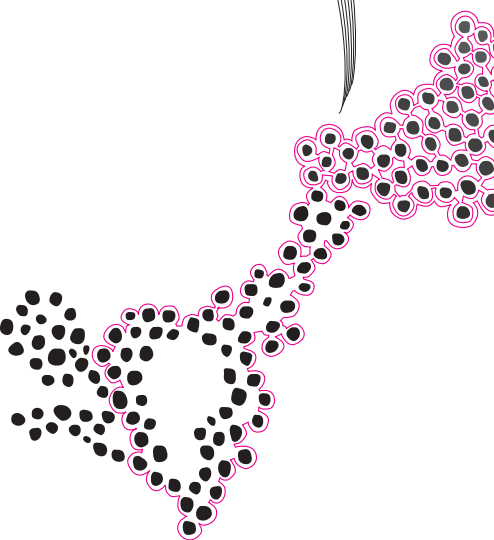


Master Thesis Interaction Technology
Faculty of Electrical Engineering, Mathematics, and
Computer Science

**Exploring lexical alignment
for understandable
information from trusted
healthcare chatbots**



Keara Schaij
August, 2024

Graduation committee:

dr. M. Theune
Faculty of Electrical Engineering, Mathematics and Computer
Science University of Twente

dr. rer. nat. D. Braun
Faculty of Behavioural, Management and Social Sciences,
Industrial Engineering and Business Information Systems
University of Twente

ABSTRACT

This research investigates the impact of lexical alignment in a healthcare chatbot on the user's understanding of the provided information and their trust in the chatbot providing this information during an information-seeking task. Previous research shows alignment to be essential for successful communication and leads to enhanced interaction and evaluation of agents in human-agent dialogue, therefore, this research explores the potential benefits of lexical alignment in addressing discrepancies in language use between healthcare professionals and patients due to varying health literacy levels (referring to the ability to obtain, understand, and use health information and services to make appropriate decisions regarding one's health).

In order to do so, a chatbot was developed to answer questions related to a hypothetical diagnosed disease and proposed treatment, with its responses either lexically aligning with the user or not. An experiment was conducted with a total of $n = 24$ participants, randomly assigned to either the lexically aligning chatbot or the non-aligning chatbot. Lexical alignment was realised by replacing predefined placeholders in the chatbot's responses with the participant's preferred term, obtained through a questionnaire done at the start of the experiment. The interaction between participant and chatbot was based on a predefined scenario explaining the hypothetical diagnosed disease and proposed treatment for the participant, about which they could ask questions to the chatbot. After this, an understanding test on the content discussed with the chatbot as well as a questionnaire on perceived trust were conducted.

Results showed that participants often preferred different terms than the original terms used as placeholders, highlighting the relevance of alignment. However, no statistically significant differences in understanding and trust between conditions were found. This can be explained by several limitations of this research, including the educational background of the participants, which could have minimised health literacy level discrepancies between chatbot and participant. Future research should consider the limitations and explore alternative alignment strategies to confirm or reject the findings of the current research. Post-experiment interviews revealed that participants in the non-aligning condition perceived the chatbot's language as difficult and expressed not having asked for clarification in case they needed it, whereas those in the aligning condition perceived the chatbot's language as less difficult and indicated they asked for clarification if needed. The lack of lexical alignment was suggested to potentially decrease trust in chatbots based on a study indicating lower health literacy levels lead to decreased trust in professional health-information sources and the observed difficulties in understanding and obtaining information without lexical alignment in the current research. Additionally, trust in the chatbot was found to be closely related to epistemic authority, which is attributed to doctors based on their qualifications, titles, and expertise. This was demonstrated by participants preferring to use healthcare chatbots over platforms such as Google to ask health-related questions based on the perceived reliability and accuracy of the sources used by the chatbot.

In conclusion, while the research did not find significant effects of lexical alignment on understanding and trust, valuable insights into the perception of language use, its influence on the interaction with a healthcare chatbot, trust, and the concept of epistemic authority were gained.

ACKNOWLEDGEMENTS

With this thesis, my journey at the University of Twente will come to an end. After studying Interaction Technology for the last two years, the final few months have been devoted to writing Research Topics and my final thesis. I want to thank everyone who has supported, encouraged, and assisted me throughout this journey.

First and foremost, I would like to thank my supervisors, dr. Mariët Theune and dr. rer. nat. Daniel Braun, for their support and guidance throughout this journey. I am especially grateful to dr. Mariët Theune for her valuable contributions, feedback, and expertise, which have helped to shape my research into its current form. Her understanding and kindness during this research, when circumstances made working on this thesis hard, helped me greatly to keep working on it and be able to complete it. All of this shows her outstanding supervision, which I deeply appreciate having had during both Research Topics and the final thesis. Additionally, I would like to express my gratitude to Sumit Srivastava, who was deeply involved in the journey of this thesis as well. His contributions, feedback, and expertise have helped to lift my research to its current form, and I am very thankful for his interest and willingness to help guide this research.

I would also like to especially thank my family. Over the past few months, we have had to cope with a very stressful, hectic, and emotional time, which continues to influence our everyday lives. Despite this, my parents and sister have always been there, encouraging and supporting me to fulfil this last part of my master's journey. A special thanks to my sister, who was always there to listen to my struggles, doubts, and little proud moments and who was a true help in completing this thesis. Despite the difficult times, I am very proud of how we are handling everything as a family. I am grateful to have been able to finalise this last step of my master's journey with the help, support, and encouragement of my family, as this was and is not always a given.

Additionally, I would like to thank all my friends who supported me throughout this process. They kept me motivated during setbacks or difficulties, listened to my endless stories about this project, and helped me find distractions when needed. I am grateful for all their support and could not have completed this thesis without them.

Finally, I would like to thank all the participants who generously contributed their time and insights to this research. With their involvement, this research was possible and the overall quality improved. The enthusiasm they showed during the experiments renewed my own enthusiasm to complete the final steps in this research.

CONTENTS

1	Introduction	7
2	Background	9
2.1	Chatbots	9
2.1.1	History of chatbots	9
2.1.2	Challenges and considerations	10
2.2	Chatbots in healthcare	11
2.2.1	Current landscape of healthcare chatbots	11
2.2.2	Challenges and considerations	12
2.3	Health literacy	13
2.3.1	Understanding	13
2.3.2	Trust	14
2.4	Epistemic Authority	15
2.5	Conclusion	15
3	Related work	17
3.1	Alignment in human-human dialogue	17
3.2	Alignment in human-agent dialogue	17
3.2.1	Presence of alignment in human-agent dialogue	18
3.2.2	Types of alignment in human-agent dialogue	18
3.2.3	Effects of alignment in human-agent dialogue	19
3.3	Conclusion	21
4	Methods	22
4.1	Research design	22
4.1.1	Questionnaire on (medical) terminology	23
4.1.2	Summary report on the diagnosis and treatment	23
4.1.3	Diagnosis and treatment	24
4.2	Conditions	25
4.3	Participants	25
4.4	Professionals	26
4.5	Procedure	26
4.6	Data collection	27
4.7	Measures	28
4.7.1	Quantitative measures	28
4.7.2	Qualitative measures	30
4.8	Conclusion	30
5	Realisation and testing	31
5.1	Platform	31
5.2	Interface	32
5.3	Chatbot as expert	33

5.4	Dialogue design	34
5.4.1	Intents	34
5.4.2	Training data	35
5.4.3	Responses	36
5.5	Implementing alignment	37
5.5.1	Alignment strategy	37
5.5.2	Terminology repository	38
5.5.3	Substitution	39
5.5.4	Sentence refinement	41
5.6	Pilot testing	41
5.6.1	Questionnaire on (medical) terminology	42
5.6.2	Non-aligning chatbot	43
5.6.3	Aligning chatbot	45
5.6.4	Understanding test	45
5.6.5	Online setup	46
5.7	Conclusion	47
6	Results	48
6.1	Quantitative results	48
6.1.1	Understanding test	48
6.1.2	Trust scale	49
6.1.3	Placeholder terms	50
6.2	Qualitative results	52
6.2.1	Future use of healthcare chatbots	52
6.2.2	Perception of terminology difficulty	54
6.3	Conclusion	55
7	Discussion	57
7.1	Analysis of the results	57
7.1.1	Understanding	57
7.1.2	Trust	59
7.1.3	Seeking clarification	59
7.1.4	Implications of the research	60
7.2	Limitations	62
7.2.1	Research design	62
7.2.2	Alignment strategy	63
7.2.3	Functionality of the chatbot	64
7.3	Future work	65
8	Conclusion	67
	References	68
A	Questionnaire to infer terminology preference (Final version)	76
B	Summary report	84
B.1	First version of the summary report	84
B.2	Second version of the summary report	85

C	Expert Interview	86
	C.1 Introduction	86
	C.2 General questions	86
	C.3 Question related to acute acoustic trauma	89
	C.4 Questions related to hyperbaric oxygen therapy	89
	C.5 Questions related to language use	91
D	Quantitative measurements	93
	D.1 Understanding test (Final version)	93
	D.2 Adjusted HCTM scale	99
E	Qualitative measurement	100
F	Flowcharts	101
	F.1 Non-aligning chatbot	101
	F.2 Aligning chatbot	101
G	Frequency of chosen terms for placeholders	106

1 INTRODUCTION

Chatbots, and especially human-agent dialogue, are a popular field of research in which a computer programme is designed to have a conversation with a human being, usually in written form over the internet. Currently, chatbots are used in a variety of different fields, from well-known applications in e-commerce as customer service chatbots such as Billie from Bol.com [1] to various applications in other fields, with a rapid increase in using chatbots in the field of healthcare [2, 3, 4].

Despite the broad applications of chatbots, they are not perfect yet. One area of interest to potentially improve human-agent dialogue is alignment, a well-researched concept in human-human dialogue that has gained interest in human-agent dialogue as well. Alignment describes how interlocutors in a conversation naturally agree on using similar linguistic representations [5]. This can be on the lexical level, which means two interlocutors start using similar words or phrases; the syntactic level, meaning two interlocutors start using similar speech patterns; and the semantic level, meaning two interlocutors start to share higher levels of representations, such as dialogue acts [6]. Alignment has been argued to be an essential aspect of successful communication, leading to an increased interest in implementing this feature in human-agent dialogue (e.g., [6, 7, 8, 9, 10, 11]). Various studies on the implementation of lexical alignment in human-agent dialogue yield promising results, enhancing both the interaction between humans and agents as well as the evaluation of the agents.

Considering the various domains in which chatbots are employed, challenges arising due to the use of chatbots can differ. Considering the rapid increase in the use of healthcare chatbots, challenges related to the exchange of information between healthcare professionals and patients become important to consider in the development of chatbots in the domain of healthcare [2, 3, 4]. A common problem present in the exchange of health-related information between healthcare professionals and patients includes patients experiencing problems understanding the information provided by healthcare professionals, which can often be attributed to the difference in health literacy levels between patients and professionals [12]. More specifically, the patient is unable to properly understand the specialised vocabulary and complex language used by the healthcare professional in the information that is conveyed [12, 13, 14]. Research focusing on this specific issue suggests adjusting the language used to convey health-related information to match the everyday language used by the patient [12, 13]. However, by doing this, difficulties arise related to whether the everyday language is the same across different cultural groups. More specifically, the interpretation of such everyday language can be different across various cultural groups, influencing the understanding of the provided information [13]. Furthermore, trust has shown to be a relevant concept regarding the problems in the health-related information exchange between professionals and patients, as research has shown a trustful relationship between healthcare professionals and patients can enhance information exchange [15, 16]. Additionally, trust is positively linked to better access and understanding of information, as well as influencing how individuals use various health-related information sources, further highlighting the important role of trust in the context of health-related information exchange [17, 18]. Another important aspect to consider related to trust is the limited trust

both healthcare professionals and patients are known to already have regarding the potential use of healthcare chatbots [2, 19, 20].

Based on that, this research will explore how the lexical alignment of a healthcare chatbot with the user during an information-seeking task impacts both the user's understanding of the provided information as well as their trust in the chatbot providing the information. This leads to the following research question:

How does the lexical alignment of a healthcare chatbot with the user during an information-seeking task impact the user's understanding of the information provided and their trust in the chatbot providing the information?

In order to answer the research question, an experiment was conducted in which participants were asked to chat with a chatbot to obtain information about a hypothetically diagnosed disease and proposed treatment. The experiment had a between-subject design in which participants interacted with either the chatbot lexically aligned with their language or with the chatbot responding with template answers that were similar to how medical information is usually conveyed by healthcare professionals. Participants took part in two questionnaires. One questionnaire is designed to test the participant's understanding of the information provided by the chatbot, and another is designed to measure the participant's perceived trust in the healthcare chatbot. By answering the research question, new insights are obtained to help with the development of chatbots, especially in the context of a healthcare chatbot. Moreover, this research provides insights about the way health-related information could be conveyed.

In the upcoming chapters, a detailed description of the process of this research will be provided. Chapter 2 offers background information on chatbots, including their history as well as the common challenges and considerations. In addition to this, the chapter has a particular focus on the use of chatbots in healthcare, describing the current landscape as well as the specific challenges and considerations in this context. Moreover, the concepts of health literacy and epistemic authority are introduced. Chapter 3 reviews related work in the on alignment, examining both alignment in human-human dialogue as well as alignment in human-agent dialogue. Chapter 4 discusses the relevant aspects of the experiment, followed by the details of the chatbot implementation and pilot testing in Chapter 5. Then, the experimental results are presented in Chapter 6, followed by a discussion on the main contributions and suggestions for future research in Chapter 7. Finally, Chapter 8 provides the conclusion of this research.

2 BACKGROUND

In this chapter, relevant background information for this research is provided. To begin with, a general introduction to chatbots, including their history, challenges, and considerations, is provided. Following this, an overview is provided of chatbots specifically used in healthcare, starting with a description of the current landscape, after which the challenges and considerations are addressed. Additionally, the concepts of health literacy as well as epistemic authority are being discussed.

2.1 Chatbots

Over the years, people have been interacting more and more with chatbots [2, 3, 21]. Within the domain of e-commerce, the integration of customer service chatbots has become particularly prevalent [2]. Some well-known examples include the IBM Watson assistant and Billie the chatbot from Bol.com [1, 22]. The Cambridge dictionary defines a chatbot as a computer programme designed to have a conversation with a human being, usually over the internet [23]. They can differ in the purpose they are designed to fulfil. Chatbots with the goal of providing the appropriate information to the user's questions are called informative chatbots [2, 24, 25]. This can, for example, include providing a schedule of the movies that play in the cinema when the user asks for information on movies in the cinema. Chat-based or conversational chatbots are designed to have more humanlike conversations, hence, to behave more as a human companion during the conversation [2, 24]. Lastly, task-based or transactional chatbots are created to help the user carry out certain tasks, for example, buying a cinema ticket [2, 25]. These diverse purposes of chatbots highlight the large range of applications chatbots can be used for, besides the well-known e-commerce applications.

To accomplish communication with the user, the chatbot relies on natural language processing (NLP) [2]. NLP can be seen as the part of artificial intelligence that focuses on making machines understand human languages [26]. NLP consists of natural language understanding (NLU) and natural language generation (NLG) [2, 26]. NLU allows the machine to understand the provided human language by extracting the meaning and relevant information [2, 26]. Subsequently, the extracted meaning and relevant information obtained by NLU is used in NLG processes to provide the user with meaningful human-like responses [2, 26].

2.1.1 History of chatbots

The first chatbot, ELIZA, was already established in 1966 by Joseph Weizenbaum [27]. ELIZA is a chatbot designed to simulate text-based conversations between humans and agents [3, 27]. It made use of simple pattern matching and a template-based response mechanism, enabling a comparison of the given input with predefined patterns stored in its database to then select an appropriate response from a predefined set of responses [27, 28].

Already, with this rather rudimentary chatbot, people seemed to believe they were conversing with a real human. This, in combination with the increased popularity of the Turing Test, resulted in a rise in building more advanced chatbots [28]. In the Turing Test, developed in 1950 by Alan Turing, a participant interacts with a human and a machine. By asking written questions, without knowing which is which, the participant must guess which is the machine and which is the human [29]. In the case where the machine "passes" the Turing Test, it is good enough to trick the participant and, hence, mimic human behaviour [28, 29].

With the increased interest in building more advanced chatbots, A.L.I.C.E., inspired by ELIZA, was built by Wallace in 1995 [30]. A.L.I.C.E. is like ELIZA in the way that it makes use of pattern-matching to understand what the user is saying [27, 30]. However, it is extended with a new language called Artificial Intelligence Markup Language (AIML) [30]. AIML can be seen as a collection of rules in the NLU in order to let the chatbot reply appropriately to the input [2, 31].

Since then, advances in artificial intelligence and NLP have made notable improvements possible since ELIZA and A.L.I.C.E. [2, 3, 21]. This resulted in smart personal assistants such as Amazon's Alexa, Apple's Siri, or Microsoft's Cortana that are well-known nowadays [3, 21]. Moreover, in the last couple of years, another significant advancement occurred: OpenAI's Generative Pre-trained Transformer (GPT), emerging as a widely used chatbot made by OpenAI [32].

Currently, GPT-4 is capable of human-level performance on several professional and academic assessments [32]. GPT-4 is a large-scale multimodal model that provides users with a textual answer to either text- or image-based inputs. By being able to understand context, coherent and contextually relevant responses are generated, resulting in a high level of mimicking human conversation [32].

2.1.2 Challenges and considerations

The evolution of chatbots over the years also comes with challenges and considerations. In the following, some of them will be highlighted to provide the reader with insights into issues that are still relevant within the context of this research.

As new technologies, such as chatbots, become more widely used, trust emerges as an important concept [15, 33, 34, 35]. Research in human-computer interaction highlights the fact that people tend to attribute human characteristics to computers and form trusting relationships with them [35]. This trusting relationship plays an important role in the acceptance and adoption of novel technologies such as chatbots [33, 34, 36, 37]. Considering the challenge associated with the adoption of chatbots, namely, the expected behavioural shift of users, the importance of this trusting relationship is emphasised. Such a behavioural shift expected of users may include a transition from well-known behaviours such as sending emails or making phone calls to chatting with a chatbot instead [2]. It is necessary to consider that it might take time for someone to adjust to the advancements chatbots bring along. Furthermore, the "old" way of doing things should not disappear completely, instead, one should be encouraged to make use of the new type of interaction (chatbot), but should be able to fall back on the well-known way of doing things (e.g., making a phone call to the helpdesk) [2].

For a chatbot to be successful, the ability to effectively communicate with humans is essential [2, 21]. Despite the enormous amount of research done in the field and the technological advancements, quite some chatbots are not able to comply with the expectations users have

regarding effective communication, which results in frustration and dissatisfaction [2, 21]. An important aspect here is that most current chatbots still suffer from issues regarding intent recognition [2, 21]. Intent recognition is an important aspect of a successful conversation, as it is the part in which the chatbot extracts the relevant meaning and information from the input provided by the user in order to understand what the user wants and what should be responded to [2, 26]. In the case that the chatbot recognises the wrong intent or no intent at all, it is not able to respond appropriately to the user. This leads to frustration and, depending on the domain in which the chatbot is used, other detrimental effects [21]. For example, in the case of customer service, when the chatbot is not able to help the customer properly with his or her complaint or question, the chance of the customer coming back might decrease compared to a helpful interaction.

Furthermore, in human-human conversation, one can talk with informal and non-standard language, which still causes problems such as misunderstandings during or breakdown of the dialogue between human and agent [8]. Often, agents struggle to understand and generate the inherent natural dialogue that humans possess, including aspects such as situatedness, context awareness, expressing social functions, and non-standard language [8]. Situatedness, for instance, is when one asks a chatbot about movies playing in the nearby cinema, the chatbot provides a list of movies in the cinema relevant to the user's location. Context awareness allows the chatbot to adjust its response to previously provided information by the user. For example, when the user has shared a preference for Pathé cinemas over Kinopolis cinemas, the chatbot can adjust its search accordingly. Chatbots being able to express social functions, are for example, those in customer service designed to express empathy during interactions. Lastly, the use and understanding of non-standard language includes the use of emojis, slang, or other informal language.

2.2 Chatbots in healthcare

In Section 2.1, the large range of applications in which chatbots can be used was already discussed. While chatbots might be well-known in the realm of e-commerce, where they fulfil a role in customer service, they can also play an important role in healthcare [2]. At present, the use of chatbots in healthcare is experiencing a rapid increase [3, 4]. The current main use of chatbots in healthcare is focused on alleviating the workload of healthcare professionals (including physicians, nurses, and other clinicians) in various ways [2, 4].

2.2.1 Current landscape of healthcare chatbots

The chatbots used in the healthcare domain serve various purposes. Each type of chatbot is designed to fit a specific need or challenge a user could encounter in the field of healthcare. A rather large group of chatbots is designed to assess the symptoms of the user. These chatbots are using medical databases in combination with healthcare professionals to evaluate the health of a user based on the specific symptoms the user is experiencing [38, 39, 40, 41, 42]. Taking into account the various types of chatbots discussed in Section 2.1 (informative, chat-based, or task-based), these healthcare chatbots will be typically described as informative chatbots. The main aim of these chatbots is to provide the user with relevant information and guidance related to the assessment of the user's health, which fits the aim of an informative chatbot [2, 25].

Another group of healthcare chatbots focuses solely on accessing health-related information. These chatbots are designed to make information access easier and, in this way, help cer-

tain patients or groups of people, for example, cancer patients or cancer survivors [43]. By making relevant information for a user more accessible and actionable, the constant reliance on healthcare professionals is decreased, which will increase independence in making health-related decisions. Similarly to the previously discussed types of chatbots, these chatbots fit the type of informative chatbots, as the primary goal is to provide the user with the information they are looking for [2, 25].

Yet another group of chatbots takes a more coach-like role regarding the user's health, this includes increasing emotional health or reaching a certain body weight goal [44, 45]. This is done by having a conversation between chatbot and user in which several psychological techniques are applied to, for example, improve the user's emotional and psychological health [44, 45]. These chatbots are an example of chat-based chatbots as the user is engaged in human-like dialogues where the chatbot takes on the role of a human companion [2, 25].

It is important to note that many of these chatbots were not designed to replace the entire role of a professional, but rather as an addition. For example, often a feature is incorporated to be able to talk to a healthcare professional every hour of the day [43, 46]. Another commonly implemented function is to help the user find the closest relevant medical resources (e.g., a doctor's office, a pharmacy, or relevant apps) [39, 42, 47]. These chatbots clearly share informative elements that are key for informative chatbots, however, the primary focus is not to provide information to the user. The goal of these chatbots is to provide the relevant information necessary to assist the user in completing certain actions, such as being able to communicate with a healthcare professional or going to the pharmacy, which fit with task-based or transactional chatbots [2, 25].

2.2.2 Challenges and considerations

The use of chatbots in healthcare is a factor that increases challenges arising from both healthcare professionals and patients [2]. Well-known challenges or difficulties, as discussed in Section 2.1.2, will also apply to chatbots used in healthcare. However, similar or new difficulties are even more relevant in this domain.

In Section 2.1.2, the importance of the concept of trust related to the overall acceptance and adoption of novel technologies such as chatbots was already discussed. With the use of a chatbot in the context of healthcare, additional challenges related to trust play an important role [15, 18, 35, 48]. Both healthcare professionals and patients experience difficulties regarding trust in the use of healthcare chatbots. Healthcare professionals lack confidence in chatbots taking on roles involving decision-making for which typically medical advice is necessary, whereas patients often lack trust in the information provided by the chatbot. More specifically, they worry about the possibility of the chatbot making unreliable predictions, e.g., providing incorrect or inaccurate information [2, 19, 20]. Additionally, empirical research done by Clayman et al. [48] looking into the differences in use of health information and trust in this information highlighted that people base their attitudes, beliefs, and behaviours on the health information sources that they trust [48].

Where healthcare professionals have served as the primary source of health information for a long time, the rise of internet use has diversified the sources of health information available to patients [16]. The various types of healthcare chatbots that are currently available for patients show that chatbots play a role in the transfer of health information, see Section 2.2.1. With this, the important aspect of communication in healthcare, especially the exchange of health-related

information, becomes relevant in chatbots as well [12, 49]. Problems in the communication of health-related information can adversely impact a patient's health outcomes [12, 50, 16]. These problems can often be attributed to patients misinterpreting or misunderstanding the provided information [12, 49, 50, 51]. This is generally a result of a disparity in the level of communication the healthcare professional utilises and the patient's level of understanding, emphasising a gap between communication strategies used in health-related information exchange between healthcare professionals and patients and the successful communication of such information [12, 14].

2.3 Health literacy

Considering the importance of communication and especially the information exchange in health-care and the problems associated with it, health literacy emerges as an important concept as it can facilitate effective communication between healthcare professionals and patients [12, 14, 50, 51]. Health literacy can be defined as the ability to obtain, understand, and use health information and services to make appropriate decisions regarding one's health [12, 14, 50, 51]. Several studies demonstrate the detrimental effects of limited health literacy on health outcomes, such as worse health status, increased hospitalisation rates, medication errors, or missed appointments [12, 14, 52]. The likelihood that healthcare providers will encounter problems related to the health literacy of patients is highlighted by a systematic review conducted in the United States that discovered that 26% of subjects had an insufficient health literacy level and 20% had marginal (minimal) health literacy [14].

2.3.1 Understanding

A person's health literacy level depends on various factors, including their general literacy skills (ability to read, write, and understand written material), cultural factors, the complexity of the information provided, and their experience in the healthcare system, with general literacy skills being the most important factor [12, 50, 51]. While most people with rather low general literacy skills often have limited health literacy as well, it is important to note that limited health literacy is not exclusive to individuals with low general literacy. Individuals with average literacy levels and even highly literate individuals experience difficulties understanding information provided by healthcare professionals. These difficulties are often attributed to the use of vocabulary and concepts that are unfamiliar to those outside the medical profession or study, as well as the more complex language with which the information is written [12, 13, 14]. This results in individuals often not being able to properly understand the provided health-related information, leading to poor health outcomes [12, 49, 50, 51].

Several studies related to health literacy and health information exchange propose to adjust the language used in conveying information to the patient in order to limit these difficulties [12, 13]. Egbert and Nanna [13] propose to do this by using plain language in information exchange, which means substituting medical and technical terms with everyday language. However, by applying this strategy, the question arises whether the everyday language of one cultural group is the same for another cultural group. Egbert and Nanna [13] underscore the significance of recognising the differences in meaning of words across different ethnic and age groups. With this, the cultural sensitivity of communication strategies and the importance of considering adjusting communication strategies in order to be able to effectively convey information across different cultural groups are highlighted. A manual for clinicians on health literacy proposes basic steps to improve communication between healthcare professionals and patients [12]. These include conveying the information at a slower pace, using plain or non-medical language, making use of pictures or drawings, limiting the amount of information provided at once, asking

questions to confirm whether the patient understands what was said, and making sure the patient feels comfortable asking questions. Here, the plain or non-medical language is explained as "living-room language" or "the language of the family", which is often referred to as the way in which one would explain something to a family member rather than a colleague [12]. However, in this case, similar problems as with the solution of Egbert and Nanna [13] will arise, namely, what can be seen as "living-room language" or "the language of the family" across different cultural groups. More specifically, challenges related to whether the meaning of certain words is consistent across different groups and whether these groups eventually have the same understanding of the concept explained arise. Where both of these methods try to improve the patient's understanding of medical information provided by a healthcare professional by using plain, non-medical language, both highlight the possible problems in understanding this information due to possible variable interpretations of words used across different cultural groups.

2.3.2 Trust

As discussed in Sections 2.1.2 and 2.2.2, trust is an important concept in the application of chatbots in general as well as in the field of healthcare. As described in the domain of human-computer interaction, trust consists of both interpersonal trust as well as technological trust [33]. Interpersonal trust is related to the social connections between the trustor and trustee [33, 53]. For example, in the case of a chatbot, interpersonal trust refers to the bond between the user (the trustor) and the chatbot (the trustee), which is shaped based on the experiences during the interaction. Technological trust, on the other hand, is related to the technological structure underlying the chatbot that is required to have successful outcomes during the interactions [33, 53]. In the case of a chatbot, technological trust refers to the confidence the user has in the underlying technological infrastructure that is necessary for the chatbot to be able to provide the service that is expected.

Within the field of healthcare, interpersonal trust between doctors and patients plays an important role. If this relationship is good, it allows the patient to rely on the doctor as, among other things, a main source of information regarding health-related matters [15, 16]. By creating a trusting relationship between doctor and patient, information exchange can be enhanced [15]. Besides this relationship, the trust one has in the provided health-related information is positively associated with the level of ease in accessing, locating, and understanding the information [17].

Accessing, locating, and understanding are closely related to the concept of health literacy, which is explained as the ability to obtain, understand, and use health information and services to make appropriate decisions regarding one's health [12, 14, 50, 51]. Accordingly, Chen et al. [18] investigated whether there is an association between an individual's health literacy level and the individual's use and trust in several health information sources. The results indicate that a lower health literacy level was associated with less trust in the information provided by healthcare professionals. On the other hand, this lower health literacy level was associated with increased trust in information provided by less professional platforms such as television or social media [18]. This, in combination with the notion that a trustful relationship can increase the success of information exchange, shows the connection between trust, health literacy, and the exchange of health-related information.

2.4 Epistemic Authority

In the context of communication in healthcare, particularly in the exchange of health-related information, epistemic authority should be understood. Epistemic authority refers to the assessment of the knowledge a certain source has within a specific domain, which limits possible uncertainties in the abilities of that source [54, 55]. For example, the knowledge a cardiothoracic surgeon has on heart and chest surgery [56]. Sources that have this epistemic authority can often be seen as sources on which individuals tend to rely [55]. Related to the relationship between doctor and patient, patients will rely most on the doctors that have this authority in order to acquire medical information or knowledge [55]. The concept of epistemic authority is not limited to doctors, any source can become an epistemic authority [54]. Moreover, various characteristics contribute to the perceived epistemic authority of a source [54, 55]. Within the field of healthcare, the fact that doctors are often perceived as possessing this authority can be attributed to their qualifications, titles, and expertise [54, 55]. This is in line with the original paternalistic approach known in the field of healthcare, where it is assumed that the doctor always knows what is best for the patient [57]. In this approach, the doctor is perceived to have a high level of authority, resulting in having a large role in decision-making processes related to the patient's health, whereas the role of the patient is very limited [54, 55, 58]. However, over the last few years, there has been a shift towards a more patient-centred approach [54, 55]. In this approach, the patient has a more prominent role in decisions related to their health compared to the paternalistic approach [54, 55]. With the patient having more authority in decision-making regarding their own health, the role of the doctor transitions to one focused on providing information instead of making the decisions. The perceived epistemic authority of the doctor remains an important aspect even in this more patient-centred approach, as it shapes the way in which the individual will rely on the doctor to obtain health-related information.

2.5 Conclusion

In this chapter, a general overview of the evolution, applications, and prevalent challenges of chatbots was provided, with a particular focus on their integration within the healthcare sector. Due to the advancements in chatbot development, the application possibilities have increased, going beyond the well-known applications in e-commerce. However, despite the advancements made, several challenges are still present. Difficulties related to the concept of trust, which can be linked to the acceptance and adoption of chatbots, as well as effective communication, including intent recognition, and the ability of people to use non-standard language, are still challenges that need consideration in the adoption and utilisation of chatbots.

Given the general increase in usage of chatbots, an increase in the adoption of chatbots in the field of healthcare is present. Various types of healthcare chatbots were discussed, each with the purpose of being informative, chat-based, or task-based. Considering the challenges that are present in the use of chatbots, the importance of the concept of trust is underscored by its application within the field of healthcare. More specifically, trust issues on the side of the healthcare professional as well as the patient are discussed. With the prevalence of chatbots in the field of healthcare taking on the role of informative chatbots, the relevance of considering the difficulties present in the exchange of health-related information between healthcare professionals and patients in the use of chatbots in the field of healthcare is highlighted. This leads to the relevance of understanding the concept of health literacy, which is known to facilitate effective communication between healthcare professionals and patients in cases of similar health literacy levels.

