What Did the Doctor Say?
Empowering Patient Comprehension with Generative AI

Completed Research Paper

Gero Strobel, University Duisburg-Essen, Essen, Germany, gero.strobel@uni-due.de
Leonardo Banh, University Duisburg-Essen, Essen, Germany, leonardo.banh@uni-due.de

Abstract

As global challenges, such as pandemics, population growth and widespread illnesses, continue to rise, healthcare systems are facing greater strain, resulting in a shortage of resources and increased demands for medical care. Effective communication between healthcare professionals and patients is essential for the provision of good services to prevent confusion and induced anxiety of patients, particularly when medical jargon is employed and not understood. Generative AI (GAI) presents a chance to transform healthcare communication by providing language processing capabilities that enhance patient-centered services. This paper examines how GAI-based conversational agents for explaining medical jargon in healthcare should be designed. We derived eleven design principles from a systematic literature review and evaluated them with nine clinical cardiological scenarios through a prototypical instantiation of an LLM-based conversational agent. The results provide insights for researchers and healthcare providers in form of prescriptive design knowledge to improve patient communication using GAI.

Keywords: Generative AI, Conversational Agent, Healthcare, Design Science Research.

1 Introduction

While historically a fundamental human necessity, the perceived security of physical integrity has proven to be illusory in recent times (Kumar et al., 2020). The landscape of global events has unmasked the delusion of this security, revealing a reality burdened by pandemic crises, escalating population growth, an aging demographic, and a surge in widespread diseases (Baker et al., 2017; United Nations, 2017). This growing medical demand exerts an overwhelming strain on resources—be it infrastructure, skilled workforce, or essential medications—thus pushing the boundaries of conventional treatment modalities to an impending breaking point. Consequently, these escalating burdens have pushed the current healthcare system to its limits, particularly during times of growing demands and global crises (Kumar et al., 2020; Roy et al., 2020). The implications extend beyond inconveniences and lead to life-threatening implications even in highly developed nations. Furthermore, in middle to low-income countries, the situation worsens, leaving treatment inaccessible due to geographical distances, overwhelmed infrastructure, or its complete absence (Pathinarupothi et al., 2016; Roy et al., 2020). Even with access to medical resources, healthcare staff must carefully manage their time as their workloads increase (Portoghese et al., 2014). Effective communication between doctors and patients is an essential element in ensuring the provision of good medical services, avoiding misunderstandings, and strengthening patients' trust. Besides ongoing communication challenges, a good communication is further hindered by the use of medical jargon that most patients have difficulties to understand (Sevinc et al., 2005). When patients are not receiving the required clarification by the doctors, they are left with uncertainty, fear, and despair that could lead to unwanted consequences.

With the advent of generative AI (GAI), novel possibilities emerge that leverage advanced natural language processing and a vast amount of information depth to provide human-centered services for good (Sai et al., 2024; Strobel et al., 2024). Designing systems that support patients in their recovering

process by explaining medical results and complex facts understandably and providing an asynchronous, always available channel for further questions without human judgement can relieve healthcare providers and improve the communication quality because only the unresolved questions remain to be discussed. Although research activities have focused on generative AI as a disrupting phenomenon (Susarla et al., 2023; Thirunavukarasu et al., 2023) as well as on the development of better GAI models (e.g., Wang et al., 2023), current works have not yet investigated the integration of GAI into conversational agents that are targeted for healthcare patients to answer their medical questions, regardless its promising medical capabilities (Peng et al., 2023; Singhal et al., 2023). Against this backdrop, this paper investigates GAI-based conversational agents for healthcare from a design perspective to support researchers and practitioners in enabling better healthcare offerings to patients. We, therefore, ask the following research question:

How should generative AI-based conversational agents be designed for explaining medical jargon in healthcare?

To answer our question, we conduct a design science research project and derive design principles (DP) from an extensive systematic literature review. By instantiating the DPs in a prototypical conversational agent, we can assess its performance based on machine evaluation. The remainder of this paper is structured as follows: First, we outline the theoretical background of GAI and present related literature for its potential in healthcare. Second, we elaborate on our research methodology, i.e., how we apply the design science research methodology (DSR) by conducting a systematic literature review (SLR) and developing a prototype to derive design knowledge. Third, we present our findings in form of design principles and outline the evaluation results of our GAI-based prototype. Last, we conclude our work, state limitations, and give an outlook on future research prospects.

2 Generative AI in Healthcare

Discriminative AI-based systems are widely associated within the healthcare domain, tackling diverse tasks encompassing pattern recognition in both medical and organizational applications (e.g., disease detection, drug selection, documentation, etc.) (Davenport and Kalakota, 2019; Yu et al., 2018). Recent technological advancements have not only introduced an entirely new class of artificial intelligence with generative AI but have also extended the classical domain of classification and prediction towards novel generative tasks that are indistinguishable from human-generated outcomes (Strobel et al., 2024).

Deep generative models (DGMs), rooted in artificial neural networks, underpin the current landscape of GAI systems, and showcase architectural paradigms such as generative adversarial networks (GANs) or generative pre-trained transformers (GPTs). Instead of focusing solely on processing data to determine decision boundaries (e.g., classifying images or predicting revenues), GAI models delve into probabilistic data generation, enabling a wide array of applications (Jebara, 2004; Weisz et al., 2023). Employing statistical methodologies, DGMs are trained to comprehend high-dimensional probability distributions by using extensive training datasets, generating novel samples that closely emulate the underlying class of the original training data (Tomczak, 2022). For example, large language models (LLMs) are trained on expansive text corpora, thereby capable of generating context-specific texts by predicting the most probable token (i.e., text segment-like characters or words) to follow the prior tokens in a sentence (Schramowski et al., 2022). Furthermore, the multi-modal capabilities of various GAI models extend beyond text generation to include images, audio, or even complex data types like proteins (Hie et al., 2023; Strobel et al., 2024). Equipped with novel capabilities and user-friendly interaction paradigms (e.g., natural language prompting for instructions and engagement), GAI applications facilitate opportunities to augment and automate traditionally challenging processes (Banh and Strobel, 2023; Schmidt et al., 2023).

These advancements encompass human-like reasoning and empathetic interactions (Pelau et al., 2021), which are crucial aspects within the healthcare domain. Therefore, several research articles have commented on the use of GAI in healthcare (Clusmann et al., 2023; Peng et al., 2023; Varghese and Chapiro, 2023). The focus, however, often lies on the challenges of implementing GAI into the highly regulated domain of healthcare and medicine, with proposals for regulators to approach a safe GAI

application (Jindal et al., 2024; Meskó and Topol, 2023; Reddy, 2024). Only a few works deal with GAI use for accessing medical services (Peng et al., 2023; Sai et al., 2024; Varghese and Chapiro, 2023). Thus, GAI, serving as the technological foundation for conversational agents (CA), holds significant promise within healthcare, supporting patients and caregivers along their recovery journey. It bridges the gap in scenarios where human expertise is limited, aiding in answering queries and processing information where human resources are scarce or insufficient. Generating prescriptive knowledge for how to design GAI-based CAs can accelerate researchers and practitioners in providing better healthcare offerings to patients.

3 Research Design

The central aim of this research is to derive scientifically substantiated and evaluated design knowledge for the development and utilization of GAI-based CAs in the healthcare domain, considering the existing knowledge base. Within the research process, we adapt the Design Science Research methodology as outlined by Peffers et al. (2007) (see Figure 1).

