Conversational AI tool utilized in this research to generate healthcare responses. The specific functionalities and constraints of this tool may not accurately reflect those of all AI models and algorithms, thus misrepresenting the outcomes. Subsequent studies should explore the incorporation of diverse artificial intelligence (AI) tools to conduct a thorough examination of the capabilities and constraints inherent in various healthcare Conversational AI technologies.

Furthermore, this study did not investigate a diverse array of prompts, including those that include visual stimuli, which might have significant effects on the caliber and relevance of AI-generated replies. The lack of research with different image-based prompts may have resulted in missing the opportunity to fully evaluate the range of Quality criteria that may be used to evaluate AI responses.

5.4 Future Scope

This study provides valuable insights on the utilization of artificial intelligence (AI) in the healthcare sector. Additionally, it paves the way for further extensive and in-depth research. An essential focus for future study is in the broadening of participant diversity. Future research should strive to encompass a more extensive range of participants, encompassing individuals from varied geographical areas and a variety of healthcare systems. This extension would offer a broader international outlook on the use and effectiveness of AI technologies in healthcare, facilitating an understanding of cultural variations that might impact their value and effectiveness.

Another crucial area for future study is the advancement of research procedures that focus on the needs and perspectives of clients. It is important to prioritize the customers' requirements and preferences in healthcare environments. Directly involving patients and other healthcare receivers in gathering their feedback on AI interactions might yield useful information on how to enhance customization and happiness with these technologies.

Furthermore, this study did not examine prompts specifically related to visual pictures, which might have a substantial impact on the quality and importance of AI-generated responses. Incorporating multimodal prompts in future studies would enhance the dataset and offer more profound insights into the processing and response mechanisms of AI tools when dealing with intricate inputs. This research has the potential to facilitate the development of advanced and adaptable AI systems that can effectively manage the intricacies of healthcare settings in the real world.

References

[1] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu et al., “Summary of chatgpt-related research and perspective towards the future of large language models,” Meta-Radiology, p. 100017, 2023.

[2] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu et al., “Prompt engineering for healthcare: Methodologies and applications,” arXiv preprint arXiv:2304.14670, 2023.

[3] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., “A survey of large language models,” arXiv preprint arXiv:2303.18223, 2023.

[4] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172-180. https://doi.org/10.1038/s41586-023-06291-2

[5] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and Applications of Large Language Models. *ArXiv, abs/2307.10169*.

[6] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M.A., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey. *ArXiv, abs/2402.06196*

[7] Shanahan, M. (2024). Talking about Large Language Models. *Communications of the ACM*, *67*(2), 68-79. https://doi.org/10.1145/3624724

[8] A. Borji. 2023. A Categorical Archive of ChatGPT Fail ures. ArXiv:2302.03494 [cs].

[9] P. Henderson, M. S. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky and D. E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. In Thirty-sixth Confer ence on Neural Information Processing Systems Datasets and Benchmarks Track.

[10] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl et al. 2023. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

[11] Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., & Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, *2*(4), 255-263. https://doi.org/10.1002/hcs2.61

[12] Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, *9*, e45312. https://doi.org/10.2196/45312

[13] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering, 23*(5), 649–685. https://doi.org/10.1017/S1351324916000383

[14] Peek, N., Combi, C., Marin, R., & Bellazzi, R. (2015). Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artificial intelligence in medicine*, *65*(1), 61–73. https://doi.org/10.1016/j.artmed.2015.07.003

[15] Yang, X., Chen, A., PourNejatian, N. *et al.* A large language model for electronic health records. *npj Digit. Med.* **5**, 194 (2022). https://doi.org/10.1038/s41746-022-00742-2

[16] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus*, *15*(6), e40895. https://doi.org/10.7759/cureus.40895

[17] Xu, C., Guo, D., Duan, N., & McAuley, J. (2023). Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*

[18] Ben Abacha, A., Demner-Fushman, D. A question-entailment approach to question answering. *BMC Bioinformatics* **20**, 511 (2019). https://doi.org/10.1186/s12859-019-3119-4

[19] Karabacak, M., & Margetis, K. (2023). Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus*, *15*(5), e39305. https://doi.org/10.7759/cureus.39305

[20] Briganti, G. (2023). A clinician's guide to large language models. *Future Medicine AI*. https://doi.org/10.2217/fmai-2023-0003

[21] Biswas, S. (2023). ChatGPT and the Future of Medical Writing. *Radiology*, *307*(2). https://doi.org/10.1148/radiol.223312

[22] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S. B., & Zhang, X. (2023). Chatdoctor: a medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus. https://doi.org/10.7759/cureus.40895

[23] Juhi, A., Pipil, N., Santra, S., Mondal, S., Behera, J. K., & Mondal, H. (2023). The Capability of ChatGPT in Predicting and Explaining Common Drug-Drug Interactions. *Cureus*. https://doi.org/10.7759/cureus.36272

[24] Wang, L., Wan, Z., Ni, C., Song, Q., Li, Y., Clayton, E. W., Malin, B. A., & Yin, Z. (2024). A Systematic Review of ChatGPT and Other Conversational Large Language Models in Healthcare. *medRxiv : the preprint server for health sciences*, 2024.04.26.24306390.

https://doi-org.tudelft.idm.oclc.org/10.1101/2024.04.26.24306390

[25] Montagna, S., Ferretti, S., Klopfenstein, L. C., Florio, A., & Pengo, M. F. (2023). Data decentralisation of llm-based chatbot systems in chronic disease self-management. Proceedings of the 2023 ACM Conference on Information Technology for Social Good. https://doi.org/10.1145/3582515.3609536

[26] Jo, E., Epstein, D. A., Jung, H., & Kim, Y. (2023). Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3544548.3581503

[27] Van Bulck, L., & Moons, P. (2023). What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. *European Journal of Cardiovascular Nursing*, *23*(1), 95-98. https://doi.org/10.1093/eurjcn/zvad038

[28] Liu, X., McDuff, D.J., Kovács, G., Galatzer-Levy, I.R., Sunshine, J., Zhan, J., Poh, M., Liao, S., Achille, P.D., & Patel, S.N. (2023). Large Language Models are Few-Shot Health Learners. *ArXiv, abs/2305.15525*

[29] S, A., A, S. A., R, A., S, D., & Sekar, R. (2023). A novel ai-based chatbot application for personalized medical diagnosis and review using large language models. 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication. https://doi.org/10.1109/rmkmate59243.2023.10368616

[30] Ummara Mumtaz, Awais Ahmed, Summaya Mumtaz. LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties. AIH 2024, 1(2), 16–28. https://doi.org/10.36922/aih.2558

[31] Balloccu, S., Reiter, E., Kumar, V., Recupero, D. R., & Riboni, D. (2024). Ask the experts: sourcing high-quality datasets for nutritional counselling through Human-AI collaboration. *arXiv preprint arXiv:2401.08420*

[32] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922

[33] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1-38. https://doi.org/10.1145/3571730

[34] Liu, J.M., Li, D., Cao, H., Ren, T., Liao, Z., & Wu, J. (2023). ChatCounselor: A Large Language Models for Mental Health Support. *ArXiv, abs/2309.15461*

[35] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering, 23*(5), 649–685. https://doi.org/10.1017/S1351324916000383

[36] University Student Counselees' Attitudes and Experiences Towards Online Counseling During the Covid-19 Pandemic: A Mixed Methods Study. (2023). *Journal of Higher Education Theory and Practice*, *23*(4). https://doi.org/10.33423/jhetp.v23i4.5902

[37] Jungwirth, D., & Haluza, D. (2023). Artificial Intelligence and Public Health: An Exploratory Study. *International journal of environmental research and public health*, *20*(5), 4541. https://doi-org.tudelft.idm.oclc.org/10.3390/ijerph20054541

[38] Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2023). Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. *ArXiv, abs/2307.11991*.

[39] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. *ArXiv, abs/2303.17564*.

[40] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G.V., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., & Sifre, L. (2022). Training Compute-Optimal Large Language Models. *ArXiv, abs/2203.15556*.

[41] Liu, R., & Shah, N.B. (2023). ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. *ArXiv, abs/2306.00622*.

[42] Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *382*(2270), 20230254. https://doi.org/10.1098/rsta.2023.0254

[43] Trebble, T. M., Hansi, N., Hydes, T., Smith, M. A., & Baker, M. (2010). Process mapping the patient journey: an introduction. BMJ (Clinical research ed.), 341, c4078. https://doi.org/10.1136/bmj.c4078

[44] Gualandi, R., Masella, C., Viglione, D., & Tartaglini, D. (2019). Exploring the hospital patient journey: What does the patient experience?. PloS one, 14(12), e0224899. https://doi.org/10.1371/journal.pone.0224899

[45] McCarthy, S., O'Raghallaigh, P., Woodworth, S., Lim, Y. L., Kenny, L. C., & Adam, F. (2016). An integrated patient journey mapping tool for embedding quality in healthcare service reform. Journal of Decision Systems, 25(sup1), 354-368. https://doi.org/10.1080/12460125.2016.1187394

[46] Panagoulias, D. P., Palamidas, F. A., Virvou, M., & Tsihrintzis, G. A. (2023). Rule-augmented artificial intelligence-empowered systems for medical diagnosis using large language models. 2023 IEEE 35th International Conference on Tools With Artificial Intelligence (ICTAI). https://doi.org/10.1109/ictai59109.2023.00018

[47] Balogh, E.P.; Miller, B.T.; Ball, J.R. (2015). Improving Diagnosis in Health Care National Academies Press. https://doi.org/10.17226/21794

[48] Panagoulias, D. P., Virvou, M., & Tsihrintzis, G. A. (2024). Augmenting Large Language Models with Rules for Enhanced Domain-Specific Interactions: The Case of Medical Diagnosis. *Electronics*, *13*(2), 320. https://doi.org/10.3390/electronics13020320

[49] Peng, C., Yang, X., Chen, A. et al. A study of generative large language model for medical research and healthcare. npj Digit. Med. 6, 210 (2023). https://doi.org/10.1038/s41746-023-00958-w

[50] Gerke, S., Minssen, T., Yu, H., & Cohen, I. G. (2019). Ethical and legal issues of ingestible electronic sensors. Nature Electronics, 2(8), 329-334. https://doi.org/10.1038/s41928-019-0290-6

51] Schramowski, P., Turan, C., Andersen, N. et al. Large pre-trained language models contain human-like biases of what is right and wrong to do. Nat Mach Intell 4, 258–268 (2022). https://doi.org/10.1038/s42256-022-00458-8

[52] Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A.J., Krishna, R., Shen, J., & Zhang, C. (2023). Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. ArXiv, abs/2306.15895.

[53] Gillibrand, N., & Draper, C. (2023). Informational Sovereignty: A New Framework For AI Regulation.

[54] D. Oba, M. Kaneko, and D. Bollegala, "In-contextual bias suppression for large language models," arXiv preprint arXiv:2309.07251, 2023.

[55] Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J., & Daneshjou, R. (2024). Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. *Annals of internal medicine*, *177*(2), 210–220. https://doi.org/10.7326/M23-2772

[56] Thapa, S., & Adhikari, S. (2023). ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. *Annals of biomedical engineering*, *51*(12), 2647–2651. https://doi.org/10.1007/s10439-023-03284-0

[57] Li, H., Moon, J. T., Purkayastha, S., Celi, L. A., Trivedi, H., & Gichoya, J. W. (2023). Ethics of large language models in medicine and medical research. *The Lancet. Digital health*, *5*(6), e333–e335. https://doi.org/10.1016/S2589-7500(23)00083-3

[58] Weidinger, L., Mellor, J.F., Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S.M., Hawkins, W.T., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W.S., Legassick, S., Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from Language Models. *ArXiv, abs/2112.04359*.

[59] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science (New York, N.Y.)*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

[60] Schwartz, I. S., Link, K. E., Daneshjou, R., & Cortés-Penfield, N. (2024). Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *78*(4), 860–866. https://doi.org/10.1093/cid/ciad633

[61] Hadi, M.U., tashi, A., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M., Akhtar, N., Wu, J., Mirjalili, S., Al-Tashi, Q., Muneer, A., Al-garadi, M.A., Cnn, G., & RoBERTa, T. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects.

[62] Nazi, Z.A., & Peng, W. (2023). Large language models in healthcare and medical domain: A review. *ArXiv, abs/2401.06775*.

[63] Antaki, F., Touma, S., Milad, D., El-Khoury, J., & Duval, R. (2023). Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmology science*, *3*(4), 100324. https://doi.org/10.1016/j.xops.2023.100324

[64] Haemmerli, J., Sveikata, L., Nouri, A., May, A., Egervari, K., Freyschlag, C., Lobrinus, J. A., Migliorini, D., Momjian, S., Sanda, N., Schaller, K., Tran, S., Yeung, J., & Bijlenga, P. (2023). ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of

a doctor in the tumour board?. *BMJ health & care informatics*, *30*(1), e100775. https://doi.org/10.1136/bmjhci-2023-100775

[65] Guidelines for The Diagnosis, Prevention and Management of Cryptococcal Disease in HIV-Infected Adults, Adolescents and Children: Supplement to the 2016 Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection. Geneva: World Health Organization; 2018 Mar. Available from: https://www.ncbi.nlm.nih.gov/books/NBK531449/.

[66] McCoy, R.T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. Annual Meeting of the Association for Computational Linguistics.

[67] Weston, J., Dinan, E., & Miller, A.H. (2018). Retrieve and Refine: Improved Sequence Generation Models For Dialogue. *ArXiv, abs/1808.04776*.

[68] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. ArXiv, abs/1803.05355.

[69] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models," arXiv preprint arXiv:2304.03738, 2023.

[70] Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A.T., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J., Shafey, L.E., Huang, Y., Meier-Hellstern, K.S., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G.H., Ahn, J., Austin, J., Barham, P., Botha, J.A., Bradbury, J., Brahma, S., Brooks, K.M., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C.A., Chowdhery, A., Crépy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., D'iaz, M.C., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., García, X., Gehrmann, S., González, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A.R., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W.H., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J.Y., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, O., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A., Roy, A., Saeta, B., Samuel, R., Shelby, R.M., Slone, A., Smilkov, D., So, D.R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L.W., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., & Wu, Y. (2023). PaLM 2 Technical Report. *ArXiv, abs/2305.10403*.

