master's thesis

Developing of Q&A bots for medicinal disclosure for CKD-patients

Zurich University of Applied Science
School of Management and Law
Master of Science in Business Information Management
Spring semester 2023

Jan Berchtold

Matriculation number 16-570-301
1st Supervisor: Prof. Dr. Thomas Keller
2nd Supervisor: Prof. Dr. Alexandre de Spindler

Winterthur, 28th May 2023

Abstract

This thesis was a subproject of the RealCo project by Prof. Dr. Thomas Keller at ZHAW which provides information about a medication called SGLT2 inhibitors to patients with chronic kidney disease (CKD). While the goal of the project is to improve patient health literacy and compliance, this thesis rather focused on developing a chatbot to answer questions related to the use of SGLT2 inhibitors to CKD patients. Chatbots, as software components that communicate with users via natural language, are considered as an appropriate instrument for improving health literacy. The developed chatbots were implemented using natural language understanding (NLU) platforms, which, due to their structure, enable rapid prototyping, deployment and simple integrations. This thesis addressed the question of which NLU platform is most suitable for the use case.

In this thesis, two artefacts were built with over 800 training questions about SGLT2 inhibitors to answer the question above. The developed chatbots were tested with physicians and pharmacists for correctness. The results showed that DialogFlow and Watson Assistant are the most popular and widely used NLU platforms were therefore selected for the chatbot development. The tests conducted and the feedback gathered from physicians and pharmacists showed that the answers were medically correct and the chatbot was perceived as friendly and appealing. Also, in the majority of cases, users received an answer that was relevant to their question. The implementation of the chatbots in these two platforms demonstrated that Watson Assistant was superior to DialogFlow in terms of latency as well as the delivery of the correct answer to the question asked. Future studies using the existing chatbots within the RealCo project should involve patients for testing and further development.

Table of content

Abstract..I

List of figures IV

List of tables IV

1 Introduction 1

2 Relevance 3

3 Related work 5

3.1 Chatbots in healthcare for patient education.. 5

3.1.1 Chatbots based on large language models in healthcare 7

3.2 Evaluation of third-party platform 8

3.3 Algorithm and techniques implemented in chatbots.................................. 10

3.3.1 Data preprocessing 10

3.3.2 Feature extraction 13

3.3.3 Intent Classification 14

3.3.4 Named Entity Recognition (NER)... 15

3.3.5 Response generation 15

3.4 Evaluation methods of chatbots 16

4 Current thesis.. 19

4.1 Objectives 19

4.2 Research questions.. 19

5 Methods 21

5.1 Evaluation of NLU platforms 21

5.1.1 Literature Review 21

5.1.2 Case Study 21

5.2 Artefact development and design of the bot 22

5.2.1 Preparations for development... 22

5.2.2 Chatbot design 23

5.2.3 Development of the DialogFlow-Bot 24

5.2.4 Development of the Watson Assistant-Bot 26

5.3 Testing the chatbots 26

6 Results 28

6.1 Evaluation of software 28

6.2 Developing and testing the artefacts 31

6.2.1 Developing the prototype 31

6.2.2 Testing the prototype with testing-software 31

6.2.3 Testing the prototype with physicians 32

6.2.4 Second testing of the prototype with physicians 33

6.2.5 Development of the second prototype in Watson Assistant 34

6.3 Experimental comparison of NLU platforms 34

7 Discussion 39

7.1 Experimental comparison of NLU platforms 39

7.2 Quality of the artefacts 41

7.3 Limitations 42

7.4 Next steps and future work 43

8 Conclusion 45

9 References 46

10 Appendix 60

10.1 Exports from chatbots 60

10.2 Questionnaire and answers from third testing 60

List of figures

Figure 1 Illustration by Brucker-Kley et al. (2021) of the different research fields 4

Figure 2 NLU-Pipeline 10

Figure 3 High-level chatbot architecture 24

Figure 4 Intent with training questions in DialogFlow 25

Figure 5 Dialog-Nodes in Watson Assistant 26

Figure 6 AI-powered data generator by Botium 31

Figure 7 Chatbot in DialogFlow 31

Figure 8 Chatbot in Watson Assistant 31

List of tables

Table 1 Possible patient questions and classified intents to the questions 22

Table 2 Questionnaire for second testing with physicians and pharmacists 27

Table 3 Result from literature search to NLU platforms 29

Table 4 Updated taxonomy according to Canonico & Russis (2018) 30

Table 5 Comparison of SGLT2-Bots in DialogFlow and Watson Assistant 36

Table 6 Summary of the test results 37

1 Introduction

One of the most challenging aspects of medicine is low compliance (Horne et al., 2005). Compliance is defined as the willingness of a patient to actively participate in therapeutic measures. If the patient has low compliance, this could lead to negative consequences for his or her health. One reason for hesitation in taking medication is believed to be patient's low health literacy (Buehrig et al., 2020). To provide patients with health information about their disease and therapy options, chatbots offer a potential communication channel. According to the research of Reis et al. (2020), the use of chatbots for medication counseling is one of its many application in the healthcare area and is considered as an appropriate method according to physicians. This paper aims to address the problem of low health literacy by developing a chatbot that provides patients with chronic kidney disease with additional information about sodium dependent glucose co-transporter 2 (SGLT2) inhibitors. SGLT2 inhibitors have been prescribed primarily for patients with type 2 diabetes, but are more frequently recommended for patients with chronic kidney disease. It is expected that the additional information provided by the developed chatbots in this thesis will improve health literacy and therefore compliance. This will have a positive impact on disease progression. In addition to increasing the health of the patient, chatbots can relieve physicians.