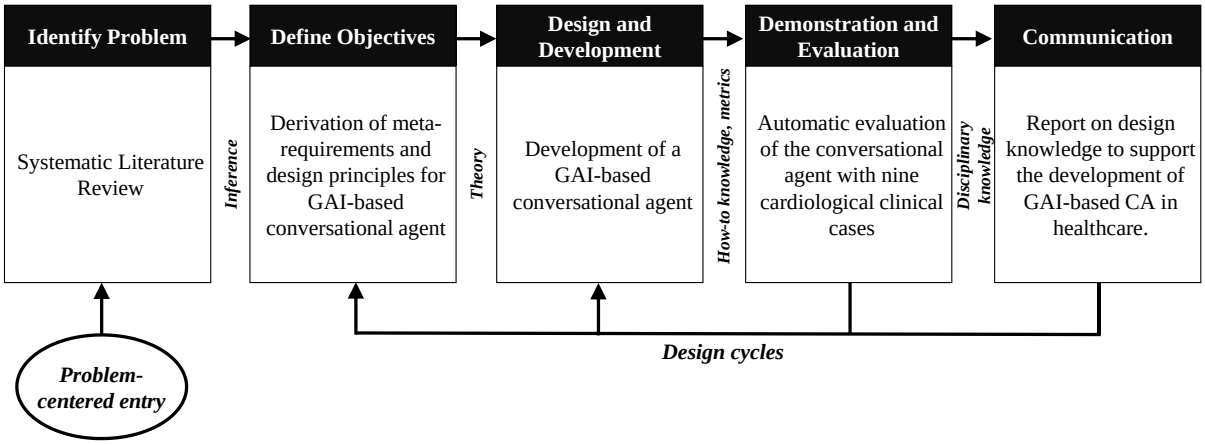

Figure 1. Research process adopted by Peffers et al. (2007).

The starting point of this research is the **identification of the problem** and the associated exploration of the problem space, as guided by a SLR following the approach of Watson and Webster (2020). Based on this assumption, the search string *"(Health AND (Chatbot OR Conversational Agent))"* is used for seven known databases (e.g., AISeL, IEEE, PubMed, etc.) in the search fields of "title", "abstract", and "keywords". To explore the broadest and most diversified range of potential solutions, the systematic literature search was conducted cross-sectionally, encompassing both papers from the field of information systems and related domains such as healthcare, computer science, and psychology.

A total of 4,801 publications were initially identified within the databases. However, after screening based on title, abstract, keywords, and removing duplicates, only 85 publications were deemed relevant (see Table 1). To maintain a high level of quality, additional exclusion criteria were applied during the full-text screening, beyond the initial search string. Consequently, only publications that align with a novel perspective on GAI or CA and their application in the specified domain were considered. Publications with a sole technical focus lacking a systemic perspective were excluded. Given the evolving nature of artificial intelligence, particularly in the generative aspect, and the goal of the literature search to encompass a comprehensive range of requirements and existing literature, a forward and backward search was performed based on the final sample. This led to the identification of two more publications, resulting in a total of 44 publications for the final literature sample. The publications of the final literature corpus were iteratively open-coded in the context of goal definition to create design requirements as the basis for forming meta-requirements and design principles.

Databases	Initial	Title/Abstract incl. Duplicates	Full-Paper Screened	Final Sample
AISeL	45	14	12	10
ACM	819	25	5	3
PubMed	479	34	7	5
IEEE	984	69	5	2
ScienceDirect	395	28	3	1
Scopus	1500	148	32	13
Web of Science	579	63	21	8
Sum	**4801**	**381**	**85**	**42**
			Forward Search	0
			Backward Search	2
			Sum	**44**

Table 1. Literature review process.

Within the first step of the iterative coding process, 165 design requirements were identified across the 44 publications. As design principles, by definition, encapsulate the formulation of design knowledge (Chandra et al., 2015), addressing not just an instance of an artifact but the artifact class, we correspondingly elevated the set of design requirements to a higher order of meta-requirements (MRs) (Walls et al., 1992). The development of design principles based on meta-requirements ensures value grounding, signifying that no design principles exist without fulfilling at least one requirement (Goldkuhl, 2004). In this regard, duplicates and all irrelevant design requirements were initially removed, and the remaining requirements were then axial coded to distil the most relevant requirements of the artifact class into meta-requirements (Thoring et al., 2020). Based on this logical content aggregation (Koppenhagen et al., 2012), three meta-requirements were derived: **faithfulness** (i.e. the exactness and depth of generated answers), **human-centricity** (i.e., the factors for humans to efficiently use the CA), and **adaptiveness** (i.e., the customizability to personalize output for an individual patient). The derived meta-requirements serve as guidelines for the final selective coding step to derive eleven design principles. Various templates for formulating design principles can be found in the literature (e.g., Goldkuhl, 2004; van Aken, 2004). Within this publication, we adhere to the approach outlined by Chandra et al. (2015) both structurally and linguistically. A complete alignment of the literature corpus, meta-requirements, and design principles can be found in the online appendix (https://bit.ly/47Fhms3).

For evaluating the meta-requirements and design principles derived from the literature, they were instantiated in the form of a prototype during **the design and development phase**. The foundation of the prototype is based on the Retrieval-Augmented Generation Architecture (RAG), which allows the combination of LLMs with additional information sources (i.e., retrievers) (see Figure 2). While LLMs possess the ability to store vast amounts of knowledge, they marginally address knowledge in a pinpoint and accurate manner, crucial for knowledge-based activities such as communicating specialized content in healthcare. The RAG architecture enables the integration of parametric memory, a pre-trained seq2seq model, with non-parametric memory, facilitating the dynamic provision of pertinent knowledge (Lewis et al., 2020). This empowers the LLM to generate technically sound responses, reducing the degree of hallucinations. Moreover, it allows for tracing the information sources, forming the basis for the agent's responses. However, the increased information density and transparency provided by this architectural approach has negative implications for performance due to the extended search process in external data sources and the processing time required by the LLM.

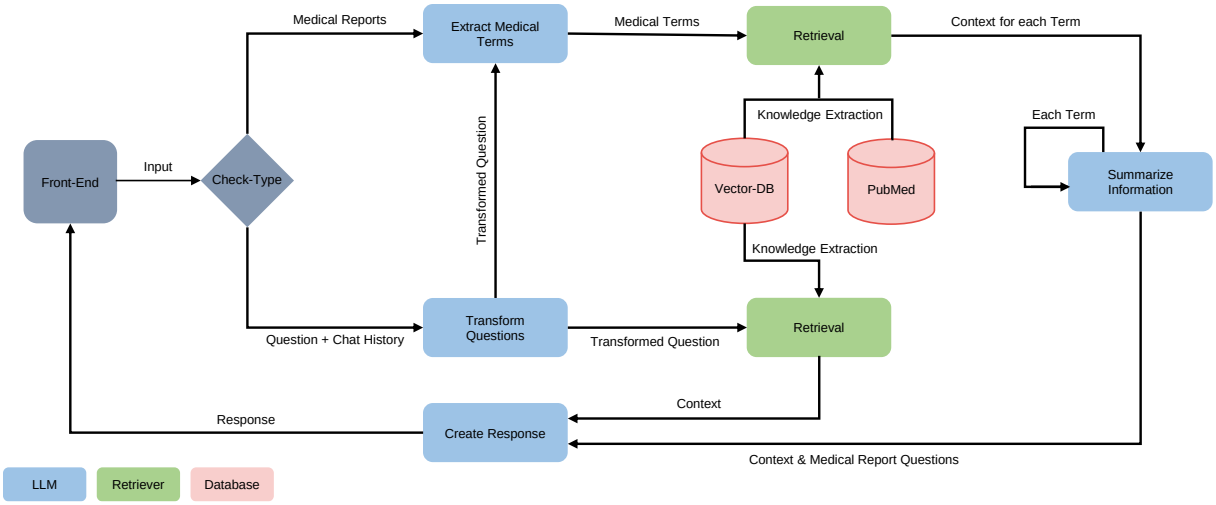

Figure 2. Instantiation of the prototypical architecture.