[71] ASHER R. (1960). Clinical sense. The use of the five senses. *British medical journal*, *1*(5178), 985–993. https://doi.org/10.1136/bmj.1.5178.985

[72] Liu, Y., He, H., Han, T., Zhang, X., Liu, M., Tian, J., Zhang, Y., Wang, J., Gao, X., Zhong, T., Pan, Y., Xu, S., Wu, Z., Liu, Z., Zhang, X., Zhang, S., Hu, X., Zhang, T., Qiang, N., Liu, T., & Ge, B. (2024). Understanding LLMs: A Comprehensive Overview from Training to Inference. *ArXiv, abs/2401.02038*.

[73] Yun, H., Marshall, I.J., Trikalinos, T.A., & Wallace, B. (2023). Appraising the Potential Uses and Harms of LLMs for Medical Systematic Reviews. *Conference on Empirical Methods in Natural Language Processing*.

[74] Kamalloo, E., Dziri, N., Clarke, C.L., & Rafiei, D. (2023). Evaluating Open-Domain Question Answering in the Era of Large Language Models. *ArXiv, abs/2305.06984*.

[75] Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.

[76] Kirchherr, J., & Charles, K. (2018). Enhancing the sample diversity of snowball samples: Recommendations from a research project on anti-dam movements in Southeast Asia. *PloS one*, *13*(8), e0201710. https://doi.org/10.1371/journal.pone.0201710

[77] Sadler, G. R., Lee, H. C., Lim, R. S., & Fullerton, J. (2010). Recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. *Nursing & health sciences*, *12*(3), 369–374. https://doi.org/10.1111/j.1442-2018.2010.00541.x

[78] Cooke, R., & Jones, A. (2017). Recruiting adult participants to physical activity intervention studies using sport: a systematic review. *BMJ Open Sport &Amp; Exercise Medicine*, *3*(1), e000231. https://doi.org/10.1136/bmjsem-2017-000231

[79] Bans-Akutey, A., & Tiimub, B. M. (2021). Triangulation in Research. *Academia Letters*. https://doi.org/10.20935/al3392

[80] FRANKLIN, C., & BALLAN, M. (2001). Reliability and validity in qualitative research. In The Handbook of Social Work Research Methods (pp. 273-292). SAGE Publications, Inc., https://doi.org/10.4135/9781412986182

[81] Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin, B., Gevaert, O., Li, L.-J., Jain, R., & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01074-z

[82] Novikova, J., Dusek, O., Curry, A.C., & Rieser, V. (2017). Why We Need New Evaluation Metrics for NLG. *Conference on Empirical Methods in Natural Language Processing*.

[83] Wang, S., Zhao, Z., Ouyang, X., Wang, Q., & Shen, D. (2023). ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models. *ArXiv, abs/2302.07257*.

[84] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS digital health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

[85] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., ... & Wang, Y. (2024). A Literature Review and Framework for Human Evaluation of Generative Large Language Models in Healthcare. *arXiv preprint arXiv:2405.02559*.

[86] Riedel, M., Kaefinger, K., Stuehrenberg, A., Ritter, V., Amann, N., Graf, A., Recker, F., Klein, E., Kiechle, M., Riedel, F., & Meyer, B. (2023). ChatGPT's performance in German OB/GYN exams - paving the way for AI-enhanced medical education and clinical practice. *Frontiers in medicine*, *10*, 1296615. https://doi.org/10.3389/fmed.2023.1296615

[87] Wang, R., & Strong, D.M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst., 12*, 5-33.

[88] Sallam, M., Barakat, M., & Sallam, M. (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), e49373. https://doi.org/10.7759/cureus.49373

[89] Chatman, S. (1975). Towards a theory of narrative. New Literary History, 6(2), 295. https://doi.org/10.2307/468421

[90] Ethics and governance of artificial intelligence for health. Geneva: World Health

Organization; 2021 (https://www.who.int/publications/i/item/9789240029200, accessed

on 12 August 2024).

[91] Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. FDA. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (accessed on 12th August 2024).

[92] Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ et al.

Foundation models for generalist medical artificial intelligence, Nature.

2023;616(7956):259–65. doi:10.1038/s41586-023-05881-4.

[93] Organization, W. H. (2024). Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance. World Health Organization.

[94] Sovrano, F., Sapienza, S., Palmirani, M., & Vitali, F. (2022). Metrics, Explainability and the European AI Act Proposal. J, 5(1), 126-138. https://doi.org/10.3390/j5010010

[95] AI Act. (2024, July 30). Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed on 12th August 2024).

[96] Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K., & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, *154*, 102900. https://doi.org/10.1016/j.artmed.2024.102900

[97] OpenAI, R. (2023). GPT-4 technical report. *ArXiv*, *2303*, 08774.

[98] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H.J., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P.A., Prakash, S., Green, B., Dominowska, E., Arcas, B.A., Tomašev, N., Liu, Y., Wong, R.C., Semturs, C., Mahdavi, S.S., Barral, J.K., Webster, D.R., Corrado, G.S., Matias, Y., Azizi, S., Karthikesalingam, A., & Natarajan, V. (2023). Towards Expert-Level Medical Question Answering with Large Language Models. *ArXiv, abs/2305.09617*.

[99] Yang, X., Chen, A., PourNejatian, N. *et al.* A large language model for electronic health records. *npj Digit. Med.* **5**, 194 (2022). https://doi.org/10.1038/s41746-022-00742-2

[100] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *ArXiv, abs/2206.07682*.

[101] Wornow, M., Xu, Y., Thapa, R. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *npj Digit. Med.* **6**, 135 (2023). https://doi.org/10.1038/s41746-023-00879-8

[102] Clusmann, J., Kolbinger, F.R., Muti, H.S. *et al.* The future landscape of large language models in medicine. *Commun Med* **3**, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1

[103] Kripalani, S., LeFevre, F., Phillips, C. O., Williams, M. V., Basaviah, P., & Baker, D. W. (2007). Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *JAMA*, *297*(8), 831–841. https://doi.org/10.1001/jama.297.8.831

[104] Agarwal, R., Sands, D. Z., & Schneider, J. D. (2010). Quantifying the economic impact of communication inefficiencies in U.S. hospitals. *Journal of healthcare management / American College of Healthcare Executives*, *55*(4), 265–282.

[105] Singhal, K., Azizi, S., Tu, T. *et al.* Large language models encode clinical knowledge. *Nature* 620, 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2

[106] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS digital health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

[107] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2206–2222. https://doi.org/10.1145/3531146.3534637

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

[109] Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. In Proceedings of the ACM on Web Conference 2024 (WWW '24). Association for Computing Machinery, New York, NY, USA, 2627–2638. https://doi.org/10.1145/3589334.3645643

[110] Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q.N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., & Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. *ArXiv, abs/1606.06031*.

[111] Silfen E. (2006). Documentation and coding of ED patient encounters: an evaluation of the accuracy of an electronic medical record. *The American journal of emergency medicine*, *24*(6), 664–678. https://doi.org/10.1016/j.ajem.2006.02.005

[112] Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J.A., Wornow, M., Swaminathan, A., Lehmann, L.S., Hong, H.J., Kashyap, M., Chaurasia, A.R., Shah, N.R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M.A., & Shah, N.H. (2024). A Systematic

Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs). *medRxiv*.

[113] Glaser, B., & Strauss, A. (1999). Discovery of Grounded Theory: Strategies for Qualitative Research (1st ed.). Routledge. https://doi.org/10.4324/9780203793206

[114] Ibrahim, L., Huang, S., Ahmad, L., & Anderljung, M. (2024). Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks. *ArXiv, abs/2405.10632*.

[115] Reddy, S. (2023). Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, *41*, 101304. https://doi.org/10.1016/j.imu.2023.101304

[116] Ziems, C., Held, W.B., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can Large Language Models Transform Computational Social Science? *Computational Linguistics, 50*, 237-291.

[117] Awasthi, R., Mishra, S., Mahapatra, D., Khanna, A., Maheshwari, K., Cywinski, J., Papay, F., & Mathur, P. (2023). HumanELY: Human evaluation of LLM yield, using a novel web-based evaluation tool. *medRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.12.22.23300458

[118] Turbek, S. P., Chock, T. M., Donahue, K., Havrilla, C. A., Oliverio, A. M., Polutchko, S. K., Shoemaker, L. G., & Vimercati, L. (2016). Scientific Writing Made Easy: A Step-by-Step Guide to Undergraduate Writing in the Biological Sciences. *The Bulletin of the Ecological Society of America*, *97*(4), 417-426. https://doi.org/10.1002/bes2.1258

[119] Heijne, K., & Meer, H. van der. (2019). Roadmap for creative problem solving techniques : organizing and facilitating group sessions. Boom uitgevers Amsterdam

Appendix A. Creative Facilitation Session

Session 1

Before the workshop started, a short introduction briefing was provided of the aim of the study, the researchers involved, followed by the introduction of the mock-up of digital application concept to the group. The problem as given (PaG) in the Session 1 was the: *"How was the experience of Healthcare experts with families in vulnerable situations?"*

The discussion further continued the further sub-questions like,

"How do you experience supporting the healthy development of families in vulnerable situations?

"Wishes & concerns for digital health tools for families in vulnerable situations such as the 'Buddy'"

In the Converging stage the responses were noted by a simple exercise to identify the perspective of healthcare experts and their views on the "Wish" and "Concern" about the digital mock-up of application which basically highlighted the desirability and vulnerability of the service.

Session 2

In the 2nd Session the workshop the Diverging stage started with a case study from the *"Families in IJsselmonde (a city in Netherlands)"*, where factors like stressors and isolation affects the family and how the Conversational AI a.k.a. "Buddy" as a team platform can support the family in reduction of stress and address the lifestyle choices.

The Healthcare experts were sensitized by the buddy Concept through a journey map of a user who wants to integrate the Conversational AI a.k.a. "Buddy" into their daily lifestyle and discussion were held around three important topics which acted as the Reverging stage

Buddy as a connector

Here questions were asked to probe further from the Session 1 beyond experience and concerns towards "How" and "What"

The discussion revolved around questions like,

"How could a Buddy who connects families assist you in your work?"

"What role could you play in a Buddy that connects families?"

Buddy for question formulation

Discussions also happened around how Conversational AI a.k.a. “Buddy” could bring in value that would assist users in their daily workflow by understanding more of their situation and their values.

The discussion revolved around questions like,

“How could a Buddy who helps parents ask a question about their situation and values assist you in your work?”

“What role could you play in a Buddy that assists parents to ask a question about their situation and values?”

Buddy for offer (suggestions)

Discussions further happened around how Conversational AI a.k.a. “Buddy” could bring in value that would assist users in their daily workflow by understanding more of their situation and their values.

The discussion revolved around questions like,

“How could a Buddy who helps parents ask a question about their situation and values assist you in your work?”

“What role could you play in a Buddy that assists parents to ask a question about their situation and values?”

In the Converging stage of the second session the participants were said to note down their views on the questions like,

“What are the Buddy's points that can add value to you?

When would you promote the 'Buddy' in or around a consultation?”

This exercise helped to identify the perspective of healthcare experts and their views on the “Values” the OSFB has in their healthcare practices.

The discussion further continued the further sub-questions like,

“How do you experience supporting the healthy development of families in vulnerable situations?

“Wishes & concerns for digital health tools for families in vulnerable situations such as the 'Buddy'”

The responses were noted by a simple exercise to identify the [perspective of healthcare experts and their “Wish” and “Concern” about the Buddy in OSFB.

Session 3

The final session was the most crucial one as the results of this session helped us to design the Second Research Method which was Semi-Structured one-on-one interviews. The session initiated with healthcare experts interacting with a digital application which had two version of it to conduct a within-subject design method, which is also known as repeated measures experiment design where the participants test out and evaluates both the types of products or services or interfaces. The two experiments were as follows

User Interaction Design 1: Users Interacted with a digital application in healthcare

Here participants scanned a QR code provided which redirected to the digital application and interacted with the application where there was no chatbot or Conversational AI present. The reason for this step was to understand the perspective differences of participants interacting with a digital application in healthcare with and without the intervention of AI.

The discussion in this session revolved around the question:

How useful would your patients find this intervention?

User Interaction Design 2: Users Interacted with a digital application in healthcare with the integration of Conversational AI or a Chatbot

Participants went through the same process again of interacting with the digital application and now they were asked to interact with the chatbot mockup. The discussion in this session revolved around the questions like:

"Would you trust your patients with this chatbot? "

"How would you improve the trustworthiness of this prototype? "

The entire session was Converged with a last stage where all the participants were provided sheets which consisted of responses of Conversational AI on the topics of

- Nutrition
- Illness
- Stress
- Sleep

The responses of AI were evaluated by participants who were Healthcare experts on the basis of factors like **Accuracy, Relevance, Appropriateness and Trustworthiness** with their experience with clients, their knowledge base, the way they provide the response, and the best practices of lifestyle advice consulting.

Appendix B. User Interview Guide

Questionnaire guide:

Step 1: Introduction of me and my team and the project.

Step 2: Sign the consent forms and ask permission to record the interview with an audio recorder.

Step 3: Sensitizing with the UI UX mock-up of OSFB application.

Session 1 : Figma Mockup

Background/Demographic Information:

Ask about the background information of the person.

Questions/Discussions on Session 1:

- **Opinions and Values Questions:**

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?

- **Feeling and Emotional Responses:**

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?

- **Knowledge and Factual Information:**

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?

Session 2: OSFB live chatbot

Questions/Discussions on Session 2:

- **AI Trust and Reliability: Questions probing towards value trade-offs while interacting with the live chatbot:**

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?

- **AI Transparency and Accountability: Questions probing towards perception of responses by chatbot what values or knowledge it lacks or has that's good:**

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?