Chatbots, also called virtual assistants or conversational agents, are software components which are able to communicate in natural language with users. They are used in different industries today. A major sector, in which the first chatbot named ELIZA was developed over 50 years ago, is the medical sector (Weizenbaum, 1966). A chatbot can be developed based on different algorithms. In most cases, natural language processing (NLP) is crucial. The development in recent years also shows that more and more chatbots use machine learning techniques and rule-based approaches are used as a complement (Adamopoulou & Moussiades, 2020b). Which algorithms are most suitable always depends on the use case of the chatbot. It is therefore important to have a clear idea of what the chatbot must be able to do, and what the limitations are.

Due to the recent growth in the use of chatbots, vendors for development, deployment and maintenance of such bots have also emerged. These offer libraries, frameworks and low to no-code environments which simplify the development and provide easier maintenance and scalability of a bot (Canonico & Russis, 2018). Furthermore, the vendors who offer natural language understanding (NLU) platforms enable faster prototyping of chatbots. Due to this and the time constraints of this thesis, a bot was built based on such a third-party system. The vendors will be presented in more detail later, and a taxonomy was utilized to compare the NLU platforms and to identify the most suitable solution. Although comparisons of NLU platforms exist, due to the existing literature and the non-transparency of the vendors, no platform was able to massively differentiate from the competitors, therefore instead of one chatbot, two chatbots were developed on the most widely used platforms according to the literature. Then, based on the use case for SGLT2 inhibitors, these two platforms were compared. To ensure the quality, the developed chatbots were also be tested with physicians and pharmacists for correctness and trained to answer questions they receive from patients on a daily basis about SGLT2 inhibitors.

2 Relevance

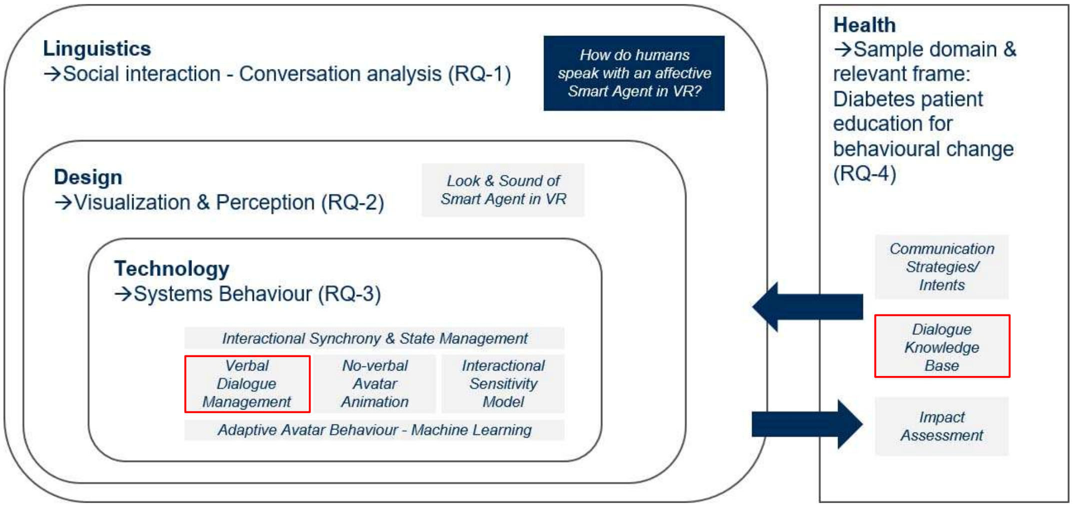
According to Forni Ogna et al. (2016), CKD affects approximately one in ten adults in Switzerland. The risk is supposed to be particularly high for individuals over the age of 60. Chronic kidney disease cannot be cured and worldwide, it was the tenth leading cause of death in 2019 (WHO, 2020). However, the most effective way to improve the situation of CKD patients is to slow the progression of renal function loss and delay renal failure as long as possible (Martin, 2017). Recent clinical trials have shown that SGLT2 inhibitors can effectively slow down the progression of CKD (Heerspink et al., 2020; Perkovic et al., 2019). In addition to the great benefit to patients by preventing progression of the disease, such interventions are also an effective way to reduce the financial burden of renal failure. According to Martin (2017), dialysis costs up to CHF 250,000 per patient could be saved by preventing progression to dialysis. However, Maddox et al. (2022) found that communicating information about SGLT2 inhibitors to patients is a major challenge, limiting the potential cost savings. Nephrology care providers may struggle to effectively communicate their expertise to patients, resulting in dissatisfaction among patients (Maddox et al., 2022) and leads to patients who are insufficiently educated. Further, patients with CKD frequently face challenges in comprehending their condition and the available treatment options. The educational materials they receive are often insufficient to fully grasp their illness and treatment possibilities. According to Vernon et. al (2007), 7-17% of the total U.S. healthcare expenditures can be related to low health literacy. Research has shown that an improvement in health literacy can effectively enhance one's quality of life and reduce stress levels (Mokmin & Ibrahim, 2021). Furthermore, they stated that chatbots have potential for education patient and increase their health literacy. The same result is also reached by Nadarzynski et al. (2019). They conducted a survey to investigate the perceived utility of health chatbots. The result presented that more than 70% of the participants perceived a utility for receiving information about medications. Another research by Ayers et al. (2023) scrutinized the quality and empathy of responses to user questions with ChatGPT towards physicians. They concluded that chatbots provide better answers to user questions than physicians in terms of quality and empathy. Hence, these results support the use case of chatbots for patients with chronic kidney disease and their medications and shows that they are of a great relevance. This is why this thesis has the purpose to develop such a chatbot and promote its implementation in the care of CKD patients. The chatbot should be able to answer questions about SGLT2 inhibitors. These are the medication that can be used for type 2 diabetes and for patients with CKD (Perkovic et al., 2019). However, the correct use of the medication is essential to ensure that the active substance acts as desired. In this context, patients often have questions about the application, use and side effects of the medication after the first face to face consultation. Instead of additional face to face medical consultations, information about the medication is communicated via chatbot, and the physician is consulted in emergencies and for follow-up appointments. This is intended to relieve the scarce resources of physicians.