Understanding is a prominent aspect of the definition of health literacy, which is the ability to

obtain, *understand*, and use health information and services to make appropriate decisions regarding one's health [12, 14, 50, 51]. By explicitly exploring the reasons behind the lack of understanding patients have of the health-related information provided by a healthcare professional, an important issue is highlighted. Specifically, patients often experience problems understanding the provided information due to the healthcare professional using words, concepts, and overall language that is unfamiliar to people outside the healthcare field. In order to address this issue, researchers propose to use plain language in communication between healthcare professionals and patients. However, this strategy encounters challenges, particularly regarding what can be seen as plain language across various cultural groups. More specifically, research recognised cultural sensitivity in communication strategies as various cultural groups might have different interpretations of words or attribute different meanings to words. Diving into the concept of trust, the interpersonal trust between doctor and patient is highlighted. Especially the positive impact a trusting relationship can have on the information exchange. Besides this, it is highlighted how trust can be related to the concept of health literacy by discussing the influence of one's health literacy level on the trust one has in the provided information and the influence of the sources providing the information on this trust.

In the exchange of health-related information, doctors have traditionally been the main source of information and decision-making authority regarding a patient's health. The epistemic authority a doctor is recognised to have based on their expertise shows to be an important aspect in this view, as patients tend to rely on doctors that have this authority. The shift towards a more patient-centred approach changes the way in which patients rely on doctors, as patients are becoming more active in the decision-making process, shifting from the doctor having a decisive role to a more informative role regarding the patient's health.

This chapter showed the way in which chatbots can be used, especially in the field of healthcare, and the common challenges that accompany their application in this field. With this overview, the relevance of this research is highlighted by the increased use of healthcare chatbots in combination with the challenges related to trust and information exchange. Considering these challenges, the concept of health literacy is discussed as well as epistemic authority. In the next chapter, an overview of related work will be given.

3 RELATED WORK

This chapter will focus on the related work on alignment. In order to do this, this chapter is divided into two parts. The first part focuses on the concept of alignment in human-human dialogue, and the second part will focus on the relevant literature on alignment in human-agent dialogue.

3.1 Alignment in human-human dialogue

In a conversation between two interlocutors, it is well established that both interlocutors will align on their ways of speaking by starting to use similar linguistic representations [5]. For example, when one interlocutor refers to an object as *sofa*, the other interlocutor will also refer to the object as *sofa*, even if normally the word *couch* would have been their reference for that object. This is an example of lexical alignment, which is one of the linguistic levels at which alignment can occur.

In 2004, Pickering and Garrod [5] proposed the interactive alignment model that underlies dialogue. This model describes how interlocutors in a conversation naturally align on different levels of linguistic representations with the ones of the other interlocutor. This results in having shared linguistic representations, which is an essential feature of successful communication [5, 8]. Pickering and Garrod made the assumption that alignment is an automatic process happening unconsciously [5]. This means that alignment does not occur due to pre-existing knowledge or assumptions interlocutors have about each other, but rather through priming processes. Priming processes entail the activation of certain representations that are associated with the perceived utterance. For example, if one uses the word *bike*, then this might activate the following representations: *bicycle*, *cycle*, *ride*, *ride a bike*.

Alignment can be observed across different linguistic levels, including the example of lexical alignment, in which two interlocutors begin to use the same words or phrases, such as the use of the word *sofa* instead of *couch* [6]. In addition to lexical alignment, alignment can also occur at the syntactic and semantic levels. Syntactic alignment occurs when two interlocutors start to use similar speech patterns. For example, one interlocutor says, *I want to watch Star Wars*, and then the second interlocutor says, *I am excited to see Star Wars*, instead of *Watching Star Wars seems exciting* [6]. Third, semantic alignment is when the interlocutors start to share similar higher levels of representations, e.g., dialogue acts [6]. It is important to note that alignment on one of these three levels promotes alignment on other levels [5].

3.2 Alignment in human-agent dialogue

While alignment is a well-established phenomenon in human-human dialogue and, according to the interactive alignment model proposed by Pickering and Garrod [5], considered an essential aspect for successful communication, it is also a relevant concept in human-agent dialogue.

The following will dive into the presence of alignment in human-agent dialogue, the different types of alignment in human-agent dialogue, as well as the effects of alignment by examining relevant research.

3.2.1 Presence of alignment in human-agent dialogue

The mechanism of alignment in human-agent interactions has been studied by Koulouri et al. [7] using a Wizard-of-Oz experiment. Participants took part in a visual task with the goal of navigating a robot in an urban environment using a text-based interface. In the Wizard-of-Oz approach, another participant without predefined scripts or guidelines was pretending to be the robot and had to communicate with the other participant to obtain navigation guidelines. The results showed the presence of alignment in human-agent dialogue, resulting in stabilising and gradually reducing vocabulary, highlighting its reciprocal nature. Moreover, little alignment during the dialogue resulted in less successful interactions. Later, Sinclair et al. [6] looked into alignment between students and a tutor in a language learning context in both human-human and human-agent dialogue. More specifically, it was researched whether students learning a second language would align with an automated dialogue agent (taking the role of a tutor). In addition, Sinclair et al. [6] looked at the nature of this human-agent alignment to see whether there exist differences compared to alignment in human-human dialogue, especially focusing on the degree of engagement the student had during the dialogue. Results indicated the presence of alignment in human-agent dialogue, but not stronger than in human-human dialogue, confirming the presence of alignment in human-agent dialogue found by Koulouri et al. [7]. In addition to the evidence provided for the presence of alignment in human-agent dialogue, the results indicated differences in alignment across dialogues, showing increased variability in human-agent alignment compared to human-human alignment. Therefore, Sinclair et al. [6] suggest different levels of alignment can be related to various levels of student engagement.

3.2.2 Types of alignment in human-agent dialogue

In addition to the research on the presence of alignment in human-agent dialogue, several studies focusing on human-agent alignment further investigate the types of alignment present.

Suzuki and Katagiri [59] looked into the alignment of prosody in human-agent interactions. More specifically, the alignment of the user to the prosodic features of a computer was examined by means of an experiment in which participants took part in a question-and-answer session with a computer. The prosodic features that were investigated include the loudness of voice as well as the response latency (the duration of a pause between the end of the utterance of the participant and the start of the utterance of the computer, or vice versa). The results showed unidirectional prosodic alignment of the participant's speech during the interaction with the computer. In cases of increased loudness from the computer, the participant also increased the volume of their speech. However, this adjustment did not occur in reverse. Similar results were found for the latency response, where, in the case of a shorter latency response from the computer, the participant adjusted their latency response as well, but no similar adjustment was found in the case of a longer latency response from the computer. These results provide evidence for the presence of prosodic alignment in human-agent dialogue [59].

Stoyanchev and Stent [60] examined both lexical and syntactic alignment in human-agent dialogue. They did this by means of an experiment using the *Let's Go!* telephone-based spoken dialogue system that participants interacted with in order to get route information for the bus [60]. Lexical alignment was examined by analysing how the choice of verbs used by the sys-

tem influenced the verb form used by the participant in response. Similarly, syntactic alignment was explored by examining whether participants used action verbs and prepositions in their responses to the system in case the system used them. The results showed that the participants were both syntactically and lexically aligned with the system, which provides evidence for the presence of syntactic and lexical alignment in human-agent dialogue [60].

Research by Cowan et al. [61] examined the presence of syntactic alignment in human-agent interaction, together with the effect of the conversational partner being a computer or a human, as well as the influence of voice anthropomorphism on the degree of syntactic alignment. This was done with a controlled experiment in which participants took part in a game where the participant and the conversational partner needed to describe and match pictures. The results showed the presence of syntactic alignment, as well as the alignment being independent of the conversational partner being a human or a computer, or the level of voice anthropomorphism. Hence, this research provides evidence for the presence of syntactic alignment in both human-human dialogue as well as human-agent dialogue [61].

Spillner and Wenig [8] investigated three different levels of alignment: no alignment, lexical alignment, and lexical alignment together with syntactic alignment. This was done by creating a chatbot capable of simple information retrieval in the film domain. The chatbot used in the research utilises templates and rule-based methods to provide answers to the user. To achieve this, the chatbot either responded with a static answer or an answer where the requested information was inserted. Three different levels of alignment were implemented: baseline (no alignment), substitution (lexical alignment), and transformation (lexical alignment as well as sentence structure alignment). The overall results showed alignment to be beneficial for the level of frustration and perceived workload. This is in line with research in human-human dialogue showing alignment decreases the perceived workload [62].

3.2.3 Effects of alignment in human-agent dialogue

The review of the existing research so far underscored the presence of alignment in human-agent dialogue and highlighted the different types of alignment that can be implemented. However, the impact of the implementation of alignment should also be considered.

The research discussed by Spillner and Wenig [8] showed some specific effects of lexical alignment on interaction outcomes (frustration level and perceived workload). Similarly, research by Huiyan and Min [9] investigated what influence lexical alignment in human-agent interaction has on the evaluation of the conversational partner and the interaction itself. This research employed a text-based interaction where participants had to name and match pictures. In the aligned condition, the same word as the participant was used, whereas in the misaligned condition, a different word was used. The findings demonstrated an improved assessment of the interaction in terms of perceived cognitive demand and response accuracy. Furthermore, the aligned condition showed an increased liking for the conversational partner. Additionally, Levitan et al. [63] highlighted the effects of alignment after the validation of an architecture to incorporate acoustic-prosodic alignment in spoken dialogue systems. Validation of this architecture through experimental contexts and pilot testing revealed participants to perceive increased reliability and likability of the system in cases where it showed alignment. In addition to the previously discussed studies, Nuñez et al. [10] specifically investigated the influence of lexical alignment on the side of the agent in human-agent dialogue. This was done by means of an object-naming-matching game in which the virtual agent aligned with the words used by the user in the first round. However, in the second round of the game, the virtual agent did

not align with the user. Note that the reversed order was also investigated: the first round of no alignment and the second round of alignment. The influence of alignment on the perceived likability, competence, and autonomy of the agent was examined, and the results showed the competence of the agent to be rated more positively when the agent displayed alignment in the second round of the game as opposed to only in the first round of the game [10].

Another effect that has been researched is the influence of alignment on the trust between two interlocutors. Scissors et al. [64] looked into the influence of lexical alignment on the level of trust between interlocutors in a social dilemma investment game. Participants played several rounds of an investment game, and after every five rounds, they could communicate with their partner. In the study, lexical alignment was defined as the recurrence of a word or word phrase in the dialogue by both the participant and their partner. The results showed that partners with greater trust exhibited a higher level of lexical alignment compared to partners with lower trust [64]. Succeeding research by Scissors et al. [65] explored the different forms of linguistic similarity that influence trust between interlocutors in a social dilemma investment game. This research extended the previous research by highlighting, for example, that similarity in terms having positive content, such as *happy*, had a different influence on trust than the similarity between terms with more negative content, such as *hate*. Specifically, the results showed that the use of similar positive terms during dialogue between interlocutors enhanced trust, whereas using similar terms with negative content during dialogue diminished trust [65]. Drawing inspiration from this research on alignment and trust, Hoegen et al. [66] developed a voice-based conversational agent capable of naturalistic multi-turn dialogue with the user as well as the ability to align with the conversational style of the user. Participants interacted with this agent by discussing several tasks that were provided to them, for example, scheduling a lunch meeting with the agent. In order for the agent to align with the conversational style of the participant, the word choice and utterance length, as well as the prosody and speech rate, were adjusted to match those of the participant. Results showed that participants exhibiting a conversational style characterised by higher levels of consideration perceived the agent as more trustworthy in cases where the conversational styles matched. On the other hand, participants exhibiting a conversational style characterised by high levels of involvement did not perceive the agent to be more or less trustworthy, regardless of whether the conversational styles matched or not [66].

Aside from research investigating the effects of lexical alignment on the perception of the agent and the interaction, research done by Srivastava et al. [11] investigated the effects of lexical alignment on the understanding of an explanation provided by a conversational agent. The explanation provided by the conversational agent was in the context of causes and effects of lung cancer [11]. In the first stage of the experiment, the users got to see images representing the concepts used in dialogue in a later stage of the experiment. For each image, the participants had to choose the most appropriate word describing it from a drop-down menu. This word (prime) was used during the dialogue stage of the experiment. The participants either interacted with an agent aligning with the user, an agent misaligning with the user, or a control condition in which no dialogue was present. Alignment with the user was implemented by using the primes the user chose in the first stage of the experiment to describe the specific concept. In the case of misalignment, on the other hand, the answers did not include this prime. In the case of alignment and misalignment, the users interacted with the agent by having a dialogue in order to obtain information on the context and causes of lung cancer. In order to draw conclusions about the understanding of the explanation, measures of recall were used. The results indicated positive effects of alignment in conversational agents; participants interacting with the agent that incorporated alignment showed a higher information recall compared to the participants who interacted with the misaligning agent or did not take part in dialogue [11].

3.3 Conclusion

The discussed research on alignment in human-agent interaction revealed promising results regarding its influence on human-agent dialogue. More specifically, alignment has been shown to lead to more successful interactions as well as an enhanced evaluation of both the conversational agent and the conversation itself. Additionally, research was discussed in which alignment was shown to aid in the participant's understanding of explanations provided by a conversational agent about a health-related topic by means of greater information recall.

Despite the limited research done on alignment in the context of healthcare, alignment offers a valuable concept to address challenges in the exchange of health-related information arising from differences in health literacy levels. More specifically, alignment makes it possible to create tailored communication strategies for the dialogue between healthcare professionals and patients by aligning with the health literacy level of the patient. This is expected to limit problems in understanding the provided information due to a disparity in health literacy levels between healthcare professionals and patients or problems related to cultural sensitivity in the current proposed solutions for the problems in health-related information exchange related to differences in meanings of words across various cultural groups.

The research discussed on trust and alignment underscored that both concepts are related, as increased trust between two interlocutors leads to an increased degree of alignment between the two interlocutors. Additionally, trust has been shown to be closely related to the concept of health literacy by means of the health literacy level influencing the trust one has in the source providing the health-related information based on the source's professionalism, as discussed in Section 2.3.2. Both of these findings highlight the potential positive influence of alignment on trust.

Based on the discussed literature, hypotheses are formed for the understanding of the provided information and the perceived trust in the chatbot. The following hypothesis is set regarding understanding:

- Integrating lexical alignment with the user in a healthcare chatbot will increase the understanding the user has about the provided information by the chatbot.

And the hypothesis for the perceived trust is:

- Integrating lexical alignment with the user in a healthcare chatbot will increase the perceived trust in the chatbot.

4 METHODS

The discussed research on alignment in human-agent dialogue yields promising results regarding communication. However, there is limited research focusing on the implementation of alignment in chatbots in the healthcare domain. Given the challenges that exist regarding information exchange in the field of healthcare, the combination of both appeared to be a valuable area for research. Therefore, this research aimed to evaluate the effect of lexical alignment in a healthcare chatbot during an information-seeking task on the understanding the participant has of the provided information and the perceived trust in the chatbot providing the information. More specifically, the information provided to the user was adjusted to the user's vocabulary as opposed to being written at a high health literacy level typically used by healthcare professionals. This was done by implementing lexical alignment in the chatbot, as in most of the research done on alignment in agents (see the discussed research in Section 3.2), lexical alignment yields promising results related to enhanced communication between humans and agents.

4.1 Research design

The participants that have agreed to take part in the experiment were first asked to participate in a questionnaire on (medical) terminology. More information about this questionnaire can be found in Section 4.1.1. After completing the questionnaire, the participants received a document explaining the hypothetical diagnosed disease and proposed treatment, which is explained in Section 4.1.2. After this, they were asked to chat with a healthcare chatbot to ask possible questions they had about the hypothetically diagnosed disease and proposed treatment. See Section 4.1.3 for more information on this disease and treatment. Participants either chatted with a chatbot using lexical alignment with the user or with a chatbot that did not. After the participant felt sufficiently informed about their new treatment or stopped their conversation alternatively, they were asked to take part in two questionnaires. One questionnaire was designed to test the participant's understanding of the information provided by the chatbot, and the other questionnaire was designed to measure the participant's perceived trust in the healthcare chatbot. Then, after this, the participants took part in a short, semi-structured interview to get a better understanding of the experience they had with the chatbot. An in-depth description of the measurements will be provided in Section 4.7.

Note that the participant was not told beforehand that there would be a small test on their understanding of the provided information at the end of the conversation. On the one hand, telling the participant about this would potentially increase their motivation to ask questions during the interaction. On the other hand, this would potentially change the nature of the interaction towards a more learning-like setting in which the participant would pay more attention to certain details or try to remember the information way more in depth compared to the normal information-seeking task the experiment was aiming to represent. Therefore, it was chosen to not tell the participants, which was the case for both conditions (interacting with the aligning or non-aligning chatbot). Note that a soft threshold of a minimum number of questions, to be at

least three, was incorporated. In the case where less than three questions were asked, a message would appear clarifying that the participant asked a rather limited number of questions, and a few directions for possible questions were proposed (these were based on the defined intents of the chatbot, see Section 5.4.1). In case the participant still wanted to stop the interaction, this was allowed, and the interaction was stopped. The decision to incorporate a threshold of a minimum number of questions as a soft requirement was made because the conversation should be as natural as possible and the participants should not be forced to ask questions they do not genuinely have. In the real-life scenario of a conversation between a healthcare professional and patient, there is not a minimum number of questions to be asked either. By providing the participant with some possible question directions, a true interaction between the participant and chatbot is encouraged, even if the participant lacks ideas for questions.

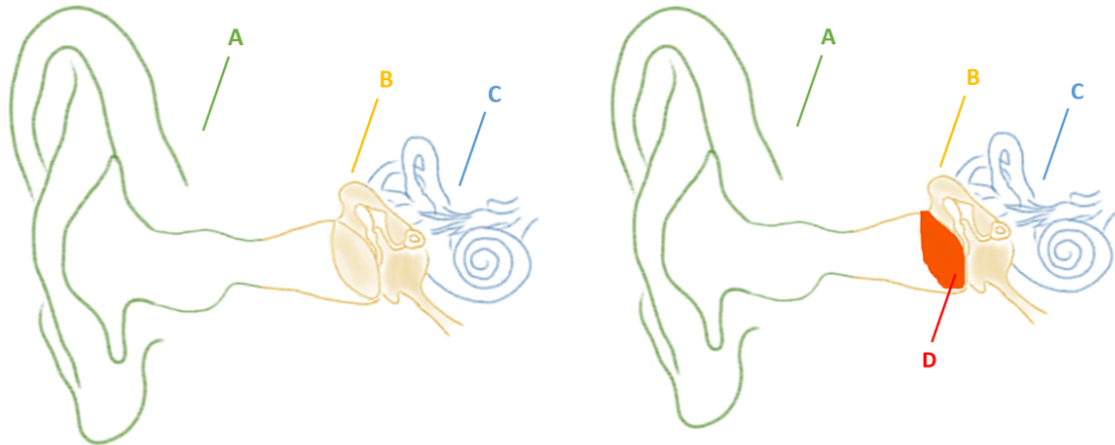
4.1.1 Questionnaire on (medical) terminology

The first part of the experiment consisted of filling out a questionnaire, which was used to infer the terms used by the participants for relevant concepts or terminology that might be the topic of conversation with the chatbot in the later stage of the experiment. Note that the participant was not aware of the specific purpose of this questionnaire. This questionnaire consisted of 25 questions in which the participant was asked to select the term they would most likely use to describe the concept in the provided picture or to fill in the *[blank]* in the sentence. The participant was able to pick a term from a set of possible answers, including the options *I do not know* and *Other*. For the last option, they were required to enter the term they would use alternatively. It is important to note that all the answer options were ordered in alphabetical order. This was chosen to increase the neutrality of the answer options and to avoid bias or favouritism towards a certain position in the answer list when creating the questionnaire. The questions were accompanied with images that are free stock images (e.g., from Pexels) [67]. In addition to these stock images, the image used to infer the term used for oronasal mask was obtained from Aviv Cilicis using this image in an explanation on hyperbaric oxygen therapy [68]. Moreover, two figures were created to support the questions relating to the terminology of the ear anatomy. These figures were created based on several images of the anatomy of the ear and used colouring and arrows to highlight relevant parts of the ear for the accompanying question. These two images can be found in Figure 4.1 [69]. The first (left) figure is accompanying questions 22 to 24, inferring the term for the outer ear (A), middle ear (B), and inner ear (C). The second (right) figure accompanies question 25, inferring the term for the eardrum (D). The questionnaire was filled out on an online webpage using Qualtrics [70]. It was chosen to have a different interface for the questionnaire(s) and the chatbot, as the interaction with the chatbot should be focused on the information the participant wants to obtain related to the disease and proposed treatment. Additionally, the chatbot was represented as a professional medical source for this information, as will be discussed in Section 5.3. If the questionnaire had been in the same environment as the interaction with the chatbot, this could interfere with the role of the chatbot as a medical expert. An overview of the questions asked in the questionnaire can be found in Appendix A.

4.1.2 Summary report on the diagnosis and treatment

On the side of the participant, the goal of the interaction with the healthcare chatbot was to obtain all the information they wanted about a new treatment that was proposed to them for their diagnosed disease by a healthcare professional. Therefore, the participants first received a document, referred to as *summary report*, in which their diagnosis was described along with the new proposed treatment. This document represents a summary of the findings and ob-

Figure 4.1: Overview of the ear



servations of the healthcare professional and was created in collaboration with a professional in the field, see Section 4.4. This ensures that the document contains correct information, is written using the correct and relevant medical terms and concepts, and is written in equivalent language as would be the case in real life. The document described the participant's hypothetical complaint, physical examinations that were already done, diagnosis, treatment plan, and follow-up. This provided the participant with a global idea about the scenario while leaving enough room for questions related to the disease and proposed treatment, as it was not an in-depth explanation about the disease or the proposed treatment and included medical terminology. The document can be found in Appendix B.1. In addition to the information provided in the interview, several sources were consulted in order to create this document and as a source of inspiration for what elements and content should be in it. These include a database for guidelines on care for people with hearing loss and the positioning of the Dutch National Health Care Institute on the treatment of hyperbaric oxygen therapy [71, 72]. More detailed information on the chosen diagnosis of acute acoustic trauma and the treatment of hyperbaric oxygen therapy can be found in Section 4.1.3. The choice of this particular treatment was based on its relative novelty, which increases the likelihood that participants truly required the chatbot to explain certain aspects of it [73, 74]. The diagnosis, on the other hand, is rather easy to interpret when understanding the information properly. Consequently, it was expected that participants would genuinely need to interact with the chatbot in order to acquire information about the treatment in combination with the diagnosis.

4.1.3 Diagnosis and treatment

As already explained, the diagnosis and proposed treatment were carefully chosen in order to create a proper interaction between the participant and the chatbot, given the fictitious nature of the scenario. The chosen diagnosis is called acute acoustic trauma, and the chosen treatment is called hyperbaric oxygen therapy. Acute acoustic trauma is the result of a brief exposure to extreme high levels of noise, such as noise coming from an explosion or a shooting firearm [73, 74]. Aside from acute acoustic trauma due to an acoustic trauma such as explained before, one can experience sudden loss of hearing in one of the ears. Both of these problems share issues with blood flow in the inner ear [74]. The use of hyperbaric oxygen therapy has proven to be a successful treatment for these issues. Hyperbaric oxygen therapy was first discovered by the Amsterdam surgeon professor Ite Boerema and is further investigated in collaboration

with the Dutch military [73]. Currently, this treatment is solely available for military personnel and policemen that suffer from severe hearing loss after an acoustic trauma, such as described before [73, 74].

Here, a short description of the disease and treatment will be given in order to provide the reader with a basic understanding of the treatment the participants in the experiment will converse about with the chatbot.

Acute acoustic trauma is a type of hearing loss due to a so-called acoustic trauma [71, 72, 73, 74, 75, 76, 77]. This acoustic trauma refers to an injury to the inner ear that is caused by a loud noise. Note that this can be due to (even a brief) exposure to loud noise such as an explosion or gunshot but also to prolonged exposure to, e.g., music. The treatment options for acute acoustic trauma are limited and often include a combination of corticosteroids (to limit the swelling in the ear) and hyperbaric oxygen therapy (see Appendix C). Patients allowed to make use of this hyperbaric oxygen therapy must go to one of several locations where a recompression chamber (hyperbaric chamber) is available. Such a room is well-known in the diving world and now also more and more in the healthcare world. The patient enters the room, after which the doors will be closed and the pressure in the chamber will slowly become equivalent to that experienced by a diver at a depth of 14 metres (pressurisation) [73]. Then, the patient will inhale pure oxygen for approximately two hours. This treatment consists of ten consecutive sessions in order to have the most effective results of the treatment [74].

4.2 Conditions

There were two conditions in the experiment. Condition A entailed that the participant interacted with the chatbot aligned with the participant, whereas in condition B, the participant interacted with the chatbot without alignment. In condition B (without alignment), the answers provided by the chatbot were template answers that were created together with healthcare professionals. In condition A (the aligning condition), the chatbot adjusted these template answers to the vocabulary of the user. The experiment had a between-subject design. A between-subject design was chosen to see how the two different conditions were separately perceived by the participants. This means that each participant was randomly assigned to either one of the conditions. Aside from this, the setup of the experiment was identical.

4.3 Participants

The research's participant group consisted of $n = 24$ men and women between the ages of 18 and 30, who had the ability to speak, read, and write in English and were recruited via study acquaintances, friends, and acquaintances via friends. The age group was chosen based on studies showing that, compared to older age groups, this age group has more experience with computers and is more skilled [78, 79]. This is relevant to the research as the participant was required to interact with a chatbot via a web page. Note that the age is irrelevant for the described illness and treatment that the participant conversed about in the experiment [73, 74]. Within the research, there were two conditions, i.e., the chatbot aligned with either the user or not, resulting in 12 participants for each condition. Beforehand, the participants were informed about the level of English required for the experiment: being able to properly converse in English. Note that none of the participants taking part in the experiment were native English speakers. Furthermore, participants did not require any previous knowledge in order to be able to participate in the experiment.

4.4 Professionals

In several parts of the research, the knowledge or experience of healthcare professionals was an important aspect, considering I do not have the required background in medical studies to be able to replicate the way in which medical information is usually provided to patients myself. First, the document provided to the participant explaining the scenario of their hypothetically diagnosed disease and proposed treatment was adjusted and improved with the help of a professional. Secondly, in order to be able to generate template answers for the chatbot, the experience of a professional in the field was essential.

The professional assisting in this research was a military nurse of the Royal Netherlands Navy with relevant expertise in the treatment and disease [80]. The nurse that assisted in this research brings extensive expertise, having fulfilled roles as both a nurse for divers as well as a diving medical supervisor for the Royal Netherlands Navy for three years at the time of this research. Within these three years, the nurse has gained experience and knowledge in the application of hyperbaric oxygen therapy, not only for military divers but also for patients suffering from acute acoustic trauma. In the case of applying hyperbaric oxygen therapy, the nurse serves as the patient's first point of contact and creates a treatment plan in close collaboration with the diving doctor. Additionally, the nurse needs to be within the chamber during treatment to ensure safety and conduct medical checks when needed.

In order to obtain the relevant information needed in the design of the chatbot as well as the summary report discussed in Section 4.1.2, an interview was designed and conducted. The interview consisted of four subsections. The first consisted of some general questions related to the steps that need to be taken when someone gets diagnosed with acute acoustic trauma and the treatment of hyperbaric oxygen therapy is proposed. The second consisted of some questions related to the disease. The third consisted of questions related to the treatment. Finally, the fourth consisted of questions related to the word use and language style used by the professional during the conversation with the patient. The transcription of this interview can be found in Appendix C.

4.5 Procedure

The procedure of this experiment was the same for both conditions in order to limit the differences between the two. The procedure could be described by the following steps, where the steps in *italics* are for the researcher:

1. *In case of the online setup: Setup online environment using Microsoft Teams or Google Meet*
2. Information letter and consent
3. Questionnaire on (medical) terminology
4. Reading summary report
5. *In case of the aligning chatbot: Export and import the correct file for alignment in the code (results of the questionnaire on (medical) terminology)*
6. *In case of the online setup: Setup online connection and provide link to interact with chatbot (see Section 5.1 for more information on this)*
7. Interaction with MedWiseBot

8. *Select the appropriate questions for understanding test*
9. Test on understanding
10. Questionnaire on trust
11. Semi-structured interview
12. Debriefing

At first, the participant was asked to read and sign the consent form accompanied by the information letter. Then, the participant was asked to take part in the questionnaire on (medical) terminology via a weblink leading to the questionnaire. After this, the participant received the document explaining the scenario, including the disease and proposed treatment, at the start of the experiment (referred to as the summary report, which can be found in Appendix B). Then, after the participant finished reading this document, they were asked to interact with the chatbot, MedWiseBot, via a web page that they could access on the same laptop as the questionnaire. The chatbot started with some small introductory messages to clarify the goal of the interaction, which was to ask all the questions the participant had regarding the diagnosed disease and proposed treatment. During the dialogue, the participant could end the conversation at any given time by typing "stop" which was indicated at the start of the conversation. Note that in the case the participant asked less than three questions, a message appeared indicating this to the participant and proposing possible question directions as explained in Section 4.1. After the interaction was stopped by the participant, participation in two questionnaires was asked. For the questionnaire on understanding, the researcher first had to select the relevant questions, as will be explained in Section 4.7.1. To enter the questionnaires, the participant had to simply open the link that was provided to them by the researcher. At the end of the questionnaires, a short, semi-structured interview was done. This required the experiment to take place either at the same location as the researcher, in a public space where the participant was able to focus properly, or via an online meeting (using Microsoft Teams or Google Meet). The second, an online meeting, allowed for more flexibility in conducting the experiment, which increased the number of people wanting to participate. Note that in the case of the online setup, the researcher had to set up the online environment of either Microsoft Teams or Google Meet, as well as the online connection for the chatbot. After the questionnaires and interview, the participant got a small debriefing in which the purpose of the experiment was explained as well as whether they interacted with the aligning chatbot or not. Next to this, the participant was asked whether they had any questions related to the experiment or research.

4.6 Data collection

During the experiments, several data were collected. All the data were anonymised to limit possible biases during the analysis of the results. Anonymisation was done by using a unique ID per participant in order to make data analysis possible during the analysis of the results later in the research. This unique ID per participant was created by combining a randomly generated number between zero and 100 with a random letter of the alphabet. At the start of the experiment, participants received a consent form that required some personally identifiable information. However, during the rest of the experiment, no personally identifiable information was required. This included the answers to the questionnaire on (medical) terminology, dialogues from the interaction with the chatbot, the answers from the two questionnaires after the interaction, as well as the answers from the short, semi-structured interview.

4.7 Measures

This research used both quantitative and qualitative measures to obtain an insight into the influence of the independent variable of the implemented lexical alignment on the dependent variables of the understanding the user has of the information provided by the chatbot as well as their perceived trust in the chatbot. After the research was conducted, statistical tests were utilised for hypothesis testing, which will be discussed in Chapter 6. In the following, more information on the quantitative and qualitative measures is discussed.

4.7.1 Quantitative measures

Here, the relevant quantitative measurements for this research based on literature are discussed. First, the quantitative measurements to assess the participant's understanding of the provided information by the chatbot will be explained. Second, a scale to measure the perceived trust participants have in the chatbot will be discussed.

Understanding

In order to obtain an insight into the understanding the participant has of the information provided by the chatbot, a test was created to verify whether the participant grasped the key aspects of the provided information. In order to do this, methods to check for understanding (CFU) were used [81]. CFU offers various strategies to check for understanding, mainly used in educational settings. These strategies can be categorised into three main types: asking questions that need to be answered verbally, asking questions that need actions or demonstrations, and asking questions that need to be answered in a written format. For the questions that need to be answered verbally, an answer type includes *choral responses*, where a question is asked and after a given sign, everyone is asked to respond simultaneously [81]. The questions that are answered by actions or demonstrations include *response cards*, where one needs to show the card with the appropriate answer to the question (e.g., yes-or-no cards or A, B, C answers) [81]. Finally, the questions that need to be answered verbally include *quick write*, where one needs to write a short paragraph on the discussed subject [81]. For this research, inspiration was drawn from the response card method, in which one must choose between the responses true or false, yes-or-no, several multiple-choice answers, or which is the odd one out. Several questions related to the treatment and disease were defined and presented in a questionnaire format where the participant needed to choose the correct answer. Besides this, fill-in-the-blank questions, Cloze test, were created where participants needed to choose the appropriate answer from a set of options. This is based on the research of Srivastava et al. [11], where the Cloze test was employed to investigate participants' information recall and understanding gain when interacting with an aligning agent, misaligning agent, or having no interaction at all, as already explained in Section 3.2.3. Additionally, Paus and Jucks [82] employed the Cloze test in their research on human-human communication to gain insights into participants' understanding, in terms of learning gain, after dyadic conversations focused on either aligning or misaligning learning materials.