- **Past Experiences with AI Chatbots: Questions probing towards value trade-offs while interacting with the live chatbot**

In your past experiences 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?

Short Creative Activity Human Centric Evaluation: Questions and activities that probe further into getting insights from participants regarding what are they looking for in a chatbot response, also there are few activities that provide the insights much further ahead?

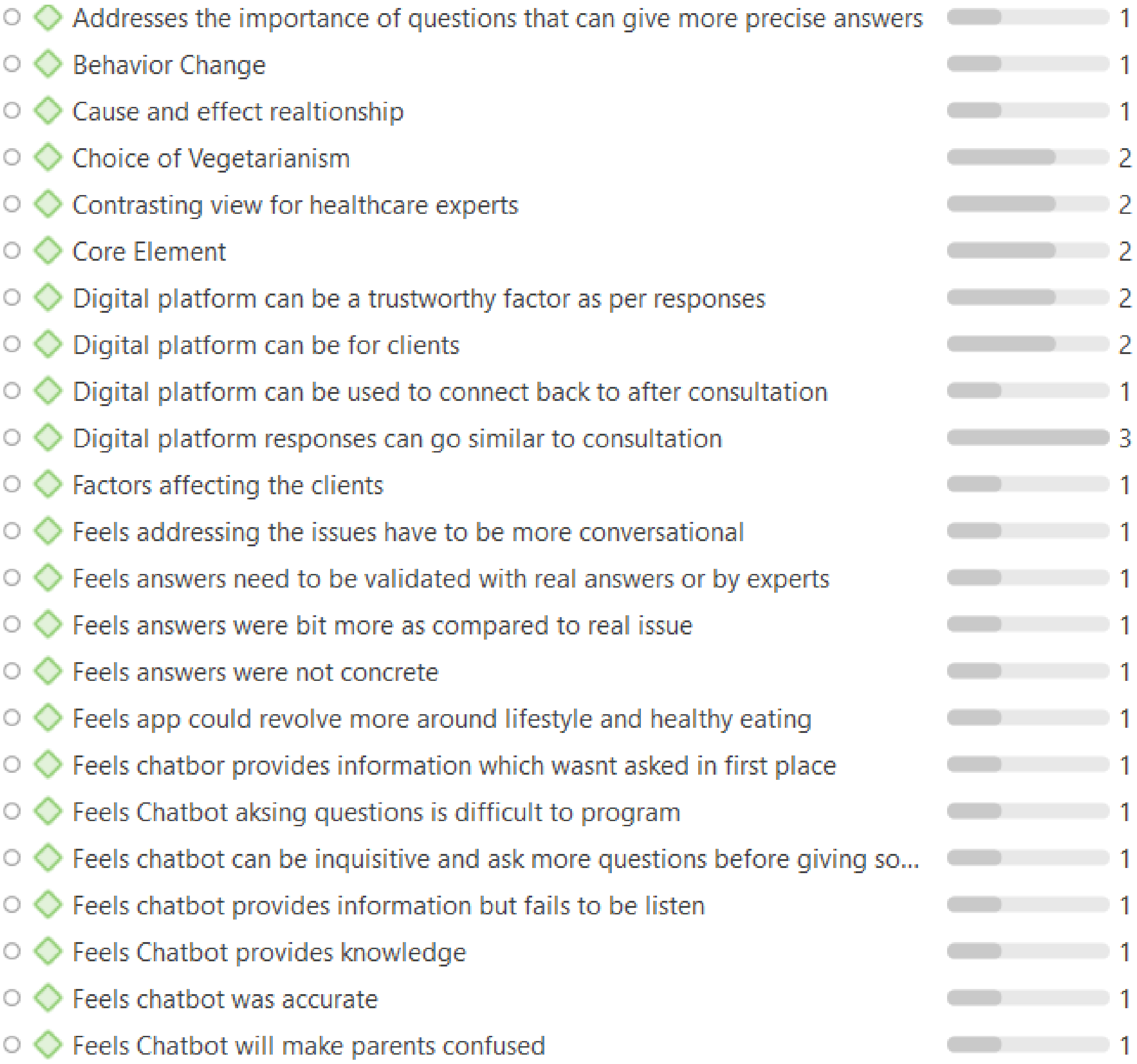
In this section, provide the sheet-with the print of an existing framework (hide few complex wordings), with the existing metric groups and let them sensitize on it for a quick bit.

Then provide a rough idea of a framework (DIY sheet of make your ideal framework) which have different permutation of evaluators as per the image, and then

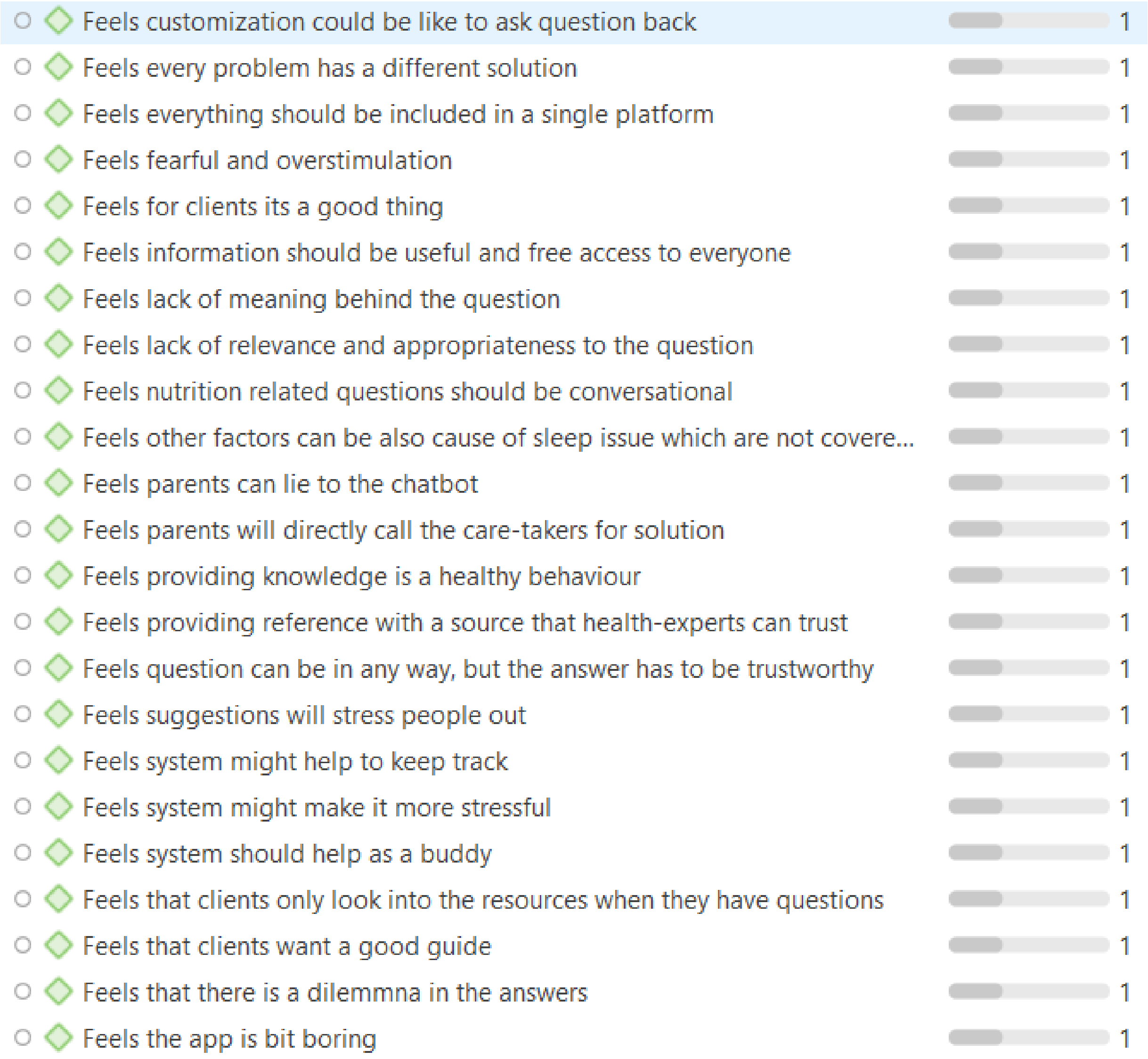
Ask them to think of ways that they would judge the conversation on a expert level or patient level. Also ask is there is something they would take from existing judgement criterias,

Bunch of words are provided: words collected from literature review, mainly transcript of facilitation, their comments on previous evaluation sheet, they can also write something that comes to their mind.
When they select a word, probe further a level and ask why do they think this way

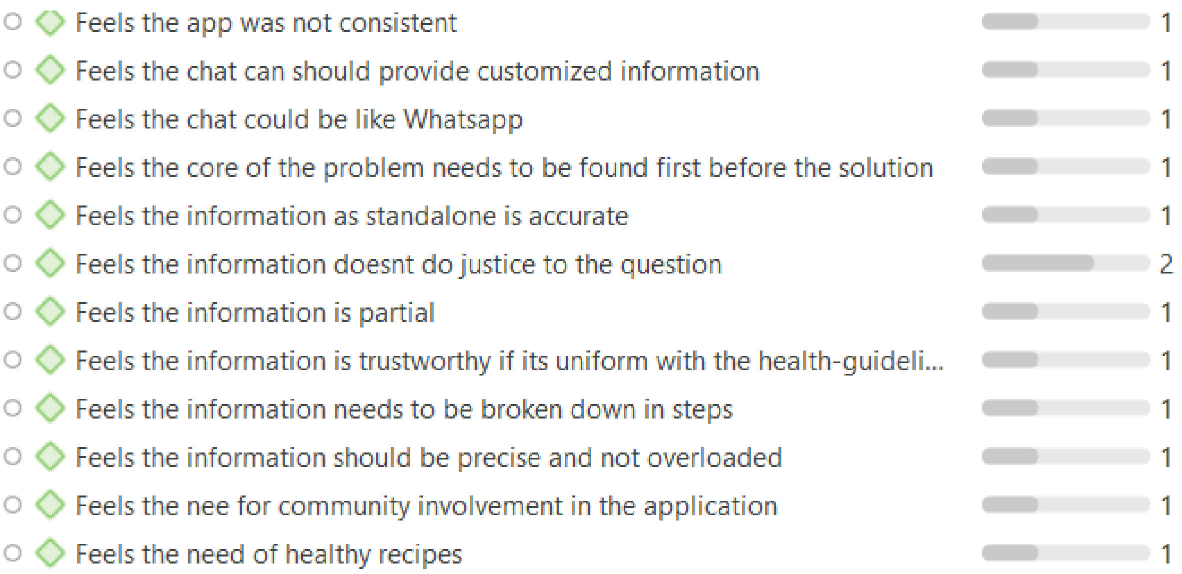
Appendix C. Creative Facilitation Codebook



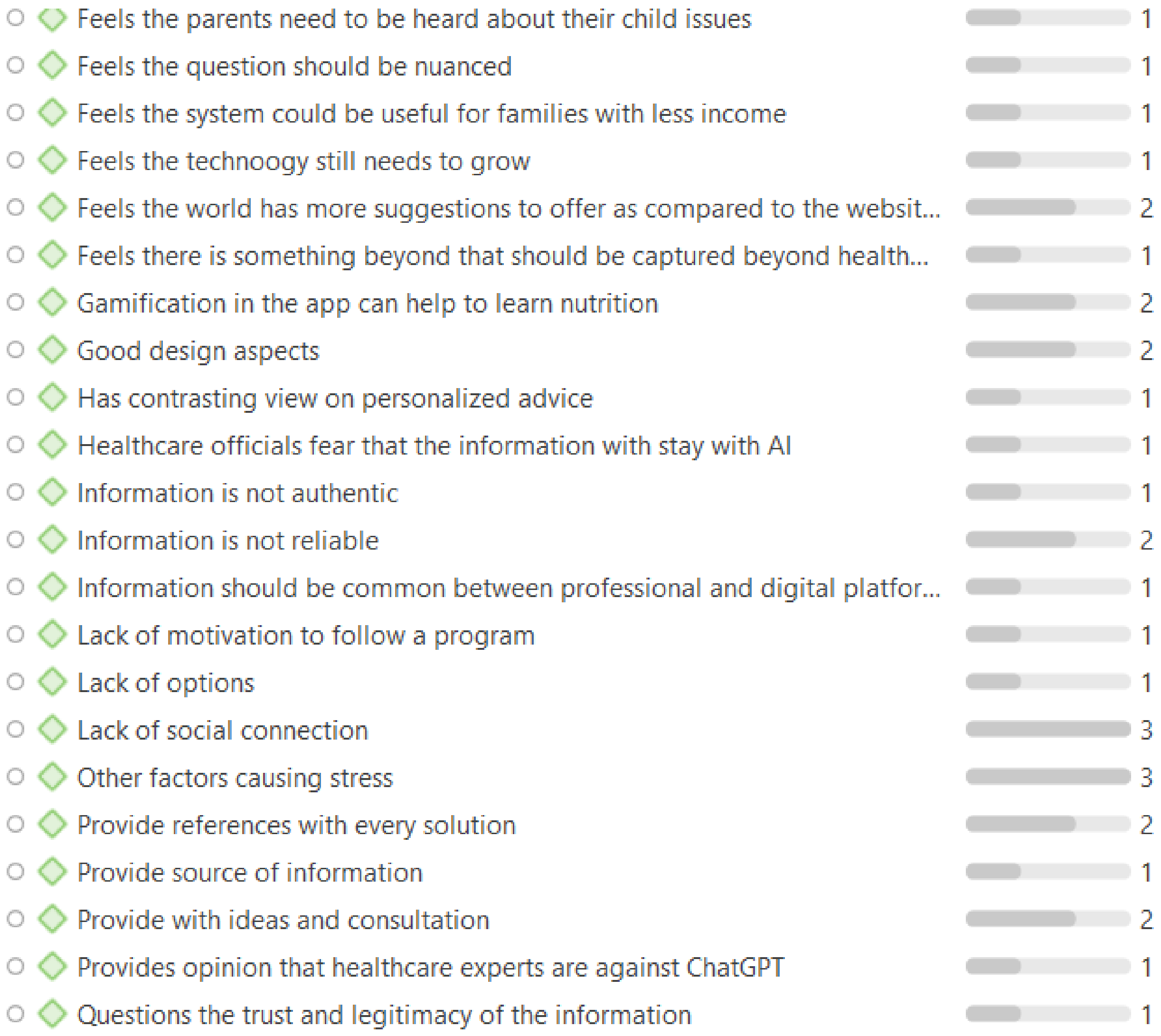
Addresses the importance of questions that can give more precise answers	1
Behavior Change	1
Cause and effect realtionship	1
Choice of Vegetarianism	2
Contrasting view for healthcare experts	2
Core Element	2
Digital platform can be a trustworthy factor as per responses	2
Digital platform can be for clients	2
Digital platform can be used to connect back to after consultation	1
Digital platform responses can go similar to consultation	3
Factors affecting the clients	1
Feels addressing the issues have to be more conversational	1
Feels answers need to be validated with real answers or by experts	1
Feels answers were bit more as compared to real issue	1
Feels answers were not concrete	1
Feels app could revolve more around lifestyle and healthy eating	1
Feels chatbor provides information which wasnt asked in first place	1
Feels Chatbot aksing questions is difficult to program	1
Feels chatbot can be inquisitive and ask more questions before giving so...	1
Feels chatbot provides information but fails to be listen	1
Feels Chatbot provides knowledge	1
Feels chatbot was accurate	1
Feels Chatbot will make parents confused	1