The bot created in this thesis is embedded as part of the ZHAW RealCo project. It forms the knowledge base for answering questions from patients with CKD. The goal of the project is to offer counselling to patients with CKD in the metaverse. Instead of brochures, educational leaflets or guides, the use of an avatar with a verbal language bot is intended to provide information in a more interactive way. This is expected to increase patient literacy and thereby increase compliance, as this is one of the major problems in public health today.

The RealCo project is an interdisciplinary project consisting of people from linguistics, healthcare and computer science. As with the paper by Brucker-Kley et al. (2021) there are different research questions to be answered. The illustration of Brucker-Kley et al. (2021) in **Figure 1** shows the different areas. Instead of diabetes patients, the current project is working with patients with chronic kidney disease. This thesis contributes to address the subjects within verbal dialogue management (marked in red in **Figure 1**).

Figure 1

Illustration by Brucker-Kley et al. (2021) of the different research fields


Linguistics
→Social interaction - Conversation analysis (RQ-1)
How do humans speak with an affective Smart Agent in VR?
Design
→Visualization & Perception (RQ-2)
Look & Sound of Smart Agent in VR
Technology
→Systems Behaviour (RQ-3)
Interactional Synchrony & State Management
Verbal Dialogue Management
No-verbal Avatar Animation
Interactional Sensitivity Model
Adaptive Avatar Behaviour - Machine Learning
Health
→Sample domain & relevant frame: Diabetes patient education for behavioural change (RQ-4)
Communication Strategies/ Intents
Dialogue Knowledge Base
Impact Assessment

Note. Red boxes are part of this thesis.

3 Related work

This chapter reviews the existing literature. It discusses different approaches to algorithms, examines comparisons of NLU platforms, and presents the importance of proper evaluation methods for the testing of chatbots.

3.1 Chatbots in healthcare for patient education

The ability of chatbots to deliver healthcare information in the language of the patient with high availability and low costs is a promising tool. In this part, related papers on chatbots in healthcare for patient education are reviewed.

According to Bickmore et al. (2016), search engines with chatbots as interface are a valuable substitute compared to traditional search engines, particularly for patients with limited health literacy or computer expertise. Several papers confirm this observation that simple keyword-based interfaces deliver for many users unsatisfying results. Interfaces like chatbots, which deliver results in the language of the user could help to provide health information (Aula & Käki, 2005; Gossen et al., 2012). When healthcare information is tailored to the needs of individual patients in a way that they can understand it easily, it gives them a feeling of control and confidence in their healthcare decisions. With a better understanding of their healthcare needs, patients can be empowered to make informed decisions that lead to faster recovery, reduced burden on the healthcare system, and ultimately lower costs for both the patient and the healthcare system (Biro et al., 2023). Another research by Boren (2009) reached the same conclusion as the study by Biro et. al. (2023). Having adequate health literacy is essential for individuals to take responsibility for their own health (Boren, 2009). The study by Boren (2009) showed a correlation between health literacy and diabetes outcomes, but there is a requirement to develop and evaluate methods that can enhance diabetes-related health outcomes while also considering health literacy. The author sees opportunities provided by information and communication technology to mitigate the impact of limited health literacy on diabetes-related health outcomes. To close this gap, chatbot could be a suitable technology. Further, chatbots can serve as a new source of information and enable health knowledge to be communicated to people with low health literacy, as this patient group according to Chen et al. (2018) currently seeks health information on channels where low-quality health knowledge is available. Chen et al. (2018) have investigated health literacy and the trust in health information and the used sources. Individuals with lower health literacy were found to have lower chances of using medical websites for health information, but higher chances of relying on television, social media, and celebrity webpages. They were also less likely to trust information from doctors, but were once again more likely to trust sources such as social media, celebrity webpages, friends, and pharmaceutical companies. However, such sources might contain health information with lower quality compared to information from healthcare professionals (Chen et al., 2018). One solution for this issue can be the use of chatbots in the healthcare area. Contrary to websites with low information quality, the chatbot's knowledge can be based on the professional data and consequently provide supervised medical information. Additionally, a chatbot has the potential to find the patient's language, which could make it easier for patients to understand the given information. This is why it can be seen as an improved future communication channel. There is further reason why chatbots could be suitable in the healthcare sector is the anonymity and non-judgmental space for sensitive topics. Using chatbots in health promotion has great potential to connect with a diverse range of people and give them information and guidance on sensitive topics such as sex, drugs, and alcohol (Crutzen et al., 2011).