Within the prototype framework, the seq2seq GAI model LLaMA2, comprising 70 billion parameters, functions as the LLM, while a vector database containing medical literature from cardiology, as well as an interface to the PubMed literature database serves as the retriever. Through a web interface, users can engage in conversations with the agent on medical queries, either through simple questions or by providing a medical report for the CA to base the conversation on. In both scenarios, relevant medical keywords are extracted and forwarded to the retriever. The retriever then utilizes these keywords as a basis for extracting knowledge from external data sources, providing pertinent information to the LLM. The LLM formulates responses to the user's queries based on this information.

4 Design Principles for GAI-based Assistants in Healthcare

This section reports on the design principles for GAI-based conversational agents in healthcare that we developed based on a systematic literature review (see Figure 3).

Meta-Requirement	Design Principle	Description
MR1: Faithfulness	Information Superiority	DP1: Provide the latest medical, regularly updated knowledge to enable the best specialized communication.
	Explainability	DP2: Provide sources and other explainable methods to support comprehension of the information origins.
	Boundary Disclosure	DP3: Provide functional and technical limitations to disclose the capability boundaries.
MR2: Human-Centricity	Anthropomorphism	DP4: Provide adaption mechanisms to facilitate human-like interaction.
	Information Symmetry	DP5: Provide information about medical knowledge in plain language to foster easy understanding by everyone.
	Multimodality	DP6: Provide multimodal communication channels to enable rich, accessible, and natural communication for all users.
	Patient Centricity	DP7: Provide a personalized communication based on the user's individual life situation and medical history to give most relevant and useful results.
	Modulation	DP8: Provide mechanisms that enable a personalized communication based on user needs and education to enable comprehensive and accessible answers embedded in a patient context.
MR3: Adaptiveness	Error-Proneness	DP9: Provide robustness against user errors to ensure an efficient and effective user experience.
	Intuition	DP10: Provide reinforcement capabilities to continuously adapt to the user behavior, learn and improve from feedback, and generate personalized answers.
	Escalation	DP11: Provide education, further information, and emergency options when professional assistance is needed to ensure that users get the correct medical support from human doctors and healthcare providers

Figure 3. Design principles for GAI-based conversational agents.

DP1 – Information Superiority: *Provide the latest medical, regularly updated knowledge to enable the best specialized communication*. Users must be able to trust that the information provided by the assistant is correct, valid, and comes from reputable sources (Sallam, 2023). Therefore, regular updating knowledge from validated external sources is crucial in contributing to the accuracy of responses during user dialogues and optimizes the overall consultation quality (Al-Nazer and Helmy, 2012; Rahman Khilji et al., 2020). This dynamic process not only supports in answering queries but also plays a vital role in advancing healthcare (Sheth et al., 2019). Implementing information superiority effectively in GAI-based systems requires both the technical connection to validated, external data sources and the implementation of an adaptive decision-making process to determine from which data sources information is obtained depending on the situation (Janssen et al., 2021; Thimmanayakanapalya et al., 2022). These data sources serve as a basic building block for retrieving information and verifying the assistant's medical responses (Dhinagaran et al., 2022). Moreover, the trustworthiness of health-related information provided by conversational agents is paramount (Sallam, 2023). Considering that users, particularly seniors, prioritize information quality over empathic capabilities, credible sources are crucial in impacting the perceived professionalism of CAs, thus increasing trustworthiness (Mesbah and Pumplun, 2020; Moilanen et al., 2022; Moilanen et al., 2023; You et al., 2023). To ensure the reliability of the information, the evidence-based approach draws from extensive literature reviews, clinical guidelines, and trustworthy organizations such as the world health organization, the centers for disease control and prevention (CDC) in America, or the national health service (NHS) in the United Kingdom (Denecke, 2023; Dhinagaran et al., 2022; Siangchin and Samanchuen, 2019).

DP2 – Explainability: *Provide sources and other explainable methods to support comprehension of the information origins.* Understanding algorithmic systems and AI-based outcomes are important factors for users to successfully use and accept CAs (Abdulrahman and Richards, 2019). This is amplified in medical contexts where patients seek comprehensible and explainable solutions to mitigate missing transparency of systems and unclear information (Mozafari et al., 2021; Sallam, 2023). GAI-based assistants in healthcare should be able to showcase the sources of information and their reasoning behind generating certain answers (Abdulrahman and Richards, 2019). A lack of transparency towards users reduces the level of trust, thus leading to doubts about the system reliability and ultimately a reduction in acceptance of CAs (Benbasat and Wang, 2005; Janssen et al., 2021; Su et al., 2020). In the context of CAs, transparency means revealing to users that they are interacting with a GAI-based chatbot. It is important that users understand that they are not communicating with a human doctor, but with an AI model trained on prior knowledge and data. By disclosing this information, users can better assess what kind of support and information they can expect from the app (Abdulrahman and Richards, 2021). This can be facilitated by presenting the information about algorithmic processing in a clear and simple manner, e.g., by implementing a disclaimer upon starting the application (Siemon et al., 2022; Zhang et al., 2020). This offers the opportunity to increase trust in the assistant as users are empowered to individually make informed decisions (Lai et al., 2023; Zhang et al., 2021). Moreover, transparency regarding information collection is of paramount importance, meaning that users should be informed of the data types being collected, their intended use, and how they are safeguarded (Ahmad et al., 2022; Parmar et al., 2022). Especially in healthcare with medical data, highly sensitive data needs particular data protection and security mechanisms to ensure confidentiality and integrity, thus ultimately mitigating stigmatization and discrimination (Laumer et al., 2019; Xu et al., 2021). To achieve this goal, the assistant must provide privacy policies that clearly state the data collected and its purposes in simple language (Dhinagaran et al., 2022; Polignano et al., 2020). Additionally, technical measures such as local storage and processing as well as end-to-end encryption can help ensuring users that data is not maintained centrally and data is not shared unintentionally, fostering data sovereignty (Siemon et al., 2022). Furthermore, a welcome dialog can be made available to provide users with an overview of the information collected about them (Boucher et al., 2021). This promotes user satisfaction, trust, and contributes to their perception of the app as a useful tool for simplifying access to medical information (Curtis et al., 2021; Zhang et al., 2020).

DP3 – Boundary Disclosure: *Provide functional and technical limitations to disclose the capability boundaries.* The use of LLM-based CAs in a medical environment can lead to incorrect or even false

statements, causing harm to patients (Schachner et al., 2020). Indicators such as training data, bias, and alignment can contribute to these inaccuracies and even hallucinations (Banh and Strobel, 2023). Patients often overestimate the capabilities of these CAs and accept their recommendations without question. To address this issue, it is important to disclose the limitations of the assistant and implement alignment mechanisms that allow it to recognize when the CA cannot provide a correct answer (Sharma et al., 2022; Sweeney et al., 2021). These limitations should be communicated to the user to reduce the probability of incorrect recommendations and minimize potential harm (Mozafari et al., 2021). Additionally, users should be periodically informed about the capabilities and limitations of the assistant before using it (Sweeney et al., 2021). By taking these precautions, the risk of harmful recommendations can be reduced, and patient safety can be improved.