Since the topics covered and discussed in the interaction with the chatbot varied among participants and the aim was to gain an insight into the understanding of the information provided by the chatbot, the test was tailored according to the topics discussed during the interaction. This was done by keeping track of the information provided by the chatbot (in terms of recognised intents) and incorporating only those questions into the test for understanding that could be answered based on the provided information. In total, 35 questions were created, consisting of yes-or-no questions, multiple-choice questions where one answer was correct, and fill-in-the-

blank questions. For all types of questions, the answers included the option "I do not know" and all the answer options were ordered alphabetically. This was chosen to increase the neutrality of the answer options and to avoid bias or favouritism towards a certain position in the answer list when creating the questionnaire. For the fill-in-the-blank questions, the participant was presented with the text with blank(s) as well as the answer options. In the case of multiple blanks in one question, the answer options contained answers for all the blanks, and a separate question was asked for each blank.

Note that the language used in the questions did not incorporate the alignment or non-alignment strategy used by the chatbot, and therefore, the questions may not always match the specific terms for placeholders participants have seen during the interactions. However, the answer options typically included the more common term for the placeholders (e.g., *Eardrum* instead of *Tympanic membrane*). Additionally, some answer options included a brief clarification, such as for the term *Pressurisation* where the text "*increasing the pressure*" was added to ensure participants from both conditions would understand the answer option. Additionally, participants could have asked for clarification of terms used by the chatbot during the interaction in case they did not understand them, which leads to the assumption that participants understand the more common terms for placeholders used in the test. Therefore, any differences in terminology between the questions in the test and the terms used during the interaction were not expected to pose a significant problem. An overview of all the questions defined for the test can be found in Appendix D.1. Note that these are all the questions, and only the relevant questions based on the provided information by the chatbot were selected and presented to the participant, again using Qualtrics to create the questionnaire [70].

Trust

In order to be able to draw conclusions about the perceived trust in the chatbot with which the participant interacted, the Human-Computer Trust Model (HCTM) scale to measure trust in human-computer interaction developed by Gulati et al. [34] was used. Empirical results obtained using the HCTM scale showed its validity, reliability, and predictive power, suggesting its usefulness in human-computer research with trust as a primary outcome [34, 83]. The HCTM scale consists of 12 questions where participants rate each item on a 5-point Likert scale (1 indicating strongly disagree, 5 indicating strongly agree) [34, 35, 83]. The model is based on three factors: risk perception, competence, and benevolence [34]. Risk perception is related to how likely the user thinks a problem occurs when using the chatbot and how concerned they are about this. A question related to the perceived risk is: *I believe that there could be negative consequences when using MedWiseBot*. Benevolence is related to the chatbot being able to provide the user with the appropriate responses to eventually help the user reach their goal of the interaction. A question related to benevolence is: *I believe that MedWiseBot will act in my best interest*. Competence is related to the perception that the chatbot has all the necessary qualities and functionalities to achieve the desired outcome. A question related to competence is: *I think that MedWiseBot is competent and effective in offering support* [34]. Note that the ratings of the first three questions of the HCTM scale were inverted such that the higher the participant rates them, the less risk is perceived by the participant. An overall trust score was calculated using all the items on the scale. This was based on research done by Pesonen [35] that made use of the HCTM scale to obtain an overall trust score in order to measure the trust of students in a chatbot that proactively offers academic and non-academic support. The adjusted scale can be found in Appendix D.2. All original items are used, where in each item the ellipsis (...), created by Gulati et al. [34] to make the scale easy to use, is filled in with the information fitting the current research. An example of this is the following question: *I believe that (...) will act in my best interest*, where the ellipsis was filled in by the name of the chatbot

in this research, *MedWiseBot*.

4.7.2 Qualitative measures

In order to find the influence of alignment with the user during the interaction with the chatbot, qualitative measures were also used. The qualitative measures included a small, semi-structured interview after the interaction. This provides a more complete understanding of the participants' experience during the experiment. In turn, this could give valuable insights about the factors that were investigated by using the quantitative measurements (understanding and trust). The questions for this semi-structured interview can be found in Appendix E. The first few questions were based on the User Experience Questionnaire, which has proven to be a reliable questionnaire to gain insights into the user experience of interactive products [84]. The next five questions were based on statements from the Chatbot Usability Questionnaire, which is a questionnaire designed to measure the usability of chatbots specifically [85]. The final questions were added to gain some general additional information about the interaction the participant had with the chatbot as well as their opinion about the use of a healthcare chatbot. Specifically, the question asked related to the potential future use of healthcare chatbots was included in the interview in order to indirectly obtain the participant's overall perceptions and satisfaction with their interaction with the chatbot. By asking about possible future use, leading questions were avoided, and a genuine reflection on the most important aspects, according to the participant, of such a healthcare chatbot was obtained.

4.8 Conclusion

This chapter discussed the method used in this research to investigate the impact of lexical alignment on participants' understanding of the provided information and their trust in the chatbot providing this information. The choice of disease and treatment was intentionally chosen to facilitate meaningful interactions between participants and the chatbot, simulating real-world scenarios in which a patient seeks information about a medical condition and treatment. With close collaboration with an expert in the field, it could be ensured that the information provided to the participants resembled real-world scenarios.

The identical experimental procedures across both conditions using a between-subject design guaranteed increased control of confounding variables such as the overall functionality of the chatbot. During the experiments, quantitative measures such as questionnaires for the understanding of the provided information and perceived trust in the chatbot were used. Aside from these quantitative measurements, qualitative insights were also obtained through a semi-structured interview.

5 REALISATION AND TESTING

In the experiment, participants got to interact with a chatbot (aligning either with the participant or using the template answers created with medical professionals) in order to get the information they wanted about the hypothetical disease and treatment. To account for the confounding variable of the general functionality of the chatbot, the only difference between the chatbots used in groups A and B was whether the chatbot aligned with the participant (condition A) or not (condition B). In the following, details on the realisation of the chatbot will be discussed, ending with pilot testing of the various steps in the procedure of the experiment to ensure a smooth execution of the experiment.

5.1 Platform

To start, the base of the chatbot developed in this project was chosen to be Rasa, which is a widely used open-source platform that allows the creation of chatbots [86]. It is capable of robust NLP techniques that facilitate precise intent recognition and entity extraction. Besides the fact that Rasa is open-source, the ability to integrate pre-trained models and packages is another advantage over other platforms that can be used for the creation of chatbots, such as Dialogflow, where the freedom to design certain behaviours using, for example, Python code is limited or not possible at all [87]. The use of external coding within the framework of the chatbot is an important aspect of the implementation of alignment in this project.

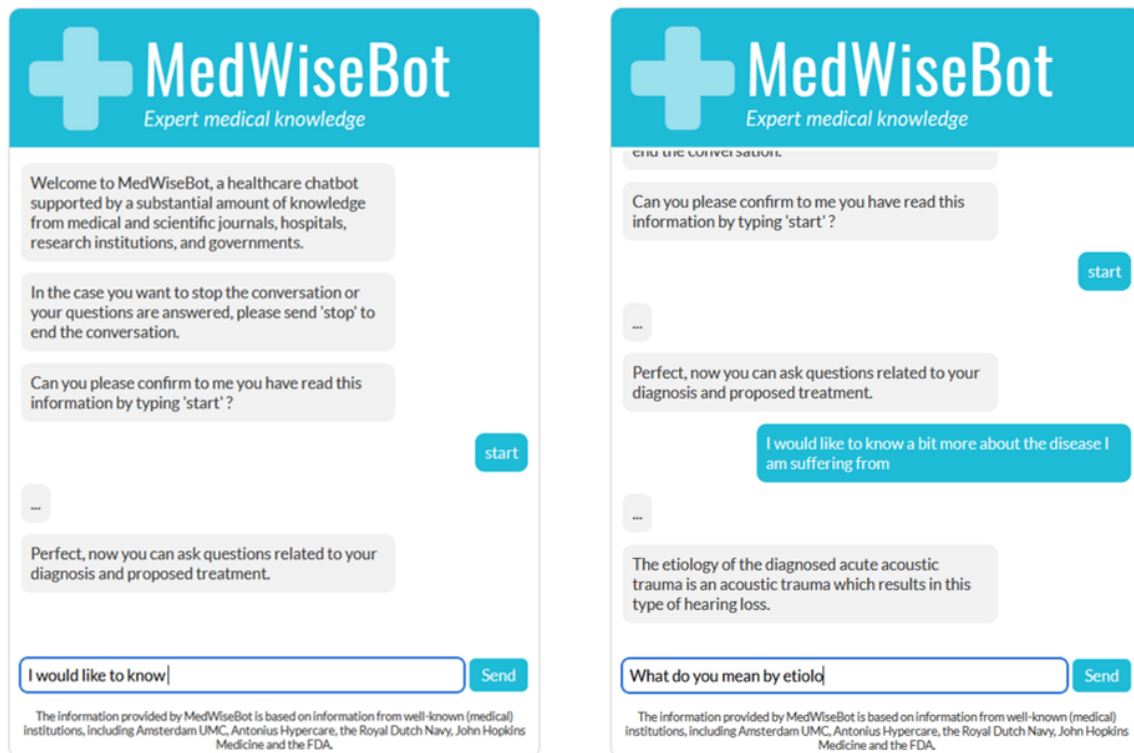
The internal configurations of the chatbot were handled by Rasa (.yml files). Additionally, several Python files were employed to account for the actions of the chatbot, including the implementation of alignment. Aside from this, an interface was created for the participant to interact with the chatbot. This interface was built using JavaScript, CSS, and HTML.

For the participants to be able to interact with the chatbot through a designated link provided by the researcher, the chatbot needed to be able to be deployed on the web. In order to achieve this, Netlify in combination with GitHub was used [88, 89]. Additionally, to ensure communication between the chatbot and the Rasa server as well as the actions.py file, ngrok was used [90]. In this setup, the Rasa server and the actions.py file run locally on the researcher's machine. Ngrok creates a secure tunnel that makes these local services public and allows interaction between the chatbot and the local server. In this way, the participant could be provided with a link from the researcher and be able to interact with the chatbot by opening the link, despite the backend components running locally during the experiment. It is important to note that this only allows interaction with the chatbot at times when the researcher is able to run the appropriate files locally. However, this did not conflict with the design of the experiment, as explained in Sections 4.1 and 4.5.

5.2 Interface

The interface of the chatbot was designed to be a rather simple and basic interface, for which several design choices were made related to colour choice, layout of the interface, and the name of the chatbot. In Figure 5.1, the initial interface of the chatbot is shown, and in the following, the design choices are discussed.

Figure 5.1: Interface of MedWiseBot



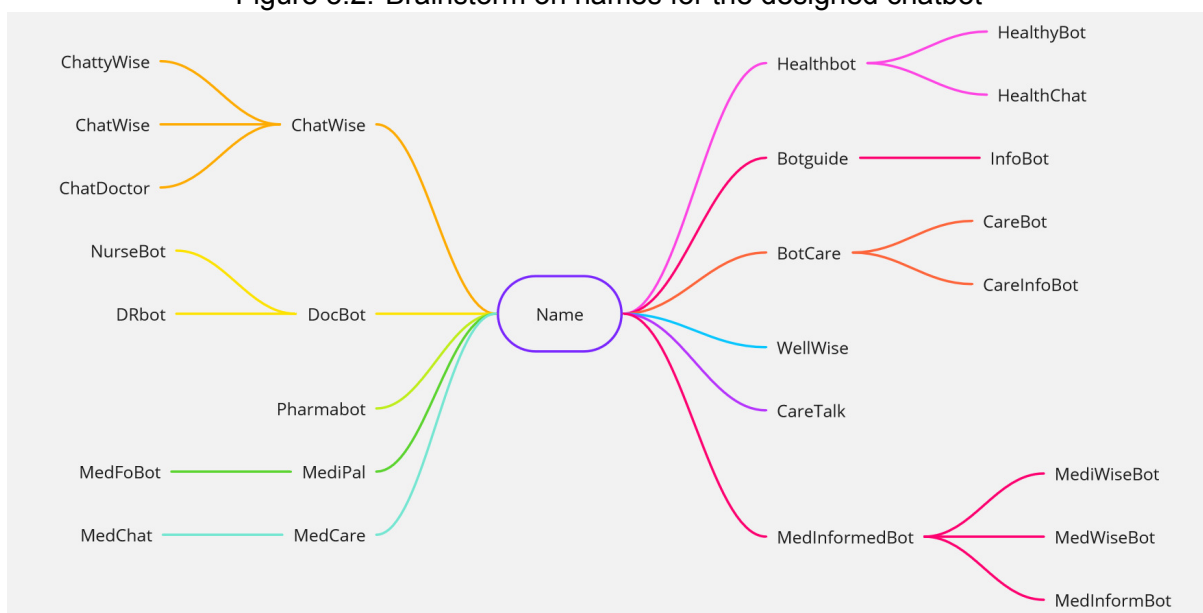
Looking into the properties of colours, the colour blue was chosen as the main colour used in the interface. Various studies looking at the attributes of colours have indicated that the colour blue has a calming effect [91, 92]. This makes blue a well-thought-through choice for the interface in the context of a healthcare-related conversation, as it increases the feeling of being at ease while interacting with the chatbot, which will contribute to its overall effectiveness.

The layout of the interface was based on various chat(bot) interfaces that were available at the time of creating the chatbot to limit difficulties in use [40, 41, 47, 46, 93]. The text input bar was positioned at the bottom of the interface, which is the place where the user could type their message and dispatch it to the chatbot. The input bar was accompanied by a "send" button that enabled the user to dispatch their message. Note that pressing the enter key had the same functionality. Within the chat window, the user could see their history of messages exchanged with the chatbot. In case the conversation became too long to fit on the screen, a scroll bar appeared to the right in order to be able to navigate to older messages. The messages of the chatbot were left aligned, having a light grey colour, whereas the messages of the user were right aligned, having a blue colour.

A name was also created for the chatbot shown at the top of the communication window and used in the first few sentences the chatbot provided to the user. In these sentences, the chatbot

introduced itself and its functions. Various chatbots in the realm of healthcare that were already discussed in Section 2.2 all have a name attached to them. Examples include *Ada Health*, *Buoy Health*, *Infermedica*, and *Healthility* [38, 40, 41, 42]. From these names, it becomes clear that most of them include some reference to the healthcare domain. After a short brainstorm session (see Figure 5.2), the following name was created and chosen for the chatbot in this research: *MedWiseBot*. This is a combination of the words *medical*, *wise*, and (*chat*)*bot*. Obviously, the word *medical* is a reference to the domain. The word *wise* is a synonym for *intelligent* or *informed* which are all relevant terms for the context in which this chatbot was used: to help the patient obtain the information they need about the proposed treatment and diagnosis. The word *bot* is a reference to the fact that the application is a chatbot. Furthermore, the name *MedWiseBot* is gender neutral, as it does not explicitly suggest any gender but rather the goal of the chatbot. This is an important aspect, as it limits gender stereotyping and increases inclusivity.

Figure 5.2: Brainstorm on names for the designed chatbot



5.3 Chatbot as expert

It is important to highlight that despite the primary focus of this research being on the effects of lexical alignment within healthcare chatbots, the concept of epistemic authority was also considered, as discussed in Section 2.4. Recognising the role a doctor’s perceived epistemic authority has on the way patients rely on the provided information, the chatbot is chosen to be presented as an expert on medical knowledge. In order to achieve this, several aspects were taken into account. It is important to note that according to literature, the notion of perceived expertise a doctor has can often be attributed to their qualifications, title, and expertise [54, 55]. In order to achieve something similar in the chatbot, introductory sentences were used to showcase the chatbot’s expertise by relying on medical resources. The following introductory sentence was used for this purpose:

- *Welcome to MedWiseBot, driven by state-of-the-art technology supported by a substantial amount of knowledge from medical and scientific journals, hospitals, research institutions, and governments.*

Additionally, there was a disclaimer added at the bottom of the interface. This was chosen as, after several questions asked and answers provided during the interaction, the first message from the chatbot will move to the top and will not be in the direct view of the participant anymore. The following text was used in the disclaimer:

- *The information provided by MedWiseBot is based on information from well-known (medical) institutions, including Amsterdam UMC, Antonius Hypercare, Royal Dutch Navy, John Hopkins Medicine and the FDA.*

5.4 Dialogue design

The two versions of the chatbot in the proposed research share the same intents, however, the responses generated vary based on whether they align with the user or not, as will be discussed in Section 5.5. The subsequent sections will provide more information on defining the various intents and the process of creating responses (template answers).

5.4.1 Intents

The various intents the chatbot should recognise and use were a crucial aspect of building and testing this chatbot. In order to define the intents, common queries or information provided regarding the disease and treatment were accumulated. In order to find those, several informative websites were investigated. The sources used include pages of medical information centres, governmental information pages (United States and the Netherlands), hospital information pages, patient brochures, as well as scientific papers, all focusing on the proposed disease and treatment [73, 74, 75, 76, 77, 94, 95, 96, 97, 98].

Overall, the sources contained similar structures in what information was provided. Besides this, some of the visited websites contained a "frequently asked questions" page that was used to find other common user queries. After a general layout of topics present in the various sources, the information and queries were categorised in order to find potential intents that should be incorporated into the chatbot. This process is visualised in Figure 5.3. It is important to note that the goal of the interaction for the participant was to obtain information regarding the treatment. However, it might be possible that one needs some more information related to the disease in order to understand the treatment properly, or wants to know more about the disease in general. Therefore, a similar visualisation was created for the intents related to the disease in Figure 5.4. Note that the intents that are split into several other (more detailed) intents were still incorporated in the chatbot individually. This means that when one asks for the procedure of the treatment, one will get a general explanation of the treatment. However, when asking more specific questions related to, e.g., clothing that one is allowed to wear during the treatment, an answer related to clothing in more detail is provided. Furthermore, several more general intents were created that are part of terminology related to the anatomy of the ear, related to aspects of physical examinations during the treatment, described in the scenario document the participant received before interacting with the chatbot, or related to the general dialogue structure. These are visualised in Figure 5.5. A fallback intent was created to account for the participants asking something to which no intent was coupled. The fallback intent tells the participant the chatbot was not able to understand what was said and asks the participant to rephrase the question asked. The general intent *Question ideas* was created mostly in order to account for the soft threshold of the minimum number of questions as explained in Section 4.1, however, participants might also want to know what frequently asked questions are or whether they have missed some important questions, for which this intent could also account.

Figure 5.3: Specified intents based on common queries related to the treatment

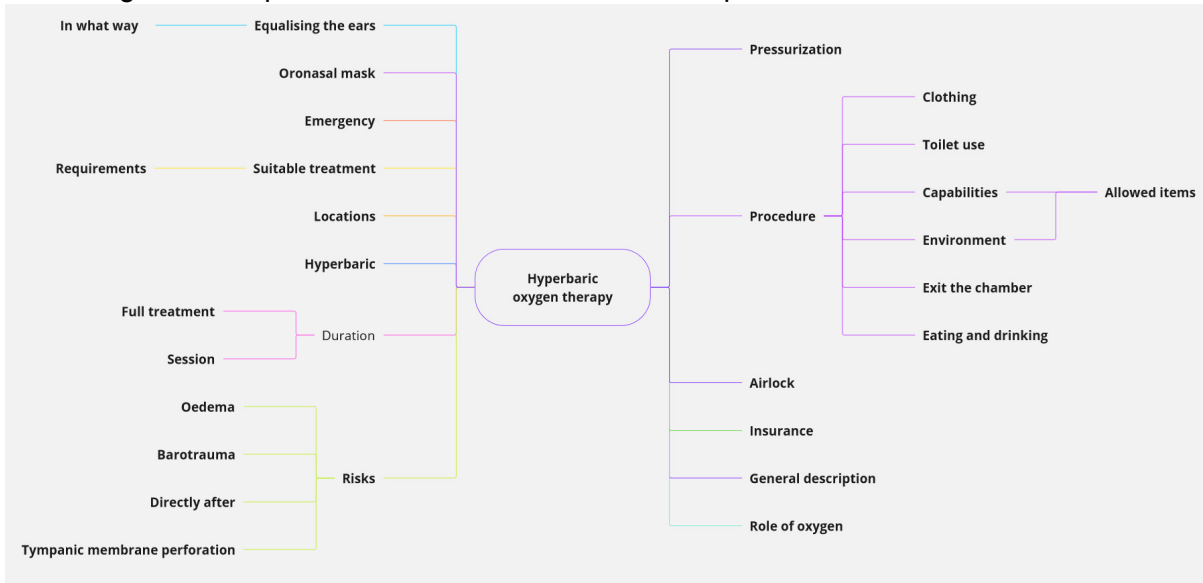
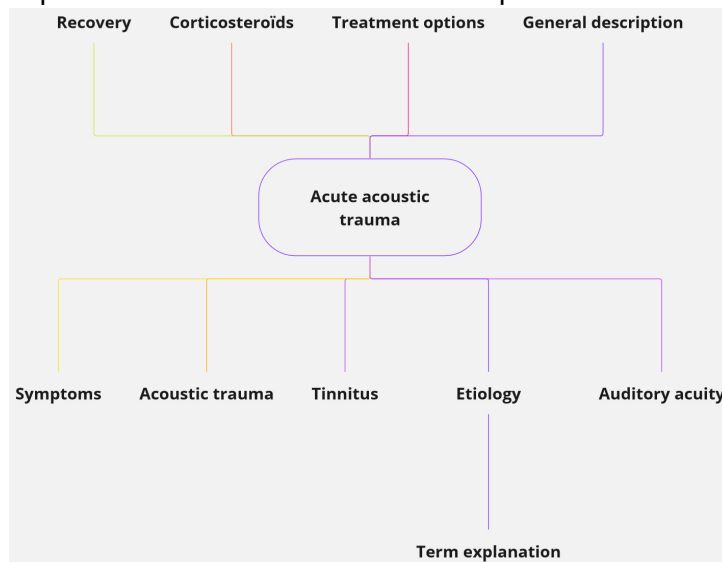


Figure 5.4: Specified intents based on common queries related to the disease

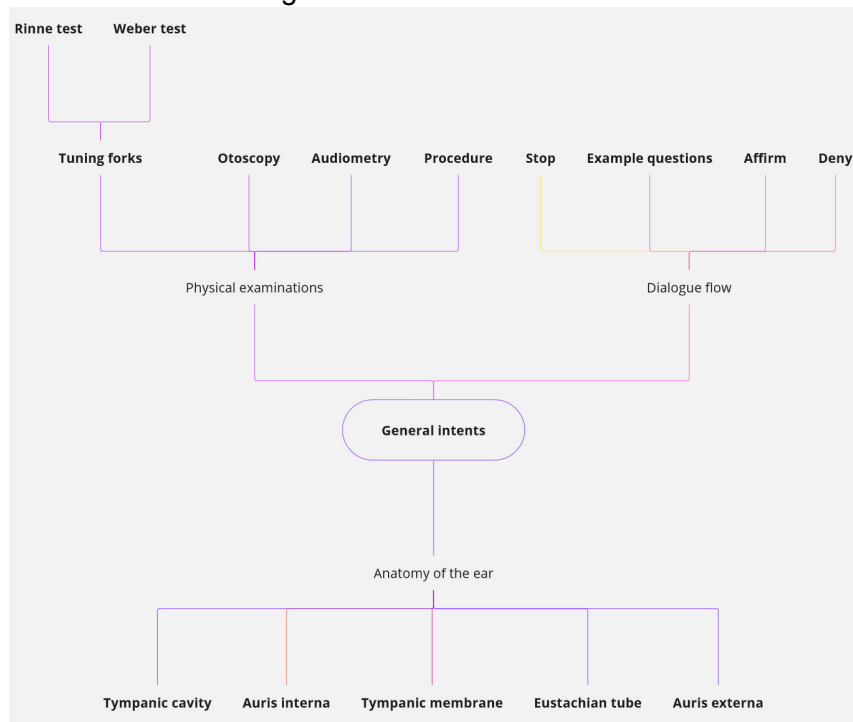


5.4.2 Training data

The training data used for the several intents specified in the chatbot was generated using a combination of methods. Initially, a set of samples was defined, after which ChatGPT [32, 99] was used to identify additional samples that were still missing. Following this, paraphrasing and synonym replacement techniques were applied to increase samples even further. Despite the possibility of using real user queries or crowdsourcing, the decision to use ChatGPT instead was based on factors such as time efficiency and the notion of the expert interview that was done in the research. The expert interview also provided insights into common user queries and question patterns. Moreover, preliminary testing was conducted, which provided additional insights into the way intent specification was implemented (see Section 5.6).

ChatGPT was used by first outlining the scenario of the proposed treatment and disease, followed by the intent explanation. After outlining the scenario and intent, ChatGPT was prompted to generate a list of possible questions a patient might ask related to the intent. The generated

Figure 5.5: General intents



samples were thoroughly reviewed, and relevant samples were incorporated into Rasa with the appropriate use of placeholder(s), which will be elaborated on in Section 5.5.3. Additionally, paraphrasing and synonym replacement techniques were used to further expand the training data. The website used for this was QuillBot [100], which is an online writing platform that employs AI technologies and offers various tools, including paraphrasing [101]. When using the paraphrasing tool, one must enter a sentence and QuillBot paraphrases this by adjusting certain words or by applying structural changes to the sentence. It is important to mention that the paraphrased sentences were, again, carefully checked and only added to the training samples in case they were truly different compared to the other samples, to limit overfitting. Moreover, they were checked to still fit the context and domain of the intent.

5.4.3 Responses

For each intent, an answer template was created using the information from the sources used to define the intents as well as the sources used to create the summary report in Section 4.1.2. An initial outline of the answer template was made, integrating accurate and contextually correct medical terminology. To ensure the correct use of medical terminology in the template answers, several resources were employed. Firstly, the expert interview included questions related to commonly used medical terminology when conversing about the disease and treatment. Secondly, MedlinePlus, specifically the medical encyclopaedia, was used [102]. MedlinePlus is a service of the National Medical Library (NML), which is part of the National Health Institutes (NIH) [102]. By combining the information from both of these sources, medical terminology could be accurately used in the answer templates. Finally, a final review of the template answers was conducted to ensure they were fluent, accurate, and fitting with the intent.

Below, a few template answers are shown to provide the reader with an insight into the types of answers the chatbot provided to the participant during the interaction. In these template answers, the placeholders are shown in brackets [], which were replaced with the preferred term

of the participant in the aligning condition, as will be discussed in Section 5.5.

For the intent of *Hyperbaric*, the following answer template was created:

- Hyperbaric is the use of greater than normal pressure, in case of the [hyperbaric oxygen therapy] proposed for the [diagnosed] [acute acoustic trauma], this is about the increased pressure of [oxygen].

For the intent of *Toilet use*, the following answer template was created:

- The [patient] undergoing the [hyperbaric oxygen therapy] may need to use the [restroom] when being in the [hyperbaric chamber]. During the [hyperbaric oxygen therapy], one is, in essence, not able to make use of the [restroom].

In case it is not possible to wait till after the process of [depressurisation] of the [hyperbaric chamber], one can use a provided bucket during the five-minute break between the [sessions] of [oxygen] administration to “use the [restroom]”. This can be done in a separate small compartment where the [patient] has privacy from the others in the [hyperbaric chamber].

For the intent *Procedure*, the following template answer was created:

- First is the preparation in which the [patient] changes to the appropriate [attire] before entering the [hyperbaric chamber].

Then the process of [pressurisation] starts which will start when the [patient] gives the “thumbs up” sign. During the process of [pressurisation], the [patient] needs to equalise the [tympanic membrane]. In case the [patient] is not able to do this, the “thumbs up” sign needs to be stopped and a [nurse] provides assistance.

After the process of [pressurisation], the [patient] will be wearing an [oronasal mask] to facilitate [inhaling] 100% (pure) [oxygen].

At the end of each of the [sessions], the process of [depressurisation] is started. After this, the [patient] is allowed to [exit] the [hyperbaric chamber].

Following each of the [hyperbaric oxygen therapy] [sessions], otoscopy is [conducted] to assess the [tympanic membrane], as well as the Weber and Rinne test to keep track of the degree of [hearing loss].

5.5 Implementing alignment

In Chapter 3 alignment in human-human interaction was already discussed, as were the effects of implementing alignment in human-agent interaction. The following will discuss the implementation of alignment in this research.

5.5.1 Alignment strategy

The chatbot used in the experiment employed lexical alignment, utilising the substitution method described by Spillner and Wenig [8]. The chatbot had template answers that were formulated similarly to how information is usually conveyed by medical professionals. This was done in

cooperation with a professional on the topic of hyperbaric oxygen therapy and the disease of acute acoustic trauma (see Section 4.4). The substitution method from Spillner and Wenig [8] entails that the template answers the chatbot has are adjusted to the terms used by the user. In this research, this was done by defining several placeholders in the answer templates that could be substituted with the input from the user. For example, *disease* could be swapped with *illness*. The placeholders used in the substitution method were defined based on their relevance to the medical language found in the sources used to create the answer templates, as discussed in Section 5.4.3. The placeholders were selected based on their frequent occurrence in the medical information sources and their multiple common synonyms or alternative terms. Examples include *diagnosed*, *indications*, and *hearing loss*. In order to account for the immense variability of terms a user can use for the placeholders, Spillner and Wenig [8] make use of word embeddings in combination with similarity testing in their substitution method.

Word embeddings are vector representations of words that can be used to look at semantic similarity. The research by Spillner and Wenig [8] makes use of SpaCy's 300-dimensional GloVe word vectors [103]. GloVe is an algorithm used to obtain a vector representation of a word. This vector representation allows similarity testing with other words. The embeddings of GloVe are based on large-scale text corpora capturing the way words are used in general language use (for example, based on data from Wikipedia or Twitter) [103]. Although GloVe word vectors could have been used in the current research, the context in which the chatbot was applied should be taken into account. Given that the conversation between chatbot and participant was of a medical nature and the template answers included medical terminology, the suitability of GloVe is questionable. Considering the training data of GloVe is based on general language use, problems might arise related to the limited representativeness of medical terminology or domain specificity. To address these limitations, BioWordVec embeddings were used instead [104, 105].

BioWordVec is a set of biomedical word embeddings consisting of 200-dimensional vectors computed using the fastText algorithm [104]. By using the fastText algorithm to compute the word embeddings, BioWordVec integrates subword information from unlabelled biomedical data, incorporating the widely-used biomedical ontology Medical Subject Headings (MeSH) [104, 105]. MeSH is a controlled English vocabulary thesaurus that can be used to index, catalogue, and search biomedical and health-related information [106]. The data used in this research was the embedding file *bio_embedding_intrinsic*, which can be applied for similarity calculations and consists of words from MeSH as well as PubMed [104, 105]. PubMed is an open interface that can be used to search through the MEDLINE database, which is a bibliographic database from the National Library of Medicine's, containing references to biomedical literature as well as science journals and books [107, 108].

5.5.2 Terminology repository

Before discussing the substitution method used in the alignment strategy in more detail, it is essential to understand how the chatbot gathers the terms for this substitution. Ideally, the alignment strategy would rely solely on the input provided by the user during the interaction with the chatbot. However, this would lead to limited alignment if the user did not provide alternative terms for the specified placeholders. To address this problem in the current research, a terminology repository was created. This repository serves as a backup source of terminology that can be used in the substitution method to align the chatbot's template responses with the preferred terminology of the participants without requiring the participant to explicitly mention specific terms.