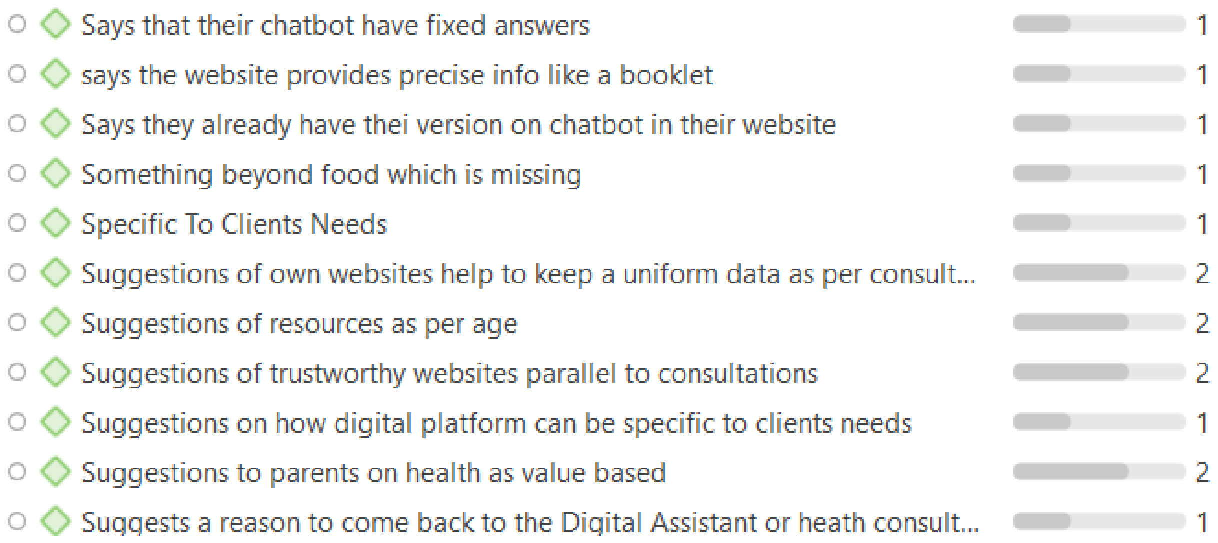
Feels customization could be like to ask question back	1
Feels every problem has a different solution	1
Feels everything should be included in a single platform	1
Feels fearful and overstimulation	1
Feels for clients its a good thing	1
Feels information should be useful and free access to everyone	1
Feels lack of meaning behind the question	1
Feels lack of relevance and appropriateness to the question	1
Feels nutrition related questions should be conversational	1
Feels other factors can be also cause of sleep issue which are not covere...	1
Feels parents can lie to the chatbot	1
Feels parents will directly call the care-takers for solution	1
Feels providing knowledge is a healthy behaviour	1
Feels providing reference with a source that health-experts can trust	1
Feels question can be in any way, but the answer has to be trustworthy	1
Feels suggestions will stress people out	1
Feels system might help to keep track	1
Feels system might make it more stressful	1
Feels system should help as a buddy	1
Feels that clients only look into the resources when they have questions	1
Feels that clients want a good guide	1
Feels that there is a dilemmna in the answers	1
Feels the app is bit boring	1



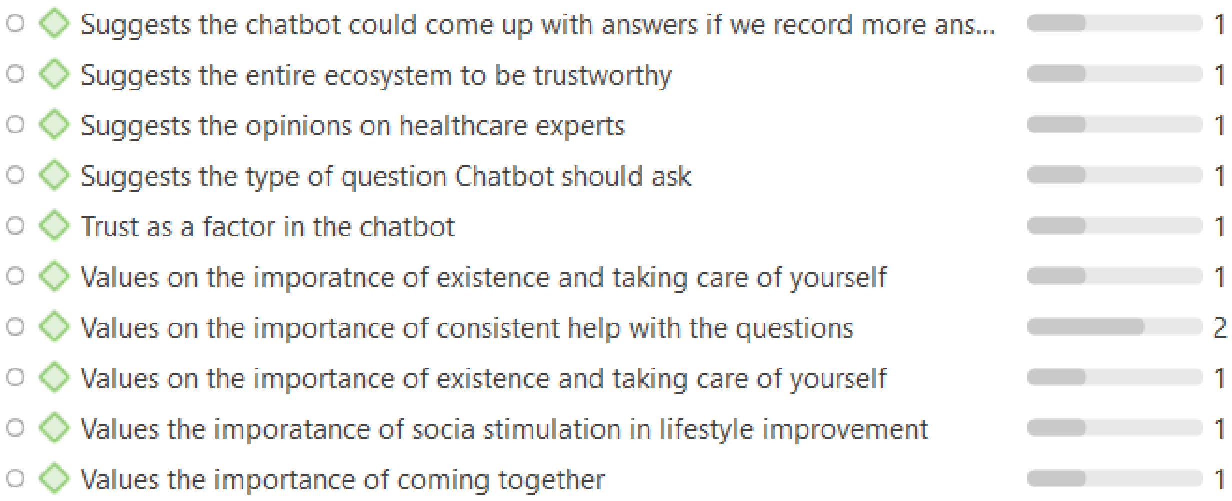
Feels the app was not consistent	1
Feels the chat can should provide customized information	1
Feels the chat could be like Whatsapp	1
Feels the core of the problem needs to be found first before the solution	1
Feels the information as standalone is accurate	1
Feels the information doesnt do justice to the question	2
Feels the information is partial	1
Feels the information is trustworthy if its uniform with the health-guideli...	1
Feels the information needs to be broken down in steps	1
Feels the information should be precise and not overloaded	1
Feels the nee for community involvement in the application	1
Feels the need of healthy recipes	1



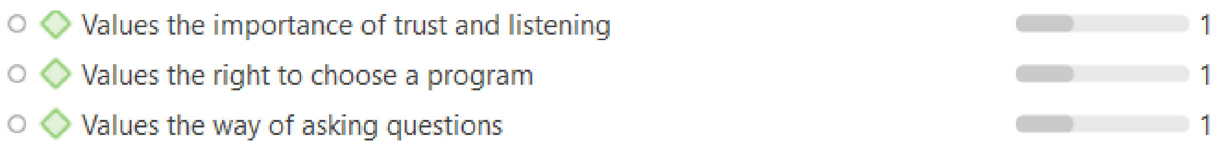
○ ◇ Feels the parents need to be heard about their child issues	1
○ ◇ Feels the question should be nuanced	1
○ ◇ Feels the system could be useful for families with less income	1
○ ◇ Feels the technoogy still needs to grow	1
○ ◇ Feels the world has more suggestions to offer as compared to the websit...	2
○ ◇ Feels there is something beyond that should be captured beyond health...	1
○ ◇ Gamification in the app can help to learn nutrition	2
○ ◇ Good design aspects	2
○ ◇ Has contrasting view on personalized advice	1
○ ◇ Healthcare officials fear that the information with stay with AI	1
○ ◇ Information is not authentic	1
○ ◇ Information is not reliable	2
○ ◇ Information should be common between professional and digital platfor...	1
○ ◇ Lack of motivation to follow a program	1
○ ◇ Lack of options	1
○ ◇ Lack of social connection	3
○ ◇ Other factors causing stress	3
○ ◇ Provide references with every solution	2
○ ◇ Provide source of information	1
○ ◇ Provide with ideas and consultation	2
○ ◇ Provides opinion that healthcare experts are against ChatGPT	1
○ ◇ Questions the trust and legitimacy of the information	1



○ ◇ Says that their chatbot have fixed answers	1
○ ◇ says the website provides precise info like a booklet	1
○ ◇ Says they already have thei version on chatbot in their website	1
○ ◇ Something beyond food which is missing	1
○ ◇ Specific To Clients Needs	1
○ ◇ Suggestions of own websites help to keep a uniform data as per consult...	2
○ ◇ Suggestions of resources as per age	2
○ ◇ Suggestions of trustworthy websites parallel to consultations	2
○ ◇ Suggestions on how digital platform can be specific to clients needs	1
○ ◇ Suggestions to parents on health as value based	2
○ ◇ Suggests a reason to come back to the Digital Assistant or heath consult...	1



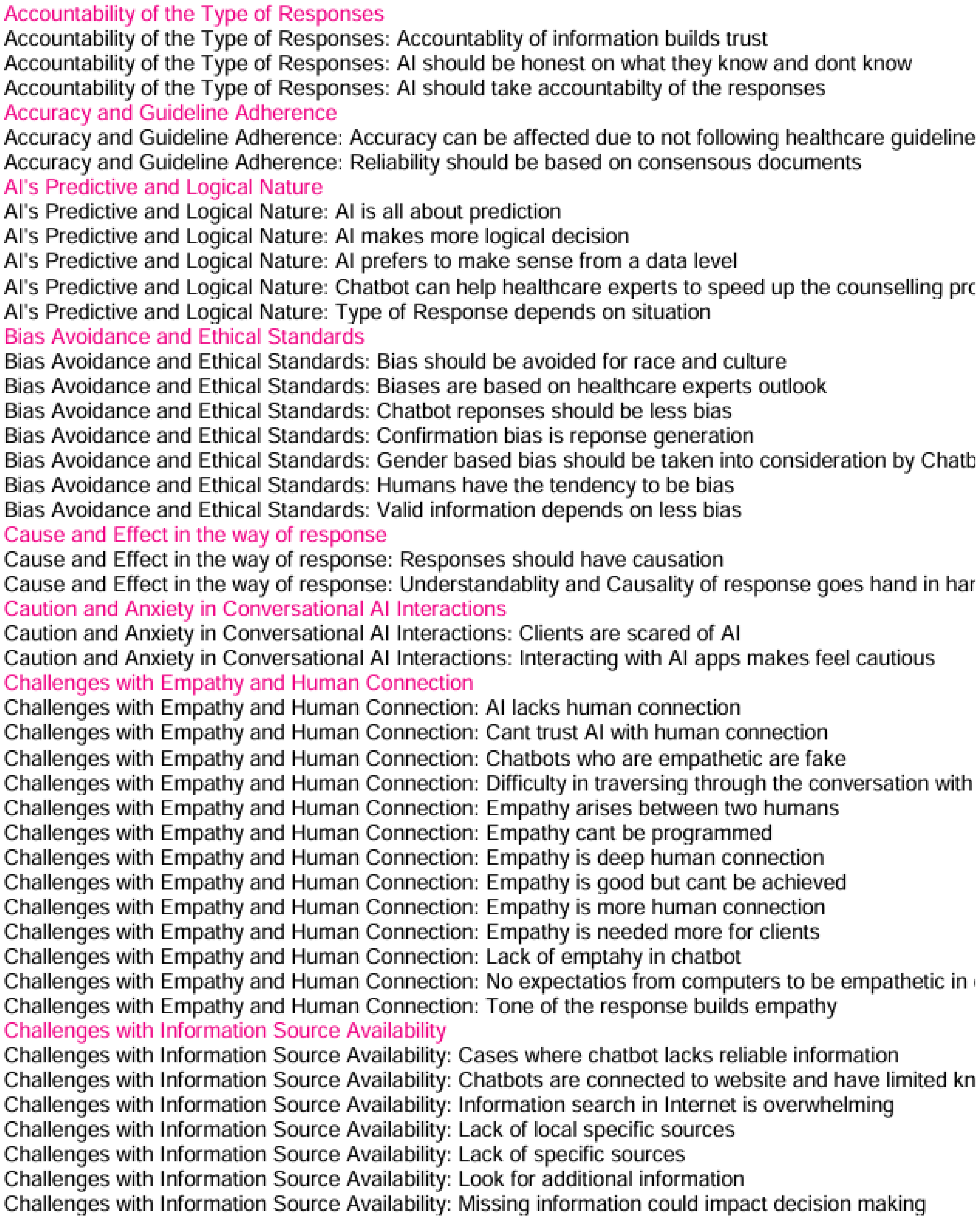
○	◇ Suggests the chatbot could come up with answers if we record more ans...	1
○	◇ Suggests the entire ecosystem to be trustworthy	1
○	◇ Suggests the opinions on healthcare experts	1
○	◇ Suggests the type of question Chatbot should ask	1
○	◇ Trust as a factor in the chatbot	1
○	◇ Values on the imporatnce of existence and taking care of yourself	1
○	◇ Values on the importance of consistent help with the questions	2
○	◇ Values on the importance of existence and taking care of yourself	1
○	◇ Values the imporatance of socia stimulation in lifestyle improvement	1
○	◇ Values the importance of coming together	1