DP4 – Anthropomorphism: *Provide adaption mechanisms to facilitate human-like interaction.* In personalized environments like healthcare, adopting a human-like interaction style is crucial to increase users' willingness to interact with a CA, foster a natural and intuitive connection, and enhance trust in provided resources and information (Moilanen et al., 2023; Shah et al., 2022). By emulating human interactions, the user experience becomes more natural and pleasant, leading to increased satisfaction and a sense of better understanding (Schuetzler et al., 2020; Sharma et al., 2022; Su et al., 2020). Imitating human behavior aids communication and fosters user trust in the CA, e.g., by personalizing the CA with a name and profile picture to contribute to users feeling addressed and perceiving the CA as an individual entity (Dhinagaran et al., 2022; Moussawi et al., 2021). However, maintaining a neutral, factual language style is essential to establish a professional relationship between the user and the assistant (O' Connor et al., 2021; Shan et al., 2022).

DP5 – Information Symmetry: *Provide information about medical knowledge in plain language to foster easy understanding by everyone*. Addressing the information asymmetry between medical experts and patients or caregivers is an essential challenge in healthcare provision and extends to CAs that are developed to support the users (Müller et al., 2019; Sharma et al., 2022). The CA should offer precise and comprehensible information, determining the appropriate level of detail during communication and validating it through continuous user feedback loops (Lai et al., 2023; Sheth et al., 2019). Beyond a horizontal knowledge base, the assistant should also be able to provide detailed and understandable explanations for technical terms and recommendations (Holzinger et al., 2017). Simplicity and clarity are vital in conveying information, especially for older users, enhancing their willingness to engage with the assistant. The goal is to simplify medical texts for users of all education levels, avoiding technical terms unless desired (Denecke, 2023; Mesbah and Pumplun, 2020; Nguyen et al., 2021). Emphasizing readability and user understanding ensures that the CA is accessible to all users and applicable across knowledge levels (Dhinagaran et al., 2022; Moilanen et al., 2022). Ultimately, the focus should be on conveying necessary medical information in an understandable form to assist users without generating unnecessary questions, contributing to a reduction in the understanding deficit (Denecke, 2023).

DP6 – Multimodality: *Provide multimodal communication channels to enable rich, accessible, and natural communication for all users*. With advancements in the development of GAI models, multimodal capabilities are introduced that allow the processing of multiple data types, e.g., text, images or audio (Banh and Strobel, 2023). This enables GAI-based CAs to ensure dynamic responsiveness by accommodating various input methods that are crucial in a medical context, for instance, where imaging data (e.g., x-ray images) plays a central role to the patient information and medical history (Janssen, 2020; Sheth et al., 2019). Multimodal capabilities, surpassing text-based communication, offer advantages in information acquiring, processing, and presentation, thus enhancing user-friendliness (Abdulrahman and Richards, 2021; Scholten et al., 2019). By eliminating barriers with natural language interactions, vulnerable groups like the elderly or impaired individuals who may struggle with long, complex text inputs are enabled in using the CA, hence overall accessibility is improved (Bharti et al., 2020; Sharma et al., 2022). Voice input is particularly essential in healthcare, enhancing the user experience for impaired patients who might have difficulties typing on small smartphone keyboards (Baldauf et al., 2018). Integrating voice interaction often involves text-to-speech and automatic speech recognition systems, similar to popular commercial voice assistants so users might already be familiar with the interaction style (Motger et al., 2023). Besides interaction paradigms, technical considerations

include the integration of multilingual support or sentiment analysis that are in scope of GAI's possibilities (Perez-Soler et al., 2021; Shum et al., 2018; Siu, 2023).

DP7 – Patient Centricity: *Provide a personalized communication based on the user's individual life situation and medical history to give most relevant and useful results.* To ensure patient-centric communication and assessment, fusing existing patient data with medical expertise is one essential requirement (Kocielnik and Hsieh, 2018; Stegemann et al., 2023). Enriching this information with context-sensitive external parameters, such as regional specificity or latest medical findings, allows the formation of a holistic patient profile (Sheth et al., 2019). The goal is to offer context-specific recommendations based on the individual patient information as well as historical data from previous interactions to facilitate a continuous improvement and a thorough understanding of the patient (Reddy et al., 2020; Reis et al., 2020; Su et al., 2020). Therefore, the CA should adeptly respond to user-provided information, incorporating details like previous illnesses or symptoms to maximize the correctness of outputs (Prayitno et al., 2021). Users benefit from contextual and detailed responses in which the CA references past statements for extended conversations. Effective responses to user queries with user-specific relevant information contribute to a patient-centric approach (Boucher et al., 2021; Nguyen et al., 2021). By integrating (return) questions, CAs can extract valuable user information like age or medical history to enhance the conversation depth and context available to generate answers.

DP8 – Modulation: *Provide mechanisms that enable a personalized communication based on user needs and education to enable comprehensive and accessible answers embedded in a patient context*. Effective communication between the CA and the user relies on adapting language to the user's proficiency and knowledge (Bharti et al., 2020; Sokolaj et al., 2023). For instance, technical terms should be presented in universally understandable language or accompanied by explanations (Al-Nazer and Helmy, 2012). By personalizing the communication style that origins from past interactions' data or direct user feedback, user motivation can increase and the assistant's reusability enhanced (Paul et al., 2021; Thimmanayakanapalya et al., 2022). The CA should tailor language to the various users' demographics, considering cultural and age-related aspects for improved engagement and acceptance (Reis et al., 2020). Preferences for communication styles vary among age groups, with young users potentially favoring a more informal approach (e.g., using slang and emojis), while older users lean towards more factual and dialog-oriented communication (Dosovitsky and Bunge, 2023). Personalizing conversations based on demographic, social, educational, and cultural backgrounds enhances effectiveness and engagement (Polignano et al., 2020; Zhang et al., 2020). Moreover, analyzing a current conversation to consider the emotional state, personality, or cultural sensitivity further contributes to user trust and helps designing a more effective dialogue (Kocaballi et al., 2020). Conversing in the native language reduces language-related errors and increases user engagement, adding a familiar touch (Dhinagaran et al., 2022; Nguyen et al., 2021). To prevent monotonous conversations, the CA should use encouraging, friendly, polite, and slightly humorous language, considering users' preferences for formal or informal formulations (Dhinagaran et al., 2022; Moilanen et al., 2022). Recognizing and responding to emotions positively impacts the dialogue, fostering user willingness to share information and enhancing the overall conversation experience (Nadarzynski et al., 2019; Zhang et al., 2020). Thus, enabling the modulation of a CA by integrating emotion-based communication offers potential for improving the dialogue quality and aiding users in conveying information more understandably.

DP9 – Error-Proneness: *Provide robustness against user errors to ensure an efficient and effective user experience*. Errors in CAs can restrict user options and impair the conversation, thus requiring errors to be either prevented or easily recoverable (Denecke, 2023). For instance, spelling and grammatical errors made by users should be automatically corrected or completely ignored by the CA. Establishing error-proneness ensures that the conversation remains unaffected. The goal of the CA is to provide meaningful answers to user questions. Hence, it requires the CA to clarify certain questions when needed to obtain a correct answer. If a prompt is incorrect or the provided information is not enough, the CA should point it out to the user before continuing the conversation. It is also crucial to provide factually correct answers to avoid hallucinations (see DP3). Poor input can lead to inadequate answers and the CA mimicking falsehoods. Therefore, user should be guided during input prompting to facilitate efficient communication with the CA (Au Yeung et al., 2023).

DP10 – Intuition: *Provide reinforcement capabilities to continuously adapt to the user behavior, learn and improve from feedback, and generate personalized answers*. As the use with the CA increases, users benefit from more personalized communication that stems from previous behaviors, their language, and preferences (Bharti et al., 2020; Schlimbach et al., 2023). Special attention is directed towards users in vulnerable situations, ensuring accessibility and inclusivity for individuals with disabilities or those not fluent in the CA's language (Sasseville et al., 2022). The goal achieving intuition is to foster a closer connection to the user, reducing inhibitions and increasing motivation for use. To evaluate the efficiency of the CA's responses, a feedback mechanism from the user is needed (Das et al., 2022). By eliciting user feedback, insights can support a learning process, resulting in recognizing potential errors, improving for future interactions, and ensuring user satisfaction (Ayanouz et al., 2020; Shan et al., 2022). Requesting feedback at the end of a chat is also effective for gathering user evaluations, impressions, and insights, contributing to the system's continuous adaptation (Shah et al., 2022). Users are given the opportunity to share their thoughts, suggestions, or concerns, fostering an interactive and collaborative feedback environment. This feedback loop enables customization and improvement of the CA to user needs and preferences, ultimately elevating the quality of provided healthcare services.