The terms chosen by the participant in the questionnaire on (medical) terminology discussed in Section 4.1.1 were stored in a file, which was saved. The terms were loaded into the Python action file to check whether each term was suitable to replace its placeholder using the two strategies of substitution explained in the following, Section 5.5.3. To load the terms from the questionnaire into the Python action file, an additional intent, *begin_conv*, was created. This intent was activated when the participant entered the term "start" at the beginning of the interaction. Participants were prompted to enter "start" in order to begin the interaction with the chatbot, as can be seen in Figure 5.1. This was similar for the aligning and non-aligning versions of the chatbot. In the case of the non-aligning chatbot, there was no action attached to the intent, and there would only be the response from the chatbot, as shown in Figure 5.1. Note that if the participant used a different term during the interaction with the chatbot compared to the one saved in the repository and used in the answers, this term replaced the saved term (in case this term was regarded as suitable based on the discussed strategies). It is important to note that the participants' word usage was not derived for all the placeholders present in the template answers. Not all placeholders were incorporated in the questionnaire due to the large number of placeholders present in the template answers and the likelihood that participants would naturally refer to some of the placeholders in their own terminology in their input. For instance, in the case of the placeholder being *hyperbaric oxygen therapy*, it is likely the participant would refer to it in their own words when asking a question about the treatment, like, *Can you tell more about the treatment?*

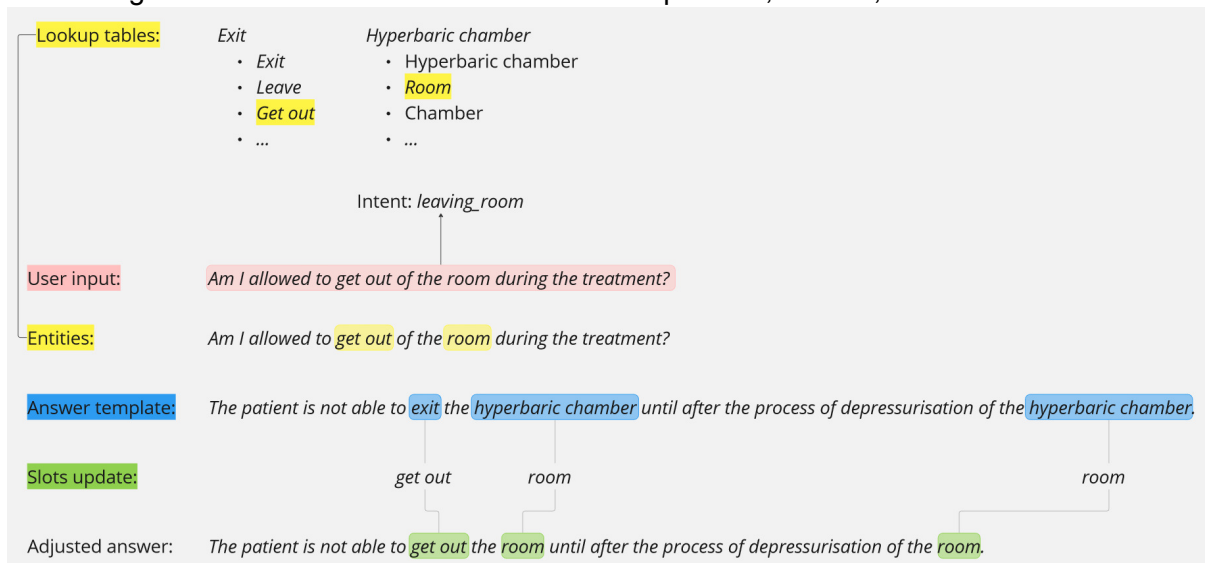
5.5.3 Substitution

For the implementation of substitution, two strategies were employed. Firstly, during the design of the intents in the chatbot, several placeholders in the template answers were defined. In Rasa, this was facilitated using the lookup tables, entities, and slots. Lookup tables can be used to extract entities by providing a list of possible values for that specific entity. The terms in these lookup tables included common synonyms or alternative terms for the placeholders [109]. These terms were taken from the terminology often used in the various sources used to create the intents and answer templates discussed in Sections 5.4.1 and 5.4.3. Entity slots in Rasa act as the memory of the chatbot, which means that the value for a certain entity (the placeholder in the answer template of the chatbot) was saved at the appropriate slot and used throughout the conversation until a new value replaced it. In this way, the answer templates of the chatbot were dynamically adjusted to align with the participant's terminology.

Initially, the answers saved from the questionnaire on (medical) terminology, serving as a terminology repository, were checked against the lookup tables. If the user chose a term for a placeholder that was present in the lookup table, it was identified and stored in the corresponding entity slot. The stored term was then used in the answer by replacing the original value (placeholder) in the template answer. This adjusted version of the answer was then outputted onto the screen of the user. A visualisation of this process using lookup tables, entities, and slots in Rasa is shown in Figure 5.6. Note that this is not an exhaustive example, and only a subset of the lookuptables, entities, and slots is shown to maintain clarity and oversight.

The second strategy was used to account for the enormous variety of terms a user can choose for certain words and was largely based on the research from Spillner and Wenig but made use of the BioWordVec embeddings instead of the GloVe embeddings [8]. Note that this strategy only became active when there was no match between the entered terms in the input of the user and the values that were part of the lookup table defined for that placeholder. First, the

Figure 5.6: Visualisation of the use of lookup tables, entities, and slots in Rasa



punctuation was removed from the user input, after which the input was tokenized based on whitespaces to obtain a list of terms used by the user. It was chosen to remove the punctuation to increase the extraction of meaningful information from the input. More specifically, the last term of the original sentence *Can you tell more about the disease?* would be *disease* instead of *disease?*. Additionally, to reduce unnecessary computations and increase efficiency, each term was checked to be part of the set of NLTK stopwords for the English language (words such as "a", "the", or "in"). If the term was part of the set of stopwords, this term was excluded in the subsequent steps. After this, each of the relevant terms was transformed into a vector representation using the BioWordVec word embeddings. Then, the cosine similarity between each term and the placeholder was calculated using these embeddings. Finally, it was checked whether the similarity of the term(s) and the placeholder was above a defined threshold. If this was the case, the placeholder in the template answer was substituted with the term of the user and this term was saved in the slot corresponding to the entity of the placeholder. The threshold can be seen as a precision insurance such that in the case the word used by the user was not similar enough (according to the cosine similarity between the two word vectors), it was not replaced. This threshold was set to 0.74 based on testing with the various terms present in the lookup tables and terms that should not be replaced. Note that the similarity measure based on the vector representation of the placeholder and entered term was also used to check whether manually entered terms, obtained from questionnaire on (medical) terminology saved in the terminology repository, were similar enough to be saved in the corresponding entity slot and used in the substitution. To account for multi-word expressions that were not present in the lookup tables but could potentially be similar enough to be used in the answer template, combinations of a term with the preceding, succeeding, and both preceding and succeeding terms in the input were also considered. While this implementation would not capture all types of multi-word expressions, the expectation of participants using such expressions was relatively low. Additionally, pilot testing would determine if this basic implementation for checking similarity with multi-word expressions required further refinement.

Note that in cases where the chatbot was not aligning, both of these strategies were skipped and the template answer using the placeholder was used as a response from the chatbot. A visualisation of the explained processes for both the non-aligning chatbot and the aligning chatbot, including the use of the terminology repository, can be found in Appendix F.

5.5.4 Sentence refinement

After the substitution of the placeholder(s) in the template answer with the term entered by the user, the resulting sentence may not always be grammatically correct. In order to address this, functions were created to rectify grammatical mistakes that arose due to the substitution method used to achieve lexical alignment. Only after these functions were called to refine the sentence the chatbot intended to output, the refined answer was put on the screen for the user.

The first aspect checked was the use of the articles "a" or "an". It might occur that the term that was substituted requires a different article than the one being inserted. For example:

Template answer: ... *wearing **an** [oronasal mask] to ...*

Term entered for placeholder: *breathing mask*

After substitution: ... *wearing **an** [breathing mask] to ...*

In order to achieve grammatical correctness, a Python function was created to check the basic rules on the use of the articles "a" and "an" in the English language. The function determines whether the term following the article starts with a consonant (requiring the article "a") or a vowel (requiring the article "an") and adjusts the article if necessary. Note that this method of checking does not include exceptions, such as the use of "an" before nouns like "hour" which start with a consonant but sound like a vowel and therefore require the article "an". However, the expectation was that this function would be sufficient in the current research and pilot testing would show whether additional adjustments were necessary.

The next aspect checked was whether an entered noun by the user should be in plural or singular form. It might occur that the user has entered a form of the noun that is not the form that the answer template is expecting. In order to make sure this influences neither the similarity measure used to decide for substitution with that term, as explained in Section 5.5.3, nor the grammatical structure of the adjusted template answer, a function was created to address this issue. An example of this issue is presented here, where the entered term is in singular, whereas the template sentence expects the term to be in plural form:

Template answer: ... ***are several** [complications] that might occur ...*

Term entered for placeholder: *side effect*

After substitution: ... ***are several** [side effect] that might occur ...*

To make sure the term is adjusted to the correct form, either plural or singular, the form of both the placeholder and the entered term by the user were compared. In the case that these were not the same, the entered term was adjusted to match the form of the placeholder. After this, the term (adjusted to the form of the placeholder if necessary) was used in the similarity measure discussed in Section 5.5.3. To check whether a term is in plural or singular form, the *WordNetLemmatizer* from NLTK in Python was used [110]. Additionally, the *inflect* Python module was used to obtain the plural or singular form of a term [111].

5.6 Pilot testing

After the creation of the chatbot, pilot testing was done to make sure everything functioned as it was expected to. Aside from the chatbot itself, the additional steps in the experiment procedure, such as the use of the different questionnaires designed for this research, were also incorporated into the pilot testing to ensure smooth execution of the experiment after testing.

5.6.1 Questionnaire on (medical) terminology

To ensure the functionality of the questionnaire on (medical) terminology was as expected, testing of this aspect was conducted separately. Results were examined to ensure the clarity and understanding of all the questions, as well as the objectives of the questionnaire (it being for the participant to fill in the terminology they would use for the asked concept). Additionally, in cases where the participant entered a term that was not a predefined answer, these terms were tested in the substitution strategies explained in Section 5.5.3.

The tests showed that the majority of the terms chosen for specific concepts in the questionnaire varied across different participants. This is an important aspect, as it shows the relevance of aligning these terms per participant. Additionally, no participants filled in different terms than the predefined answers, so no additional testing of substitution for the terms could be done at this stage of pilot testing. However, this strengthens the expectation that the basic implementations to account for grammatical correctness for the article use (“a” or “an”) or the use of multi-word expressions were sufficient.

The tests also showed participants were taking approximately 9.19 minutes ($n = 6$) to fill out the questionnaire, which seems reasonable considering the number of questions and the estimated time for the full experiment.

Only a minor adjustment was made to the questionnaire, which is related to the introductory message at the start of the questionnaire. During the pilot tests, it was noticed that several participants had the feeling that the questionnaire was a “test” on medical or English terminology, even though the questions were formulated in such a manner that participants were asked to choose the term or concept that *they* would use. Therefore, the introductory message was changed to: *Try to choose the answer that fits your own word use as best as possible, try to answer quickly and not think too in depth. There are no wrong answers!*, to try and limit this stress effect of it being a test.

After another pilot test with this new adjustment, participants still often tried to choose the “most correct” answer, while this was not the goal of the questionnaire, instead, the questionnaire was used to try to infer the preferred terminology used by the participant. Therefore, another small adjustment was made in the phrasing of the questions. Instead of only adjusting the introductory message to include the notification that there were no wrong answers, which participants could not read after they started the questionnaire, the phrasing of the questions was adjusted to include the following: ... *you would prefer to ...* instead of only ... *you would ...*. For example: *Please select the term you would prefer to use to refer to the people in the picture* instead of *Please select the term you would use to refer to the people in the picture*.

Another finding during the pilot testing related to the questionnaire was that participants often did not ask for further clarification when encountering difficult (medical) terms used as placeholders in the chatbot’s responses. They stated to “remember the context of the term” from the questionnaire because the placeholders were among the answer options in the questionnaire. These priming effects led participants to recognise the terms without a full understanding of the concept behind them, as was shown in their answers provided in the understanding test afterwards. After some further testing, this finding was confirmed, as more participants had similar experiences. Therefore, it was decided to remove certain placeholder terms from the answer options in the questionnaire. Note that this adjustment applied particularly to the concepts that are less known or medical terminology that participants might not be familiar with. However, participants who would prefer this more complex terminology could still use it by choosing the answer option “other” and entering their preferred term. In this way, the priming

of the more difficult terminology or unknown concepts was limited, without the participants being limited in the use of such language. The following placeholders were removed from the answer options: *Oronasal mask*, *Pressurisation*, *Depressurisation*, *Etiology*, *Auditory acuity*, *Auris externa*, *Tympanic cavity*, and *Tympanic membrane*. Additionally, the answer option *Decreased auditory acuity* was also removed for the placeholder *Hearing loss*, as this was very similar to the placeholder auditory acuity. The adjusted questionnaire can be found in Appendix A.

Main adjustments made after pilot testing with the Questionnaire on (medical) terminology:

- *Adjusted introductory message of questionnaire*
- *Adjusted phrasing of questions*
- *Adjusted answer options to limit priming*

5.6.2 Non-aligning chatbot

The test for the non-aligning chatbot was mainly done to assess the fundamental functionality of the chatbot as well as the flow of the designed experiment. Participants were instructed in a similar manner as in the actual experiment. The participants were asked to fill out the questionnaire on (medical) terminology, and the summary report on the diagnosis and treatment was provided to them in order for them to become familiar with the topic. Then, the participant was asked to interact with the chatbot.

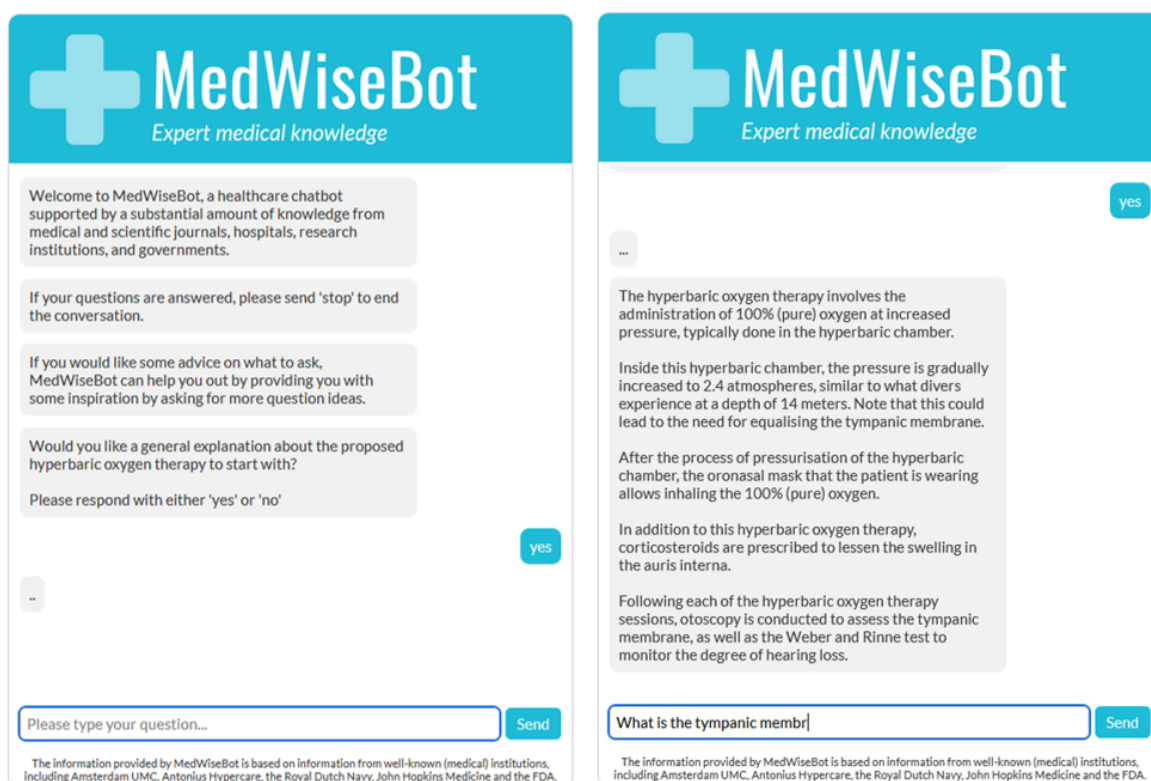
The first testing round indicated some possible improvements. At first, the fact that participants were only allowed to read the summary report at the start and then immediately needed to interact with the chatbot seemed to increase the stress and pressure on remembering the terminology used in the summary report. As a result, wrong terminology was used, which led to misunderstandings of the chatbot or the chatbot outputting the fallback intent instead of being able to properly understand the question asked. Examples of wrong terminology used were several combinations of partial terms from the disease combined with partial terms from the treatment. Additionally, during the interaction with the chatbot, it was observed that people were struggling to come up with questions to ask and did not really know what to think of. Despite the option to ask the chatbot for question options, this was not used as people were not thinking about this as an option.

Based on this, the summary report was adjusted to include a more basic story about the context of the interaction with the chatbot and a shorter overview of the more difficult part, which was the medical examination report. In this way, the participant was still confronted with more medical terminology, but at the same time, the document was easier to interpret and guided the interaction with the chatbot afterwards. Note that this version of the document was also checked and approved by the expert, as was the original version discussed in Section 4.1.2. The adjusted version can be found in Appendix B.2.

Additionally, to limit the stress participants experienced in the feeling of needing to remember the difficult terminology or aspects discussed in the summary report, it was chosen to incorporate a starting message from the chatbot to ask whether the participant would like a short summary of the proposed treatment. The participant could accept this explanation by typing "yes" after which the explanation of certain aspects present in the summary report was shown. Additionally, this explanation potentially included some new concepts or aspects of the treatment or disease, leading to additional questions on the side of the participant. In case the

participant did not want this explanation and typed "no", they could ask their own questions immediately. Note that this question replaced the original confirmation question, which had to be answered with "start" as discussed in Section 5.5.2. Aside from this, the possibility for the chatbot to provide example questions or question ideas was also added to the introductory messages to make sure participants were aware of this feature. These adjustments can be seen in Figure 5.7. The fact that in the pilot test, participants could not read the summary report when interacting with the chatbot also resulted in increased stress amongst the participants, so an additional line was added in the summary report explaining the chatbot could provide assistance during the interaction. This is the following line: *When you have read this document, please notify the researcher so that the interaction with MedWiseBot can start. Do not worry about remembering all the content of this document, as MedWiseBot will be able to offer assistance if needed.*

Figure 5.7: Adjusted introductory messages of MedWiseBot



Pilot testing with the adjusted version showed that participants were much less scared by the summary report and understood the context of the interaction with the chatbot better. Regarding the question from the chatbot to provide a short summary of the proposed treatment, this was perceived well, as participants did indeed want to have this overview and were happy they did not have to "remember" everything from the summary report. Moreover, this led to a natural flow of asking questions and providing answers from the chatbot.

Regarding the functioning of the chatbot itself, no issues were present during pilot testing. Some questions were not recognised, but after rephrasing, they were answered correctly. The additional inputs obtained during pilot testing were added as training data for the aligning and non-aligning chatbots.

Main adjustments made after pilot testing with the non-aligning chatbot:

- *Simplified summary report, see Appendix B.2*
- *Adjusted the introductory messages of the chatbot to also include the explanation of asking for question ideas*
- *Adjusted the introductory messages of the chatbot to ask for a short explanation on the treatment*
- *Added some sentences of reassurance and guidance in the summary report, see Appendix B.2*

5.6.3 Aligning chatbot

The test for the aligning chatbot was mainly done to evaluate the functionality and effectiveness of the alignment strategy implemented in this research. This test followed a similar structure as the test for the non-aligning chatbot, however, participants were also asked to fill out the questionnaire on (medical) terminology beforehand, following the procedure of the actual experiment. This was a necessary aspect to be able to test the alignment strategy, as the answers to this questionnaire were used for the terminology repository as discussed in Section 5.5.2.

Testing the alignment of the chatbot showed the strategies used and implemented for the alignment worked as expected. When asking the participants whether they noticed the alignment during the interaction, they often responded that they did not really notice this, which shows the smooth integration of the word substitution strategies involved in this research.

Regarding terms used by the participants in their questions, these were rather limited. Often, short questions were asked or questions were asked related to the information provided in the previous question by the chatbot. This showed the need for the questionnaire to infer the terminology preferred by the participants. Moreover, there were no urgent indications of the chatbot being too slow in responding during the pilot tests (taking into account the ideal circumstances in terms of power supply for the laptop, internet connections, etc.).

Looking into the implementation to account for multi-word expressions, it was decided to remove this part of the substitution method. Given the necessary adjustments made to the summary report that were mostly related to simplifying what was in there, as well as the limited input of alternative terminology for placeholders by participants during the interaction with the chatbot in the pilot tests, the likelihood of participants entering a multi-word expression was even further reduced. Additionally, as the lookup tables already include the most well-known terminology for the placeholders (including multi-word expressions), the expectation that a possible entered multi-word expression would be similar enough according to the similarity threshold to be substituted in the template answer was rather small. By summarising all these points, this part of the substitution method could be considered redundant in the current research and was therefore skipped. Removing this part of the code also improved its efficiency and improved the response time even further.

Main adjustments made after pilot testing with the aligning chatbot:

- *Removal of (basic) multi-word expression in the substitution method*

5.6.4 Understanding test

The test for understanding was evaluated in order to draw conclusions about the degree of difficulty of the questions asked. To assess this, participants were asked to complete the un-

derstanding test without any interaction with the chatbot. This helped to determine whether the questions are sufficiently challenging to draw conclusions about the level of understanding participants have after interacting with the chatbot and then filling out the test in the actual experiment. Note that the participant has had access to the summary report on the diagnosis and treatment in order for them to have a global understanding of the topic. Moreover, this document was also provided in both conditions of the actual experiment, influencing the baseline knowledge each participant has about the topic. Consequently, this must be considered when assessing the difficulty level of the test and drawing conclusions based on it. Note that in both pilot tests, for the non-aligning chatbot and the aligning chatbot, participants were asked the created test questions from the understanding test that were relevant to the interaction they had, as discussed in Section 4.7.1, assessing whether the questions were not overly difficult to answer even after interacting with the chatbot.

After the first round of testing, it came to light that the yes-or-no questions were a bit difficult to interpret. More specifically, terms such as "confusing" or "feeling unnatural" were used to describe the answers accompanying these questions. This led to adjusting the yes-or-no answers to true or false answers. Moreover, the fill-in-the-gap questions often led to confusion as the answers included the answer options for several gaps in case there was more than one gap to be filled in. Therefore, this was adjusted to only include the relevant answer options for each gap that needed to be filled in. Note that this led to some rephrasing or refinements in these specific questions.

Finally, the evaluation of the created understanding test did not show problems related to the fact that the questions and answers did not have the same language use as the chatbot the participants interacted with (aligning or non-aligning).

Main adjustments made after pilot testing with the understanding test:

- *Adjusted the yes-or-no answers to true/false answers*
- *Adjusted the fill-in-the-gap questions to only have the relevant answer options*

5.6.5 Online setup

In order to be sure the online setup as discussed in Section 5.1 was working properly, this was also tested. In order to test this, the aligning chatbot was tested in an online setting. It was chosen to specifically test the aligning chatbot as the implementation of the alignment required more steps and could potentially lead to slowing down the process or other problems.

The initial pilot test showed the online setup to be successful. However, responding to the participant's questions proved to be slower compared to the normal setup in which the participant could interact with the chatbot on the laptop of the researcher. Therefore, a second pilot test was done using the online setup, but now with the non-aligning chatbot. By running the non-aligning chatbot in the online setup, no big difference in the timing of the responses from the chatbot could be noticed compared to the normal setup. To ensure consistent response time during interactions, it was decided to only employ the online setup for the non-aligning chatbot, limiting the aligning chatbot to only be used in the offline setup. Note that the non-aligning chatbot was also used in the offline setup. Any potential influences of the online setup were assessed after reviewing the results in Chapter 6. Additionally, it was anticipated that the number of experiments conducted online would remain relatively limited.

Main adjustments made after pilot testing with the online setup:

- *Only use the non-aligning chatbot in online setup and use the aligning chatbot solely in normal setup*

5.7 Conclusion

This chapter provided an in-depth explanation of the various aspects considered during the development and realisation of the chatbot. The choice to use the Rasa platform as a start for creating the chatbot was based on its robust NLP techniques and open-source nature. Additionally, details on the implementation of the chatbot for use via an online web link were provided utilising Netlify, GitHub, and ngrok. Several design choices were discussed, including the interface, name, and way of presenting the chatbot as an expert. The creation of the various intents for the chatbot, the training data used, and the creation of the answer templates were also considered. Furthermore, a comprehensive explanation of the alignment method used for the aligning chatbot was provided, concluding with the results and refinements made after the pilot tests for the various steps in the procedure of the experiment.

6 RESULTS

In order to investigate the impact of the implementation of lexical alignment on users' understanding of information provided by a healthcare chatbot and their trust in the chatbot during an information-seeking task, both quantitative and qualitative data were collected. Quantitative measures included a questionnaire to assess the participant's understanding of the provided information as well as a scale to evaluate their perceived trust in the chatbot. Additionally, a questionnaire was conducted at the start of the experiment to infer the participant's word use. These quantitative data will be analysed in the first section of this chapter. The qualitative measurements involved a semi-structured interview conducted at the end of the interaction with the chatbot to gain an in-depth understanding of the participants' general opinions about the interaction, the aspects they consider relevant for healthcare chatbots, and their perception of the terminology used by the chatbot. The results of the interview will be analysed in the second section of this chapter.

6.1 Quantitative results

The independent variable in this research was categorical, indicating whether the chatbot made use of lexical alignment or not. The dependent variables in this research were understanding and trust, measured through a questionnaire including test questions for understanding and the HCTM scale for trust. The Shapiro-Wilk and Kolmogorov-Smirnov tests confirmed the normality of both datasets, and Levene's test indicated equal variances could be assumed for both datasets. Consequently, an Independent Samples t-test with a 95% confidence interval was employed to compare the results across conditions. The outcomes of the measures will be discussed in the following sections.

The total number of participants taking part in (and finishing) the experiment was $n = 24$ (hence, $n = 12$ per condition). One participant did not fully comply with the protocol of the experiment, failing the soft requirement of asking at least three questions during the interaction with the chatbot and intentionally asking unnecessary, difficult questions to the chatbot. Analyses of the results were conducted both including and excluding this participant's data. Given the minimal impact of the inclusion of this data and the limited sample size ($n = 24$), it was decided to keep this participant's data in the final analysis. Additionally, one experiment was conducted online. The results of this experiment did not show any discrepancies with the other results and were therefore kept in the final analysis.

6.1.1 Understanding test

The first variable analysed is understanding. Figure 6.1 shows the boxplot of the mean understanding score per condition. Note that the questionnaire to test participants' understanding of the provided information was adjusted based on the topics they discussed with the chatbot, resulting in variations in length and content between participants (see Section 4.7.1). To

ensure comparability, the scores obtained from these tests were normalised and presented as percentages. The scoring system used to calculate the final score per participant was as follows:

- Correct answer: +1 point
- Incorrect answer: -1 point
- "I do not know" response: 0

The formula for calculating the test score was:

- $\text{test_score} = (\text{raw_score} + \text{total_number_of_questions}) / (2 * \text{total_number_of_questions} * 100)$
- where $\text{raw_score} = (\text{questions_correct} * 1) + (\text{questions_incorrect} * -1)$

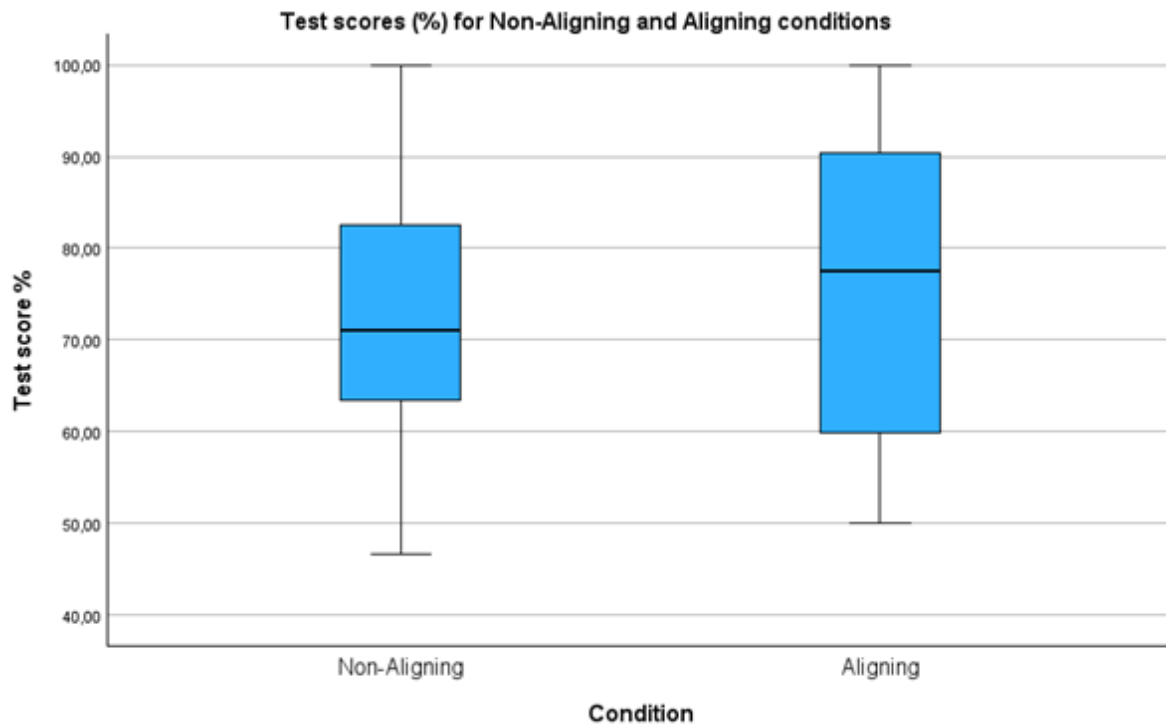
Note that in this scoring system, different scores were assigned to the answer option "I do not know" and a wrong answer. This is based on the assumption that acknowledging uncertainty reflects a certain level of understanding, whereas providing an incorrect answer indicates a misunderstanding. This approach is supported by a report on testing and test analysis of closed questions [112]. This report states that a scoring system of 1, 0, and -1 points in questions with two answer options would lead to more trustworthy results compared to a binary scoring system using only 0 or 1 points [112]. The research by Srivastava et al. [11] also supports this difference in scores, as their research made a similar distinction between partially correct answers and incorrect answers. In their scoring system, partially correct answers were awarded one point, incorrect answers zero points, and correct answers two points.

The boxplot in Figure 6.1 indicates that no outliers were identified by SPSS. Outliers were defined as data points beyond 1.5 times the interquartile range (IQR) below the first quartile or above the third quartile. The mean understanding scores were 71.73% for the non-aligning condition and 76.50% for the aligning condition. Both of these values exceed the chance level, confirming that the test questions were not overly challenging. Additionally, only one participant per condition scored 100%, indicating the test questions were sufficiently challenging. Statistically, the Independent Samples t-test found no significant difference in understanding scores between the aligning condition ($M = 76.50$, $SD = 17.09$) and the non-aligning condition ($M = 71.73$, $SD = 15.37$), $t(22) = .718$, $p = .480$. These findings do not support the hypothesis that lexical alignment with the user leads to an increase in understanding of the provided information by the chatbot. The effect sizes accompanying the Independent Samples t-test were Cohen's $d = 0.293$ and Hedges' $g = 0.283$, indicating a small effect. Typically, values around 0.2 indicate a small effect, around 0.5 indicate a moderate effect, and around 0.8 or higher indicate a strong effect. Considering Hedges' g is often proposed for smaller datasets, Cohen's d and Hedges' g were both considered and showed similar results. In conclusion, the difference in test scores across conditions was too small to reach statistical significance, despite the trend observed in the mean test scores and the boxplot in Figure 6.1.