A research project in South Africa created a chatbot for patients with diabetes during the COVID19 pandemic (Mash et al., 2022). They were able to determine that chatbots added value in that patients noted an improvement in their self-confidence and self-management regarding their diabetes. The chatbot demonstrated significant promise in augmenting conventional healthcare methods for individuals with diabetes, providing more extensive patient education. It was also noted that further research with chatbots in the area of patient education is needed (Mash et al., 2022). The chatbots in this thesis aim to contribute to this. Similarly to Mash et al. (2022), Anastasiadou et al. (2020) formed an artifact of a chatbot with RASA as the NLU platform. The chatbot is trained on diabetic information and was tested during 6 months with over 3200 questions. The results showed that only 2% of the questions could not be answered. This indicates that a long testing period is key to collecting more and more user questions. With more user questions, the chatbot gets more variety in how to handle the input, thus increasing robustness. Consequently, also in the current thesis, various testing with physicians, but also with patients who eventually use the chatbot, are considered essential for the chatbot. If the chatbot is designed in a variety of ways and can effectively answer the patient's questions, this can also relieve the workload on physicians. Bibault et al. (2019) has investigated this in more detail. They stated in their research that chatbots in healthcare, including oncology, had shown their potential by saving patients with minor problems from unnecessary physician consultations. However, rigorous quality assessment and access to large datasets are critical to their effectiveness. Further, they conclude, chatbots should not be considered as a replacement for physicians, rather as a complement to them. For safe and effective integration into healthcare systems, challenges such as regulatory compliance must be overcome (Bibault et al., 2019). In order to integrate chatbots into the treatment of physicians, chatbots must pass the quality assessment from the medical side as well as be accepted by the patient. While quality is also reviewed from the regulatory perspective, patient acceptance must not be forgotten. Only if the chatbot is able to communicate in the simple language of the patient without too many jargon terms, the patient will use the chatbot. Chatbots have the potential to close the gap in the accessibility of health information by providing easy-to-understand content for people with limited health literacy. This could contribute to solving the existing inequalities concerning the access to health information (Biro et al., 2023). In the studies presented above, it has been stated that the design of the chatbot and the perception as well as interaction between chatbot and user are to be monitored in future study. By using chatbots in a virtual reality environment with an avatar physician, the RealCo project, to which the current thesis belongs, addresses this issue. As already mentioned, the present thesis is in charge of providing the chatbots for the dialogue management in the virtual reality environment.

3.1.1 Chatbots based on large language models in healthcare

In recent months, significant advances have been published in the field of generative AI related to large language models. As these models will have a major impact on the development and use of chatbots in the future, this chapter takes a closer look at the current literature around large language models in healthcare. Nakhleh et al. (2023) evaluated the value of ChatGPT, a large language model (LLM) chatbot from OpenAI, in answering the 24-DKQ, a diabetes knowledge assessment questionnaire. ChatGPT responded accurately to all questions and provided understandable explanations. Large language models have according to Nakhleh et al. (2023) the potential to automate and personalize educational materials for diabetes patients, but its effectiveness is still uncertain. Collaboration between researchers, developers, and healthcare professionals is critical to develop validated and reliable programs tailored to the individual needs of diabetes patients (Nakhleh et al., 2023). This is also considered in the RealCo project and in the development of the chatbots in this thesis. In the RealCo project, scientists from the fields of linguistics, computer science and healthcare as well as physicians from the field are involved.

General medical knowledge was also analyzed in a recent study. Gilson et al. (2023) examined the impact of LLM's on medical training. They investigated how well ChatGPT-3 performed on the United States medical licensing examination. They conclude that

ChatGPT is a strong improvement in natural langue processing models. ChatGPT performs as well as a third-year medical student. They perceived potential for ChatGPT as a tool for medical education. No finetuning was applied in the study and ChatGPT's data is limited to 2021 and earlier.

Similar results were also obtained by Thirunavukarasu et al. (2023) that large language models such as ChatGPT are getting closer to the performance of human experts, but further developments are needed to match the performance of qualified general medical practitioners.

In various LLMs such as FlanPaLM, ChatGPT-3.5/4, MedPaLM2 an increasing accuracy in medical tests or similar can be observed (Gupta & Waldron, 2023; Nori et al., 2023; Thirunavukarasu et al., 2023). Therefore, it can be concluded that the knowledge of LLM about medical subjects is rising, with many studies arguing for further research in this area before it is applied in practice. Patients also need to be protected from getting misinformation from LLMs, as the accuracy of LLM answers does not yet reach the physicians. As patients cannot distinguish whether the LLM is giving correct or incorrect information (Nov et al., 2023), their use in practice is delayed because this is considered an essential requirement. More confidence in the accuracy and trust in the answers of the application must be first established (Buck et al., 2022; Koman et al., 2020). To improve general LLMs like ChatGPT, it requires fine tuning to the medical themes. Currently, the most promising large language model in healthcare is MedPaLM2 by Google, which is however in a test phase and not yet publicly available (Gupta & Waldron, 2023). Since accurate response delivery is one of the objectives of the SGLT2 bots, the use of large language models is omitted. Nonetheless, studies on LLMs should be considered with caution due to their recent publication and thus possible lack of peer review.