DP11 – Escalation: *Provide education, further information, and emergency options when professional assistance is needed to ensure that users get the correct medical support from human doctors and healthcare providers*. Educating the user about professional help is an important design principle to ensure that people with specific needs, especially in critical situations, receive the appropriate support (Nayar et al., 2022). The CA should be able to provide helpful information so that the user can make informed decisions about managing their condition. It is also necessary to identify resources to assist the user in seeking professional help. Once the user is aware of the resources available, the CA should provide additional support and encouragement. It is particularly important to provide empathetic and guiding responses to at-risk patients, such as those with depression or suicidal thoughts (Kocaballi et al., 2020; Park and Lee, 2020). Additionally, implementing an emergency detector that can escalate the conversation by providing information about professional help can prevent users from self-harm or critical danger if responses might indicate such behavior (Anjum et al., 2023; Rathnayaka et al., 2022).

5 Evaluation

To validate the developed design principles, we instantiated them into a prototypical web application by developing a conversational agent that allows users to leverage GAI and LLMs for understanding medical texts. Specifically, the prototype provides a tangible evaluation basis and baseline for the formulation of design knowledge for the entire artifact class (Peffers et al., 2007).

There are different approaches to evaluate conversational agents, with manual and automatic methods requiring more or less human effort in testing (Meier et al., 2019). Manual evaluation includes experimental user studies that requires the gathering of experts (e.g., regarding output quality) and end-users (e.g., regarding usability) to collect qualitative and quantitative feedback (Denecke et al., 2021; Kowatsch et al., 2017). Automatic evaluation, on the other hand, relies on machine-based feedback and has created a new wave of attention with current capabilities of GAI models like GPT-4 and LLaMA (Lin and Chen; Liu et al., 2023). LLMs show promising capabilities to evaluate data on a similar level to humans, e.g., by generating test sets with question-answer pairs or by following instructions to rate texts based on certain criteria (Liu et al., 2023). In this work, we focus on the latter approach and used GPT-4 as an evaluation tool to assess the quality of our prototype because of its streamlined automation potential and its availability in comparison to medical experts.

Regarding the context of our evaluation, we developed three scenarios that revolve around the prototype domain cardiology. We opted for stroke, cardiac arrest, and cardiac arrhythmia as common heart diseases and derived three case studies for each disease from medical literature (see examples in Table 2). Each scenario was prompted to the CA as an input and subsequent follow-up questions were asked (see online appendix for more details: https://bit.ly/47Fhms3). The generated answers of our prototypical CA were then automatically evaluated with the GPT-4 LLM. We divided the evaluation into three parts to assess the quality of the retriever (i.e., how accurate are the results), the quality of the

summary (i.e., how relevant are the results), and the quality of the explanation (i.e., how understandable are the generated answers). Further details are found in the online appendix.

Scenario	Description	Reference
Stroke	**#1**: An 82-year-old woman with coagulopathy and a history of anticoagulant therapy presented with severe headache, vomiting, and third cranial nerve palsy, leading to a diagnosis of pituitary apoplexy. Due to her unsuitability for surgery, conservative treatment was chosen, and she made a full recovery with outpatient follow-up.	Drissi Oudghiri et al. (2021)
Cardiac Arrest	**#6**: A 51-year-old man with exertional dyspnea was diagnosed with mitral regurgitation and had tortuous coronary arteries, which can cause myocardial ischemia and infarction. Surgery and bypass were performed, and he was symptom-free at the follow-up.	Xing et al. (2017)
Cardiac Arrhythmia	**#9**: A 41-year-old woman with CMT, a group of inherited diseases that affect the peripheral nervous system, experienced ventricular fibrillation after taking prescribed sumatriptan, suggesting a potential association between triptans and arrhythmias in people with degenerative neuropathies.	Rubinstein et al. (2004)

Table 2. Evaluation scenarios to assess the digital health CA.

Overall, our system achieved performance scores that were rated on average across all scenarios with 75.11 % regarding the retriever quality, 87.33 % in summary quality and 90.44 % in explanation quality. Noticeably, several results were scored with 9/10 instead of a perfect score, although only positive feedback was noted. Reasons for bad scores, according to the LLM, included that too much information was left out or missing data was made up. The lower quality of the retriever could be due to the inclusion of irrelevant information, such as limitations in the vector database or insufficiently informative PubMed abstracts. Because our CA summarizes the retriever results in a next step, irrelevant information is filtered out and the evaluation score increases. Finally, an answer is generated by fusing the original contextual information with the summarized information, improving the overall explanatory power for the user in a comprehensible way.

6 Conclusion, Limitations, and Outlook

AI has become an essential part of our lives, and we can no longer imagine life without it. Recent technological advancements have resulted in generative AI, a completely new class of AI that can create data almost indistinguishable from human-produced content. Particularly in the healthcare sector, this ability, combined with user-friendly interaction options, creates a wide range of potential applications for supporting patients and their relatives in their medical treatment. Despite its promising prospects, GAI has limitations, particularly in high-risk, personal domains like healthcare. The ability to imitate human communication almost perfectly can quickly create a basis of trust that is not justified. Therefore, the objective of this paper is to develop design knowledge for the development of GAI-based conversational agents in healthcare.

Despite the methodological grounding through the use of extant literature and an evaluation with nine medical scenarios, our insights are not without **limitations**. **First**, our research focuses on cardiology as a specialized medical domain because of its tangibility, relevance, and availability of scientific reports. Future research could expand the domain in focus and integrate further medical specialties or domains outside healthcare to gain a broader and more generalized insight into GAI-based CAs. **Second**, the instantiated prototype only has access to a limited amount of medical data, as we use the general-purpose GAI model LLaMA2 in combination with abstracts from PubMed and selected specialist literature on cardiology. Although the results show promising answers, the quality could be improved by incorporating a larger and more refined information sources pool. We suggest following the training curriculum of medical students and including its literature that should provide a broad and extensive source of information for retrieving knowledge. **Last**, our work employs an automatic evaluation method to measure the CAs performance. We believe that machine-based evaluation approaches can

contribute with objective criteria for assessing the results but propose to complement the evaluation with manual methods and human opinions to gain deeper understandings. By considering feedback from end-users, **future research** can examine usability factors and focus on the comprehensibility of the answers as well as the interaction behavior of the CA. By including medical experts, the domain-specific content (e.g., correctness and transparency of answers) can be further evaluated by a deeper examination of our derived design principles and the prototype. Semi-structured interviews and experiment studies might provide suitable methodologies for extended data collection and analysis in that regard. Closing with this year's conference theme, we hope that the provided design knowledge will serve as a starting point to support researchers and practitioners in "putting people first" and leveraging GAI for conversational agents, thereby contributing to the digital transformation of the healthcare industry.

References

Abdulrahman, A. and Richards, D. (2019). "Modelling Therapeutic Alliance Using a User-Aware Explainable Embodied Conversational Agent to Promote Treatment Adherence," in: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, New York, NY, USA: ACM, pp. 248–251.

Abdulrahman, A. and Richards, D. (2021). "In search of embodied conversational and explainable agents for health behaviour change and adherence," *Multimodal Technologies and Interaction* 5 (9).