6.1.2 Trust scale

The second variable analysed is perceived trust. Figure 6.2 displays the boxplot of the mean score for the perceived trust per condition as measured by the HCTM scale discussed in Section 2.3.2. To derive a total score for perceived trust from the HCTM scale results, the scores of the first three questions were inverted. This was achieved by using the following formula: $((\text{maximum_score} + 1) - \text{obtained_score})$. For instance, if a participant rates the first question

Figure 6.1: Test scores on the understanding test



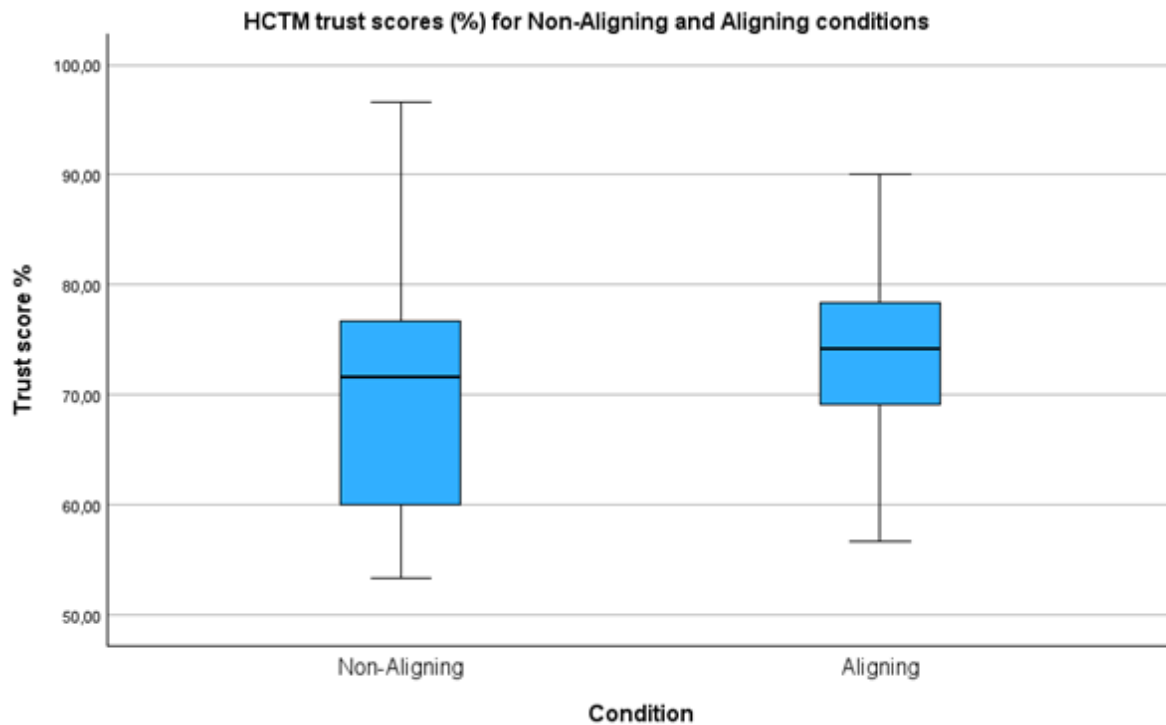
I believe that there could be negative consequences when using MedWiseBot with a rating of 4, indicating a perception of risk associated with the use of MedWiseBot, the inverted score would be $6 - 4 = 2$, to be used to obtain the overall trust score. This approach is based on the literature that utilises this same scale [34]. Additionally, the scores have been transformed into percentages to enhance interpretability.

The boxplot in Figure 6.2 indicates that no outliers were identified by SPSS. Again, outliers were defined as data points falling beyond 1.5 times the IQR below the first quartile or above the third quartile. The mean perceived trust scores per condition were 70.28% for the non-aligning condition and 74.31% for the aligning condition. Statistically, the Independent Samples t-test found no significant difference in perceived trust score between the aligning condition ($M = 74.31$, $SD = 8.54$) and the non-aligning condition ($M = 70.28$, $SD = 13.01$), $t(22) = .718$, $p = .380$. These findings do not support the hypothesis that lexical alignment with the user leads to increased trust in the chatbot providing the information. The effect sizes accompanying the Independent Samples t-test were Cohen's $d = 0.366$ and Hedges' $g = 0.353$, indicating a slightly higher than small effect. In conclusion, the difference in perceived trust score between both conditions was too small to reach statistical significance, despite the observed trend in mean perceived trust scores and the boxplot in Figure 6.2.

6.1.3 Placeholder terms

The relevance of the questionnaire administered at the start of the experiment, which aimed to infer participants' word use, was assessed by analysing the variation in terms chosen by different participants. To maintain brevity, only a few placeholders are discussed in detail here. For a comprehensive overview of the frequency of different terms chosen for each placeholder, please refer to Appendix G. Figure 6.3 shows pie charts illustrating the distribution of terms selected for four placeholders present in the initial message delivered by the chatbot, which all

Figure 6.2: HCTM trust scores

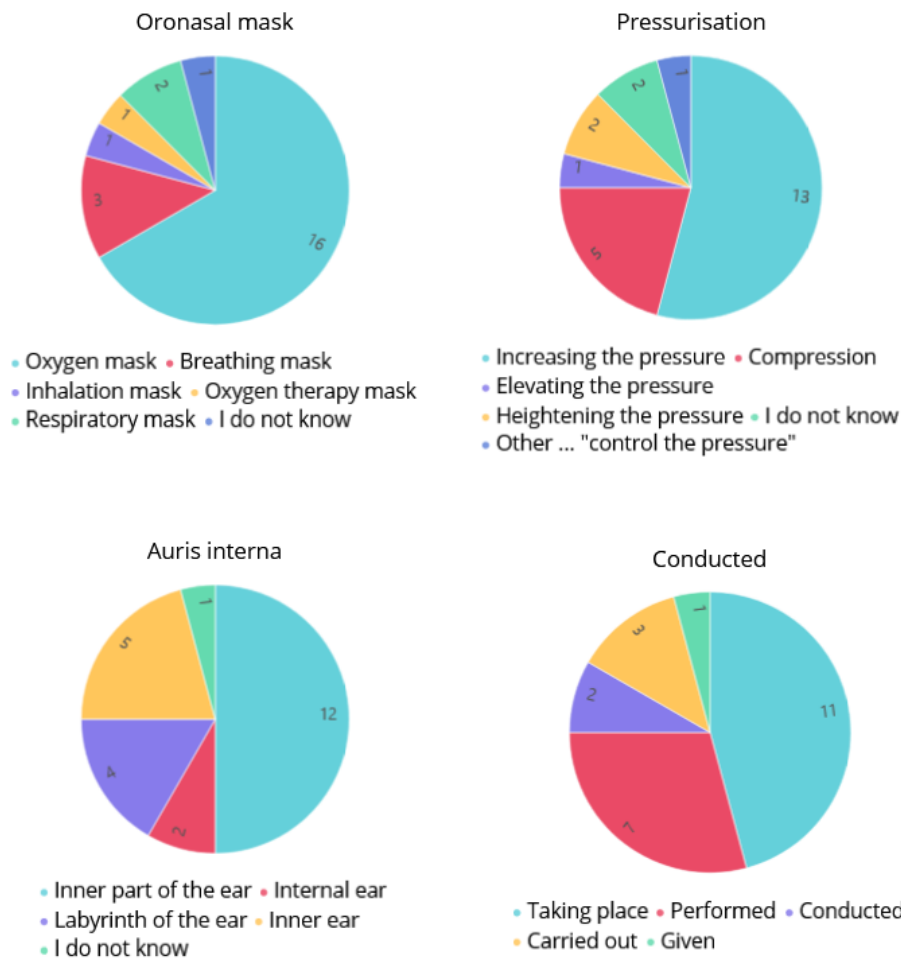


participants chose to read, see Section 5.6.2 for further explanation about this.

The different pie charts in Figure 6.3 depict the number of times each term was chosen by participants, with a total of 24 participants. Note that terms that were not chosen by any of the participants are not shown in the pie chart. Notable variation in the terms selected for the different placeholders can be observed. Specifically, for the placeholder *Conducted*, although the term itself was included in the answer options of the questionnaire, it was only selected by two participants as the preferred term; the other 22 participants preferred a different term for the placeholder. Note that an in-depth explanation of the rationale behind including or excluding the placeholder itself in the answer options can be found in Section 5.6.1. Figure 6.3 shows the diversity in word preferences for these specific placeholders, highlighting the importance of lexical alignment, especially as the most frequently chosen terms differed from the original terms used as placeholders.

The results of these few placeholders already suggest the relevance of alignment. However, to have an overall understanding of the relevance of the placeholders used in this research, the total average percentage of alternative terms chosen and used that were different from the original placeholder was calculated. This was done by counting the number of times participants chose a term different than the placeholder itself and dividing this by the total number of participants. Note that the option *I do not know* was not considered a different term than the original term used for the placeholder. Similarly, the entered terms in the *Other ...* option that were not deemed similar enough according to the cosine similarity calculations explained in Section 5.5.3 were also not considered as a different term as, similar to the *I do not know* response, no replacement of the placeholder would take place. For example, for the placeholder *Pressurization* in Figure 6.3, the calculation would be $(24 - 2 \text{ (from the answer } I \text{ do not know)} - 1 \text{ (from the } Other \dots \text{ option that was not similar enough)}) / 24$. The average percentage of alternative terms chosen for all the placeholders was calculated to be 60.03%, which is above

Figure 6.3: Pie charts illustrating the variation in terminology chosen for placeholders



50%, indicating a preference for participants to choose an alternative term instead of choosing at random. This confirms the trend observed in the pie charts in Figure 6.3, indicating the relevance of aligning on these terms as the original placeholder terms may not fully resonate with the participants' understanding or usage. Interestingly, some placeholders showed much less variability in responses, such as *Nurse* or *Patient*, with drastically lower percentages of alternative terms chosen (4.17% and 0.00%, respectively), which will be a point for discussion in the next chapter.

6.2 Qualitative results

The qualitative measures in this research involved conducting an interview at the end of the experiment. Participants were asked several questions, yielding interesting insights that can be categorised into two main themes: the future use of healthcare chatbots and the perception of the terminology and language used by the chatbot. Within these categories, common themes, patterns, and insights emerged.

6.2.1 Future use of healthcare chatbots

The analysis of the interview showed that 22 out of 24 participants indicated that they were likely to use a healthcare chatbot similar to the one used in the experiment in the future, re-

ardless of whether they interacted with the aligning or non-aligning chatbot. With the expected use of a healthcare chatbot in the future, trust emerged as an important factor. The importance of the chatbot relying on trustworthy sources was specifically mentioned by nine out of the 22 participants as being an important aspect for them to consider using it in the future. Examples of trustworthy sources cited included the websites of hospitals or other health institutions. Additionally, four participants mentioned they would only use a healthcare chatbot if it were available on a trusted website, such as that of their regular doctor. Another aspect mentioned as important in the potential future use of a healthcare chatbot by nine out of 22 participants was the need to consult their own doctor before relying solely on a healthcare chatbot. Specifically, these participants indicated they would prefer to use the chatbot combined with information from their doctor or other information sources on the internet. This shows that many participants see the use of a healthcare chatbot as a supplementary source of healthcare information, which is consistent with the intended role of the chatbot in the current research. However, two out of 24 participants did not share the view of potentially using a healthcare chatbot in the future. For example, participant 68G from the non-aligning condition stated, *"I would still use Google in order to compare multiple sources"* and participant 52X from the aligning condition mentioned, *"I am not really such a person, I would use what I have always been using and not change my methods, maybe only for very small questions"*.

Related to the question about the potential future use of healthcare chatbots was the question of whether participants would prefer this type of information-seeking platform over the well-known and often-used platforms such as Google. Irrespective of whether they interacted with the aligning or non-aligning chatbot, a majority of the participants (15 out of 24) expressed a preference for using a healthcare chatbot or a combination of the chatbot with information from the doctor or other medical sources over the use of Google. The reasons cited for this preference were related to the perceived trustworthiness of the information provided by the chatbot compared to the often inaccurate, dramatic, or worrisome results returned by Google. Participant 70V from the non-aligning condition said, *"Yes, I would definitely prefer such a chatbot over Google. When asking a basic health-related question on Google, it will always end in the diagnosis that you are going to die or something like that"*. Similarly, participant 8o, also from the non-aligning condition, commented, *"I would definitely use such a chatbot, this provides information that is way more rational compared to Google. Google is always so extreme; I do not want to use that anymore"*. Participant 40t from the aligning condition shared this view and commented, *"Yes, in a real scenario, such a chatbot would definitely help with decreasing my stress and insecurities regarding the disease and treatment as it provides more precise information compared to other pages"*. These responses underscored the important role of trust in the willingness to use a healthcare chatbot, especially compared to well-known information-seeking platforms such as Google.

In summary, these results highlighted the pivotal role that trust plays in the use and potential future use of a healthcare chatbot. Specifically, it emerged as a key factor influencing the willingness of participants to use such a chatbot in the future, as well as their preference for using such a chatbot over platforms such as Google. Participants emphasised the importance of trustworthy sources of information used by the chatbot as well as using the chatbot as an addition to the information provided by their own doctor. Furthermore, no significant differences between the two conditions were observed, indicating that lexical alignment did not influence the aspects of trust that participants recognised as motives for willing to use the chatbot.

6.2.2 Perception of terminology difficulty

Another aspect evaluated during the interview was the participant's perception of the difficulty of the terminology used by the chatbot. In the non-aligning condition, all participants unanimously agreed that the chatbot used difficult terminology. Responses included statements like, *"Yes definitely, very much"* from participant 68G or *"Yes very complicated. I would prefer an easier language"* as said by participant 30. However, two out of the 12 participants in the non-aligning condition specifically mentioned that they understood the information provided by the chatbot despite the challenging language. Participant 13F noted, *"Yes it used difficult terminology, mostly medical terminology. I did understand what was said"*. This was confirmed by this participant's perfect score of 100% on the understanding test. The other participant who expressed understanding, participant 39o, only scored 46.67% on the test, indicating a discrepancy between perceived and actual understanding. In contrast to the unanimous perception of difficult terminology used by the chatbot in the non-aligning condition, responses in the aligning condition varied. Five out of 12 participants stated that the chatbot did not use difficult terminology, which was confirmed by their scores on the understanding test being above the chance level. Another participant in the aligning condition, 47E, described the language used by the chatbot as *"in between difficult and easy"*. The remaining six participants stated that the chatbot was using difficult terminology, however, the chatbot did use the terms they had chosen in the questionnaire for the specified placeholders. Only three participants encountered a placeholder term that did not align with their preference as they answered *I do not know* in the questionnaire conducted at the start of the experiment, resulting in the original placeholder term used in the chatbot's responses. This happened twice for the placeholder *Tympanic membrane* and once for the placeholder *Tympanic cavity*. The participants (in the aligning condition) that did not perceive the language used by the chatbot as difficult provided answers such as the response of participant 30J, *"Not really, I am used to it due to my studies"*, and from participant 8h, *"No, I could definitely follow what was said"*. The participants who did state that the language was difficult gave similar responses to those in the non-aligning condition. Overall, three main reasons for perceiving the language used by the chatbot as difficult were identified across both conditions: the responses used a high level of English, used medical terminology, or were rather long. The chatbot using medical, or unknown, terminology (such as the anatomy of the ear, medical concepts, or medical processes) was mentioned by 12 out of 19 participants, emerging as the primary reason for the perception of difficult terminology.

Related to the perception of the difficulty of the terminology used by the chatbot was the degree to which participants asked for clarification on these terms or concepts. Participants' responses to whether they asked for clarification varied, from which three main patterns could be distinguished. The first was that participants asked for clarification and indicated this was helpful. The second was that participants did not ask for clarification, as they expected the chatbot to not have this functionality. The third was that participants did not have a clear reason but indicated they would do this in the future or in a more realistic situation. In the aligning condition, five out of the seven participants who indicated the language used by the chatbot to be difficult reported that they had asked for clarification when needed and mentioned that the information provided was helpful. Responses included *"Some terms or concepts were complicated, but I could ask for clarification, which was helpful"*, said by participant 40t. In the non-aligning condition, substantially more participants—eight out of 12 participants compared to two out of seven for the aligning condition—responded to not having asked for clarification. Responses included *"I did not ask for further clarification as I was not expecting it to have this possibility"*, from participant 68G, and *"I did not ask for clarification, but I would definitely do this in a real-life scenario"*, from participant 39o. Participant 8o added, *"No, I did not ask for clarification. Now that I am thinking about it, I should have probably done that"*.

However, this difference in asking for clarification between both conditions was not clearly reflected in the analysis of the dialogues. It is challenging to draw conclusions on the difference in asking for clarification between the two conditions because the chatbot's responses differed among participants depending on the questions they asked. Overall, the analysis showed all participants asked for clarification on several concepts discussed in earlier responses from the chatbot. Examples included asking for more information about the Weber and Rinne tests or what corticosteroids are, which were concepts mentioned in the introductory summary the chatbot provided (which all participants chose to read). Interestingly, none of the participants in the aligning condition asked for clarification on the term they had seen for the placeholder *Auris interna*, present in the summary provided by the chatbot, whereas five out of 12 participants in the non-aligning condition asked for clarification on this term. Similarly, five out of twelve participants in the non-aligning condition asked for clarification on the term *Tympanic cavity* also present in the summary provided by the chatbot, whereas none of the participants in the aligning condition asked for clarification on the term used by the chatbot for that placeholder.

The analysis of difficulty perception of the terminology used by the chatbot revealed notable differences across the conditions. The most prominent finding was that in the aligning condition, fewer participants perceived the language to be difficult, whereas in the non-aligning condition, all participants agreed that the chatbot used difficult terminology. This was expected due to the lexical alignment with the preferred terms of the participant in the aligning condition (see Section 5.5). Regarding the behaviour of asking for clarification on complex terms or concepts present in the information provided by the chatbot, different patterns emerged. An interesting finding was the expressed lack of asking for clarification in the non-aligning condition. In contrast, in the aligning condition, most participants expressed that they asked for clarification if they needed it. This shows that the implemented lexical alignment had a positive effect on the perceived difficulty of the provided information by the chatbot (resulting in perceiving the provided information to be less difficult), as well as on the willingness of participants to ask for clarification when needed. However, the analysis of the dialogues did not clearly reflect the difference in asking for clarification between the two conditions as expressed by participants in the interview.

6.3 Conclusion

In this chapter, an analysis of the various results obtained from this research was provided. Firstly, the quantitative results were discussed, including the results of the understanding test and trust scale, as well as an analysis of the placeholder terms chosen in the questionnaire to infer participants' word use. The results obtained from the understanding test suggested that the participants in the aligning condition scored slightly better compared to the participants in the non-aligning condition. However, the Independent Samples t-test demonstrated the observed difference was not statistically significant, and therefore, the hypothesis of lexical alignment having a positive effect on the understanding of the provided information by the chatbot was rejected. Similarly, the results obtained from the trust scale suggested that participants in the aligning condition had an increased perceived trust compared to the participants in the non-aligning condition, however, this was not supported by the Independent Samples t-test. Hence, the hypothesis of lexical alignment increasing the perceived trust in the chatbot providing the information was rejected.

The analysis of the distribution of chosen terms for the various placeholders highlighted the relevance of lexical alignment to use different terms for the defined placeholders, as the overall percentage of a different term chosen, rather than the original placeholder, was above 50%, indicating a preference for the alternative terms of the placeholders. Only two placeholders had

limited to no variation in the alternative terms chosen. Overall, these results show the variation in preferred terms for the specified placeholders in this research.

Then the qualitative results were discussed, including the analysis of the interviews conducted at the end of the experiment. Several themes, patterns, and insights were obtained regarding the future use of healthcare chatbots and perceptions of terminology difficulty. Trust emerged as an important factor in the future use of healthcare chatbots, primarily linked to the sources of information used by the chatbot to generate its responses. This finding was confirmed during the comparison of using a healthcare chatbot or another well-known information-seeking platform such as Google, where participants strongly indicated a preference for a healthcare chatbot over Google. Regarding the perception of terminology difficulty, a difference between conditions was found. Participants in the aligning condition did not always perceive the language to be difficult, whereas participants in the non-aligning condition all agreed on the chatbot using difficult terminology. Additionally, participants in the aligning condition expressed being more willing to ask for clarification on the concepts they did not understand compared to the participants in the non-aligning condition. However, this was not clearly reflected in the dialogue analysis. Overall, a positive influence of lexical alignment on the perceived difficulty of the provided information as well as on the perceived willingness of participants to ask for clarification when needed was demonstrated.

7 DISCUSSION

This research provides insights into the effects of lexical alignment on the understanding of the information provided by a healthcare chatbot and the trust one has in this chatbot during an information-seeking task. While the results did not show a statistically significant effect of the use of lexical alignment on the understanding of the provided information and the perceived trust in the chatbot, several other observations were made. This chapter will analyse the results further and discuss the implications of the findings. Additionally, the limitations of the research are highlighted, and potential directions for future work will be discussed.

7.1 Analysis of the results

The results of the research showed there was no statistically significant difference between the understanding scores participants obtained on the understanding test in the two conditions. Despite the observation suggesting that the aligning condition outperformed the non-aligning condition, the Independent Samples t-test demonstrated no statistically significant difference. Similar results were obtained for the perceived trust participants had in the chatbot. However, there are some interesting aspects to consider and discuss in more detail.

7.1.1 Understanding

The hypothesis that lexical alignment would increase the understanding participants have of the provided information by the chatbot was largely based on literature focusing on improving health-related information exchange. This literature, discussed in depth in Section 2.3.1, emphasised that the comprehension of health-related information could benefit from reducing the specialised vocabulary and concepts that are unfamiliar to those outside the medical profession or study, as well as employing more accessible language in conveying the information in general [12, 13, 14]. Central to these approaches is the substitution of medical terminology with the terminology known to the patient, referred to as "plain language" or "living-room language" [12, 13]. In the present research setup, this was incorporated through lexical alignment by having participants complete a questionnaire at the start of the experiment to determine their preferred terms for the placeholders used in the alignment strategy. This questionnaire showed the relevance of aligning on these placeholders, as the majority of terms chosen for the specified placeholders were different from the original term used as a placeholder. A potential explanation for the limited differences across conditions on understanding could be related to the use of lexical alignment, which only substitutes words and does not consider overall sentence structure. While lexical alignment accounts for the substitution of terms as proposed in literature to adjust the language to that known by the patient, the substitution of mere terms might not be enough to make the information more understandable. In that case, syntactic alignment, which adjusts entire sentences or sentence segments, may have an increased impact on the understanding one has of the information compared to lexical alignment alone [6]. This potential advantage of syntactic alignment over lexical alignment will be discussed in more

detail in Section 7.2.2.

An additional point of discussion emerged from the analysis of the questionnaire used to infer the participant's word preference for the placeholders, as not all placeholders appeared to be as relevant to use in the alignment strategy. For the placeholders *Nurse* and *Patient*, alternative terms were chosen only once or never at all. A possible explanation for this limited preference to use an alternative term could be the prevalent use and familiarity of the original placeholder terms. This suggests that certain terms may not be as relevant to adjust within the alignment strategy. Specifically, when the preference for terminology for a certain word or concept is similar among a large number of people, this could result in limited alignment possibilities and reduce the potential influence of this alignment. In order to have a better insight into the frequency of the placeholders used, the word frequency of the placeholders and potential alternative terms should be taken into account in order to determine whether they are useful in lexical alignment strategies and which placeholder terms are already widely understood and used. For the two placeholder terms that stood out in this research, the alternative terms in the questionnaire had lower word frequencies compared to the original placeholders, as shown by the Corpus of Contemporary American English (COCA), which is an American English corpus consisting of more than one billion words [113]. For instance, the alternative term *Medic* had a frequency of 3104, compared to the original placeholder *Nurse* having a frequency of 43524. Additionally, as most of the participants in this research were Dutch, the word frequencies of the alternative terms for these placeholders were also evaluated using the OpenSoNaR corpus, which contains more than 500 million Dutch words [114]. Again, the alternative terms showed lower frequencies than the original term, with, for example, *Client* having a frequency of 1720 versus the original placeholder *Patient* having a frequency of 4176. Only the alternative term *Specialist* for the placeholder *Nurse* did not show a lower frequency, likely due to *Specialist* being a common word in Dutch as well [114]. Overall, this suggests that future research should consider the word frequencies of terms to be used in the lexical alignment strategy.

An important observation is that the mean understanding scores in both conditions exceeded 70%, indicating performance above the chance level. This suggests that the non-aligned information may have been better understood than anticipated, potentially limiting the influence of lexical alignment in this experiment. A possible explanation for this is related to the participant demographics. The participants taking part in this research were either currently enrolled in university or holding bachelor's or master's degrees. This educational background increases the possibility of the participants being more used to encountering more difficult terminology or complex textual material than people who do not have such an educational background. The interview analysis indicated that participants overall perceived the terminology used by the chatbot as difficult, with this observation being more frequent for the non-aligning condition. However, several responses in the interviews confirmed the suggestion that the educational background of the participants potentially limited the difficulties experienced in understanding the terminology used by the chatbot. Participant 30J in the aligning condition said, *"In my studies, I am used to such challenging wording"*, similarly, participant 33B in the non-aligning condition said, *"Yes it was advanced language, I think when I did not have my current level of education, this would be a problem"*. Literature reviewed in Section 2.3.1 indicated individuals across all levels of general literacy could experience difficulties in understanding health-related information, primarily due to the use of unknown or difficult medical terminology and complex language. However, this may not fully apply to the current participant group, as they potentially have an overall increased health literacy level instead of solely an increased general literacy level due to their educational background, limiting the discrepancies in language used by the chatbot and understood by the participant. This could explain why the effects of alignment on the understanding of the provided information by the chatbot were limited in this research.

7.1.2 Trust

Previous research by Scissors et al. [64] demonstrated the positive effects of lexical alignment on trust within the context of a social dilemma investment game. They found that partners exhibiting higher levels of lexical alignment also exhibited increased trust [64]. However, the significant contextual difference between a social dilemma investment game and the current health-related information-seeking task suggests that the impact of lexical alignment on trust may vary across different contexts. In a social dilemma investment game, the need for partners to cooperate, reach a shared agreement, and understand each other might lead to an amplified influence of alignment on the perceived trust in each other [64]. In the context of the current research, trust might be built on other factors that will be impacted less or differently by lexical alignment, such as whether the chatbot comes across as trustworthy or the perceived accuracy and reliability of the provided information, which are well-known aspects of trust related to the potential use of healthcare chatbots on the side of patients [2, 19, 20]. This was also reflected in the analysis of the interview, where trust appeared to play a pivotal role in the willingness of participants to use a healthcare chatbot in the future. In this willingness, the trustworthiness of the sources used by the chatbot was shown to be an important factor. Additionally, the majority of participants expressed a preference for the use of a healthcare chatbot over Google in a scenario such as presented in the experiment, which stemmed from the fact that the information provided by the chatbot was perceived as more reliable and accurate compared to the information provided by Google. This emphasises the suggestion that the way in which trust is built varies across contexts and that this influences the impact of lexical alignment on it, which could be a possible explanation of the limited effects of lexical alignment on trust in the current research.

The observation that trust in this research was primarily related to the perceived accuracy and reliability of the provided information can be linked to the concept of epistemic authority (see Section 2.4). The definition of epistemic authority in the context of healthcare states that patients tend to rely on doctors who possess this authority, attributed to their qualifications, titles, and expertise [54, 55]. The design of the chatbot also considered this concept of epistemic authority by incorporating elements to present it as an expert on the diagnosed disease and proposed treatment of the experiment, see Section 5.3. Participants in both conditions indicated a preference to use a healthcare chatbot over platforms such as Google based on a perceived increased accuracy and reliability of the information used by the chatbot to provide answers, suggesting that participants attributed epistemic authority to the chatbot in a similar manner as to how epistemic authority is usually attributed to doctors [54, 55]. This implies that the trust participants had in the chatbot in this experiment was mainly based on the attribution of epistemic authority, which was based on perceived expertise and was not influenced by the implemented lexical alignment.

7.1.3 Seeking clarification

An interesting observation from the analysis of the interview was that participants in the non-aligning condition expressed more often that they had not asked for clarification when needed compared to those in the aligning condition. However, this was not clearly reflected in the dialogue analysis. Additionally, participants in the non-aligning condition perceived the language used by the chatbot to be more difficult than those in the aligning condition more frequently, which would suggest an increased need for clarification. In the following, these matters will be discussed in more detail.

Participants in the non-aligning condition expressed a lower willingness to ask for clarification at the end of the interaction. However, the dialogue analysis did not clearly reflect this matter. This suggests that these participants might have felt an increased effort to understand the information provided by the chatbot, leading to the perception of needing more clarification without actually asking for it. The notion of them experiencing an increased effort to understand the provided information by the chatbot is underscored by the findings that the participants in the non-aligning condition more often perceived the chatbot's language use as difficult. In contrast, those in the aligning condition less often perceived the language as difficult, limiting the feeling of an increased effort to understand the provided information. This may have led to a decreased feeling of needing to ask for clarification, resulting in them feeling as if they had asked for clarification when truly necessary. An interesting observation is that some of the participants in the non-aligning condition asked for clarification on certain placeholders present in the chatbot's responses, whereas those in the aligning condition did not. This can be explained by the use of lexical alignment. Specifically, participants in the non-aligning condition encountered placeholders that were not adjusted to their preferred terms, potentially increasing the difficulty of understanding them. Despite the expressed reluctance to ask for clarification, some participants might have felt like asking for clarification on this difficult terminology was a natural part of the interaction and may not have interpreted their questions as asking for special clarification but rather as triggering a normal feature of the chatbot. Further research exploring these possible explanations for the observations discussed is needed to fully understand them and their implications.

In addition, research by Chen et al. [18] indicated an association between an individual's health literacy level and their use of and trust in various health information sources. Specifically, individuals with lower health literacy levels tend to trust information from healthcare professionals less and information from less professional platforms more. Important here is the definition of health literacy: the ability to obtain, understand, and use health information and services to make appropriate decisions regarding one's health [12, 14, 50, 51]. In the current research, it was found that participants in the non-aligning condition experienced a lack of seeking clarification. This behaviour might be influenced by factors associated with a lower health literacy level, such as difficulties in understanding and obtaining information potentially due to an increased effort required to understand the non-aligned information provided in response. This, in turn, could lead to participants having a decreased trust in the chatbot, as suggested by the association found in the research of Chen et al. [18]. Therefore, an indirect relationship between the use of lexical alignment and trust in chatbots mediated by health literacy is suggested. Future research should investigate this suggested relationship further in order to understand it in more detail as well as its potential implications.

7.1.4 Implications of the research

Previous research highlights that patients' health outcomes can greatly benefit from an increased ability to obtain, understand, and use health information, but several challenges in achieving this still exist [12, 14, 16, 50, 51]. In this context, the current research investigated the impact of lexical alignment on understanding health-related information provided by a healthcare chatbot and participants' trust in the chatbot. The objective was to potentially enhance the understanding of the provided information as well as trust in the chatbot through the use of lexical alignment. Although the results did not yield statistically significant effects, several important trends and observations could be observed. These findings have a variety of valuable implications for the design and development of healthcare chatbots, as well as suggestions for

future research in this domain.

Regarding the understanding of health-related information provided by healthcare chatbots, the observed trends in this research suggest a potential beneficial role of lexical alignment. However, no statistically significant effects of lexical alignment on enhanced understanding were found, possibly explained by the higher-than-expected health literacy level of participants. This underscores the need to investigate the effects of lexical alignment on the understanding of individuals with varying health literacy levels in order to confirm or reject the potential influence of lexical alignment on understanding. Despite this, the finding of lexical alignment decreasing the perceived difficulty of the language used by the chatbot supports the literature suggesting to adjust health-related information to the language used by the patient to improve information exchange between patients and healthcare professionals [12, 13]. Furthermore, the discussion on the difficulty perception of the language used by the chatbot and the participant's willingness to seek clarification when needed further emphasises the relevance of adjusting the chatbot's language to that of the user, as the participants in the aligning condition did not experience the same reluctance in asking for clarification as was expressed by the participants in the non-aligning condition. Overall, the obtained results in this research, combined with existing literature on improving health-related information exchange by using plain language and replacing medical terminology with terms familiar to the patient, highlight the need to consider language adjustments in the development and deployment of healthcare chatbots used to provide health-related information to patients [12, 13].

Regarding trust in healthcare chatbots, the current research emphasises the importance of considering the concept of epistemic authority in the development of healthcare chatbots. Participants expressed being willing to use healthcare chatbots in future real-life scenarios similar to the scenario used in this research. This willingness was explained by the attribution of epistemic authority to the chatbot in a similar manner to the attribution of epistemic authority to doctors. This finding was emphasised by participants' unwillingness, sometimes even refusal, to use well-known platforms such as Google to obtain health-related information stemming from experiences with Google providing unreliable, extreme, and often untrustworthy information, and participants' preference to use healthcare chatbots based on their assumption that these chatbots would be based on trustworthy sources such as websites from hospitals or health institutions. This underscored the importance of the perceived epistemic authority of the chatbot as a motive for participants to rely on it to obtain health-related information. Additionally, literature has suggested a shift towards a more patient-centred approach, where patients take a more prominent role in health-related decision-making instead of solely relying on their own doctor [54, 55, 57, 58]. This highlights the importance of considering and incorporating epistemic authority in the development of healthcare chatbots in order to increase the reliance of patients on these information sources in the future.