○	◇ Values the importance of trust and listening	1
○	◇ Values the right to choose a program	1
○	◇ Values the way of asking questions	1

Appendix D. User Interview Codebook

Sub-Categories and their Codes



Accountability of the Type of Responses

Accountability of the Type of Responses: Accountablity of information builds trust

Accountability of the Type of Responses: AI should be honest on what they know and dont know

Accountability of the Type of Responses: AI should take accountabilty of the responses

Accuracy and Guideline Adherence

Accuracy and Guideline Adherence: Accuracy can be affected due to not following healthcare guideline

Accuracy and Guideline Adherence: Reliability should be based on consensous documents

AI's Predictive and Logical Nature

AI's Predictive and Logical Nature: AI is all about prediction

AI's Predictive and Logical Nature: AI makes more logical decision

AI's Predictive and Logical Nature: AI prefers to make sense from a data level

AI's Predictive and Logical Nature: Chatbot can help healthcare experts to speed up the counselling prc

AI's Predictive and Logical Nature: Type of Response depends on situation

Bias Avoidance and Ethical Standards

Bias Avoidance and Ethical Standards: Bias should be avoided for race and culture

Bias Avoidance and Ethical Standards: Biases are based on healthcare experts outlook

Bias Avoidance and Ethical Standards: Chatbot reponses should be less bias

Bias Avoidance and Ethical Standards: Confirmation bias is reponse generation

Bias Avoidance and Ethical Standards: Gender based bias should be taken into consideration by Chatb

Bias Avoidance and Ethical Standards: Humans have the tendency to be bias

Bias Avoidance and Ethical Standards: Valid information depends on less bias

Cause and Effect in the way of response

Cause and Effect in the way of response: Responses should have causation

Cause and Effect in the way of response: Understandablity and Causality of response goes hand in har

Caution and Anxiety in Conversational AI Interactions

Caution and Anxiety in Conversational AI Interactions: Clients are scared of AI

Caution and Anxiety in Conversational AI Interactions: Interacting with AI apps makes feel cautious

Challenges with Empathy and Human Connection

Challenges with Empathy and Human Connection: AI lacks human connection

Challenges with Empathy and Human Connection: Cant trust AI with human connection

Challenges with Empathy and Human Connection: Chatbots who are empathetic are fake

Challenges with Empathy and Human Connection: Difficulty in traversing through the conversation with

Challenges with Empathy and Human Connection: Empathy arises between two humans

Challenges with Empathy and Human Connection: Empathy cant be programmed

Challenges with Empathy and Human Connection: Empathy is deep human connection

Challenges with Empathy and Human Connection: Empathy is good but cant be achieved

Challenges with Empathy and Human Connection: Empathy is more human connection

Challenges with Empathy and Human Connection: Empathy is needed more for clients

Challenges with Empathy and Human Connection: Lack of emptahy in chatbot

Challenges with Empathy and Human Connection: No expectatios from computers to be empathetic in

Challenges with Empathy and Human Connection: Tone of the response builds empathy

Challenges with Information Source Availability

Challenges with Information Source Availability: Cases where chatbot lacks reliable information

Challenges with Information Source Availability: Chatbots are connected to website and have limited kn

Challenges with Information Source Availability: Information search in Internet is overwhelming

Challenges with Information Source Availability: Lack of local specific sources

Challenges with Information Source Availability: Lack of specific sources

Challenges with Information Source Availability: Look for additional information

Challenges with Information Source Availability: Missing information could impact decision making

Challenges with Information Source Availability: Quantity of referred resources

Challenges with Information Source Availability: Reliable information is difficult to find

Challenges with Information Source Availability: Responses are good but lacks resources

Challenges with Information Source Availability: Transparency on the resources used

Challenges with Mental Health and AI Responses

Challenges with Mental Health and AI Responses: Chatbots response has higher impact to clients with

Challenges with Mental Health and AI Responses: Facts have a bigger impact on mental health cases

Challenges with Mental Health and AI Responses: Mental health disucssions depend on back and forth

Challenges with Reliability and Trustworthiness

Challenges with Reliability and Trustworthiness: Annoyed of wrong answers

Challenges with Reliability and Trustworthiness: Cherry picking affects reliablity

Challenges with Reliability and Trustworthiness: Direct response leads to doubt the accuracy and reliab

Challenges with Reliability and Trustworthiness: Direct responses are a big red flag

Challenges with Reliability and Trustworthiness: Factors causing trust issues

Challenges with Reliability and Trustworthiness: Fixation in response is harmful

Challenges with Reliability and Trustworthiness: Incorrect responses give rise to trust issues of next rep

Challenges with Reliability and Trustworthiness: Lack of reliability

Challenges with Reliability and Trustworthiness: Most of the information is not reliable and truth

Challenges with Reliability and Trustworthiness: Reliablity of information proves the worth of AI

Challenges with Reliability and Trustworthiness: Responses feel just like information retrival

Challenges with Reliability and Trustworthiness: Responses make you doubt on reliability

Challenges with Reliability and Trustworthiness: Responses which are robotic affects trust

Chatbot shouldnt feel like a stone

Clarity and Simplicity in AI Responses

Clarity and Simplicity in AI Responses: Answers you cant make sense of it

Clarity and Simplicity in AI Responses: Avoid complex wordings

Clarity and Simplicity in AI Responses: Clarity gives more grounded response

Clarity and Simplicity in AI Responses: Clear wordings

Clarity and Simplicity in AI Responses: Clients have difficulty in understanding jargons

Clarity and Simplicity in AI Responses: Doctors understand jargons

Clarity and Simplicity in AI Responses: Importance of clarity in AI generated responses

Clarity and Simplicity in AI Responses: Information should be provided equally to everyone

Clarity and Simplicity in AI Responses: Information should be simple and not confusing

Clarity and Simplicity in AI Responses: Response feels like lost in translation

Clarity and Simplicity in AI Responses: Response in different language is not trustworthy

Clarity and Simplicity in AI Responses: Response needs to be grasped by humans

Clarity and Simplicity in AI Responses: Responses should be clear and concise from Subject matter ex

Clarity and Simplicity in AI Responses: Simple wordings

Client Expectations and Patience

Client Expectations and Patience: Clients have to be patient in responses

Client Expectations and Patience: Clients overexpect responses

Client Expectations and Patience: Clients should consume information from both sides

Client-Focused User Experience in AI

Client-Focused User Experience in AI: Chatbot should make them feel at ease

Client-Focused User Experience in AI: Clients might prefer UI and Usablity

Client-Focused User Experience in AI: Clients prefer warm and friendly interfaces

Client-Focused User Experience in AI: Design feels comfortable

Client-Focused User Experience in AI: Personalise information as per needs of clients

Client-Focused User Experience in AI: Understanding should be more friendlier

Client-Focused User Experience in AI: Usablity of the interface brings in friendliness

Client-Focused User Experience in AI: Value tension between technicality and simplicity

Completeness of responses

Completeness of responses: Dosent mind asking questions further

Completeness of responses: Information should cover complete and sufficient information

Consistency and Completeness in Responses

Consistency and Completeness in Responses: Appropriateness of responses

Consistency and Completeness in Responses: Completeness gives rise to reliablity

Consistency and Completeness in Responses: Consistent reliable information builds trust

Consistency and Completeness in Responses: Reliablity gives rise to consistency

Consistency and Completeness in Responses: Responses are not based on judgement

Consistency and Completeness in Responses: Responses should be complete

Consistency and Completeness in Responses: Responses were direct but accurate

Context-Based Understandability

Context-Based Understandability: Interpretaion of questions should be done on more contextual terms

Context-Based Understandability: Levels of deep information

Context-Based Understandability: Reliablity and understandability is needed to understand the respons

Context-Based Understandability: Understanablity depends on context based conversation

Context-Based Understandability: Understanding of issues at social and economical level

Contextual Understanding Reduces Uncertainty

Contextual Understanding Reduces Uncertainty: Direct responses should be given on context

Contextual Understanding Reduces Uncertainty: Suggestions come after dealing with uncertainties

Contextual Understanding Reduces Uncertainty: Tackling uncertainty can only be possible after context

Contextual Understanding Reduces Uncertainty: Uncertainty in information is removed when context co

Data Understanding and Presentation

Data Understanding and Presentation: Clarity provides a direction in response

Data Understanding and Presentation: People should understand the data

Data Understanding and Presentation: Presentation of content matters

Data Understanding and Presentation: Simple wordings to focus on more clarity and reliablity

Dosent believe in empathy

Dosent expect the client to read all documents but references should be there

Effective human conversation between clients and experts

Effective human conversation between clients and experts: Doctors have a way of conversation with cli

Effective human conversation between clients and experts: Effective Conversation

Effective human conversation between clients and experts: Non verbal communication between doctor

Effective Response from Information Categorization and Filtering

Effective Response from Information Categorization and Filtering: Data from external sources or word c

Effective Response from Information Categorization and Filtering: Filter good information from Internet

Effective Response from Information Categorization and Filtering: General information is sometimes ha

Effective Response from Information Categorization and Filtering: Information in the internet is not alwa

Effective Response from Information Categorization and Filtering: Quality of dataset used

Effective Response from Information Categorization and Filtering: Question based filtering

Effective Response from Information Categorization and Filtering: Redundancy gives rise to lack of hon

Effective Response from Information Categorization and Filtering: Seek additional sources of informatic

Effective Response from Information Categorization and Filtering: Seek and categorize healthcare infor

Effective Response from Information Categorization and Filtering: Sources from various websites are ir

Effectiveness of suggestive responses

Effectiveness of suggestive responses: AI chatbot can be used as a tool to help

Effectiveness of suggestive responses: AI could be suggestive to clients

Effectiveness of suggestive responses: Careful suggestions should be provided

Effectiveness of suggestive responses: Chatbot can act as a screening tool

Effectiveness of suggestive responses: Chatbot can be used as an assistant
Effectiveness of suggestive responses: Chatbots can be used for tip or advice
Effectiveness of suggestive responses: Direct commands feels harsh
Effectiveness of suggestive responses: Direct response would give rise to distrustful nature
Effectiveness of suggestive responses: Logical advices for clients during the time of uncertainty
Effectiveness of suggestive responses: Responses should not feel like commands
Effectiveness of suggestive responses: Solutions are suggested on collaborative manner
Effectiveness of suggestive responses: Suggestive responses build habit

Encouraging responses to the clients

Encouraging responses to the clients: Chatbot AI fails to visualise
Encouraging responses to the clients: Chatbots can drive behavioural change as habit building
Encouraging responses to the clients: Clients try to make connection with Chatbot
Encouraging responses to the clients: Dealing with complaints with motivation and supportiveness
Encouraging responses to the clients: Encourage the clients to follow a habit for consistency
Encouraging responses to the clients: Healthcare experts and clients both need suggestions
Encouraging responses to the clients: Optimism depends on the trust
Encouraging responses to the clients: Positive manipulation in right direction is important
Encouraging responses to the clients: Positive motivation builds habits
Encouraging responses to the clients: Steps towards good habits

Ensuring High-Quality Responses through Effective Questions

Ensuring High-Quality Responses through Effective Questions: Easy to get good answer on the basis c

Evidences of warmth and kindness in responses generated

Evidences of warmth and kindness in responses generated: Clients needs empathy
Evidences of warmth and kindness in responses generated: Empathy has layers which depend on othe
Evidences of warmth and kindness in responses generated: Empathy requires warmth and kindness in
Evidences of warmth and kindness in responses generated: Empathy should be shown to clients with d
Evidences of warmth and kindness in responses generated: Humans have understanding of emotions a
Evidences of warmth and kindness in responses generated: Implementaion of empathy is challenging c
Evidences of warmth and kindness in responses generated: Kindness leads to understandablity
Evidences of warmth and kindness in responses generated: Kindness related to humanisation
Evidences of warmth and kindness in responses generated: Recognition of human emotions
Evidences of warmth and kindness in responses generated: Response is pretty direct
Evidences of warmth and kindness in responses generated: Responses should not be rude
Evidences of warmth and kindness in responses generated: Sensitive information should be suggested
Evidences of warmth and kindness in responses generated: Tradeoff for human connection, compassic
Evidences of warmth and kindness in responses generated: Warm responses to clients struggles

Evolving AI Capabilities and Future Prospects

Evolving AI Capabilities and Future Prospects: AI has more knowledge factor
Evolving AI Capabilities and Future Prospects: AI has promising development
Evolving AI Capabilities and Future Prospects: AI in future could detect human emotions
Evolving AI Capabilities and Future Prospects: AI in healthcare is unavoidable
Evolving AI Capabilities and Future Prospects: AI is not capable for every solution
Evolving AI Capabilities and Future Prospects: AI should be constantly learning
Evolving AI Capabilities and Future Prospects: Machine should act like a human

Factors Affecting Information Trustworthiness

Factors Affecting Information Trustworthiness: Caring and Trustworthy should be taken care of
Factors Affecting Information Trustworthiness: Chatbot provides and answer and its not reliable
Factors Affecting Information Trustworthiness: Grammatical efforts results in distrust
Factors Affecting Information Trustworthiness: Incorrect answers impacts trustworthiness
Factors Affecting Information Trustworthiness: Prefer sources from trustworthy sites