Ahmad, R., Siemon, D., Gnewuch, U. and Robra-Bissantz, S. (2022). "Designing Personality-Adaptive Conversational Agents for Mental Health Care," *Information Systems Frontiers* 24 (3), 923–943.

Al-Nazer, A. and Helmy, T. (2012). "Toward a Cross-Cultural and Cross-Language Multi-Agent Recommendation Model for Food and Nutrition," in: *Proceedings of the the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03*, USA: IEEE Computer Society, pp. 245–249.

Anjum, K., Sameer, M. and Kumar, S. (2023). "AI Enabled NLP based Text to Text Medical Chatbot," in: *2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, pp. 1–5.

Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R. J. and Teo, J. T. (2023). "AI chatbots not yet ready for clinical use," *Frontiers in digital health* 5.

Ayanouz, S., Abdelhakim, B. A. and Benhmed, M. (2020). "A Smart Chatbot Architecture Based NLP and Machine Learning for Health Care Assistance," in: *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, New York, NY, USA: ACM.

Baker, S. B., Xiang, W. and Atkinson, I. (2017). "Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities," *IEEE Access* 5, 26521–26544.

Baldauf, M., Bösch, R., Frei, C., Hautle, F. and Jenny, M. (2018). "Exploring requirements and opportunities of conversational user interfaces for the cognitively impaired," in: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, New York, NY, USA: ACM, pp. 119–126.

Banh, L. and Strobel, G. (2023). "Generative artificial intelligence," *Electronic Markets* 33 (63).

Benbasat, I. and Wang, W. (2005). "Trust In and Adoption of Online Recommendation Agents," *Journal of the Association for Information Systems* 6 (3), 72–101.

Bharti, U., Bajaj, D., Batra, H., Lalit, S. and Gangwani, A. (2020). "Medbot: Conversational Artificial Intelligence Powered Chatbot for Delivering Tele-Health after COVID-19," in: *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 870–875.

Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C. and Zilca, R. (2021). "Artificially intelligent chatbots in digital mental health interventions: a review," *Expert review of medical devices* 18 (1), 37–49.

Chandra, L., Seidel, S. and Gregor, S. (2015). "Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions," in: *2015 48th Hawaii International Conference on System Sciences*: IEEE, pp. 4039–4048.

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J.-N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G. P., Wagner, S. J. and Kather, J. N. (2023). "The future landscape of large language models in medicine," *Communications medicine* 3 (1), 141.

Curtis, R. G., Bartel, B., Ferguson, T., Blake, H. T., Northcott, C., Virgara, R. and Maher, C. A. (2021). "Improving User Experience of Virtual Health Assistants: Scoping Review," *Journal of Medical Internet Research* 23 (12), e31737.

Das, A., Sen, V. and Rose, A. C. (2022). "Developing a chatbot/intelligent system for neurological diagnosis and management,". In Pillai, A. S. and Menon, B. (eds.) *Augmenting Neurological Disorder Prediction and Rehabilitation Using Artificial Intelligence*, pp. 273–291: Academic Press.

Davenport, T. and Kalakota, R. (2019). "The potential for artificial intelligence in healthcare," *Future healthcare journal* 6 (2), 94–98.

Denecke, K. (2023). "Framework for Guiding the Development of High-Quality Conversational Agents in Healthcare," *Healthcare (Basel, Switzerland)* 11 (8).

Denecke, K., Abd-Alrazaq, A., Househ, M. and Warren, J. (2021). "Evaluation Metrics for Health Chatbots: A Delphi Study," *Methods of information in medicine* 60 (5-06), 171–179.

Dhinagaran, D. A., Martinengo, L., Ho, M.-H. R., Joty, S., Kowatsch, T., Atun, R. and Tudor Car, L. (2022). "Designing, Developing, Evaluating, and Implementing a Smartphone-Delivered, Rule-Based Conversational Agent (DISCOVER): Development of a Conceptual Framework," *JMIR mHealth and uHealth* 10 (10), e38740.

Dosovitsky, G. and Bunge, E. (2023). "Development of a chatbot for depression: adolescent perceptions and recommendations," *Child and adolescent mental health* 28 (1), 124–127.

Drissi Oudghiri, M., Motaib, I., Elamari, S., Laidi, S. and Chadli, A. (2021). "Pituitary Apoplexy in Geriatric Patients: A Report of Four Cases," *Cureus* 13 (12).

Goldkuhl, G. (2004). "DESIGN THEORIES IN INFORMATION SYSTEMS - A NEED FOR MULTI-GROUNDING," *The Journal of Information Technology Theory and Application* 6, 7.

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U. J., Weidenbacher, P. A., Tang, S., Wu, W., Pak, J. E. and Kim, P. S. (2023). "Efficient evolution of human antibodies from general protein language models," *Nature biotechnology*.

Holzinger, A., Biemann, C., Pattichis, C. S. and Kell, D. B. (2017). "What do we need to build explainable AI systems for the medical domain?,".

Janssen, A. (2020). "Virtual Assistance in Any Context - A Taxonomy of Design Elements for Domain-Specific Chatbots," *Business & Information Systems Engineering* 62 (3), 211–225.

Janssen, A., Grützner, L. and Breitner, M. H. (2021). "Why do Chatbots fail? A Critical Success Factors Analysis," *ICIS 2021 Proceedings*.

Jebara, T. (2004). "Generative Versus Discriminative Learning,". In Jebara, T. (ed.) *Machine Learning*, pp. 17–60, Boston, MA: Springer US.

Jindal, J. A., Lungren, M. P. and Shah, N. H. (2024). "Ensuring useful adoption of generative artificial intelligence in healthcare," *Journal of the American Medical Informatics Association : JAMIA*.

Kocaballi, A. B., Quiroz, J. C., Laranjo, L., Rezazadegan, D., Kocielnik, R., Clark, L., Liao, Q. V., Park, S. Y., Moore, R. J. and Miner, A. (2020). "Conversational Agents for Health and Wellbeing," in: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, pp. 1–8.

Kocielnik, R. and Hsieh, G. (2018). "Facilitating Self-Learning in Behavior Change Through Long-Term Intelligent Conversational Assistance," in: *23rd International Conference on Intelligent User Interfaces*, New York, NY, USA: ACM, pp. 683–684.

Koppenhagen, N., Gass, O. and Müller, B. (2012). "Design Science Research in Action - Anatomy of Success Critical Activities for Rigor and Relevance," in: *ECIS 2012 Proceedings*.

Kowatsch, T., Volland, D., Shih, I., Rüegger, D., Künzler, F., Barata, F., Filler, A., Büchter, D., Brogle, B., Heldt, K., Gindrat, P., Farpour-Lambert, N. and l'Allemand, D. (2017). *Design and evaluation of a mobile chat app for the open source behavioral health intervention platform mobilecoach*.

Kumar, A., Rajasekharan Nayar, K. and Koya, S. F. (2020). "COVID-19: Challenges and its consequences for rural health care in India," *Public health in practice* 1, 100009.

Lai, Y., Panagiotopoulos, P. and Lioliou, E. (2023). "Empowering users with medical artificial intelligence technologies," *ECIS 2023 Research Papers*.

Laumer, S., Maier, C. and Gubler, F. (2019). "CHATBOT ACCEPTANCE IN HEALTHCARE: EXPLAINING USER ADOPTION OF CONVERSATIONAL AGENTS FOR DISEASE DIAGNOSIS," *ECIS 2019 Research Papers*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H. (eds.) *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Lin, Y.-T. and Chen, Y.-N. "LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models,". In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C. (2023). *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*.