Finally, this research highlighted the importance of trust in the potential use and adoption of healthcare chatbots and the way in which trust is established in healthcare chatbots. Additionally, an indirect relationship between lexical alignment and trust, mediated by health literacy, was proposed. Note that this relationship requires further research to confirm, reject, and elaborate upon. An essential aspect related to trust is ensuring trust in healthcare chatbots is at an appropriate level. In cases where patients have unquestioning trust in all healthcare chatbots, this could be problematic, as they could trust chatbots that are providing inaccurate information. The observations in this research related to the attribution of epistemic authority to the healthcare chatbot, contributing to trusting it, suggest that patients might naturally avoid or distrust chatbots that lack reliable sources (similar to the expressed motives of not using Google to obtain health-related information). However, an additional risk is the chatbot pretending to

be using reliable sources to provide information without actually doing so, which could mislead patients into trusting inaccurate information. Therefore, it is essential for the development and future research of healthcare chatbots to focus on having a balanced approach to foster users' trust in healthcare chatbots, considering both aspects contributing to trust. On the one hand, trust should emerge from the potential relationship with lexical alignment mediated by health literacy (the ability to obtain, understand, and use health-related information and services to make appropriate decisions regarding one's health [12, 14, 50, 51]). On the other hand, the chatbot's epistemic authority must be genuinely based on accurate and reliable information to foster trust.

7.2 Limitations

In this research, several factors may have influenced the results, some of which were already (partially) discussed in the previous sections. In the following, an overview of the limitations of this research is provided, including the various aspects that might have influenced the potential impact of lexical alignment on participants' understanding of the information provided by the chatbot and the trust they have in the chatbot.

7.2.1 Research design

Participant demographics

Regarding the research design, the first limitation to discuss is related to the participant demographics. As already discussed in the analysis of the results of the variable understanding, the participants shared similar educational backgrounds: either still enrolled in university or already holding a bachelor's or master's degree. This was a result of the way of recruitment and the context in which this research was conducted: it was a graduation project, and mostly friends and acquaintances through friends were asked to participate. It is clear that the participant group used was not well representative of the broader population, limiting the generalisability of the results. Additionally, the research sample size of $n = 24$ participants was relatively small, further limiting the generalisability of the results.

Moreover, another limitation related to participant demographics is the fact that the participants were non-native English speakers. Whereas the ability to speak, write, and read English was a requirement to take part in this experiment, the fact that participants were non-native English speakers could have influenced their general understanding of the information, potentially limiting the influence of lexical alignment on both the understanding of the provided information and perceived trust in the chatbot providing this information. This suggestion was confirmed by various statements mentioned in the interviews (across conditions). Participant 21O in the aligning condition said, *"In English, I am not too comfortable with difficult terminology, probably in Dutch, it would have been easier to understand what was said"*. Similarly, participant 39C in the aligning condition said, *"English is not my biggest strength, I prefer to read this type of information in Dutch"*.

Scenario

Aside from the participants' demographics, another significant limitation of this research is the fact that the scenario presented to participants for the experiment was of a hypothetical nature. Participants were asked to imagine themselves in the scenario of being diagnosed with acute acoustic trauma and in need of the proposed hyperbaric oxygen therapy. While this scenario was carefully designed and chosen to increase the likelihood of participants truly wanting and

needing to ask questions during the interaction with the chatbot, the hypothetical nature still posed challenges during the experiment. Only one participant did not meet the requirement of asking at least three questions, as discussed in Chapter 6, but some participants made comments such as the following said by participant 13F in the non-aligning condition, *"Okay, so I have to pretend to be very interested in this treatment, right?"*, which were answered by the researcher by statements such as, *"You should ask any questions you have regarding the treatment and disease, if you are well enough informed, you can stop"*. This is considered a limitation because it indicates that participants were potentially not paying as much attention to the responses provided by the chatbot as they would have done in a real-life situation. This was also supported by several comments made during the interview when asking about whether a participant asked for clarification on difficult or complex concepts present in the answers from the chatbot. Participant 39O in the non-aligning condition said, *"I did not ask for clarification, but I would definitely do this in a real-life scenario"*, and similarly, participant 100V in the aligning condition said, *"I did not pay that much attention when reading the answers, when I noticed it was the correct answer to my question, I just asked the next one"*. However, the test on understanding showed performance above the chance level in both conditions, indicating participants were somewhat engaged with the responses of the chatbot. Nonetheless, the limitation that the hypothetical nature of the scenario potentially influenced the participant's level of interest in the provided information by the chatbot and the way in which participants interacted with the chatbot compared to an interaction in a real-life situation should be acknowledged.

7.2.2 Alignment strategy

There are several potential explanations for the limited effects of the implemented lexical alignment on understanding and trust observed in this research. One of them includes the specific alignment strategy employed. The lexical alignment strategy, based on the research by Spillner and Wenig [8], focused on substituting predefined placeholders with the preferred terms of the participants. While this method did allow for tailoring the language used by the chatbot to that of the participant to some extent, the language may not have sufficiently aligned with that of the participants in a more general manner. Literature states that alignment at one linguistic level facilitates alignment at other linguistic levels, however, this effect might be limited due to the narrow scope of the alignment strategy implemented in the current research [5]. Specifically, the adjustments made by substituting the placeholders with the terms preferred by the participant do not address further adjustments to sentence structure or context, aside from some minor adjustments to ensure grammatical correctness after substitution, as discussed in Section 5.5.4. This could lead to a limited overall alignment of the language used by the chatbot with that used by the participant. A broader form of alignment, such as syntactic alignment, could possibly be more effective, as syntactic alignment goes beyond the replacement of individual terms by adjusting the sentence structures to those used by the other conversational partner [6].

Furthermore, an important point of discussion is whether this research is truly measuring the impact of lexical alignment or, more specifically, the effect of terminology simplification on the participants their understanding of the provided information by the chatbot and their perceived trust in the chatbot. Lexical alignment involves adjusting the language to match that of the conversational partner by substituting terms with synonyms or equivalent terminology, for example, using "sofa" instead of "couch" if that is the term used by the conversational partner [6]. In this research, however, the alternative terms used for the placeholders often represented more common or simplified forms of the original placeholders, which tends towards simplification. However, since the terms used for the substitution of the placeholders were based on the personal preferences of participants, this method goes beyond simple simplification. In the questionnaire conducted at the start of the experiment designed to obtain the participants' word

preferences, participants were asked to select the terms that they preferred to describe specific concepts in pictures or descriptions (see Sections 4.1.1 and 5.6.1 for further information). In this way, the chatbot could use terms that aligned with the participants' language use and understanding in the responses provided. Therefore, this method can be considered a form of lexical alignment rather than a simple simplification. Nonetheless, it is important to acknowledge that this approach incorporates and involves a degree of simplification. Future research could explore more comprehensive alignment strategies that do not only substitute individual terms (e.g., syntactic alignment, which also adjusts sentence structure) in order to better distinguish between simplification and alignment.

Another limitation of the alignment strategy currently employed is the practical application of the chatbot. The implementation of lexical alignment in this research, while functional for the current research purposes, was not the most efficient. The current method for substituting a placeholder with a term used by the participant involved several nested for-loops so that each term in the input sentence was evaluated. If the term was not part of the NLTK stopwords, its vector representation was obtained, and the similarity with the placeholder was calculated (for a more in-depth explanation of this process, refer to Section 5.5.3). This can be time-consuming and excessive, resulting in slow performance. This especially became clear when deploying the chatbot online, which limits its use for broader deployment. Additionally, the current implementation of alignment does not account for all language adjustments necessary after substituting placeholders when used in real-world applications. For example, the current implementation did not include the non-standard article rules like using "an" for "hour" instead of "a", which was not a problem in the current research but should be considered for a broader application of the chatbot. Moreover, the current implementation did not account for similarity checking of multi-word expressions, as this was expected and shown to not be required in the current research, however, this could become relevant when deploying in broader settings. While these limitations did not pose significant problems in the current controlled research setting, they emphasise the need to optimise the implemented lexical alignment strategy and improve the online performance of the chatbot in order to enhance the usability and effectiveness of such a chatbot in real-life scenarios.

7.2.3 Functionality of the chatbot

Another aspect to consider when discussing the limitations of this research is related to the general functionality of the chatbot. At the time of conducting this experiment, Open AI's ChatGPT was a well-known and widely used chatbot among the participants [32, 115]. Given the common use of this advanced chatbot, participants may have had different expectations regarding the chatbot's capabilities compared to its actual performance. During the interviews conducted at the end of the experiment, several participants expressed being positively surprised about the chatbot's functionalities, recognising it as part of a graduation project. For example, participant 30J in the aligning condition said, *"I was amazed by some functionalities it had, sometimes I first thought to not ask a certain question, but later I did, and it was perfectly fine in answering it!"*. Conversely, others were disappointed that the chatbot could not answer all their questions perfectly, aligning with increased expectations for the chatbots used and developed nowadays. Participant 9w from the aligning condition mentioned *"I would prefer a chatbot to be able to answer more in-depth questions as well"*, an example of an in-depth question asked by this participant that was not answered by the chatbot included *"What frequencies are the tuning forks?"*. It is important to note that participants in both conditions used the same chatbot, with the only variation being the use of alignment, and therefore the potential influence of increased expectations of the chatbot's functioning on the interaction is similar across conditions. How-

ever, this should be taken into consideration when developing a healthcare chatbot, or other chatbots, for real-life usage as the use of advanced chatbots like ChatGPT is getting more and more common nowadays, potentially changing users' expectations [115].

7.3 Future work

Various possible improvements to the current research have already been mentioned in the previous sections. In the following, interesting possibilities for future work and research will be discussed.

Several points of discussion arose related to the alignment strategy that was employed in this research. A promising area for future research is to broaden the scope of the alignment strategy beyond lexical alignment. Future research should explore alignment strategies that not only substitute terms but also incorporate sentence adjustments and overall communication styles. By doing this, the alignment of the overall language can be significantly enhanced, improving how well the language is truly adjusted to the language used by the conversational partner, potentially leading to better information exchange between healthcare professionals and patients. Therefore, future research should consider not only lexical alignment but also syntactic alignment, which focuses on shared speech patterns, as well as semantic alignment, which focuses on using similar higher levels of representations [5].

Additionally, it would be valuable to explore how a combination of both simplification of terminology and alignment strategies can be integrated to improve health-related information exchange. Future research could investigate how simplification can be used to make complex medical terminology more accessible while employing alignment strategies to adapt the overall language used in the information exchange to match the conversational style of the patient. These alignment strategies should include syntactic and semantic alignment, which go beyond the substitution of individual terms. In this way, insights can be gained into the ways in which both simplification and alignment can be used to enhance health-related information exchange.

While the current research has focused on the influence of lexical alignment on the trust one has in the chatbot that provides the information, it would be valuable to explore how alignment affects the trust one has in the information provided by the chatbot. Given the observed importance of epistemic authority, where participants trust the chatbot based on their perception of the chatbot providing responses based on trustworthy, reliable, and accurate sources, it is important to evaluate whether aligning the information with the user's language impacts the perceived trustworthiness of the provided information. This is an important distinction because trust in the chatbot as an information-providing entity might lead to users interacting with the chatbot more often and relying on it to obtain information, however, trust in the specific information provided by the chatbot is important for users to use and rely on the information itself. Future research could investigate the perceived trustworthiness and reliability of information that either aligns with the user or not, as well as compare this with the currently obtained results of the influence of lexical alignment on trust in the chatbot. This would provide insights into how alignment can impact both the trust in the chatbot providing health-related information and the trust in the information provided by the chatbot, which are important aspects of the effectiveness and adoption of such healthcare chatbots.

The difference shown in the concept of trust across different contexts highlighted in the current research proposes another interesting direction for future research: investigating whether this difference in the concept of trust varies across different health-related topics. Specifically,

the way in which people attribute trust to healthcare chatbots providing information might be different for a chatbot focusing on providing, e.g., information related to more psychological aspects, such as mental health problems, and a chatbot providing information on, for example, the course of a kidney surgery. If there are differences in the way one attributes trust to the chatbot across the various topics within the health-related domain, this should be taken into account during the development of health-related chatbots.

Another direction of future research could be to examine the ways in which alignment could contribute to making the healthcare chatbot more conversational or human-like instead of the chatbot solely providing information. In the interviews, it was mentioned by a participant, 100V in the aligning condition, that they usually preferred a more human-like, natural dialogue with chatbots. However, this participant also expressed being unsure whether that would be fitting in the context of a healthcare chatbot. Future research could investigate the role of alignment to create a more conversational, human-like role for the chatbot by adjusting its language to that of the user, as well as the influence of this on the use of the chatbot. This would provide valuable insights into the type of healthcare chatbots users prefer to obtain health-related information, the way this influences the experience with the chatbot, and the role that alignment could play in the perception of the chatbot taking on a more conversational, human-like role.

8 CONCLUSION

This research explored the impact of lexical alignment in a healthcare chatbot on the user's understanding of the provided information and their trust in the chatbot during an information-seeking task. The research question accompanying this research was: *How does the lexical alignment of a healthcare chatbot with the user during an information-seeking task impact the user's understanding of the information provided and their trust in the chatbot providing the information?*

To address this research question, a healthcare chatbot was created and an experiment was conducted where participants interacted with the chatbot in order to obtain information related to a hypothetically diagnosed disease and proposed treatment. The experiment had a between-subject design where participants were randomly assigned to either interact with a chatbot using lexical alignment or not. A total of $n = 24$ participants, 12 per condition, took part in the experiment. The experimental results showed no statistically significant differences in the understanding participants had of the provided information or in the trust participants had in the chatbot between the conditions. However, the variation in terms chosen for the various placeholders used in the chatbot's template answers showed the relevance of lexical alignment, as most often different terms than the original placeholders were chosen. Post-experiment interviews revealed that the difficulty perception of the chatbot's language varied between conditions, with the participants in the aligning condition perceiving the language used as difficult less frequently. Interestingly, participants in the non-aligning condition indicated not having asked for clarification as often compared to those in the aligning condition, despite the increased difficulty perception of the language used by the chatbot by participants in the non-aligning condition. However, this was not clearly reflected in the dialogue. Moreover, literature and observed findings suggest that the lack of lexical alignment could decrease trust in chatbots, especially for those with limited health literacy. Additionally, trust was linked to epistemic authority, with participants expressing a preference for trusting a healthcare chatbot over platforms such as Google to obtain health-related information based on the perceived trustworthiness and reliability of the information sources used by the chatbot.

Whereas various studies have investigated the concept of alignment, this had not yet been combined with the language-based challenges in health-related information exchange. Despite the lack of a statistically significant impact of lexical alignment on understanding and trust, relevant insights into the perception of the chatbot's language use and its influence on the interaction were obtained, and the importance of epistemic authority in trusting and using a healthcare chatbot was highlighted. Future research should involve a more representative participant group, considering the potential limited effect of lexical alignment due to the educational background of the participants as well as the other limitations and implications discussed. Further interesting future directions include exploring the impact of different and broader alignment strategies, such as syntactic alignment or semantic alignment, as well as combining simplification and alignment. Additionally, the influence of lexical alignment on trusting the provided information, the influence of context differences within the health domain on the concept of trust, and the role of alignment in creating a more conversational healthcare chatbot, could be investigated.

REFERENCES

- [1] Bol.com, “Klantenservice,” Bol.com, 2021. [Online]. Available: <https://www.bol.com/nl/nl/klantenservice/index.html>
- [2] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” *Machine Learning with Applications*, vol. 2, p. 100006, Dec. 2020. [Online]. Available: <https://doi.org/10.1016/j.mlwa.2020.100006>
- [3] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballı, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Lau, and E. Coiera, “Conversational agents in healthcare: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, July 2018. [Online]. Available: <https://doi.org/10.1093/jamia/ocy072>
- [4] B. Mesko, “The Top 10 Healthcare Chatbots,” *The Medical Futurist*, accessed November, 2023, Aug. 2023. [Online]. Available: <https://medicalfuturist.com/top-10-health-chatbots/>
- [5] M. J. Pickering and S. Garrod, “Toward a mechanistic psychology of dialogue,” *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.
- [6] A. Sinclair, K. McCurdy, C. G. Lucas, A. Lopez, and D. Gašević, “Tutorbot corpus: Evidence of human-agent verbal alignment in second language learner dialogues.” *International Educational Data Mining Society*, 2019.
- [7] S. L. Theodora Koulouri and R. D. Macredie, “Do (and say) as i say: Linguistic adaptation in human–computer dialogs,” *Human–Computer Interaction*, vol. 31, no. 1, pp. 59–95, 2016. [Online]. Available: <https://doi.org/10.1080/07370024.2014.934180>
- [8] L. Spillner and N. Wenig, “Talk to me on my level – linguistic alignment for chatbots,” in *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, 2021, pp. 1–12.
- [9] S. Huiyang and W. Min, “Improving interaction experience through lexical convergence: The prosocial effect of lexical alignment in human-human and human-computer interactions,” *International Journal of Human–Computer Interaction*, vol. 38, no. 1, pp. 28–41, 2022. [Online]. Available: <https://doi.org/10.1080/10447318.2021.1921367>
- [10] T. R. Nuñez, C. Jakobowsky, K. Prynda, K. Bergmann, and A. M. Rosenthal-von der Pütten, “Virtual agents aligning to their users. lexical alignment in human–agent-interaction and its psychological effects,” *International Journal of Human-Computer Studies*, vol. 178, p. 103093, 2023.
- [11] S. Srivastava, M. Theune, and A. Catala, “The role of lexical alignment in human understanding of explanations by conversational agents,” in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 423–435.

- [12] B. Weiss and AMA Foundation and American Medical Association, *Health Literacy*. American Medical Association Foundations and American Medical Association, 2006.
- [13] N. Egbert and K. Nanna, "Health Literacy: Challenges and Strategies," *Online journal of issues in nursing*, vol. 14, no. 3, Sept. 2009. [Online]. Available: <https://doi.org/10.3912/ojin.vol14no03man01>
- [14] M. K. Paasche-Orlow, R. M. Parker, J. A. Gazmararian, L. Nielsen-Bohlman, and R. Rudd, "The prevalence of limited health literacy," *Journal of General Internal Medicine*, vol. 20, no. 2, pp. 175–184, Feb. 2005. [Online]. Available: <https://doi.org/10.1111/j.1525-1497.2005.40245.x>
- [15] L. Seitz, S. Bekmeier-Feuerhahn, and K. Gohil, "Can we trust a chatbot like a physician? a qualitative study on understanding the emergence of trust toward diagnostic chatbots," *International Journal of Human-Computer Studies*, vol. 165, p. 102848, 2022.
- [16] H. Ishikawa and T. Kiuchi, "Health literacy and health communication," *BioPsychoSocial medicine*, vol. 4, pp. 1–5, 2010.
- [17] Y. Ye, "Correlates of consumer trust in online health information: findings from the health information national trends survey," *Journal of health communication*, vol. 16, no. 1, pp. 34–49, 2010.
- [18] X. Chen, J. L. Hay, E. A. Waters, M. T. Kiviniemi, C. Biddle, E. Schofield, Y. Li, K. Kaphingst, and H. Orom, "Health literacy and use and trust in health information," *Journal of health communication*, vol. 23, no. 8, pp. 724–734, 2018.
- [19] W. Wang and K. Siau, "Living with artificial intelligence—developing a theory on trust in health chatbots," in *Proceedings of the sixteenth annual pre-ICIS workshop on HCI research in MIS*. Association for Information Systems San Francisco, CA, 2018, pp. 1–5.
- [20] A. Viswanath Prakash and S. Das, "Would you trust a bot for healthcare advice? an empirical investigation," 2020.
- [21] U. Gnewuch, S. Morana, and A. Mädche, "Towards designing cooperative and social conversational agents for customer service," *International Conference on Information Systems*, Jan. 2017. [Online]. Available: <https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1312&context=icis2017>
- [22] IBM, "IBM watsonx Assistant Virtual Agent," IBM. [Online]. Available: <https://www.ibm.com/products/watsonx-assistant>
- [23] "Chatbot," *Cambridge dictionary*. Accessed November, 2023. [Online]. Available: <https://dictionary.cambridge.org/dictionary/english/chatbot>
- [24] J. H. Lee, H. Yang, D. Shin, and H. Kim, "Chatbots," *ELT Journal*, vol. 74, no. 3, pp. 338–344, July 2020. [Online]. Available: <https://doi.org/10.1093/elt/ccaa035>
- [25] P. Kucherbaev, A. Bozzon, and G. Houben, "Human-Aided bots," *IEEE Internet Computing*, vol. 22, no. 6, pp. 36–43, Nov. 2018. [Online]. Available: <https://doi.org/10.1109/mic.2018.252095348>
- [26] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, July 2022. [Online]. Available: <https://doi.org/10.1007/s11042-022-13428-4>

- [27] J. Weizenbaum, "ELIZA - a computer program for the study of natural language communication between man and machine," *Communications of The ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966. [Online]. Available: <https://doi.org/10.1145/365153.365168>
- [28] R. Dale, "The return of the chatbots," *Natural Language Engineering*, vol. 22, no. 5, pp. 811–817, Sep. 2016. [Online]. Available: <https://doi.org/10.1017/s1351324916000243>
- [29] R. M. French, "The Turing Test: the first 50 years," *Trends in Cognitive Sciences*, vol. 4, no. 3, pp. 115–122, Mar. 2000. [Online]. Available: [https://doi.org/10.1016/s1364-6613\(00\)01453-4](https://doi.org/10.1016/s1364-6613(00)01453-4)
- [30] R. Wallace, *The anatomy of A.L.I.C.E.*, Nov. 2007. [Online]. Available: https://doi.org/10.1007/978-1-4020-6710-5_13
- [31] Engati, "Artificial Intelligence Markup Language (AIML)," Engati, 2021. [Online]. Available: <https://www.engati.com/glossary/artificial-intelligence-markup-language>
- [32] OpenAI, "GPT-4 Technical Report," *arXiv (Cornell University)*, Mar. 2023. [Online]. Available: <http://arxiv.org/abs/2303.08774>
- [33] M. Yang, J. Jiang, M. Kiang, and F. Yuan, "Re-examining the impact of multidimensional trust on patients' online medical consultation service continuance decision," *Information Systems Frontiers*, pp. 1–25, 2021.
- [34] S. Gulati, S. Sousa, and D. Lamas, "Design, development and evaluation of a human-computer trust scale," *Behaviour & Information Technology*, vol. 38, no. 10, pp. 1004–1015, 2019.
- [35] J. A. Pesonen, "'Are You OK?' Students' Trust in a Chatbot Providing Support Opportunities," in *International Conference on Human-Computer Interaction*. Springer, 2021, pp. 199–215.
- [36] I. Benbasat and W. Wang, "Trust in and adoption of online recommendation agents," *Journal of the association for information systems*, vol. 6, no. 3, p. 4, 2005.
- [37] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.
- [38] Ada Health GmbH, "Health." Ada, 2023. [Online]. Available: <https://ada.com/>
- [39] Sensely, "Conversational AI to improve health and drive member engagement," Sensely, 2023. [Online]. Available: <https://sensely.com/>
- [40] Buoy Health, "Check symptoms & find the right care," Buoy Health, 2018. [Online]. Available: <https://www.buoyhealth.com/>
- [41] Infermedica, "Make healthcare decisions with confidence," Infermedica, 2023. [Online]. Available: <https://infermedica.com/>
- [42] Healthily, "Your health questions, answered," Healthily, 2023. [Online]. Available: <https://www.livehealthily.com/>
- [43] Keenethics, "Making the lives of cancer survivors easier," OneRemission, Oct. 2023. [Online]. Available: <https://keenethics.com/project-one-remission>
- [44] Youper, "Artificial Intelligence For Mental Health Care," Youper, 2023. [Online]. Available: <https://www.youper.ai/>

- [45] Woebot Health, “Relational Agent for Mental Health,” Woebot Health, Nov. 2023. [Online]. Available: <https://woebothealth.com/>
- [46] Babylon Healthcare Services [eMed], “Video call a doctor anytime, anywhere,” eMed UK, 2023. [Online]. Available: <https://www.emed.com/uk>
- [47] PACT Care BV, “Your health assistant,” Florence, 2019. [Online]. Available: <https://florence.chat/>
- [48] M. L. Clayman, J. A. Manganello, K. Viswanath, B. W. Hesse, and N. K. Arora, “Providing health messages to hispanics/latinos: understanding the importance of language, trust in health information sources, and media use,” *Journal of health communication*, vol. 15, no. sup3, pp. 252–263, 2010.
- [49] D. D. Dobrzykowski and M. Tarafdar, “Understanding information exchange in healthcare operations: Evidence from hospitals and patients,” *Journal of Operations Management*, vol. 36, pp. 201–214, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0272696315000029>
- [50] B. Powers, J. V. Trinh, and H. B. Bosworth, “Can this patient read and understand written health information?” *JAMA*, vol. 304, no. 1, p. 76, July 2010. [Online]. Available: <https://doi.org/10.1001/jama.2010.896>
- [51] J. E. Jordan, R. Buchbinder, A. M. Briggs, G. R. Elsworth, L. Busija, R. Batterham, and R. H. Osborne, “The health literacy management scale (helms): A measure of an individual’s capacity to seek, understand and use health information within the healthcare setting,” *Patient Education and Counseling*, vol. 91, no. 2, pp. 228–235, 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0738399113000414>
- [52] L. Nielsen-Bohlman, A. Panzer, and D. Kindig, “Committee on health literacy, board on neuroscience and behavioral health, institute of medicine,” *Health literacy: A prescription to end confusion*, 2004.
- [53] D. H. McKnight, V. Choudhury, and C. Kacmar, “Developing and validating trust measures for e-commerce: An integrative typology,” *Information systems research*, vol. 13, no. 3, pp. 334–359, 2002.
- [54] S. Barnoy, O. Levy, and Y. Bar-Tal, “What makes patients perceive their health care worker as an epistemic authority?” *Nursing Inquiry*, vol. 19, no. 2, pp. 128–133, July 2011. [Online]. Available: <https://doi.org/10.1111/j.1440-1800.2011.00562.x>
- [55] K. Stasiuk, Y. Bar-Tal, and R. Maksymiuk, “The effect of physicians’ treatment recommendations on their epistemic authority: the medical expertise bias,” *Journal of Health Communication*, vol. 21, no. 1, pp. 92–99, Oct. 2015. [Online]. Available: <https://doi.org/10.1080/10810730.2015.1049308>
- [56] Windsor University School of Medicine, “10 types of surgeons that perform surgery,” July 2023. [Online]. Available: <https://www.windsor.edu/10-types-of-surgeons-that-perform-surgery/>
- [57] M. S. Komrad, “A defence of medical paternalism: maximising patients’ autonomy.” *Journal of Medical Ethics*, vol. 9, no. 1, pp. 38–44, Mar. 1983. [Online]. Available: <https://doi.org/10.1136/jme.9.1.38>
- [58] MU School of Medicine, “Provider-Patient relationship.” [Online]. Available: <https://medicine.missouri.edu/centers-institutes-labs/health-ethics/faq/provider-patient-relationship>

- [59] N. Suzuki and Y. Katagiri, "Prosodic alignment in human–computer interaction," *Connection Science*, vol. 19, no. 2, pp. 131–141, 2007.
- [60] S. Stoyanchev and A. Stent, "Lexical and syntactic adaptation and their impact in deployed spoken dialog systems," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 2009, pp. 189–192.
- [61] B. R. Cowan, H. P. Branigan, M. Obregón, E. Bugis, and R. Beale, "Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue," *International Journal of Human-Computer Studies*, vol. 83, pp. 27–42, 2015.
- [62] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, and G. Mark, "Style and alignment in information-seeking conversation," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 2018, pp. 42–51.
- [63] R. Levitan, S. Benus, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar." in *Interspeech*, vol. 16, 2016. [Online]. Available: <https://doi.org/10.21437/interspeech.2016-985>
- [64] L. E. Scissors, A. J. Gill, and D. Gergle, "Linguistic mimicry and trust in text-based cmc," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 2008, pp. 277–280.
- [65] L. E. Scissors, A. J. Gill, K. Geraghty, and D. Gergle, "In CMC we trust: The role of similarity," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2009, pp. 527–536.
- [66] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski, "An end-to-end conversational style matching agent," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 111–118.
- [67] "Pexels." [Online]. Available: <https://www.pexels.com/about/>
- [68] Aviv Clinics, "Brain injury treatment [image]," 2021. [Online]. Available: https://aviv-clinics.ae/wp-content/uploads/sites/5/2021/07/Aviv-Clinics_Dubai_TBI-1.png
- [69] TeachMeAnatomy, "The ear - TeachMeAnatomy," Mar. 2024. [Online]. Available: <https://teachmeanatomy.info/head/organs/ear/>
- [70] "Qualtrics XM - Experience Management Software," Apr. 2024. [Online]. Available: <https://www.qualtrics.com/nl/?rid=langMatch&prevsite=en&newsite=nl&geo=NL&geomatch=>
- [71] Federatie Medisch Specialisten, "Perceptieve slechthoerendheid bij volwassenen - Richtlijn - Richtlijndatabase," Feb. 2016. [Online]. Available: https://richtlijndatabase.nl/richtlijn/perceptieve_slechthoerendheid_bij_volwassenen/perceptieve_slechthoerendheid_-_startpagina.html
- [72] Ministerie van Volksgezondheid, Welzijn en Sport, "Standpunt hyperbare zuurstoftherapie (HBOT)," 2019. [Online]. Available: <https://www.zorginstituutnederland.nl/publicaties/standpunten/2019/06/11/standpunt-hyperbare-zuurstoftherapie-hbot>
- [73] Ministerie van Defensie, "Genezing van acute doofheid na knal, schoten of ontploffing blijkt mogelijk," *Defensie.nl*, Nov. 15, 2022. [Online]. Available: <https://www.defensie.nl/actueel/nieuws/2022/09/21/genezing-van-acute-dooftheid-na-knal-schoten-of-ontploffing-blijkt-mogelijk>

- [74] Antonius Hypercare, “Knaltrauma,” *Hypercare*, Oct. 1, 2020. [Online]. Available: <https://hypercare.nl/knaltrauma/>
- [75] A. B. Bayoumy and J. A. de Ru, “The use of hyperbaric oxygen therapy in acute hearing loss: a narrative review,” *European Archives of Oto-Rhino-Laryngology*, vol. 276, no. 7, pp. 1859–1880, May 2019. [Online]. Available: <https://doi.org/10.1007/s00405-019-05469-7>
- [76] N. Yehudai, N. Fink, M. Shpriz, and T. Marom, “Acute acoustic trauma among soldiers during an intense combat,” *Journal of the American Academy of Audiology*, vol. 28, no. 05, pp. 436–443, 2017.
- [77] InfoNu.nl, “Knaltrauma: symptomen, oorzaken, behandeling en herstel,” Aug. 2023. [Online]. Available: <https://mens-en-gezondheid.infonu.nl/aandoeningen/197974-knaltrauma-symptomen-oorzaken-behandeling-en-herstel.html/>
- [78] S. J. Czaja, N. Charness, A. D. Fisk, C. Hertzog, S. N. Nair, W. A. Rogers, and J. Sharit, “Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (create).” *Psychology and Aging*, vol. 21, no. 2, pp. 333–352, June 2006. [Online]. Available: <https://doi.org/10.1037/0882-7974.21.2.333>
- [79] I. Hecker, S. Spaulding, and D. Kuehn, “Digital Skills and Older Workers Supporting Success in Training and Employment in a Digital World,” *Urban Institute*, Nov. 2021.
- [80] Ministerie van Defensie, “Royal Dutch Navy,” Jan. 2023. [Online]. Available: <https://english.defensie.nl/organisation/navy>
- [81] AISNSW, “Strategies to Check for Understanding (CFU),” AISNSW empowering independent education. [Online]. Available: <https://www.aisnsw.edu.au/>
- [82] E. Paus and R. Jucks, “Do we really mean the same? The relationship between word choices and computer mediated cooperative learning,” 2008.
- [83] A. Pinto, S. Sousa, A. Simões, J. Santos *et al.*, “A Trust Scale for Human-Robot Interaction: Translation, Adaptation, and Validation of a Human Computer Trust Scale,” *Human Behavior and Emerging Technologies*, vol. 2022, 2022.
- [84] A. Hinderks, M. Schrepp, J. Thomaschewski, and Team UEQ, “User Experience Questionnaire (UEQ),” 2018. [Online]. Available: <https://www.ueq-online.org/>
- [85] S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, and M. Mctear, “Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?” in *Proceedings of the 31st European Conference on Cognitive Ergonomics*, ser. ECCE ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 207–214. [Online]. Available: <https://doi.org/10.1145/3335082.3335094>
- [86] Rasa Technologies Inc, “Conversational AI Platform | Superior customer experiences start here,” Rasa, Oct. 2023. [Online]. Available: <https://rasa.com/>
- [87] Google, “DialogFlow,” Google Cloud. [Online]. Available: <https://cloud.google.com/dialogflow?hl=nl>
- [88] Netlify, “Connect everything. Build anything.” 2024. [Online]. Available: <https://www.netlify.com/>
- [89] GitHub Inc., “GitHub: Let’s build from here,” 2024. [Online]. Available: <https://github.com/>