3.2 Evaluation of third-party platform

In this thesis, a third-party software will be used to create the chatbot. Before choosing the software, the different types will be evaluated as there are two types of chatbot software. Firstly, there are bots which are coded with java, python, C++ and other programming languages and secondly, there are NLU platforms which implemented state-of-the-art technology (Adamopoulou & Moussiades, 2020a). NLU platforms are a service designed to process natural language and analyze human input, allowing machines to understand user input and respond to it contextually. These platforms facilitate the development of chatbots as they have a simple user interface. Key features include intent classification, entity extraction, context management, and integrations with external services or messaging platforms. Six leading platforms were identified by Adamopoulou and Moussiades (2020a). This includes Google's DialogFlow, Microsoft LUIS, IBM Watson Assistant, wit.ai from Meta, Amazon Lex and SAP Conversation AI. Maher et al. (2020) presented in their literature review the following bots as leading platforms: Google DialogFlow, Wit.ai from Meta, Microsoft LUIS, IBM Watson Assistant, Amazon Lex, ChatScript and Misuku. Furthermore, Abdellatif et al. (2022a) examined Google DialogFlow, IBM Watson Assistant, Microsoft LUIS, and Rasa on two software development use cases in their research. The authors justify the selection of these four platforms based on their popularity and widespread use in academic research as well as in practice (Muñoz et al., 2018; Toxtli et al., 2018). This is based on other platform comparisons in other industries (Braun et al., 2017; Gregori, 2017; Koetter et al., 2019). Canonico & Russis (2018) set up a taxonomy for their comparison of NLU platforms. Based on this taxonomy, an updated survey for two NLU platforms is also established later in this paper. In their paper, DialogFlow, wit.ai, LUIS, Watson Assistant and Amazon Lex as well as Recast.ai were examined. After creating a taxonomy for all six platforms, they evaluated three of the six NLU platforms. All three platforms have fallback intents by default, which is why they were chosen by Canonico & Russis (2018). In the performance comparison, DialogFlow, LUIS, and Watson Assistant were contrasted. The best result was achieved by Watson Assistant. While DialogFlow achieved good results, which required default responses, Watson Assistant was able to provide convincing results with its high confidence in the defined domain, which is also of great importance for the chatbot of SGLT2 inhibitors.

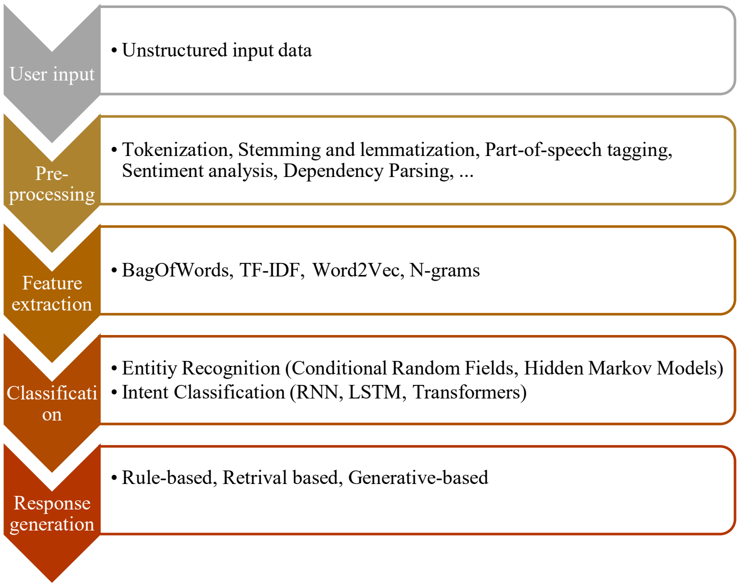
A further research by Thorat & Jadhav (2020) considered DialogFlow and Watson Assistant to be the most popular NLU platforms. The paper by Shah & Shah (2019) compared bots from DialogFlow, Watson Assistant, LUIS, Rasa, Wit.ai, Agent Bot, Pypestream, Semantic Machines, Pandorabots, Gupshup, Kitt.ai, Digital Genius. While Shah & Shah (2019) considered DialogFlow as the most comprehensive platform for chatbots, Canonico & Russis (2018), as mentioned above, highlighted IBM Watson as the best NLU platform in their study. This also coincides with the analysis of Abdellatif et al. (2022a) who considered IBM Watson Assistant as the best NLU platform. They examined intent classification, confidence scores and entity extraction. With regard to the development of a chatbot in healthcare, good results regarding the confidence score are of high significance, since false statements can have fatal consequences. Thus, this should be considered for the choice of the chatbot in this thesis. The different papers comparing NLU platforms help to classify the vendors, while also highlighting that the use case has a considerable influence. For this reason, this thesis develops two bots on different NLU platforms in order to find the most suitable solution.

3.3 Algorithm and techniques implemented in chatbots

While numerous algorithms can be applied in chatbots, this section presents different algorithm and techniques which could be used in different stages of a bot. Although there are many algorithms, the thesis explains these, which are considered as most relevant for a Q&A bot. While the NLU pipeline as presented in **Figure 2** remains non-transparent in NLU platforms, the generally known techniques are presented. However, the NLU platforms used may use other techniques in addition to those introduced.

Figure 2

NLU-Pipeline


User input
• Unstructured input data
Pre-processing
• Tokenization, Stemming and lemmatization, Part-of-speech tagging, Sentiment analysis, Dependency Parsing, ...
Feature extraction
• BagOfWords, TF-IDF, Word2Vec, N-grams
Classificati on
• Entitiy Recognition (Conditional Random Fields, Hidden Markov Models)
• Intent Classification (RNN, LSTM, Transformers)
Response generation
• Rule-based, Retrival based, Generative-based

NLU is one part of natural language processing (NLP). NLP is one major field in artificial intelligence and is about interpreting the natural language (Khurana et al., 2022). The approaches in NLP are mostly machine learning based and the goal is to extract structured data from unstructured language input (Abdellatif et al., 2022a).