Factors Affecting Information Trustworthiness: Trusting ones own knowledge base
Factors Affecting Information Trustworthiness: Trustworthiness affects due to assumed information gen
Factors Affecting Information Trustworthiness: Trustworthy is affected due to overload of information
Factors Affecting Information Trustworthiness: Trustworthy of information
Factors Affecting Information Trustworthiness: Trustworthy of information due to resources
Factors Determining Relevance and Reliability
Factors Determining Relevance and Reliability: Factors causing reliablity issues
Factors Determining Relevance and Reliability: Factors of reliablity
Factors Determining Relevance and Reliability: Factors of superficial responses
Factors Determining Relevance and Reliability: Factors of Trust and comfort
Factors Determining Relevance and Reliability: Factors that decide good and bad sources
Factors Determining Relevance and Reliability: Factors that determine relevance
Factors Determining Relevance and Reliability: Reliability comes if its accessible to everyone is society
Factors of trust
Factors of understandibility
Healthcare Expert Perspectives and Evaluations
Healthcare Expert Perspectives and Evaluations: Depends on ones own perspective
Healthcare Expert Perspectives and Evaluations: Good evaluation gives higher fairness
Healthcare Expert Perspectives and Evaluations: Healthcare experts dont need supportiveness and enc
Healthcare Expert Perspectives and Evaluations: Healthcare experts responses knows whats true and t
Healthcare Expert Perspectives and Evaluations: Healthcare from an expert perspective
Healthcare Expert-Focused User Experience in AI
Healthcare Expert-Focused User Experience in AI: Find the use of AI avatar not useful
Healthcare Expert-Focused User Experience in AI: Friendly avatar not required as its an information pro
Healthcare Expert-Focused User Experience in AI: Functionality comes first as compared to content
Healthcare Expert-Focused User Experience in AI: Healthcare experts need data not design and UI
Healthcare Expert-Focused User Experience in AI: Healthcare providers dont need logo and design
Healthcare Expert-Focused User Experience in AI: High level UI design takes away focus from key info
Healthcare Expert-Focused User Experience in AI: Use of chatbot goes down if answers are not correc
Honesty, Clarity, and Ethics in Communication
Honesty, Clarity, and Ethics in Communication: Direct and straightforward response related to honesty
Honesty, Clarity, and Ethics in Communication: Direct and transparent asnwers give rise to honesty
Honesty, Clarity, and Ethics in Communication: Healthcare experts appreciate honesty and clarity
Honesty, Clarity, and Ethics in Communication: Honesty is more than accuracy with ethics and standar
Honesty, Clarity, and Ethics in Communication: Responses have to be fair and clear
Impact of Trust and Reliability on Decisions
Impact of Trust and Reliability on Decisions: Factors of Trust and distrust
Impact of Trust and Reliability on Decisions: Impacts decision making on basis of reliablity and trust
Impact of Trust and Reliability on Decisions: Reliablity and Trustworthy depends on personal beliefs
Impact of Trust and Reliability on Decisions: Trust is affected if response is different
Importance of Accuracy
Importance of Accuracy: Accuracy can be built by asking more and more questions
Importance of Accuracy: Accuracy can be focused on my removing inaccurate ones
Importance of Accuracy: Accuracy has a higher value
Importance of Accuracy: Accuracy has more value than bias
Importance of Accuracy: Accuracy is important for Healthcare experts
Importance of Accuracy: Chatbot should provide accurate answers by default
Importance of Accuracy: Consistency depends on accurate answers to questions multiple times
Importance of Accuracy: Fully correct answer gives rise to accuracy
Importance of Accuracy: Importance of honesty gives rise to accuracy

Importance of Accuracy: Information should be sufficient enough to cover all topics

Importance of Accuracy: Informations has to be accurate in sources and responses

Importance of Accuracy: Partially correct answer to research on a case study of client

Importance of Accuracy: Responses by chatbot can be accurate and non-accurate

Importance of Action-Oriented Knowledge

Importance of Action-Oriented Knowledge: Action knowledge is very important

Importance of Action-Oriented Knowledge: Action Perspective

Importance of Action-Oriented Knowledge: Action should follow after a response

Importance of Action-Oriented Knowledge: Clients have their action perspective on suggesitveness

Importance of Action-Oriented Knowledge: Content or knowledge should be on action perspective

Importance of Action-Oriented Knowledge: Ways to acheive wanted behaviour

Importance of Context Information for Clients

Importance of Context Information for Clients: Chatbot should aim to gather context

Importance of Context Information for Clients: Chatbots can take prior information

Importance of Context Information for Clients: Context and Content of datasources is important

Importance of Context Information for Clients: Context and reasoning for suggestion

Importance of Context Information for Clients: Conversations should be restarted to understand more c

Importance of Context Information for Clients: Data is harmful if created without meaning

Importance of Context Information for Clients: Data should be context driven

Importance of Context Information for Clients: Every response should have a background to it

Importance of Context Information for Clients: Gather more context

Importance of Context Information for Clients: Information should be provided on context and not assum

Importance of Context Information for Clients: Learning happens due to context based conversations

Importance of Context Information for Clients: Reliabity gets stronger with respect to context

Importance of Context Information for Clients: Responses need to fit wider context of knowledge

Importance of Positive and Motivating Responses

Importance of Positive and Motivating Responses: Chatbots shoud focus on motivation

Importance of Positive and Motivating Responses: Encouraging and motivating

Importance of Positive and Motivating Responses: Judegement should be positive

Importance of Positive and Motivating Responses: Motivation has higher preference

Importance of Positive and Motivating Responses: Positivity should be shown at times of need

Importance of Positive and Motivating Responses: Tool should communicate positivity

Importance of Self-Research on Local Information Sources

Importance of Self-Research on Local Information Sources: Healthcare experts have udnerstanding of

Importance of Self-Research on Local Information Sources: Healthcare from a parental perspective

Importance of Self-Research on Local Information Sources: People listen to the elder in the house

Importance of Self-Research on Local Information Sources: Putting human knowledge first

Importance of Self-Research on Local Information Sources: Self research for the sources

Judgements leads to build guilt in clients

Local and Global healthcare sources

Local and Global healthcare sources: Accuracy of responses also depends on resources referred from

Local and Global healthcare sources: Ai can refer to all global healthcare resources

Local and Global healthcare sources: AI has more access to latest information

Local and Global healthcare sources: Belief of reliable information

Local and Global healthcare sources: Biggest issue is the genralized information

Local and Global healthcare sources: Chatbot should redirect it to respective websites

Local and Global healthcare sources: Clients should trust the guidelines from nutrion centre

Local and Global healthcare sources: Compare the response to the literature research being done

Local and Global healthcare sources: Cross verification with resources in provides reliablity

Local and Global healthcare sources: Data is sponsorsed by organisations

Local and Global healthcare sources: Global level of resources

Local and Global healthcare sources: Local articles and research resources

Local and Global healthcare sources: Local resources referring builds trust

Local and Global healthcare sources: Local sources have distrust

Local and Global healthcare sources: Provides global standard for healthcare resources

Local and Global healthcare sources: Real life information is converted to numbers

Local and Global healthcare sources: References to the information shared

Local and Global healthcare sources: Reliable information builds trust

Local and Global healthcare sources: Reliable resources are better than internet search

Local and Global healthcare sources: Reliablity in the global references

Local and Global healthcare sources: Sponsored sources

Local and Global healthcare sources: Suggestions also depend on doctors and their views on guideline

Necessity of Human Intervention in AI Responses

Necessity of Human Intervention in AI Responses: AI should suggest consult healthcare provider

Necessity of Human Intervention in AI Responses: Balancing the healthcare

Necessity of Human Intervention in AI Responses: Crucial questions should need human intervention

Necessity of Human Intervention in AI Responses: Healthcare expert can mitigate risks

Necessity of Human Intervention in AI Responses: Human intervention is neccessary

Necessity of Human Intervention in AI Responses: Importance of healthcare provider is more than AI in

Necessity of Human Intervention in AI Responses: Importance of human connection

Necessity of Human Intervention in AI Responses: Important disucssions should be consulted with colle

Necessity of Human Intervention in AI Responses: Knowledge comes from interacting with colleagues

Necessity of Human Intervention in AI Responses: Professional and Client expectation of response

Necessity of Human Intervention in AI Responses: Responses should be verified by doctors

Necessity of Human Intervention in AI Responses: Suggestions should be based on clients medical hist

Non-Judgmental and Kind Communication

Non-Judgmental and Kind Communication: Chatbot should never judge

Non-Judgmental and Kind Communication: Judgement builds insecurity

Non-Judgmental and Kind Communication: Judgements can be biased on situation

Non-Judgmental and Kind Communication: Responses should be more kind

Perception of AI responses

Perception of AI responses: Response of AI gives clarity to continue conversation

Perception of AI responses: Responses should be suggestive and not directive

Prompting Techniques and Interaction with Chatbot

Prompting Techniques and Interaction with Chatbot: Ask questions to get more context of the person

Prompting Techniques and Interaction with Chatbot: Chatbot answers to a good prompt

Prompting Techniques and Interaction with Chatbot: Conversation can be broad or specific

Prompting Techniques and Interaction with Chatbot: Conversation fatigue

Prompting Techniques and Interaction with Chatbot: Core idea of the query to AI

Prompting Techniques and Interaction with Chatbot: Different language has different complex words

Prompting Techniques and Interaction with Chatbot: Direction of conversation depends on the response

Prompting Techniques and Interaction with Chatbot: Generalized responses work for the masses

Prompting Techniques and Interaction with Chatbot: Good advice stems from good prompt

Prompting Techniques and Interaction with Chatbot: Right question gives a right answer

Prompting Techniques and Interaction with Chatbot: The root question which builds the conversation

Prompting Techniques and Interaction with Chatbot: Training clients from different background on how

Prompting Techniques and Interaction with Chatbot: Two-way conversation by AI Chatbot

Responses for wellbeing of clients

Responses for wellbeing of clients: Chatbot can be used for personlisation of healthy food

Responses for wellbeing of clients: Factors for better child health

Responses for wellbeing of clients: Kids and mental health responses are very sesnitive

Responses for wellbeing of clients: Wellbeing of children

Risks involved with Biased AI Responses

Risks involved with Biased AI Responses: Greater risk in using biased repsonses

Risks involved with Biased AI Responses: Prefernces for less bias

Risks of fake information in AI Responses

Risks of fake information in AI Responses: Compare it with other sources of information

Risks of fake information in AI Responses: Fake information to client is risky as they dont have knowled

Risks of fake information in AI Responses: Fake references in responses can be deceptive

Risks of fake information in AI Responses: Provides superficial answers

Risks of fake information in AI Responses: Reponses should be verified by doctors

Risks of fake information in AI Responses: Safety should be in reliablity of information

Self-Assessment and Awareness

Self-Assessment and Awareness: Self assessment is self awareness

Self-Assessment and Awareness: Self assessment of the reponse by healthcare expert

Specificity and Correctness in generated responses

Specificity and Correctness in generated responses: Asks questions to explore what kind of answers

Specificity and Correctness in generated responses: Clients are not looking for scientific correctness

Specificity and Correctness in generated responses: Correct answers give reliablity

Specificity and Correctness in generated responses: Feels annoyed when Chatbot dosent respond acc

Specificity and Correctness in generated responses: Half correct answers can be probed further to buil

Specificity and Correctness in generated responses: Need for fully correct answers

Specificity and Correctness in generated responses: Open question responses are never fully correct

Specificity and Correctness in generated responses: Probing more to chatbot becomes fluffy informatio

Specificity and Correctness in generated responses: Probing should be done till the point of reliability o

Specificity and Correctness in generated responses: Provide relevant answers to different questions

Specificity and Correctness in generated responses: Redundancy in responses

Specificity and Correctness in generated responses: Reponses are correct but going in circles

Specificity and Correctness in generated responses: Responses should be relevant to the question

Specificity and Correctness in generated responses: Specific responses feel left out

Specificity and Correctness in generated responses: Superficial answers

Strategies to give reliable information

Strategies to tackle fake information

Transparency in Information generated from sources

Transparency in Information generated from sources: Expect transparency from AI

Transparency in Information generated from sources: Lack of transparency gives rise to lack of trust

Transparency in Information generated from sources: Preferences of transparency and UI design

Transparency in Information generated from sources: Transparency in data sources and algorithms

Transparency in Information generated from sources: Transparency is affected behind attractive desigr

Trust Building through Reliability and Information

Trust Building through Reliability and Information: Good information builds trust

Trust Building through Reliability and Information: People dont accept advice once a trust is broken

Trust Building through Reliability and Information: References of resources in response builds trust

Trust Building through Reliability and Information: Reliabity and good responses go hand in hand

Trust Building through Reliability and Information: Trust and referred sources

Trust Building through Reliability and Information: Trust builds reliablity

Trust Building through Reliability and Information: Trust is due to accuracy of information

Trust Building through Reliability and Information: Value of Trust is very high in clients

Trustworthy Responses and Their Impact

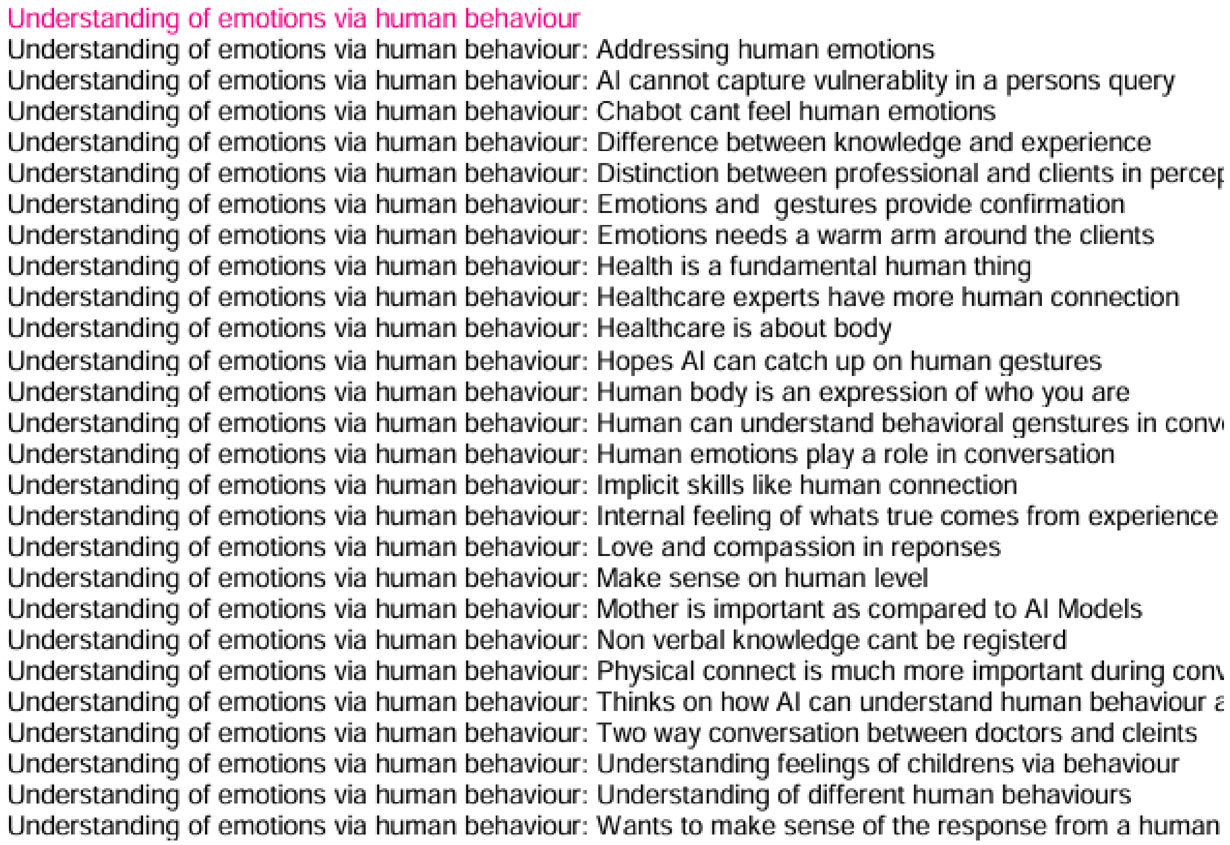
Trustworthy Responses and Their Impact: Accuracy depends on trustable information and updated reso