- [90] ngrok inc, “ngrok | Unified Application Delivery Platform for Developers,” 2024. [Online]. Available: <https://ngrok.com/>
- [91] B. Manav, “Color-emotion associations and color preferences: A case study for residences,” *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, vol. 32, no. 2, pp. 144–150, 2007.
- [92] ETKHO hospital engineering, “The importance of colour in hospitals,” Dec. 2022. [Online]. Available: <https://www.etkho.com/en/the-importance-of-colour-in-hospitals/>
- [93] WhatsApp LLC, “WhatsApp,” 2024. [Online]. Available: https://www.whatsapp.com/?lang=nl_NL
- [94] The John Hopkins University and The John Hopkins Hospital and The John Hopkins Health System, “Hyperbaric oxygen therapy,” May 2022. [Online]. Available: <https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/hyperbaric-oxygen-therapy>
- [95] FDA, “Hyperbaric oxygen therapy: Get the facts,” July 2021. [Online]. Available: <https://www.fda.gov/consumers/consumer-updates/hyperbaric-oxygen-therapy-get-facts>
- [96] M. Á. Ortega, O. Fraile-Martínez, C. García-Montero, E. Callejón-Peláez, M. A. Sáez, M. Á. Alvarez-Mon, N. García-Honduvilla, J. Monserrat, M. Álvarez-Mon, J. Buján, and M. L. Canals, “A general overview on the hyperbaric oxygen therapy: applications, mechanisms and translational opportunities,” *Medicina-lithuania*, vol. 57, no. 9, p. 864, Aug. 2021. [Online]. Available: <https://doi.org/10.3390/medicina57090864>
- [97] Amsterdam UMC Universitair Medische Centra, “Hyperbare Geneeskunde.” [Online]. Available: <https://www.amc.nl/web/specialismen/hyperbare-geneeskunde/hyperbare-geneeskunde/hyperbare-geneeskunde-polikliniek.htm>
- [98] MC Hyperbare Zuurstofftherapie, “Wat is hyperbare zuurstoftherapie?” Oct. 2020. [Online]. Available: <https://www.hyperbaarcentrum.nl/veelgestelde-vragen/>
- [99] OpenAI, “ChatGPT.” [Online]. Available: <https://chat.openai.com/>
- [100] QuillBot, a Learneo, Inc. business, “Paraphrasing Tool - Quillbot AI.” [Online]. Available: <https://quillbot.com/>
- [101] QuillBot (Course Hero), LLC. 2023, “How Does QuillBot Work?” July 2023. [Online]. Available: <https://quillbot.com/blog/how-does-quillbot-work/>
- [102] A.D.A.M., “Medical Encyclopedia: MedlinePlus.” [Online]. Available: <https://medlineplus.gov/encyclopedia.html>
- [103] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [104] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, “BioWordVec, improving biomedical word embeddings with subword information and MeSH,” *Scientific Data*, vol. 6, no. 1, May 2019. [Online]. Available: <https://doi.org/10.1038/s41597-019-0055-0>

- [105] —, “BioWordVec: Improving Biomedical Word Embeddings with Subword Information and MeSH Ontology,” Sep. 2018. [Online]. Available: https://figshare.com/articles/dataset/Improving_Biomedical_Word_Embeddings_with_Subword_Information_and_MeSH_Ontology/6882647
- [106] National Library of Medicine, “Medical subject headings,” Jan. 2024. [Online]. Available: <https://www.nlm.nih.gov/mesh/meshhome.html>
- [107] —, “PubMed Overview,” Aug. 2023. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/about/>
- [108] —, “MEDLINE,” Feb. 2024. [Online]. Available: https://www.nlm.nih.gov/medline/medline_home.html
- [109] Rasa Technologies GmbH, “NLU training data,” Mar. 2024. [Online]. Available: <https://rasa.com/docs/rasa/nlu-training-data/>
- [110] NLTK Project, “nltk.stem.WordNetLemmatizer,” Jan. 2023. [Online]. Available: <https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html?highlight=wordnet>
- [111] Python Software Foundation, “inflect 7.2.0,” Mar. 2024. [Online]. Available: <https://pypi.org/project/inflect/>
- [112] D. N. M. De Gruijter, “Toetsing en toetsanalyse,” *Rijksuniversiteit Leiden, Leiden, the Netherlands [Testing and test analysis]*, 2008.
- [113] M. Davies, “English-Corpora: COCA,” 2019. [Online]. Available: <https://www.english-corpora.org/coca/>
- [114] Instituut voor de Nederlandse Taal, “OpenSoNaR,” 2023. [Online]. Available: <http://opensonar.ivdnt.org/>
- [115] B. Thormundsson, “ChatGPT - statistics & facts,” Apr. 2023. [Online]. Available: <https://www.statista.com/topics/10446/chatgpt/#topicOverview>

A QUESTIONNAIRE TO INFER TERMINOLOGY PREFERENCE (FINAL VERSION)

Below are the questions presented that were created to infer participants' preference of terminology for the various placeholders defined in this research. The questions follow a questionnaire format, where each question requires an answer, moreover, one is not able to go back to previous questions. Note that only the visuals that are essential for the interpretation of a question are presented here, the other visuals are left out for brevity.

1. Please select the term you would prefer to use for the blank.
Your friend is telling you he is experiencing loss of taste and smell as well as fatigue and muscle aches. These are all typical [blank] for COVID-19.
 - (a) Characteristics
 - (b) Conditions
 - (c) Indicators
 - (d) Indications
 - (e) Signs
 - (f) Symptoms
 - (g) I do not know
 - (h) Other ...

2. Please select the term you would prefer to use for the blank.
After a COVID-19 test and a visit to the doctor, COVID-19 is indeed [blank].
 - (a) Confirmed
 - (b) Determined
 - (c) Diagnosed
 - (d) Discovered
 - (e) Identified
 - (f) Verified
 - (g) I do not know
 - (h) Other ...

3. Please select the term you would prefer to use to name the device in the picture. * A.1
 - (a) Breathing mask
 - (b) Full face mask
 - (c) Inhalation mask

- (d) Oxygen mask
 - (e) Oxygen therapy mask
 - (f) Respiratory mask
 - (g) I do not know
 - (h) Other ...
4. Please select the term you would prefer to use for the blank.
When starting with the new medication, he experienced some [blank] resulting in an allergic reaction and the need to stop the medication.
- (a) Complications
 - (b) Consequences
 - (c) Discomfort
 - (d) Medical risks
 - (e) Reactions
 - (f) Risks
 - (g) Side effects
 - (h) I do not know
 - (i) Other ...
5. Please select the term you would prefer to use for the blank.
On short flights, it is often asked to not make use of the [blank].
- (a) Bathroom
 - (b) Lavatory
 - (c) Loo
 - (d) Latrine
 - (e) Restroom
 - (f) Toilet
 - (g) Water Closet (WC)
 - (h) I do not know
 - (i) Other ...
6. Please select the term you would prefer to use for the blank.
In the aeroplane, certain seats have an increased price due to their location in the plane and the ability to more easily [blank] in case of emergency.
- (a) Depart
 - (b) Exit
 - (c) Get out
 - (d) Go out
 - (e) Leave
 - (f) Step out
 - (g) I do not know
 - (h) Other ...

7. Please select the term you would prefer to use to describe the following:
The process in an aeroplane where the pressure in the cabin is adjusted to maintain a comfortable environment, by adjusting the pressure to go up?
- (a) Compression
 - (b) Elevating the pressure
 - (c) Heightening the pressure
 - (d) Increasing the pressure
 - (e) Intensifying the pressure
 - (f) I do not know
 - (g) Other ...
8. Please select the term you would prefer to describe the following:
The process in an aeroplane where the pressure in the cabin is adjusted to maintain a comfortable environment, by adjusting the pressure to go down?
- (a) Decompression
 - (b) Decreasing the pressure
 - (c) Reducing the pressure
 - (d) I do not know
 - (e) Other ...
9. Please select the term you would prefer to use to describe the subject of the two pictures.
* A.2
- (a) Din
 - (b) Clamour
 - (c) Sound
 - (d) Noise
 - (e) I do not know
 - (f) Other ...
10. Please select the term you would prefer to use for the blank.
Today Bart has had his first of the eight [blank] in total with his new speech therapist.
- (a) Appointments
 - (b) Gatherings
 - (c) Meetings
 - (d) Occasions
 - (e) Sessions
 - (f) Visits
 - (g) I do not know
 - (h) Other ...
11. Please select the term you would prefer to use for the blank.
Today, the speech therapist tried to identify the [blank] of Bart his speech delay. The results show that Bart has problems in speech production as a result of a previous surgery on his teeth.

- (a) Causation
 - (b) Cause
 - (c) Determinant
 - (d) Origin
 - (e) Reason
 - (f) Trigger
 - (g) I do not know
 - (h) Other ...
12. Please select the term you would prefer to use for the blank.
The speech therapist needed 60 minutes today, however, the normal [blank] is only 30 minutes.
- (a) Duration
 - (b) Length
 - (c) Span
 - (d) Time
 - (e) Timespan
 - (f) I do not know
 - (g) Other ...
13. Please select the term you would prefer to use to talk about the following:
Someone experiences lots of difficulties in hearing sounds and conversations.
- (a) Auditory deficiency
 - (b) Hearing deficit
 - (c) Hearing impairment
 - (d) Hearing loss
 - (e) I do not know
 - (f) Other ...
14. Please select the term you would prefer to talk about the following:
How well a person hears.
- (a) Clarity of hearing
 - (b) Clearness of hearing
 - (c) Hearing ability
 - (d) Hearing sensitivity
 - (e) Hearing threshold
 - (f) I do not know
 - (g) Other ...
15. Please select the term you would prefer to use for the blank.
In the context of diving, one carries a scuba tank that is filled with 100% [blank] in order to be able to breath.
- (a) Air

- (b) O2
 - (c) Oxygen
 - (d) I do not know
 - (e) Other ...
16. Please select the term you would prefer to use for the blank.
In order to breath properly, the process of [blank] slowly and deeply through the mouth to fill the lungs completely is important.
- (a) Breathing in
 - (b) Inhaling
 - (c) Respiring
 - (d) Sucking in
 - (e) I do not know
 - (f) Other ...
17. Please select the term you would prefer to use for the blank.
After this, the process of [blank] slowly and fully is important to empty the lungs again.
- (a) Breathing out
 - (b) Exhaling
 - (c) Releasing
 - (d) Ventilating
 - (e) I do not know
 - (f) Other ...
18. Please select the term you would prefer to use for the blank.
When going to physiotherapy, you should wear sportive and comfortable [blank] such as leggings and a sports top.
- (a) Attire
 - (b) Clothes
 - (c) Clothing
 - (d) Fashion
 - (e) Outfit
 - (f) I do not know
 - (g) Other ...
19. Please select the term you would prefer to use to refer to the people in the picture. * A.3
- (a) Experts
 - (b) Health professionals
 - (c) Healthcare personnel
 - (d) Medical experts
 - (e) Medical personnel
 - (f) Medical professionals
 - (g) Medical staff

- (h) I do not know
 - (i) Other ...
20. Please select the term you would prefer to use to refer to the person standing in blue in the picture. * A.4
- (a) Caretaker
 - (b) Medic
 - (c) Nurse
 - (d) Registered nurse
 - (e) Specialist
 - (f) I do not know
 - (g) Other ...
21. Please select the term you would prefer to use to refer to the person in the picture. * A.5
- (a) Client
 - (b) Healthcare consumer
 - (c) Patient
 - (d) Recipient
 - (e) Sufferer
 - (f) I do not know
 - (g) Other ...
22. Please select the term you would prefer to use for the blank.
Soon there will be more information about the hospital where the surgery will be [blank].
- (a) Carried out
 - (b) Conducted
 - (c) Given
 - (d) Happening
 - (e) Performed
 - (f) Taking place
 - (g) I do not know
 - (h) Other ...
23. Please select the term you would prefer to use to name area A in the picture. * A.6
- (a) External ear
 - (b) External part of the ear
 - (c) Outer ear
 - (d) Outer part of the ear
 - (e) I do not know
 - (f) Other ...
24. Please select the term you would prefer to use to name area B in the picture. * A.6
- (a) Middle ear

- (b) Middle ear cavity
- (c) Middle part of the ear
- (d) Tympanum
- (e) I do not know
- (f) Other ...

25. Please select the term you would prefer to use to name area C in the picture. * A.6

- (a) Inner ear
- (b) Inner part of the ear
- (c) Internal ear
- (d) Labyrinth of the ear
- (e) I do not know
- (f) Other ...

26. Please select the term you would prefer to use to name area D in the picture. Note, this part is connected to upper part of the throat behind the nose. * A.7

- (a) Eardrum
- (b) Myringa
- (c) I do not know
- (d) Other ...

B SUMMARY REPORT

B.1 First version of the summary report

Here, the first version of the summary report is shown. Note that this was created in close collaboration with an expert in the field and more information on this report can be found in Section 4.1.2.

Introduction

Recently, you have had an appointment with your medical professional. After this appointment, you cannot remember all the details that were discussed and still have questions about what was discussed. Luckily, you have access to a summary report from your medical professional in your medical records. This report, provided below, is an overview of the appointment. However, it is possible that this does not address all your questions, or new questions arise. Therefore you will have the opportunity to interact with MedWiseBot. You can ask any (additional) questions you may have and MedWiseBot will assist you further.

Summary report on the diagnosis and treatment

Complaint

The patient reports hearing loss symptoms as well as having a ringing in the right ear. The patient experienced an explosion caused by fireworks at a ceremony at which the patient was present. The patient describes exposure to intense noise during the explosion after which ringing in the right ear started directly. After examination of the patient, the tympanic membrane is found to be normal, but there is decreased auditory acuity in the right ear.

Physical examinations

- Audiometry
 - o This is conducted to assess the patient's auditory acuity. This reveals severe hearing loss in the right ear of the patient
- Otoscopy
 - o This is conducted to assess the tympanic membrane. This reveals the tympanic membrane is intact.
- Tuning forks
 - o This is conducted to assess the patient's auditory acuity. This is done using the Weber test as well as the Rinne test. This confirms the assessment of hearing loss by audiometry.

Diagnosis

The patient is diagnosed with acute acoustic trauma, characterised by hearing loss and tinnitus resulting from exposure to an acoustic trauma.

Treatment plan

Based on the acute nature of the patient's condition and the potential of permanent hearing loss in case the condition is not treated, the medical professionals recommend starting with corticosteroids in combination with hyperbaric oxygen therapy. For the corticosteroids, 1 mg/kg with a maximum of 60 mg is prescribed for 7 days in which the dose is gradually decreased over the days. The hyperbaric oxygen therapy consists of daily sessions for 10 consecutive days, also during the weekends.

Follow-up

The patient is strictly monitored by assessing the hearing loss using the Weber test and Rinne test after each session of the hyperbaric oxygen therapy. Furthermore, otoscopy is conducted after each session of the hyperbaric oxygen therapy to assess the tympanic membrane as well as the Weber and Rinne test to keep track of the degree of hearing loss

B.2 Second version of the summary report

Here, one can find the adjusted summary report after pilot testing. This is a more "simplified" version of the original summary report. More about the reasoning behind the adjustments can be found in Section 5.6.2.

Information for the participant

Recently, you attended a ceremony where an unexpected fireworks explosion occurred close to where you were standing. Immediately, you noticed you could not properly hear anymore, especially in your right ear. After a few moments, this feeling of limited hearing was still present as well as a constant ringing sound in your right ear.

As you were scared for permanent hearing damage, you consulted your doctor the same day. After some physical examinations, the doctor provided you with a diagnosis and proposed treatment plan. After this visit you checked the report from the visit once more, but you would like to know more information about the treatment plan and diagnosis. Below, the details of the appointment with the doctor are outlined:

MEDICAL EXAMINATION REPORT	
COMPLAINT: Hearing loss and ringing in the right ear after exposure to intense noise	
PHYSICAL EXAMINATIONS	
Audiometry	Observed severe hearing loss right ear
Otосcopy	Tympanic membrane intact
Tuning forks (Weber and Rinne test)	Severe hearing loss right ear confirmed
DIAGNOSIS: Acute acoustic trauma	
TREATMENT PLAN: A combination of corticosteroids and hyperbaric oxygen therapy.	
- Corticosteroids: Dosage is 1 mg/kg with a maximum of 60 mg for 7 days with gradual decrease.	
- Hyperbaric oxygen therapy: Daily sessions for 10 consecutive days, including weekends	
FOLLOW-UP: Strictly monitored using the Weber test and Rinne test after each session of the hyperbaric oxygen therapy, as well as otосcopy to assess the tympanic membrane.	

After reading this summary, you would like to know more about the proposed treatment in order to make an informed decision about your next steps. To assist you in this process, you will chat with MedWiseBot to ask any questions you have regarding the proposed treatment and diagnosis.

When you have read this document, please notify the researcher so that the interaction with MedWiseBot can start. Do not worry about remembering all the content of this document, as MedWiseBot will be able to offer assistance if needed.

C EXPERT INTERVIEW

Date: 19/03/2024, location: Meteren

Name interviewer: Keara Schaaij, name interviewee is not presented.

C.1 Introduction

Interviewer

Thank you for participating in this interview. Your expertise and knowledge will greatly contribute to this study exploring health-related information exchange between chatbot and patient.

This research is specifically focusing on the difference in adjusting the information provided by the chatbot to the user compared to the way healthcare professionals normally communicate health-related information (related to word use, concepts, language use, etc.). The scenario for the experiment is that the participant is diagnosed with acute acoustic trauma, and a healthcare professional proposes hyperbaric oxygen therapy as treatment. The participant is asked to chat with the chatbot to ask questions related to the diagnosed disease and proposed treatment.

In order to be able to replicate the way healthcare professionals normally communicate this kind of information, this interview will dive a bit deeper into aspects of the disease, treatment, and manner of language used in the context of medical information exchange.

Before we begin, could you please introduce yourself and the background you have related to the topic of the disease acute acoustic trauma and the treatment hyperbaric oxygen therapy?

Interviewee

I am a military nurse in the Navy and also a certified hyperbaric nurse. Which means I work mostly on board navy vessels with divers who are more likely to get a decompression illness. A decompression illness has to be treated in a hyperbaric chamber, there is one on board the ship. When I am not working on board a Navy vessel, I work at the diving medical centre in Den Helder. In the Divers Medical Centre, there is a large hyperbaric chamber that is also used to treat patients with acute acoustic trauma. I have assisted in multiple sessions of hyperbaric oxygen therapy as well.

C.2 General questions

Interviewer

Do you have typical questions that are asked related to acute acoustic trauma?

Interviewee

Yes, for sure, questions such as *What can be done about it?* or *Will I ever be able to hear again or better?* are often asked.

Interviewer

What answers do you typically provide to these questions?

Interviewee

For the first question, there are not many options. The only option there is, is to start with corticosteroids and start with hyperbaric oxygen therapy. However, this is only possible if the patient checks all the requirements for the hyperbaric oxygen therapy. Otherwise, it is just the corticosteroids that reduce the swelling, and we hope for the best.

Then, for the other questions, I will explain that by starting the treatment, a great chance exists that the hearing will get a bit better, it is probably never going to be the same as it was before. But the chances are high that there will be some improvements.

Interviewer

Do you have typical questions that are asked related to hyperbaric oxygen therapy?

Interviewee

Yes, the most asked questions are things like the time the treatment takes, how it works, what happens inside the chamber, whether people can use the toilet or go out of the chamber at all, the workings of the oxygen mask, whether they can eat or drink in the chamber, what happens when they are not able to equalise the eardrums, and of course, whether there are any risks.

Interviewer

What answer do you typically provide to the question of whether people are allowed to use the toilet or go out of the chamber?

Interviewee

You will not be able to use the toilet. You will not be able to open the door when the chamber is under pressure. When you have to use the toilet, you will get a bucket. There is a small compartment that can be used for the use of the bucket, where no one can see you. When you have to do this, you will have to do it during the 5-minute "normal" oxygen time. And overall, you are not able to leave the chamber during the treatment. In case of an emergency like a fire inside the chamber, the operator will make sure the pressure is decreased again, and you will be able to leave the chamber. In case of a fire, there is a fire extinguisher inside the chamber so that the nurse is able to fight the fire and make sure everyone is safe.

Interviewer

What kind of answer do you provide to the question related to equalising the eardrums?

Interviewee

If you are not able to use the Valsalva method, there are other ways to equalise your eardrums. You can yawn or drink from a bottle of water. When those things are not working, the nurse has a xylometazoline nasal spray, which can also help to open up the Eustachian tube to eventually equalise the pressure.

Interviewer

And for the question related to eating and drinking?

Interviewee

You are not allowed to eat in the hyperbaric chamber, but you can bring a water bottle. Just make sure that the bottle is open because of the rising pressure. Otherwise, it is going to explode.

Interviewer

When you propose hyperbaric oxygen therapy to someone suffering from acute acoustic trauma, in what way is this communicated to that person?

Interviewee

During a consultation with the physician or the nurse. The treatment has to be started preferably as soon as possible. You will get the best outcome when you start within 48 hours. But it is possible to start within 7 days after the trauma.

Interviewer

More specifically, what kind of information is provided to the patient?

Interviewee

Just an explanation about the treatment and how it is going to work. Often, the risks and outcome possibilities are also discussed.

Interviewer

Do you provide them with brochures or something like that?

Interviewee

No, I do not have any brochures about the treatment. However, I am not working in a hospital either, so probably hospitals will have brochures.

At this point, the interviewer explained the idea of the research in a bit more detail, and the interviewee recommended writing a so-called medical case that the participant could read before interacting with the chatbot. The following is a summary of what should be in the medical case. Besides this, the interviewee proposed to read through the created case in order to make sure the information was correct.

Interviewee

Such a medical case needs to include some little summary about the event that has happened and for which the person needs to see the doctor, so to speak. Moreover, in this context, it will explain certain examinations done in order to check aspects of the disease, such as otoscopy and other methods. Then, there is a short aspect that explains the diagnosis, and then the treatment is proposed. This will provide the participant with a rather global understanding of the situation.

Interviewer

When someone asks you why they should choose a certain treatment or what the benefits and risks are, what type of information do you typically provide to the user?

Interviewee

Just verbal information most of the time and sometimes depending on the disease and the possible treatment options the website thisarts.nl. When telling the patient the treatment options, I will explain to them not only the most common and recommended option but also the other options. Perhaps, depending on the disease, a plan B is discussed. Next to this, I always tell them the risks of a treatment and, in what cases, they have to get back to me or the physician. Those things are very important to tell a patient.

C.3 Question related to acute acoustic trauma

Interviewer

Could you briefly describe the disease acute acoustic trauma?

Interviewee

It is an acute hearing impairment, a sort of sensorineural hearing loss, that is caused by intense noise impact.

Interviewer

What are the main causes of acute acoustic trauma?

Interviewee Very loud noise, mostly trauma. You can think of guns going off near the ear or an explosion. Just all kinds of very loud noise near the ear.

Interviewer

What are the treatment possibilities for acute acoustic trauma?

Interviewee

There are not many options. The most common one is starting with corticosteroids and, nowadays, hyperbaric oxygen therapy. At first, only corticosteroids were possible, and then we hoped for the best.

Interviewer

What is the chance of recovery?

Interviewee

Well, the hearing will probably never be the same as it was before the trauma. However, there is a great chance the hearing will improve due to the corticosteroids and the hyperbaric oxygen therapy.

C.4 Questions related to hyperbaric oxygen therapy

Interviewer

Could you explain hyperbaric oxygen therapy?

Interviewee

Normal oxygen, the air we breathe now, contains 21% oxygen. The oxygen you will breathe in the hyperbaric chamber contains 100% oxygen. The chamber is also pressurised, which helps the lungs gather and absorb more oxygen. This is due to a lot of laws in physics (Henry's law and Boyle). Due to this environment with 100% oxygen resulting in more of this oxygen in your lungs, the blood will uptake more oxygen as well. This, then, will be able to go to the tissues, which helps with wound healing, or in this case, hearing loss.

Interviewee

Could you explain how hyperbaric oxygen therapy works in general?

Interviewee

Of course, I will take you through the "rough" procedure of the treatment.

It starts with the patient, you, entering the chamber. You have to make sure you are wearing

comfortable clothing, like a t-shirt and a cardigan or sweater. You leave your shoes, and you have to leave all other things behind. Electronic devices are not allowed, like a smartphone or smartwatch. Then, inside the chamber, you lie down in a bed or sit in the chair, but do not cross your legs or arms (crossing your legs or arms is an extra risk of getting a decompression illness). You have to put on your hearing protection, which is provided in the chamber. You make a thumbs-up sign when you are ready to go. You will be able to speak with the chamber operator, but during the pressure buildup, there is a lot of noise. So that is why you have to make the thumbs-up sign during the pressure buildup. Next to this, there is a camera in the chamber, so the operator can see you at all times.

During the pressure buildup, which takes about 2 minutes, you have to equalise your eardrums. Which you can do with the Valsalva method, yawning, drinking water, or chewing gum. When you get pain in your ears or neck, you should stop the thumbs-up sign so the chamber operator can stop the pressure buildup. This gives you some more time to equalise the eardrums. When you are not able to do this, the nurse will give you a xylometazoline nasal spray to help open up the Eustachian tube.

When the pressure is equal to 14 metres under water, you can put your hearing protection off and your oxygen mask on. The oxygen mask makes sure you are breathing 100%. After those blocks, the pressure will be released from the chamber. During the pressure release, you do not have to equalise your eardrums, this will happen naturally. You do, however, need to do the thumbs-up sign and wear ear protection. During the pressure buildup at the beginning, the temperature will go up. During the pressure release, the temperature will go down. So it is nice if you can put on a cardigan or sweater. After the release of pressure, the door will be opened, and you are allowed to leave.

Interviewer

Could you explain the risks that accompany hyperbaric oxygen therapy?

Interviewee

There are several risks in hyperbaric oxygen therapy. The most common one is barotrauma of the middle ear. Barotrauma of the middle ear might occur when you are not able to equalise your eardrums during the pressure buildup. The worst-case scenario in this situation is an eardrum rupture, which is very painful.

To see if you have a barotrauma of the middle ear, the nurse will take a look at the inside of your ear after the treatment. To see if there is any indication. If it is a MacFie 1, 2, or 3, it will already be less in a few days. A MacFie 4 or 5 will take weeks to months.

The second most common barotrauma is sinus/paranasal barotrauma. In most cases, the patient also suffers from respiratory infections or allergic rhinitis. It will lead to facial pain, congestion, and edema. It will go away after the pressure is back to normal.

The third barotrauma that can occur is dental barotrauma (tooth squeeze). This results in pain in the jaw of the maxillary sinuses. It will be gone when the pressure is back to normal.

Then another well-known risk is oxygen poisoning. Because of the high pressure in the chamber, oxygen gets toxic. That is why there are blocks of 5 minutes without 100% oxygen where you have to take deep breaths. It is very unusual, but it is a risk. When this happens, the nurse will remove the oxygen mask. You will start having seizures, which is an acute situation. Because acute acoustic trauma treatment has multiple decompression moments, you will also be at risk for pulmonary oxygen toxicity. Due to long periods of inhaling 100% oxygen, there is a chance that you will experience pain in your chest, a sore throat, pain in your lungs during deep inhaling or exhaling, a cough, or shortness of breath. This can occur inside the chamber but also outside the chamber.

However, these risks are not very likely to occur. Other side effects that are more likely to happen are stomach pain and feeling tired. This is temporary and completely normal.

Interviewer

Could you explain why hyperbaric oxygen therapy is offered over other possible treatments for acute acoustic trauma? (Or how this would be communicated with a patient)

interviewee

In my field of work, it is treatment option number one. I think, due to our patient population, acute acoustic trauma is something that happens far more often compared to non-military patients. We have a lot of military physicians who specialise in hyperbaric chamber treatment, they do a lot of research in this treatment area. I know the military was the first place where these types of patients were treated this way. For one or two years, it is also used for non-military patients. This type of treatment is not for everyone, there are some requirements. The treatment has to be started within the time limits, otherwise, it will not work. The patient has to start on some corticosteroids to reduce the swelling. Plus, the patient has to be in a stable or healthy condition to avoid the risks that I have mentioned before. This treatment does have the best outcome compared to other treatments. That is why it is our first choice of treatment, based on a lot of research and evidence. However, I just want to mention that I am not the one who decides what kind of treatment the patient will get, the doctor is always the one who decides.

Interviewer

Is there a posttreatment?

Interviewee

Not really, just a follow-up. There will be a new audiogram to show the results. Plus, there will be a follow-up a few days after the treatment, just to see how it goes.

C.5 Questions related to language use

Interviewer

Looking into the communication you have with patients, what specific terminology or language is typically used when explaining acute acoustic trauma?

Interviewee

Some explanations about the ear. Acute acoustic trauma (trauma like a blast or gunshot) has an effect on the inner ear (malleus, incus, and stapes) and sometimes ruptures the tympanic membrane. These kinds of terms. Often, there are questions related to the functioning of the ear.

Interviewer

And for the treatment of hyperbaric oxygen therapy?

Interviewee

Just an explanation of the treatment and the related matters, as described before.

Interviewer

Are there any important phrases or concepts that you frequently use that might be difficult to understand for someone outside of your expertise?

Interviewee

I think everything about hyperbaric oxygen therapy is a bit difficult to understand when you have never seen it before. This is not really about the words or phrases used, but more about the concept of the chamber. Also, the explanation of the ear can be difficult because of some words used: tympanic membrane, malleus, incus, stapes, cochlea, auditory nerve, and the eustachian tube.

Interviewer

What tone do you usually use when providing information related to a disease (or treatment) to a patient?

Interviewee

It depends on the patient and how well I know him or her, of course. Overall, I always take my time with a patient to explain things in a calm and clear manner. I always ask if they have any questions. Plus, in my line of work, it is very easy to just walk by my office to ask questions or talk about things. So my door is always open if any questions come up. I am always trying to be nice and understanding, as well as trying to take my time, as people have to feel comfortable.

Interviewer

Do you make use of medical terminology when providing information to patients? If yes, do you have some examples?

Interviewee

Hyperbaric oxygen therapy, tympanic membrane, tympanic cavity, auris interna/externa, malleus, incus, stapes, cochlea, eustachian tube, the different barotrauma's mentioned before, oxygen poisoning, equalising the eardrums (tympanic membrane), oronasal mask / oxygen mask, Valsalva method, corticosteroids, seizures, acute acoustic trauma, audiometry and otoscopy, the use of tuning forks and the Weber and Rinne test, acoustic trauma, hearing loss, and more that I have already mentioned probably.

Interviewer

What do you do when someone does not understand a specific term or concept?

Interviewee

I will try to explain it again in different terms or terminology. Or show them what I am talking about using pictures, or in this case, walk to the hyperbaric chamber to show what it looks like.

Interviewer

Are there specific aspects that people often misunderstand or struggle to understand? If yes, which ones?

Interviewee

Almost everything about the hyperbaric chamber. It is a treatment that a lot of people do not know about beforehand and/or have never worked with before. When I treat divers in the chamber, they know what to expect. But patients with acute acoustic trauma do not. As you can imagine, these patients are deaf, at least in one ear, which makes communicating a bit harder. Also, when you use terminology related to the ear, this is often difficult to understand.

D QUANTITATIVE MEASUREMENTS

D.1 Understanding test (Final version)

Below are the questions that were created to be used in the test to assess the participants' understanding of information provided by the chatbot about the diagnosed disease and proposed treatment. The questions follow a questionnaire format, where each question requires an answer, moreover, one is not able to go back to previous questions. Note that the content (what questions) of the questionnaire provided to the participant depends on their conversation with MedWiseBot, as discussed in Section 4.7.1.

The correct answer is shown in bold, which is not the case in the test provided to the participant. Note that in the questionnaire provided to the participant, questions with multiple gaps were asked separately. Here, these questions are combined for brevity, and the answers to choose from per gap are separated by a blank line.