Meier, P., Beinke, J. H., Fitte, C., Behne, A. and Teuteberg, F. (2019). "FeelFit – Design and Evaluation of a Conversational Agent to Enhance Health Awareness," in: *ICIS 2019 Proceedings*.

Mesbah, N. and Pumplun, L. (2020). ""Hello, I'm here to help you" – Medical care where it is needed most: Seniors' acceptance of health chatbots," *ECIS 2020 Research Papers*.

Meskó, B. and Topol, E. J. (2023). "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *NPJ digital medicine* 6 (1), 120.

Moilanen, J., van Berkel, N., Visuri, A., Gadiraju, U., van der Maden, W. and Hosio, S. (2023). "Supporting mental health self-care discovery through a chatbot," *Frontiers in digital health* 5.

Moilanen, J., Visuri, A., Suryanarayana, S. A., Alorwu, A., Yatani, K. and Hosio, S. (2022). "Measuring the Effect of Mental Health Chatbot Personality on User Engagement," in: Döring, T., Boll, S., Colley, A., Esteves, A. & Guerreiro, J. (eds.) *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia*, New York, NY, USA: ACM, pp. 138–150.

Motger, Q., Franch, X. and Marco, J. (2023). "Software-Based Dialogue Systems: Survey, Taxonomy, and Challenges," *ACM Comput. Surv.* 55 (5), 1–42.

Moussawi, S., Koufaris, M. and Benbunan-Fich, R. (2021). "How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents," *Electronic Markets* 31 (2).

Mozafari, N., Weiger, W. H. and Hammerschmidt, M. (2021). "Resolving the Chatbot Disclosure Dilemma: Leveraging Selective Self-Presentation to Mitigate the Negative Effect of Chatbot Disclosure," *Hawaii International Conference on System Sciences 2021 (HICSS-54)*.

Müller, L., Mattke, J. and Weitzel, T. (2019). "Not Talking to Robo-Doc: A QCA Study Examining Patients' Resistance to Chatbots for Anamnesis," *DIGIT 2019 Proceedings*.

Nadarzynski, T., Miles, O., Cowie, A. and Ridge, D. (2019). "Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study," *Digital health* 5.

Nayar, A. M., Attar, Z., Kachwala, S., Biswas, T. and Wagh, S. K. (2022). "Dost-Mental Health Assistant Chatbot," in: *2022 5th International Conference on Advances in Science and Technology (ICAST)*: IEEE, pp. 252–257.

Nguyen, T.-T., Sim, K., Kuen, A. T. Y., O'donnell, R. R., Lim, S. T., Wang, W. and Nguyen, H. D. (2021). "Designing AI-based Conversational Agent for Diabetes Care in a Multilingual Context," *PACIS 2021 Proceedings*.

O' Connor, Y., Kupper, M. and Heavin, C. (2021). "Trusting Intentions Towards Robots in Healthcare: A Theoretical Framework," *Hawaii International Conference on System Sciences 2021 (HICSS-54)*.

Park, H. and Lee, J. (2020). "Can a Conversational Agent Lower Sexual Violence Victims' Burden of Self-Disclosure?," in: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, pp. 1–8.

Parmar, P., Ryu, J., Pandya, S., Sedoc, J. and Agarwal, S. (2022). "Health-focused conversational agents in person-centered care: a review of apps," *NPJ digital medicine* 5 (1).

Pathinarupothi, R. K., Ramesh, M. V. and Rangan, E. (2016). "Multi-layer architectures for remote health monitoring," in: *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*: IEEE, pp. 1–6.

Paul, S. C., Bartmann, N. and Clark, J. L. (2021). "Customizability in conversational agents and their impact on health engagement," *Human Behavior and Emerging Technologies* 3 (5), 1141–1152.

Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. (2007). "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* 24 (3), 45–77.

Pelau, C., Dabija, D.-C. and Ene, I. (2021). "What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry," *Computers in Human Behavior* 122.

Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Lipori, G., Mitchell, D. A., Ospina, N. S., Ahmed, M. M., Hogan, W. R., Shenkman, E. A., Guo, Y., Bian, J. and Wu, Y. (2023). "A study of generative large language model for medical research and healthcare," *NPJ digital medicine* 6 (1), 210.

Perez-Soler, S., Juarez-Puerta, S., Guerra, E. and Lara, J. de (2021). "Choosing a Chatbot Development Tool," *IEEE Software* 38 (4), 94–103.

Polignano, M., Narducci, F., Iovine, A., Musto, C., Gemmis, M. D. and Semeraro, G. (2020). "HealthAssistantBot: A Personal Health Assistant for the Italian Language," *IEEE Access* 8.

Portoghese, I., Galletta, M., Coppola, R. C., Finco, G. and Campagna, M. (2014). "Burnout and workload among health care workers: the moderating role of job control," *Safety and health at work* 5 (3), 152–157.

Prayitno, P. I., Leksono, R. P. P., Chai, F., Aldy, R. and Budiharto, W. (2021). "Health Chatbot Using Natural Language Processing for Disease Prediction and Treatment," in: *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, pp. 62–67.

Rahman Khilji, A. F. U., Laskar, S. R., Pakray, P., Kadir, R. A., Lydia, M. S. and Bandyopadhyay, S. (2020). "HealFavor: Dataset and A Prototype System for Healthcare ChatBot," in: *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, pp. 1–4.

Rathnayaka, P., Mills, N., Burnett, D., Silva, D. de, Alahakoon, D. and Gray, R. (2022). "A Mental Health Chatbot with Cognitive Skills for Personalised Behavioural Activation and Remote Health Monitoring," *Sensors (Basel, Switzerland)* 22 (10).

Reddy, J. E. P., Bhuwaneshwar, C. N., Palakurthi, S. and Chavan, A. (2020). "AI-IoT based Healthcare Prognosis Interactive System," in: *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pp. 1–5.

Reddy, S. (2024). "Generative AI in healthcare: an implementation science informed translational path on application, integration and governance," *Implementation science : IS* 19 (1), 27.

Reis, L., Maier, C., Mattke, J. and Weitzel, T. (2020). "CHATBOTS IN HEALTHCARE: STATUS QUO, APPLICATION SCENARIOS FOR PHYSICIANS AND PATIENTS AND FUTURE DIRECTIONS," *ECIS 2020 Research Papers*.

Roy, D., Tripathy, S., Kar, S. K., Sharma, N., Verma, S. K. and Kaushal, V. (2020). "Study of knowledge, attitude, anxiety & perceived mental healthcare need in Indian population during COVID-19 pandemic," *Asian journal of psychiatry* 51, 102083.

Rubinstein, J., Moghe, R., Mizrachi, A. and Dissin, J. (2004). "Triptan use preceding life-threatening arrhythmias in charcot-marie-tooth: a case report and review of the literature," *Clinical neuropharmacology* 27 (1), 14–16.

Sai, S., Gaur, A., Sai, R., Chamola, V., Guizani, M. and Rodrigues, J. J. P. C. (2024). "Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies, and Limitations," *IEEE Access* 12, 31078–31106.

Sallam, M. (2023). "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," *Healthcare (Switzerland)* 11 (6).

Sasseville, M., Barony Sanchez, R. H., Yameogo, A. R., Bergeron-Drolet, L.-A., Bergeron, F. and Gagnon, M.-P. (2022). "Interactive Conversational Agents for Health Promotion, Prevention, and Care: Protocol for a Mixed Methods Systematic Scoping Review," *JMIR research protocols* 11 (10).

Schachner, T., Keller, R. and Wangenheim, F. V. (2020). "Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review," *Journal of Medical Internet Research* 22 (9), e20701.

Schlimbach, R., Lapychak, A. and Robra-Bissantz, S. (2023). "Exploring the Perception of a Chatbot's Language Style in a Learning Situation," *AMCIS 2023 Proceedings*.