Trustworthy Responses and Their Impact: AI Model can be trusted on the basis of the response genera

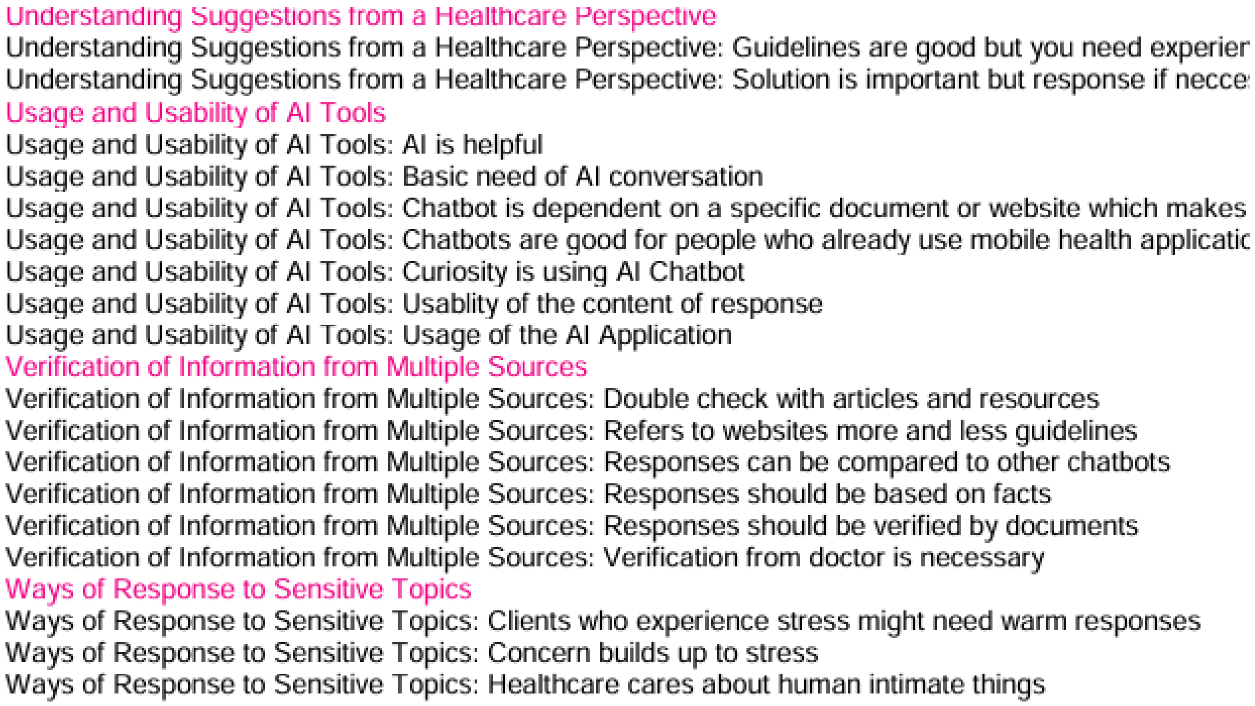
Trustworthy Responses and Their Impact: Honesty is like accuracy with morals and values

Trustworthy Responses and Their Impact: Importance of trustworthy information

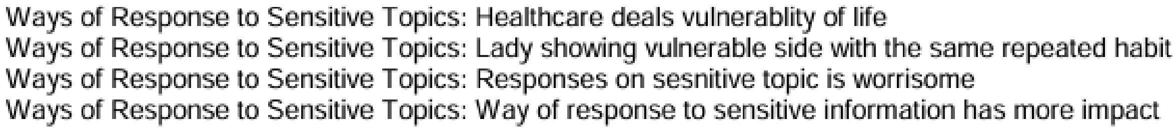
Trustworthy Responses and Their Impact: Trust plays a bigger role in AI response



Understanding of emotions via human behaviour
Understanding of emotions via human behaviour: Addressing human emotions
Understanding of emotions via human behaviour: AI cannot capture vulnerablity in a persons query
Understanding of emotions via human behaviour: Chabot cant feel human emotions
Understanding of emotions via human behaviour: Difference between knowledge and experience
Understanding of emotions via human behaviour: Distinction between professional and clients in percep
Understanding of emotions via human behaviour: Emotions and gestures provide confirmation
Understanding of emotions via human behaviour: Emotions needs a warm arm around the clients
Understanding of emotions via human behaviour: Health is a fundamental human thing
Understanding of emotions via human behaviour: Healthcare experts have more human connection
Understanding of emotions via human behaviour: Healthcare is about body
Understanding of emotions via human behaviour: Hopes AI can catch up on human gestures
Understanding of emotions via human behaviour: Human body is an expression of who you are
Understanding of emotions via human behaviour: Human can understand behavioral gerstures in conv
Understanding of emotions via human behaviour: Human emotions play a role in conversation
Understanding of emotions via human behaviour: Implicit skills like human connection
Understanding of emotions via human behaviour: Internal feeling of whats true comes from experience
Understanding of emotions via human behaviour: Love and compassion in reponses
Understanding of emotions via human behaviour: Make sense on human level
Understanding of emotions via human behaviour: Mother is important as compared to AI Models
Understanding of emotions via human behaviour: Non verbal knowledge cant be registerd
Understanding of emotions via human behaviour: Physical connect is much more important during conv
Understanding of emotions via human behaviour: Thinks on how AI can understand human behaviour a
Understanding of emotions via human behaviour: Two way conversation between doctors and cleints
Understanding of emotions via human behaviour: Understanding feelings of childrens via behaviour
Understanding of emotions via human behaviour: Understanding of different human behaviours
Understanding of emotions via human behaviour: Wants to make sense of the response from a human

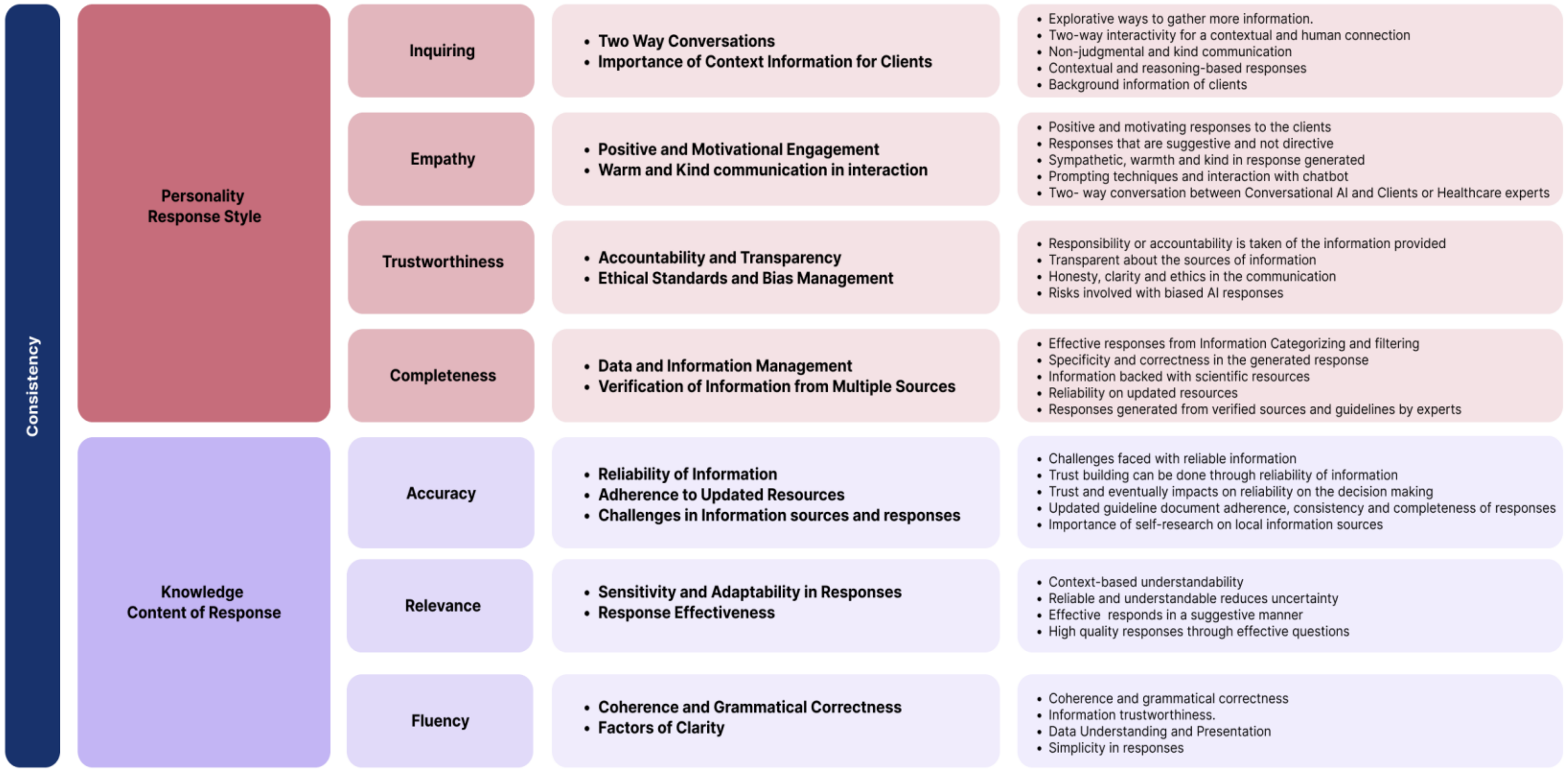


Understanding Suggestions from a Healthcare Perspective
Understanding Suggestions from a Healthcare Perspective: Guidelines are good but you need experier
Understanding Suggestions from a Healthcare Perspective: Solution is important but response if necce
Usage and Usability of AI Tools
Usage and Usability of AI Tools: AI is helpful
Usage and Usability of AI Tools: Basic need of AI conversation
Usage and Usability of AI Tools: Chatbot is dependent on a specific document or website which makes
Usage and Usability of AI Tools: Chatbots are good for people who already use mobile health applicatic
Usage and Usability of AI Tools: Curiosity is using AI Chatbot
Usage and Usability of AI Tools: Usablity of the content of response
Usage and Usability of AI Tools: Usage of the AI Application
Verification of Information from Multiple Sources
Verification of Information from Multiple Sources: Double check with articles and resources
Verification of Information from Multiple Sources: Refers to websites more and less guidelines
Verification of Information from Multiple Sources: Responses can be compared to other chatbots
Verification of Information from Multiple Sources: Responses should be based on facts
Verification of Information from Multiple Sources: Responses should be verified by documents
Verification of Information from Multiple Sources: Verification from doctor is necessary
Ways of Response to Sensitive Topics
Ways of Response to Sensitive Topics: Clients who experience stress might need warm responses
Ways of Response to Sensitive Topics: Concern builds up to stress
Ways of Response to Sensitive Topics: Healthcare cares about human intimate things