3.3.1 Data preprocessing

Data preprocessing refers to the process of cleaning, transforming, and organizing raw data to make it suitable for analysis or training machine learning models. The primary goal is to address data quality issues and prepare the data for further processing. These preprocessing methods are presented in the next section, with different NLU platforms performing this differently in detail.

Tokenization

By breaking text into smaller units called tokens, tokenization enables these NLU platforms to analyze and extract meaningful information from user input. Tokens serve as the basis for creating features such as bag-of-words representations, term frequency-inverse document frequency (TF-IDF) vectors, and word embeddings (Pai, 2020). These insights help conversational platforms generate contextually relevant responses. Tokenization is a prerequisite for named entity recognition (NER) and part-of-speech tagging. These tasks require the identification of individual words and phrases in order to tag them with appropriate categories or grammatical roles. These methods are also presented later in this chapter.

Stemming and lemmatization

Stemming and lemmatization are two major text normalization techniques in NLP (Khyani et al., 2021). These methods simplify words to their stem forms and make it easier for AI systems to analyze and understand text data. Although stemming and lemmatization share a common goal, they differ in their approaches and the results they produce. Stemming is a technique that reduces a word to its base or root form by removing inflections, prefixes, and suffixes. By simplifying words to their stems, stemming helps AI systems to recognize different forms of a word, reducing the complexity of the text and improving the system's ability to process and analyze it. Stemming algorithms typically use rule-based methods to remove affixes and reduce words to their stems. However, stemming can sometimes produce inaccurate results, as it may produce non-existent words or fail to account for irregular forms. Lemmatization is a more advanced technique that also reduces words to their base forms, but unlike stemming, it takes into account the morphological and grammatical structure of the word. Lemmatization uses linguistic knowledge to convert words into their basic forms, called lemmas, which are valid words of the language. Because lemmatization takes context and part of speech into account, it generally provides more accurate and meaningful results than stemming. Lemmatization usually relies on dictionaries or morphological analysis to determine the correct lemma for a given word. Therefore, lemmatization can be more computationally intensive than stemming, but it often leads to better performance on NLP tasks (Lang, 2022).

Part-of-speech tagging

Part-of-speech (POS) tagging is a fundamental technique in natural language processing (NLP) that assigns grammatical categories such as nouns, verbs, adjectives, and adverbs to individual words in a text. By labeling words with their corresponding part of speech, POS tagging helps AI systems better understand the structure and meaning of sentences, which improves their performance on various NLP tasks (Cutting et al., 1992).

By providing information about the grammatical structure of sentences, POS tagging enables AI systems to analyze the relationships between words, phrases, and clauses. This understanding is essential for tasks such as parsing, dependency analysis, and named entity recognition. In addition, part-of-speech tags help resolve ambiguities in word meaning by providing contextual information. POS tagging can clarify the intended meaning based on the grammatical role of the word. POS tags can be used as features in machine learning models for tasks such as text classification, sentiment analysis, and machine translation. By incorporating grammatical information, these models can make more accurate and contextual predictions. When generating natural language text, such as automatic summaries and chatbot responses, POS tagging can help AI systems produce more coherent and grammatically correct output (Chiche & Yitagesu, 2022).

Sentiment analysis

Sentiment analysis helps chatbots recognize the emotions, opinions, or attitudes that a user conveys through their text input. By detecting these sentiments, chatbots can provide personalized and emotion-aware responses that enable a more engaged and empathetic conversation. Sentiment categorization of the input as positive, negative, or neutral is performed based on the extracted features (Gavagnin, 2022).

Dependency Parsing

Dependency parsing is applied when the intention is to understand the structure of questions in order to discover relevant answers from a knowledge base or a text corpus.

It is based on the concept of dependency grammar (Tesnière, 1959), a linguistic theory that assumes that the syntactic structure of a sentence can be represented by a directed graph. In this graph, words are nodes and directed edges or dependencies indicate the relationships between them (De Spindler, 2022). Each dependency is labeled with a specific grammatical function, such as subject, object, or modifier. Dependency parsing is an essential part of understanding natural language because it provides a comprehensive representation of syntactic structure.

3.3.2 Feature extraction

Feature extraction in information retrieval involves transforming unstructured text data into structured, numerical representations that can be efficiently processed and analyzed by computer algorithms and machine learning models (Reed & Dubuf, 1993). These feature extraction techniques are not mutually exclusive, and multiple techniques can be combined to create a more comprehensive representation of the text data.

Bag of Words (BoW)

BoW is a technique for keeping track of the frequency of each word. Each input is represented as a vector in a high-dimensional space, where each dimension correspondents to a unique word in the vocabulary. In this approach, the specific order and grammatical structure of the words are disregarded, and only the word counts are considered (Great Learning Team, 2022). This is a relatively simple method and it gives some information about the frequency of words. However, it provides not much more information and uninformative information such as "and, the" occurs very often and distorts the result (Müller & Guido, 2016).

Term frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is an improvement over the Bag of Words method, as it considers not only the frequency of a word in a document but also its rarity across the entire corpus. Words that are frequent in a query document but rare in the corpus are given higher importance, while words that are frequent in both the document and the corpus are considered less important (Müller & Guido, 2016).

N-grams

N-grams are contiguous sequences of 'n' words in a given text. Instead of considering individual words as features (as in BoW), n-grams take into account word combinations and their order. This approach can capture more information about the text, such as local word dependencies and phrasal patterns, at the cost of a larger feature space (Müller & Guido, 2016). This approach helps chatbots to better grasp the context, local dependencies, and phrasal patterns in the conversation, leading to more accurate and contextually relevant responses.