Schmidt, R., Alt, R. and Zimmermann, A. (2023). "Assistant platforms," *Electronic Markets* 33 (1).

Scholten, M. R., Kelders, S. M. and Van Gemert-Pijnen, J. E. W. C. (2019). "An Empirical Study of a Pedagogical Agent as an Adjunct to an eHealth Self-Management Intervention: What Modalities Does It Need to Successfully Support and Motivate Users?," *Frontiers in psychology* 10, 1063.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. and Kersting, K. (2022). "Large pre-trained language models contain human-like biases of what is right and wrong to do," *Nature Machine Intelligence* 4 (3), 258–268.

Schuetzler, R. M., Grimes, G. M. and Giboney, J. S. (2020). "The impact of chatbot conversational skill on engagement and perceived humanness," *Journal of Management Information Systems* 37 (3).

Sevinc, A., Buyukberber, S. and Camci, C. (2005). "Medical jargon: obstacle to effective communication between physicians and patients," *Med Princ Pract* 14 (4), 292.

Shah, J., DePietro, B., D'Adamo, L., Firebaugh, M.-L., Laing, O., Fowler, L. A., Smolar, L., Sadeh-Sharvit, S., Taylor, C. B., Wilfley, D. E. and Fitzsimmons-Craft, E. E. (2022). "Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening," *The International journal of eating disorders* 55 (9), 1229–1244.

Shan, Y., Ji, M., Xie, W. X., Qian, X. B., Li, R. Y., Zhang, X. M. and Hao, T. Y. (2022). "Language Use in Conversational Agent-Based Health Communication: Systematic Review," *Journal of Medical Internet Research* 24 (7).

Sharma, D., Kaushal, S., Kumar, H. and Gainder, S. (2022). "Chatbots in Healthcare: Challenges, Technologies and Applications," in: *4th International Conference on Artificial Intelligence and Speech Technology (AIST)*.

Sheth, A., Yip, H. Y. and Shekarpour, S. (2019). "Extending Patient-Chatbot Experience with Internet-of-Things and Background Knowledge: Case Studies with Healthcare Applications," *IEEE Intelligent Systems* 34 (4), 24–30.

Shum, H., He, X. and Li, D. (2018). "From Eliza to XiaoIce: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering* 19 (1), 10–26.

Siangchin, N. and Samanchuen, T. (2019). "Chatbot Implementation for ICD-10 Recommendation System," in: *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*: IEEE, pp. 1–6.

Siemon, D., Ahmad, R., Harms, H. and Vreede, T. de (2022). "Requirements and Solution Approaches to Personality-Adaptive Conversational Agents in Mental Health Care," *SUSTAINABILITY* 14 (7).

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A. and Natarajan, V. (2023). "Large language models encode clinical knowledge," *Nature* 620 (7972), 172–180.

Siu, S. C. (2023). "Revolutionizing Translation with AI: Unravelling Neural Machine Translation and Generative Pre-Trained Large Language Models," *SSRN Electronic Journal*.

Sokolaj, U., Grundstrom, C. and Martul, A. (2023). "Addressing Uncertainty in AI Tool Development in Healthcare Through End-User Involvement," *Selected Papers of the IRIS, Issue Nr 14 (2023)*.

Stegemann, L., Gubser, R., Gersch, M., Bartschke, A., Hoffmann, A., Wagner, M. and Fürstenau, D. (2023). "FUTURE-ORIENTED AND PATIENT-CENTRIC? A QUALITATIVE ANALYSIS OF DIGITAL THERAPEUTICS AND THEIR INTEROPERABILITY," *ECIS 2023 Research Papers*.

Strobel, G., Banh, L., Möller, F. and Schoormann, T. (2024). "Exploring Generative Artificial Intelligence: A Taxonomy and Types," in: *Proceedings of the 57th Hawaii International Conference on System Sciences*.

Su, Z., Figueiredo, M. C., Jo, J., Zheng, K. and Chen, Y. (2020). "Analyzing Description, User Understanding and Expectations of AI in Mobile Health Applications," *AMIA Annual Symposium Proceedings*, 1170–1179.

Susarla, A., Gopal, R., Thatcher, J. B. and Sarker, S. (2023). "The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems," *Information Systems Research* 34 (2), 399–408.

Sweeney, C., Potts, C., Ennis, E., Bond, R., Mulvenna, M. D., O'neill, S., Malcolm, M., Kuosmanen, L., Kostenius, C., Vakaloudis, A., Mcconvey, G., Turkington, R., Hanna, D., Nieminen, H., Vartiainen, A.-K., Robertson, A. and Mctear, M. F. (2021). "Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts," *ACM Trans. Comput. Healthcare* 2 (3).

Thimmanayakanapalya, S. S., Mulgund, P. and Sharman, R. (2022). "A TOGAF Based Chatbot Evaluation Metrics: Insights from Literature Review," *AMCIS 2022 Proceedings*.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F. and Ting, D. S. W. (2023). "Large language models in medicine," *Nature medicine* 29 (8), 1930–1940.

Thoring, K., Mueller, R. and Badke-Schaub, P. (2020). "Workshops as a Research Method: Guidelines for Designing and Evaluating Artifacts Through Workshops," in: Bui, T. (ed.) *Proceedings of the 53rd Hawaii International Conference on System Sciences*: Hawaii International Conference on System Sciences.

Tomczak, J. M. (2022). *Deep Generative Modeling*. Cham: Springer.

United Nations (2017). *World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100*. URL: https://bit.ly/37Pzj88.

van Aken, J. E. (2004). "Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules," *Journal of Management Studies* 41 (2).

Varghese, J. and Chapiro, J. (2023). "ChatGPT: The transformative influence of generative AI on science and healthcare," *Journal of hepatology*.

Walls, J. G., Widmeyer, G. R. and El Sawy, O. A. (1992). "Building an Information System Design Theory for Vigilant EIS," *Information Systems Research* 3 (1), 36–59.

Wang, G., Yang, G., Du Zongxin, Fan, L. and Li, X. (2023). *ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation*.

Watson, R. T. and Webster, J. (2020). "Analysing the past to prepare for the future: Writing a literature review a roadmap for release 2.0," *Journal of Decision Systems* 29 (3), 129–147.

Weisz, J. D., Muller, M., He, J. and Houde, S. (2023). "Toward General Design Principles for Generative AI Applications," in: Smith-Renner, A. & Taele, P. (eds.) *Joint Proceedings of the IUI 2023 Workshops. Co-located with the ACM International Conference on Intelligent User Interfaces (IUI 2023)*, pp. 130–144.

Xing, Z., Tang, L., Huang, J. and Hu, X. (2017). "Woven coronary anomaly leading to silent myocardial infarction: A case report," *Medicine* 96 (44), e8302.

Xu, L., Sanders, L., Li, K. and Chow, J. C. L. (2021). "Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review," *JMIR cancer* 7 (4), e27850.

You, Y., Tsai, C.-H., Li, Y., Ma, F., Heron, C. and Gui, X. (2023). "Beyond Self-diagnosis: How a Chatbot-based Symptom Checker Should Respond," *ACM Trans. Comput.-Hum. Interact.* 30 (4).

Yu, K.-H., Beam, A. L. and Kohane, I. S. (2018). "Artificial intelligence in healthcare," *Nature biomedical engineering* 2 (10), 719–731.

Zhang, J., Oh, Y. J., Lange, P., Yu, Z. and Fukuoka, Y. (2020). "Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: Viewpoint," *Journal of Medical Internet Research* 22 (9).

Zhang, Z., Citardi, D., Wang, D., Genc, Y., Shan, J. and Fan, X. (2021). "Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data," *Health informatics journal* 27 (2), 14604582211011215.