1. The Rinne and Weber tests are conducted using an otoscope.
 - (a) True
 - (b) **False**
 - (c) I do not know

2. The Rinne and Weber tests are conducted to assess the __1__ and are conducted __2__.
 - (a) **Degree of hearing loss**
 - (b) Eardrum
 - (c) Tympanic membrane
 - (d) I do not know
 - (a) After a week of the hyperbaric oxygen therapy sessions
 - (b) **After each of the hyperbaric oxygen therapy sessions**
 - (c) After the full duration of the hyperbaric oxygen therapy (all sessions)
 - (d) I do not know

3. Audiometry is conducted to:
 - (a) **Assess the degree of hearing loss**
 - (b) Assess the eardrum
 - (c) I do not know

4. Otoscopy is conducted to:
 - (a) Assess the degree of hearing loss

- (b) **Assess the eardrum**
(c) I do not know
5. After the treatment of hyperbaric oxygen therapy, the patient will regain their level of hearing similar to what it was before the diagnosis of acute acoustic trauma.
- (a) True
(b) **False**
(c) I do not know
6. Someone suffering from acute acoustic trauma will always have an eardrum perforation.
- (a) True
(b) **False**
(c) I do not know
7. Someone suffering from acute acoustic trauma will have similar symptoms in both ears.
- (a) True
(b) **False**
(c) I do not know
8. Instead of saying someone is suffering from acute acoustic trauma, it can be said this person is suffering from tinnitus. These are the same conditions.
- (a) True
(b) **False**
(c) I do not know
9. Hyperbaric, in case of the proposed hyperbaric oxygen therapy, means the:
- (a) Decreased pressure of oxygen
(b) **Increased pressure of oxygen**
(c) Process of decreasing the pressure in the hyperbaric chamber
(d) Process of increasing the pressure in the hyperbaric chamber
10. The patient is __1__ to eat and drink in the hyperbaric chamber. Water bottles need to be opened up as otherwise there is a danger of explosion during the __2__.
- (a) Allowed
(b) Able
(c) Not able
(d) **Not allowed**
(e) I do not know
- (a) Audiometry
(b) Depressurisation (increasing the pressure)
(c) **Pressurisation (increasing the pressure)**
(d) I do not know
11. Which part(s) of the face are covered by the mask used in this therapy?

- (a) Mouth
 - (b) **Mouth and nose**
 - (c) Nose
 - (d) I do not know
12. Very likely to occur risks include oxygen poisoning.
- (a) True
 - (b) **False**
 - (c) I do not know
13. The more likely to happen risks are __1__ and can be considered __2__.
- (a) Continuing
 - (b) Long-term
 - (c) Permanent
 - (d) **Temporary**
 - (e) I do not know
- (a) Dangerous
 - (b) Insurmountable
 - (c) **Normal**
 - (d) Severe
 - (e) I do not know
14. Barotraumas are likely to occur due to the difference in pressure and will often be gone after depressurisation (decreasing the pressure) of the hyperbaric chamber.
- (a) **True**
 - (b) False
 - (c) I do not know
15. The proposed treatment is not covered by the health insurance, only if it is proposed for the diagnosed acute acoustic trauma.
- (a) **True**
 - (b) False
 - (c) I do not know
16. The increased pressure of the hyperbaric tank leads to __1__ oxygen absorption of the lungs. By __2a__ this pure oxygen, __2b__ oxygen will be in the blood plasma which helps the body heal and fight certain infections.
- (a) Compressed
 - (b) Less
 - (c) **More**
 - (d) Minimal
 - (e) I do not know

- (a) (a) Breathing in, (b) less
 - (b) **(a) Breathing in, (b) more**
 - (c) (a) Breathing out, (b) less
 - (d) (a) Breathing out, (b) more
 - (e) I do not know
17. Currently, hyperbaric oxygen therapy and the use of corticosteroids is the only possible treatment for the diagnosed acute acoustic trauma.
- (a) **True**
 - (b) False
 - (c) I do not know
18. For what reasons are corticosteroids used when hyperbaric oxygen therapy is proposed?
- (a) Corticosteroids are not used in combination with hyperbaric oxygen therapy.
 - (b) Corticosteroids are used to increase the level of oxygen in the blood to promote healing.
 - (c) **Corticosteroids are used to lessen the swelling in the inner ear.**
 - (d) Corticosteroids are used to lessen the swelling in the middle ear.
 - (e) I do not know
19. Regarding the availability of the proposed hyperbaric oxygen therapy, what statement is correct?
- (a) It is available at almost all hospitals in the Netherlands.
 - (b) **It is available at locations that are equipped with a hyperbaric chamber.**
 - (c) It is available at the hospitals in the larger cities of the Netherlands only (e.g., Amsterdam).
 - (d) It is available at the larger hospitals in the Netherlands as they are equipped with a hyperbaric chamber.
 - (e) I do not know
20. During the proposed hyperbaric oxygen therapy, the patient can wear their own clothing.
- (a) True
 - (b) **False**
 - (c) I do not know
21. At time of emergency, the patient has the opportunity to talk to the medical personnel for instructions.
- (a) **True**
 - (b) False
 - (c) I do not know
22. The treatment consists of __1__ sessions.
- (a) Biweekly
 - (b) **Daily**

- (c) Monthly
 - (d) Weekly
 - (e) I do not know
23. The treatment is stopped during the weekends.
- (a) True
 - (b) **False**
 - (c) I do not know
24. The patient is allowed to leave the hyperbaric chamber after __1__ .
- (a) Any time during the hyperbaric oxygen therapy, for example in case they need to go to the toilet.
 - (b) Putting off the oxygen mask.
 - (c) **The process of depressurisation (decreasing the pressure).**
 - (d) The process of pressurisation (increasing the pressure).
 - (e) I do not know
25. The hyperbaric oxygen therapy is conducted in a hyperbaric chamber, which accommodates one patient at a time for the therapy session.
- (a) True
 - (b) **False**
 - (c) I do not know
26. The hyperbaric chamber is designed in such a manner that it contains doors and windows in order for medical personnel to __1__ monitor the patient(s).
- (a) **Continuously**
 - (b) Intermittently
 - (c) Day-and-night
 - (d) I do not know
27. The diagnosed disease of acute acoustic trauma is a result of:
- (a) **An acoustic trauma**
 - (b) An increased pressure in the ear
 - (c) I do not know
28. What does the term "trauma" or "impact" refer to in the context of the diagnosed acute acoustic trauma?
- (a) **An injury in the inner ear due to the exposure to loud noise.**
 - (b) An injury to the outer ear due to the exposure to loud noise.
 - (c) The loud noise that causes the injury in the inner ear.
 - (d) The loud noise that causes the injury in the outer ear.
 - (e) I do not know
29. Hyperbaric oxygen therapy is proposed without any additional medication or treatment.

- (a) True
 - (b) **False**
 - (c) I do not know
30. After each of the sessions of the proposed treatment, several assessments are done related to the ear and hearing loss.
- (a) **True**
 - (b) False
 - (c) I do not know
31. The proposed hyperbaric oxygen therapy entails the administration of 100% oxygen via the:
- (a) Breathing tube
 - (b) Nasal spray
 - (c) Pressurisation of the environment (increased pressure)
 - (d) **Special mask**
 - (e) I do not know
32. During the proposed hyperbaric oxygen therapy, the pressure is __1__ similar to what divers experience at a depth of 14 metres. This might result in the need for the patient to __2__.
- (a) Gradually decreased
 - (b) **Gradually increased**
 - (c) Kept constant
 - (d) Rapidly decreased
 - (e) Rapidly increased
 - (f) I do not know
- (a) **Equalise the eardrum**
 - (b) Equalise the nasal membrane
 - (c) Leave the hyperbaric tank
 - (d) Stop the hyperbaric oxygen therapy immediately
 - (e) I do not know
33. Pressure on the ear, plugging in the ear, cracking sounds when swallowing or yawning are all results of:
- (a) Acoustic trauma
 - (b) Acute acoustic trauma
 - (c) **Difference in air pressure on both sides of the eardrum**
 - (d) Difference in airpressure on both sides of the tympanic cavity.
 - (e) I do not know
34. There are multiple manners to equalise the eardrum
- (a) **True**

- (b) False
 - (c) I do not know
35. An acoustic trauma can be caused by the __1__ exposure to a loud noise. Moreover, __2__ exposure to loud music or machinery can also lead to an acute acoustic trauma.
- (a) Low frequency
 - (b) Prolonged
 - (c) **Short**
 - (d) Varying
 - (e) I do not know
- (a) Low frequency
 - (b) **Prolonged**
 - (c) Short
 - (d) Varying
 - (e) I do not know

D.2 Adjusted HCTM scale

Here, the adjusted HCTM scale to assess the perceived trust participants have in the chatbot can be found. Note that the adjustments include the substitution of the placeholders in the original scale with the appropriate information from the current research, such as the name of the chatbot MedWiseBot [34].

1. I believe that there could be negative consequences when Using MedWiseBot
2. I feel I must be cautious when using MedWiseBot.
3. It is risky to interact with MedWiseBot.
4. I believe that MedWiseBot will act in my best interest.
5. I believe that MedWiseBot will do its best to help me if I need help.
6. I believe that MedWiseBot is interested in understanding my needs and preferences.
7. I think that MedWiseBot is competent and effective in offering healthcare information.
8. I think that MedWiseBot performs its role as a healthcare informative chatbot very well.
9. I believe that MedWiseBot has all the functionalities I would expect from a healthcare informative chatbot.
10. If I use MedWiseBot, I think I would be able to depend on it completely.
11. I can always rely on MedWiseBot for healthcare information.
12. I can trust the information presented to me by MedWiseBot.

E QUALITATIVE MEASUREMENT

Here, one can find the questions asked during the semi-structured interview that takes place after the experiment. Note that the first four questions are based on the User Experience Questionnaire [84], and the following five questions are based on the Chatbot Usability Questionnaire. The last questions are added to get a better understanding of the participant's experience with the chatbot [85] as discussed in Section 4.7.2.

1. Do you think the healthcare chatbot was understandable when providing you an answer during the interaction?
2. Do you think the healthcare chatbot was meeting your expectations when providing you an answer during the interaction?
3. Do you think the healthcare chatbot was clear when providing you an answer during the interaction?
4. Do you think the healthcare chatbot was pleasant to use?
5. Was it easy to get confused when using the chatbot?
6. Do you feel like the chatbot understood you well?
7. Do you feel that the chatbot's responses were useful, appropriate and informative?
8. Did you feel like the chatbot failed to recognise a lot of your inputs?
9. Did you feel like the chatbot responses were irrelevant?
10. Would you use a chatbot like this to help you out in the case of having questions related to a disease or treatment in the future?
 - (a) And would you prefer such a healthcare chatbot over other well-known information-seeking platforms such as Google?
11. Did you answer a lot of questions with "I do not know" in the last questionnaire (understanding test)?
 - (a) If yes, why?
12. Do you feel like the chatbot used difficult terminology?
 - (a) If yes, did you ask for explanation of the difficult terminology used by the chatbot?

F FLOWCHARTS

F.1 Non-aligning chatbot

The flowchart in Figure F.1 describes the process for the chatbot that does not apply alignment. In this case, the template answers are not adjusted, and the steps that are necessary for an appropriate answer are limited. For some intents, a certain action is created in order for the chatbot to provide a correct answer. An example of this is the intent where the chatbot provides information on whether a mentioned item by the participant can be taken into the hyperbaric chamber yes or no. In this case, the item must first be checked with the allowed items in the hyperbaric chamber in a separate function. This is incorporated into the flowchart by means of the if-statement, which checks for the presence of an action.

F.2 Aligning chatbot

The flowchart presented in Figure F.2 visualises the strategy used to save the initial entity slots based on the saved terms in the terminology repository, which are the answers in the questionnaire on (medical) terminology as explained and discussed in Section 5.5.2.

The flowchart presented in Figure F.3 visualises the first strategy used to implement alignment in the aligning chatbot, as explained in Section 5.5 and specifically Section 5.5.3. Again, the if-statement to check for a specific action, as explained in F.1, is incorporated. Note that the functions for sentence refinement discussed in Section 5.5.4 are not incorporated in the visualisation as they are not part of the alignment strategy, however, they are called before the adjusted answer is outputted to the user.

The flowchart presented in Figure F.4 visualises the second strategy used to implement alignment in the aligning chatbot, as explained in Section 5.5 and specifically Section 5.5.3. This strategy is only activated in the case that the first strategy does not result in an adjusted template answer. It is important to note that the preprocessing of the input includes the punctuation removal and tokenization, as well as the adjustment for the term to be in plural or singular form, as explained in Sections 5.5.3 and 5.5.4. Note that the function for sentence refinement related to the appropriate article discussed in Section 5.5.4 is not incorporated in the visualisation as it is not part of the alignment strategy, however, it is called before the adjusted answer is outputted to the user. To clarify the brackets surrounding "adjusted" in the *input/output* block, they denote that the template answer is only adjusted in case the threshold for the cosine similarity is exceeded. In the event that the threshold is exceeded, the slot is adjusted, and the placeholder is replaced in the template answer.

Figure F.1: Flowchart for the non-aligning chatbot

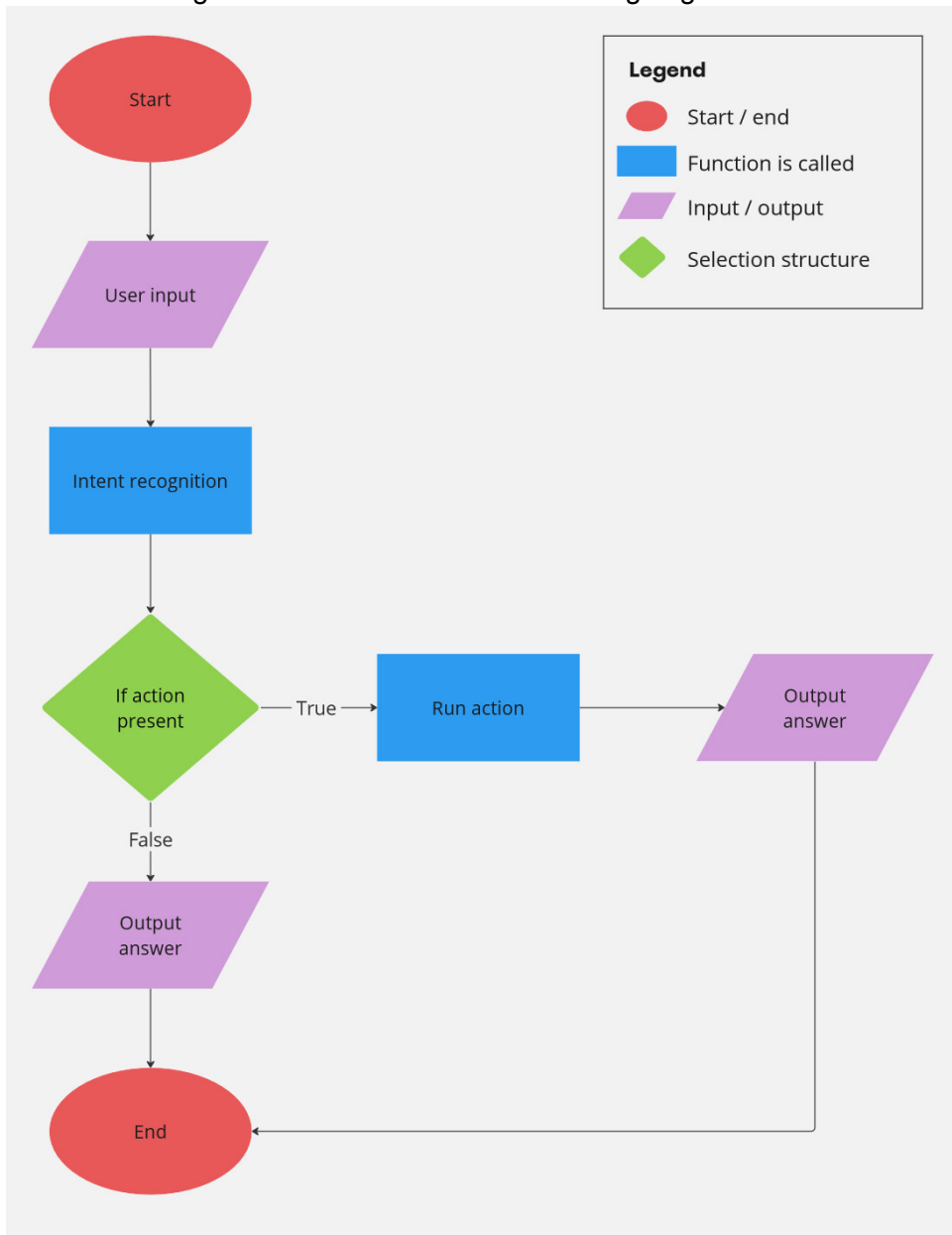


Figure F.2: Flowchart for the use of terms from terminology repository

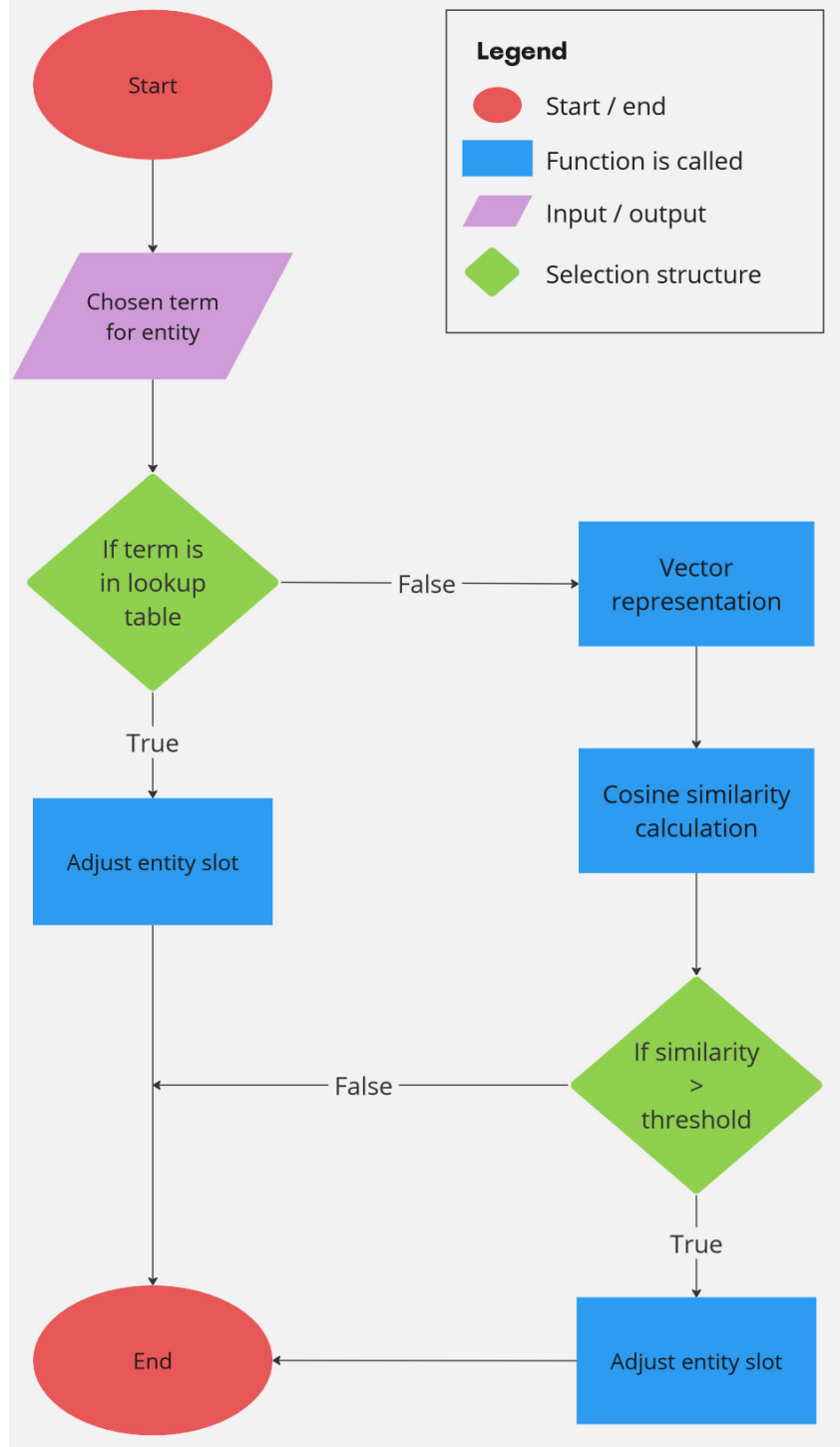


Figure F.3: Flowchart for the aligning chatbot strategy 1

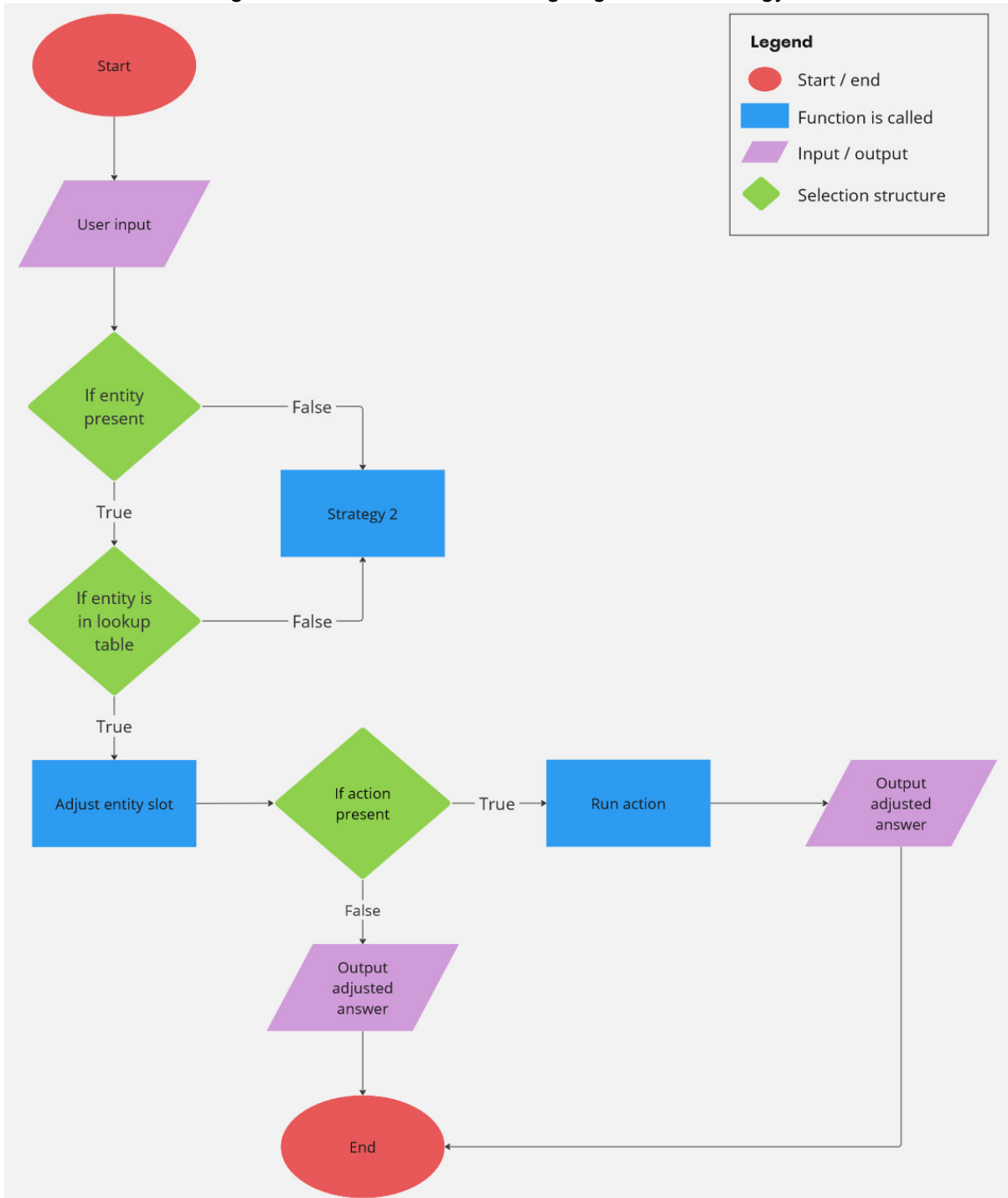
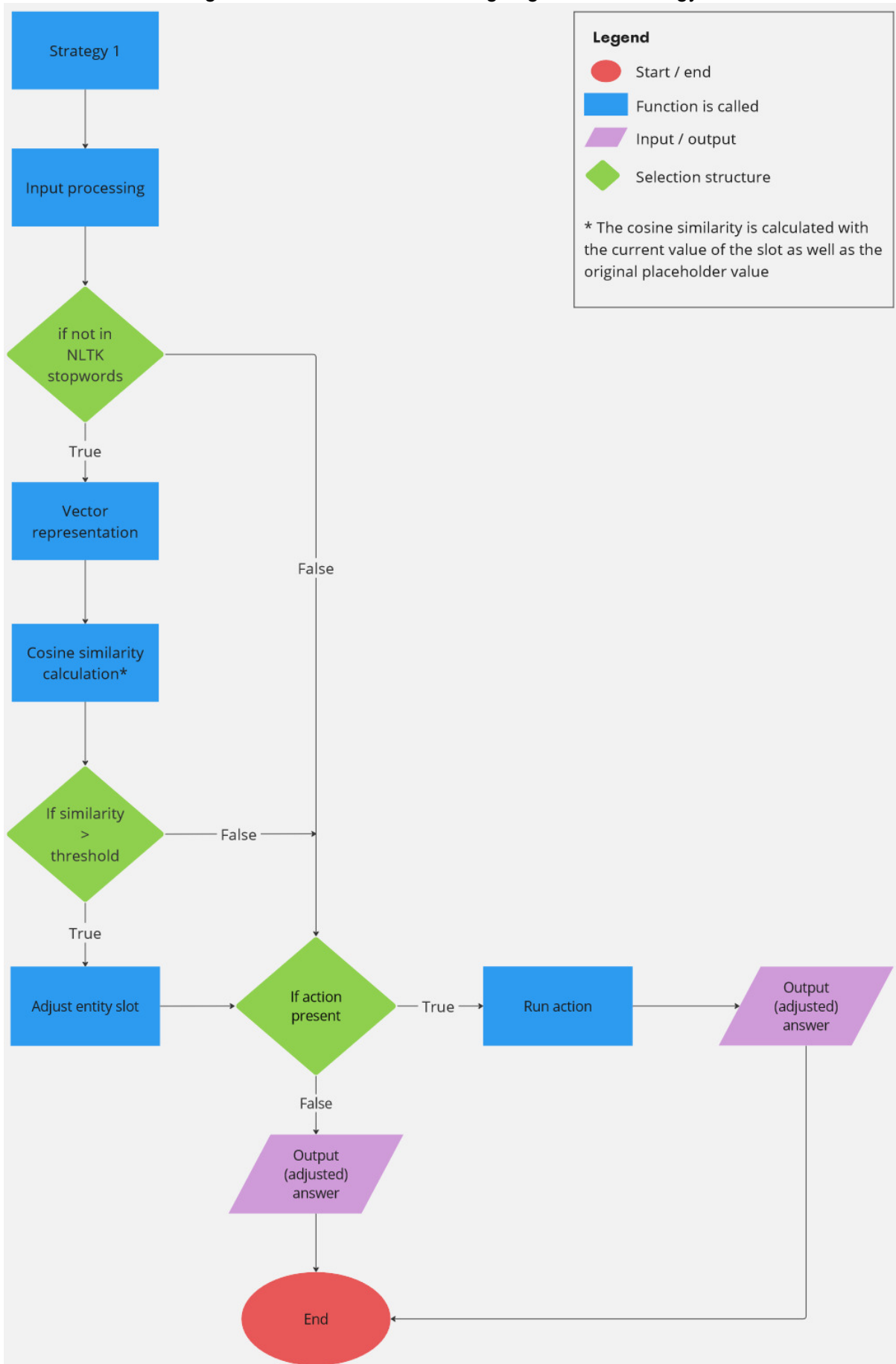


Figure F.4: Flowchart for the aligning chatbot strategy 2



G FREQUENCY OF CHOSEN TERMS FOR PLACEHOLDERS

Table F.1 below shows the frequency of terms chosen for each of the placeholders from the results of the questionnaire used to infer the participant's word use, see Section 4.1.1 and Appendix A. The percentage of alternative terms chosen than the original placeholder term (per placeholder) is shown in the last column. Note that for this percentage, the answer option *I do not know* is not considered a different term than the original placeholder term. This is based on the lack of replacement of the placeholder in the event that this option is chosen. The same holds for the answer option *Other ...* if the entered term is not considered similar enough according to their cosine similarity. Additionally, terms with a frequency of 0 are not included in the table. At the bottom of the table, the overall percentage of participants choosing a different term than the placeholder term is shown.

Table G.1: Frequency of terms chosen by participants for the placeholders

Placeholder	Answer Option	Frequency	Alternative Term %
Indications	Symptoms	21	95.83%
	Signs	1	
	Indications	1	
	Indicators	1	
Diagnosed	Diagnosed	20	16.67%
	Determined	2	
	Confirmed	1	
	Identified	1	
Oronasal mask*	Oxygen mask	16	95.83%
	Breathing mask	3	
	Inhalation mask	1	
	Oxygen therapy mask	1	
	Respiratory mask	2	
	I do not know	1	
Complications	Complications	9	62.50%
	Side effects	15	
Restroom	Restroom	5	79.17%
	Lavatory	6	
	Toilet	10	
	Bathroom	3	
Exit	Get out	2	16.67%
	Leave	2	
	Exit	20	

Continued on next page

Table G.1: (continued)

Placeholder	Answer Option	Frequency	Alternative Term %
Pressurisation*	Increasing the pressure	13	87.50%
	Compression	5	
	Elevating the pressure	1	
	Heightening the pressure	2	
	I do not know	2	
	Other ... <i>control the pressure</i> **	1	
Depressurisation*	Decompression	11	91.67%
	Decreasing the pressure	7	
	Reducing the pressure	4	
	I do not know	2	
Noise	Noise	19	8.33%
	Clamour	1	
	Sound	1	
	Other ... <i>Intens Auditory Stimulants</i> **	1	
	Other ... <i>noise pollution</i> **	1	
	Other ... <i>loud</i> **	1	
Sessions	Appointments	7	37.50%
	Sessions	15	
	Meetings	2	
Etiology*	Cause	14	95.83%
	Origin	5	
	Reason	3	
	Determinant	1	
	I do not know	1	
Duration	Duration	16	29.17%
	Length	3	
	Timespan	3	
	Time	1	
	Other ... <i>Session</i> **	1	
Hearing loss	Hearing impairment	15	66.67%
	Hearing loss	7	
	Hearing deficit	1	
	I do not know	1	
Cause	Cause	14	37.50%
	Origin	5	
	Reason	3	
	Determinant	1	
	I do not know	1	
Auditory acuity*	Hearing ability	18	91.67%
	Clarity of hearing	2	
	Hearing sensitivity	2	
	I do not know	2	
Oxygen	Oxygen	20	16.67%
	Air	2	
	O2	2	
Inhaling	Breathing in	7	
	Inhaling	16	

Continued on next page

Table G.1: (continued)

Placeholder	Answer Option	Frequency	Alternative Term %
	Respiring	1	33.33%
Exhaling	Exhaling	15	37.50%
	Breathing out	9	
Attire	Clothing	12	95.83%
	Attire	1	
	Clothes	11	
Medical personnel	Medical professionals	5	75.00%
	Medical personnel	6	
	Medical staff	6	
	Medical experts	4	
	Health professionals	2	
	Other ... <i>doctors</i>	1	
Nurse	Nurse	23	4.17%
	Specialist	1	
Patient	Patient	24	0.00%
Conducted	Taking place	11	91.67%
	Performed	7	
	Conducted	2	
	Carried out	3	
	Given	1	
Auris externa*	External part of the ear	7	100%
	Outer part of the ear	8	
	Outer ear	7	
	External ear	2	
Tympanic cavity*	Middle ear cavity	4	83.33%
	Middle part of the ear	6	
	Middle ear	8	
	Tympanum	2	
	I do not know	4	
Auris interna*	Inner part of the ear	12	95.83%
	Internal ear	2	
	Labyrinth of the ear	4	
	Inner ear	5	
	I do not know	1	
Tympanic membrane*	Eardrum	18	75.00%
	I do not know	6	75.00%
			Total: 60.03%

* This original term of the placeholder is excluded in the answer options of the questionnaire used to infer the preferred by the participant. This means that in case the participant still wants to use that specific term, they had to enter this term in the *Other ...* option. For an in-depth explanation of this, see Section 5.6.1.

** The term entered as *Other ...* option is not considered similar enough after calculating the cosine similarity between the vector representation of the placeholder and the term entered. Therefore, this term is not used to replace the placeholder term, see Section 5.5.3 for more information on this.