Word Embeddings

Word embeddings are a technique that aims to represent words as dense vectors in a high-dimensional space. Unlike the bag of words approach, word embeddings capture both the semantic and syntactic meaning of words by considering their contextual relationships.

Word embeddings are learned from large corpora of text data using neural network models such as Word2Vec, GloVe, or FastText (Gavagnin, 2022).

Contextual embeddings take word embeddings to the next level by considering not only the context in which a word appears but also the specific position of the word within a sentence or document (Lee et al., 2022). These embeddings capture the meaning of a word based on its surrounding words, allowing for a more nuanced representation of language. Contextual embeddings are typically generated using deep learning models specifically transformer-based architectures like BERT, a popular pre-trained language model developed by Google (P. Shah et al., 2022). Furthermore, there are different pre-trained biomedical word embeddings like Pub-Med-word2vec, BioWordVec and PubMed-BERT (Lee et al., 2022). In summary, word embeddings allow the chatbot to have a deeper understanding of the user's input, thus increasing the accuracy and sophistication of the chatbot's response.

3.3.3 Intent Classification

While the techniques presented so far are barely visible in NLU platforms, intent classification is more noticeable. In this chapter, the classification methods are explained in more detail.

Intent classification is a very crucial part of any chatbot. Intents refer to the goal of a user’s input, therefore, they represent the user’s intention. To recognize user intents, bots use intents classification to analyze the input und determine the most likely intent. Intents could be described as a mapping between the user’s input and an action which is triggered in the chatbot (Ramesh et al., 2017). In order to get the most appropriate response from the chatbot, the intent classification model needs to be created with manual annotation. (Motger et al., 2022). The training data (user questions e.g., in an Excel) should be as close as possible to the input of future users to get a high confidence. For intent classification different techniques can be used, for example Support Vector Machine (Mu et al., 2017). Other algorithms used for intent classification are naive bayes or logistic regression (Helmi Setyawan et al., 2018). Further, random forests as another part of machine learning algorithm for intent classification may also be used (Assayed et al., 2022). These algorithms are trained on a set of labeled data and can learn to recognize patterns and features in user queries that are indicative of specific intents. Furthermore, there are also rule-based intent classifications, which were mainly used at the beginning of chatbots. This category includes pattern matching approaches that used a standardized markup language called artificial intelligence markup language (Motger et al., 2022). Usually, different approaches are combined for the classification of intentions. Deep learning models such as Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) are often applied (Tun & Soe, 2020). RNN are used as the basis for Long Short-Term Memory (LSTM). Here, a distinction is made between short-term knowledge within the conversation and long-term general knowledge (Maroengsit et al., 2019). Deep learning models can handle more noisy data and provide more accurate results than traditional machine learning models (Tun & Soe, 2020). Approaches using transformers also belong to this category. These achieve even better results due to their ability to capture long-range dependencies and contextual information effectively. Models like BERT (Bidirectional Encoder Representations from Transformers) have achieved state-of-the-art results on various intent classification benchmarks (Devlin et al., 2019).

3.3.4 Named Entity Recognition (NER)

Entity recognition tries to identify entities e.g., location, dates, names or numbers in the user's input. It is often used in conjunction with intents to provide more context and allow the chatbot to give a more accurate response. The entity extraction is pre-trained by manually annotating entities in the user's input (Adamopoulou & Moussiades, 2020a). While intents are more focusing on verbs, entities typically are nouns. Both are widely used techniques in NLU platforms (Canonico & Russis, 2018). Named Entity Recognition has been explored through multiple methodologies, including rule-based, supervised, semi-supervised, and unsupervised techniques. According to Carstensen (2010) the statistical methods used are based mostly on already established learning methods like Hidden Markov Models (Bikel et al., 1999), Support Vector Machines (Asahara & Matsumoto, 2003) or Conditional Random Fields (McCallum & Li, 2003). There are also rule-based approaches, although these involve a lot of manual work and have become less important in recent years (Eiselen & Bukula, 2022). The use of deep learning methods in the field of NER has brought about significant improvements in the results in recent years. Supervised neural networks are combined with unsupervised models like BERT or FLAIR (Akbik et al., 2019; Devlin et al., 2019; Eiselen & Bukula, 2022).

3.3.5 Response generation

There are three approaches for the generation of suitable answers. In the following paragraph, these approaches and their (dis-)advantages will be presented. In the chapter on intent classification, various algorithms such as RNN are utilized to determine the intention behind user queries. These algorithms play a crucial role not only in classifying intents but also in generating appropriate responses. Depending on the specific implementation, the networks can provide retrieval-based answers by retrieving pre-defined responses or create generic answers based on learned patterns. Therefore, the algorithms used for intent classification are closely linked to the process of response generation, as they contribute to the overall conversational capability and effectiveness of the chatbot.

As a first approach, there are ruled based responses. These responses have been built und chosen according to a predefined set of rules. The knowledge of the bot is hardcoded and first chatbots, which were developed, had this approach implemented. However, the drawback of this model is its weak robustness, as grammatical or spelling errors in the user input can result in inaccurate or incorrect responses (Adamopoulou & Moussiades, 2020a).

Second, there are retrieval-based outputs of bots. Retrieval based systems search in a database for the most appropriate answer (Song et al., 2016). Due to the fact that the responses are taken from a database, the response possibility is restricted. Therefore, this approach is more popular when the chatbot is developed for a closed topic domain as it is done in this thesis with creating a chatbot, which provides only answers for SGLT2 inhibitors. This might be seen as a disadvantage, however, due to the fact that the answers are retrieved of the database, the output of a bot can be better monitored. This allows the answers to be returned with a higher degree of confidence (Wu et al., 2017). In addition, a higher quality and consistency can be observed when applying retrieval-based bots (Boroghina et al., 2022).