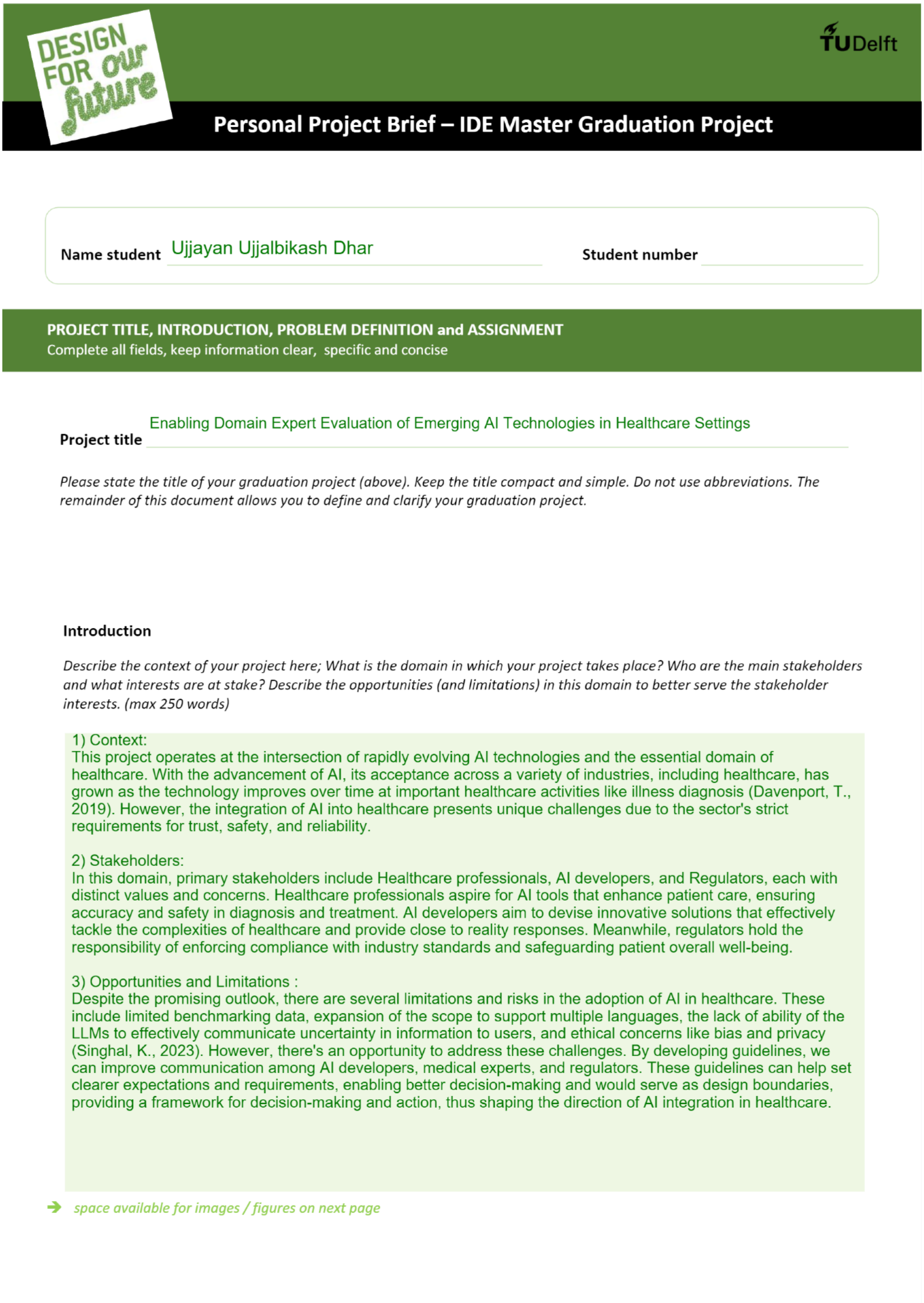
Ways of Response to Sensitive Topics: Healthcare deals vulnerablity of life
Ways of Response to Sensitive Topics: Lady showing vulnerable side with the same repeated habit
Ways of Response to Sensitive Topics: Responses on sesnitive topic is worrisome
Ways of Response to Sensitive Topics: Way of response to sensitive information has more impact

Appendix E. Evaluation Framework Extended Version



Consistency	Personality Response Style	Inquiring	• Two Way Conversations • Importance of Context Information for Clients	• Explorative ways to gather more information. • Two-way interactivity for a contextual and human connection • Non-judgmental and kind communication • Contextual and reasoning-based responses • Background information of clients
		Empathy	• Positive and Motivational Engagement • Warm and Kind communication in interaction	• Positive and motivating responses to the clients • Responses that are suggestive and not directive • Sympathetic, warmth and kind in response generated • Prompting techniques and interaction with chatbot • Two- way conversation between Conversational AI and Clients or Healthcare experts
		Trustworthiness	• Accountability and Transparency • Ethical Standards and Bias Management	• Responsibility or accountability is taken of the information provided • Transparent about the sources of information • Honesty, clarity and ethics in the communication • Risks involved with biased AI responses
		Completeness	• Data and Information Management • Verification of Information from Multiple Sources	• Effective responses from Information Categorizing and filtering • Specificity and correctness in the generated response • Information backed with scientific resources • Reliability on updated resources • Responses generated from verified sources and guidelines by experts
	Knowledge Content of Response	Accuracy	• Reliability of Information • Adherence to Updated Resources • Challenges in Information sources and responses	• Challenges faced with reliable information • Trust building can be done through reliability of information • Trust and eventually impacts on reliability on the decision making • Updated guideline document adherence, consistency and completeness of responses • Importance of self-research on local information sources
		Relevance	• Sensitivity and Adaptability in Responses • Response Effectiveness	• Context-based understandability • Reliable and understandable reduces uncertainty • Effective responds in a suggestive manner • High quality responses through effective questions
		Fluency	• Coherence and Grammatical Correctness • Factors of Clarity	• Coherence and grammatical correctness • Information trustworthiness. • Data Understanding and Presentation • Simplicity in responses

Appendix F. Project Brief



DESIGN FOR our future

TU Delft

Personal Project Brief – IDE Master Graduation Project

Name student Ujjayan Ujjalbikash Dhar **Student number**

PROJECT TITLE, INTRODUCTION, PROBLEM DEFINITION and ASSIGNMENT
Complete all fields, keep information clear, specific and concise

Project title Enabling Domain Expert Evaluation of Emerging AI Technologies in Healthcare Settings

Please state the title of your graduation project (above). Keep the title compact and simple. Do not use abbreviations. The remainder of this document allows you to define and clarify your graduation project.

Introduction

Describe the context of your project here; What is the domain in which your project takes place? Who are the main stakeholders and what interests are at stake? Describe the opportunities (and limitations) in this domain to better serve the stakeholder interests. (max 250 words)

1) Context:
This project operates at the intersection of rapidly evolving AI technologies and the essential domain of healthcare. With the advancement of AI, its acceptance across a variety of industries, including healthcare, has grown as the technology improves over time at important healthcare activities like illness diagnosis (Davenport, T., 2019). However, the integration of AI into healthcare presents unique challenges due to the sector's strict requirements for trust, safety, and reliability.

2) Stakeholders:
In this domain, primary stakeholders include Healthcare professionals, AI developers, and Regulators, each with distinct values and concerns. Healthcare professionals aspire for AI tools that enhance patient care, ensuring accuracy and safety in diagnosis and treatment. AI developers aim to devise innovative solutions that effectively tackle the complexities of healthcare and provide close to reality responses. Meanwhile, regulators hold the responsibility of enforcing compliance with industry standards and safeguarding patient overall well-being.

3) Opportunities and Limitations :
Despite the promising outlook, there are several limitations and risks in the adoption of AI in healthcare. These include limited benchmarking data, expansion of the scope to support multiple languages, the lack of ability of the LLMs to effectively communicate uncertainty in information to users, and ethical concerns like bias and privacy (Singhal, K., 2023). However, there's an opportunity to address these challenges. By developing guidelines, we can improve communication among AI developers, medical experts, and regulators. These guidelines can help set clearer expectations and requirements, enabling better decision-making and would serve as design boundaries, providing a framework for decision-making and action, thus shaping the direction of AI integration in healthcare.

➔ *space available for images / figures on next page* *introduction (continued): space for images*

click to add picture

image / figure 1

image / figure 2

Personal Project Brief – IDE Master Graduation Project

Problem Definition

What problem do you want to solve in the context described in the introduction, and within the available time frame of 100 working days? (= Master Graduation Project of 30 EC). What opportunities do you see to create added value for the described stakeholders? Substantiate your choice.
(max 200 words)

PROBLEM DEFINITION:
In the context of healthcare, the integration of AI technologies presents many-sided challenges and opportunities. Although various AI models such as Med-PaLM, evaluated on benchmarks like MultiMedQA, hold promise in real-world clinical applications, their adoption encounters several hurdles (Singhal, K., 2023).Limited benchmarking data, language barriers, and the challenge of effectively communicating uncertainty pose significant limitations.We aim to use the concepts of imaginaries to understand how healthcare experts view AI and evaluate upcoming technologies with AI-like features, such as unpredictability, limited transparency, and fragile data. Through this exploration, we seek to develop a set of guidelines and framework that enable AI developers, medical experts, and regulators to better communicate expectations and requirements, facilitating informed decision-making.

Over the course of 100 days, this project will aims an exploratory investigation on the present complexity of AI-assisted chatbots in healthcare and explore potential to provide value for stakeholders. We hope to give insights that will influence the creation of effective guidelines and decision-making frameworks by researching health experts' opinions and attitudes toward AI technology, particularly in terms of future advancements.

Assignment

This is the most important part of the project brief because it will give a clear direction of what you are heading for. Formulate an assignment to yourself regarding what you expect to deliver as result at the end of your project. (1 sentence) As you graduate as an industrial design engineer, your assignment will start with a verb (Design/Investigate/Validate/Create), and you may use the green text format:

Investigate and design a comprehensive guideline to facilitate the effective integration of AI technologies in healthcare settings, enhancing patient care and optimizing medical processes for healthcare practitioners, AI developers, and regulatory authorities in the rapidly changing healthcare technology industry.

Then explain your project approach to carrying out your graduation project and what research and design methods you plan to use to generate your design solution (max 150 words)

The main goal of this research is to explore the integration of AI technologies in healthcare settings, focusing on understanding health expert perceptions and attitudes towards AI.Through this exploration, the project will provide insights on the challenges and opportunities associated with AI adoption in healthcare
Objectives:

1. Conducting a thorough literature review to establish a theoretical foundation, supplemented by health expert interviews to shape the methodology.
2. Planning and organizing human-centered research, including participant sampling and ethical considerations.
3. Executing qualitative user research, audio transcript coding to analyze the content and identifying key Themes and Sub-Themes.
4. Organizing and analyzing data and results, summarizing findings, and deriving human-centered design guidelines.
5. Designing a framework on the basis of the design guidelines which will act as a building block for future development in this domain.
6. Conduct testing with healthcare professionals to assess usability and effectiveness with the developed guideline.

Project planning and key moments

*To make visible how you plan to spend your time, you must make a planning for the full project. You are advised to use a Gantt chart format to show the different phases of your project, deliverables you have in mind, meetings and in-between deadlines. Keep in mind that all activities should fit within the given run time of 100 working days. Your planning should include a **kick-off meeting**, **mid-term evaluation meeting**, **green light meeting** and **graduation ceremony**. Please indicate periods of part-time activities and/or periods of not spending time on your graduation project, if any (for instance because of holidays or parallel course activities).*

Make sure to attach the full plan to this project brief.
The four key moment dates must be filled in below

Kick off meeting	08/03/2024
Mid-term evaluation	07/05/2024
Green light meeting	1/07/2024
Graduation ceremony	31/07/2024

In exceptional cases (part of) the Graduation Project may need to be scheduled part-time. Indicate here if such applies to your project

Part of project scheduled part-time	
For how many project weeks	
Number of project days per week	

Comments:

Motivation and personal ambitions

Explain why you wish to start this project, what competencies you want to prove or develop (e.g. competencies acquired in your MSc programme, electives, extra-curricular activities or other).

Optionally, describe whether you have some personal learning ambitions which you explicitly want to address in this project, on top of the learning objectives of the Graduation Project itself. You might think of e.g. acquiring in depth knowledge on a specific subject, broadening your competencies or experimenting with a specific tool or methodology. Personal learning ambitions are limited to a maximum number of five.
(200 words max)

My journey from computer science engineering to strategic product design has equipped me with a unique blend of technical expertise and design thinking. Throughout my career, I've been intrigued by the intersection of human-computer interaction and AI, inspired by leaders like Don Norman, Pattie Maes and Pranav Mistry. My tenure at Dassault Systemes and subsequent studies at TU Delft have nurtured my passion for understanding how AI impacts user experiences and strategic product design. Previously, I had the opportunity to work with IBM on a project that studied the value tensions that develop when using Trusted AI in the banking industry. After having done projects at Ford, Seamless Mobility Lab, and JIP (Joint Interdisciplinary Project) with Air-France KLM, where I explored Adaptive AI in simulation trainings, I became deeply interested in applying AI and UX principles in the impactful domain of healthcare.

In this project, beyond the core objectives, I aim to pursue several personal learning ambitions. Firstly, I aspire to deepen my understanding of the ethical implications of AI adoption in healthcare, exploring how design can mitigate biases and prioritize user well-being. Secondly, I seek to broaden my competencies in conducting qualitative research and analyzing data to derive actionable insights. Additionally, I aim to experiment with innovative methodologies for designing AI-driven solutions, fostering collaboration and understanding among stakeholders. Ultimately, I aspire to contribute not only to my personal growth but also to the broader discourse on the future of AI and UX in strategic product design.

Graduation project	Masters in Strategic Product Design (30 ECTS)
Title	Enabling Domain Expert Evaluation of Emerging AI Technologies in Healthcare Settings
Chair	Dr. Evangelos Niforatos
Mentor	Professor Shatha Degachi
Student	Ujjayan Ujjalbikash Dhar
Student ID	5791073

Month	February	February	March	March	March	March	April	April	April	April	May	May	May	May	May	June	June	June	June	July	July	July	July	July	August	August	August	August
Calendar Week	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Academic Week	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	3.10	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	4.10	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9
Project Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Define	■	■	■																									
Comittee Formation	■	■																										
Developing the Brief		■	■																									
Research Questions			■																									
Finalising the brief		■	■																									
Learning Basics of Machine Learning and AI	■	■	■																									
Collect and Analyse			■	■	■	■	■	■	■																			
Literature Review			■	■	■	■																						
Desktop Research			■	■																								
Sampling of Participants					■																							
Plan and Prepare the Questionnaires					■	■																						
Pilot testing						■																						
Interviews with healthcare officials							■																					
Analyse the interviews and coding							■	■	■																			
Design and Develop										■	■	■	■	■	■	■	■											
Data Collection										■																		
Coming up with Themes and Sub-Themes											■																	
Analysing insights												■																
Develop innovation strategies													■															
Develop design guidelines														■	■													
Develop framework															■	■	■											
Test and validate																		■	■									
Conduct testing and validating workshop																		■										
Green light meeting preparation																			■									
Reflect and Revise																				■	■							
Reflection on the project outcomes																				■								
Thinking on the limitation and future scope																				■	■							
Making final additions and changes																					■							
Deliver																					■	■	■	■				
Finalising the deliverables																					■	■						
Making the presentation																						■	■					
Organizing the presentation day																							■					
Publish and Present																								■				
Milestones																												
Kick-Off			■																									
Midterm												■																
Greenlight																				■								
Graduation Report				■	■	■	■						■		■		■		■		■	■	■					
Graduation Presentation																						■	■	■				