The last model is the generative-based approach. In this approach, answers are generated based on the individual user input and the pretrained data (Motger et al., 2022). This allows the generation of a more human like response. Due to the fact that the answers are newly created, the response control is lost, which is why incomplete and incorrect answers may be possible (Kim et al., 2018). Further, a disadvantage might be that the development and training are more difficult because it requires more advanced algorithms (Adamopoulou & Moussiades, 2020a; Hien et al., 2018).

3.4 Evaluation methods of chatbots

After the development of a chatbot, its ability to provide appropriate answers has to be tested. According to Drozdal et al. (2021) there is no standard evaluation method for chatbots. Evaluation could be done automated, manually by humans or a as combination of both. The most famous evaluation method in the field of bots is the Turing test (Turing, 1950). The test is named after the scientist Alan Turing, who introduced the idea of interaction between machines and humans in the early 1950s. The Turing test examines whether a human can distinguish if a conversation is conducted by a human or a machine. It is considered to be passed when no distinction can be made between human and machine (Turing, 1950). However, there are concerns from different authors that the test is not sufficiently robust (Ramos, 2017; Wilson et al., 2017). Today, chatbots can be evaluated according to a wide variety of criteria. Motger et al. (2022) provide an overview of qualitative and quantitative evaluation criteria. For the use case with SGLT2 inhibitors, the qualitative evaluation is crucial, since the conveyed content can have a great influence on the medication compliance. In the overview based on ISO/IEC 25010, the authors have classified existing literature based on qualitative evaluation criteria. A distinction is made between functional suitability, performance efficiency, usability and security (Motger et al., 2022). Although all areas are important for an evaluation, the area of functional suitability is discussed here in more detail, since the content has a considerable influence on a Q&A bot in the medical field and is therefore of great importance for this thesis. One category of the functional suitability is the functional correctness. This is mostly considered to be effectiveness according to the literature (Casas et al., 2020; Guerino & Valentim, 2020). In addition to functional correctness, there is also functional appropriateness as a category, which also contains content evaluation (Maroengsit et al., 2019). This point is also crucial for the Q&A bot of SGLT2 inhibitors, so that correct patient education can be ensured, which is why it was considered in this thesis. For the analysis of the mentioned quality criteria above, there are different approaches. To measure criteria such as functional correctness or functional appropriateness, the measurement can be qualitative, such as interviews and questionnaires, or quantitative, as for example dialogue tracking and surveys (Motger et al., 2022).

Interviews serve as a powerful tool for evaluators to collect detailed feedback from users of the conversational agent. Hobert (2019) describes qualitative interviews as an opportunity to understand the impact of the user's interaction with the agent. In contrast, qualitative questionnaires provide an alternative approach to evaluation. The research literature reveals two dimensions of evaluation through questionnaires: goal-oriented and user satisfaction. Goal-oriented questionnaires are designed to assess specific qualities or effects of user-agent interaction. User satisfaction questionnaires, on the other hand, focus on usability and quality characteristics, covering areas such as emotional awareness, learning, and content relevance (Fitzpatrick et al., 2017). According to Maroengsit et al. (2019), user satisfaction can be evaluated on two levels: session level and turn level. Session-level questionnaires ask users to rate an entire conversation session with the agent, while turn-level questionnaires focus on individual agent responses for a more granular evaluation. For the evaluation of the chatbot by the specialists, this thesis considered the open-ended question approach in the interviews and a session-level questionnaire. Furthermore, Motger et al. (2022) observed software-based solutions that enabled automated quality testing. They identified Botium as state-of-the-art software for chatbot testing. This new possibility for the evaluation of chatbots is also used in this thesis.

Another approach to assessing the quality of the answers is the linguistic perspective. The responses of a chatbot can be divided into four areas for quality assessment (Rodríguez-Cantelar et al., 2021). First, the semantics have to be correct, meaning that the chatbot responds appropriately to the user's input. Second, the syntactic must be correct, which means that the output must be grammatically correct. Third, the answers must not only fit the input, but also be professionally correct. Finally, the answer should be specific and not too generic or neutral. The assessment of a chatbot for these characteristics is usually carried out by several test users (Rodríguez-Cantelar et al., 2021). These users evaluate the responses received for these quality features. Software-based testing enables the detection of grammatical errors or whether a certain question is followed by the desired output. This assessment was also included in the testing phase of this thesis. The main focus lay on the evaluation of syntactic as well as the semantics.

4 Current thesis

This thesis is a subproject of the project RealCo of the Institute of Business Technology at the ZHAW School of Management and Law. The RealCo project, led by Prof. Dr. Thomas Keller, aims to provide consultations on medications for CKD in a metaverse with an avatar. The goal is to increase the health literacy of the affected patients. This project is also a subproject of SHIFT, an Innosuisse Flagship Project under the direction of Prof. Dr. Alfred Angerer and Prof. Dr. Sven Hirsch. The aim is the digital transformation of hospitals into smart hospitals. The overall project runs for three years and ends in June 2025.

4.1 Objectives

The aim of the current thesis is to create and evaluate two chatbots for CKD patients concerning their questions regarding SGLT2 inhibitors. As stated above, the use of chatbots in the healthcare area may have various advantages and may even be beneficial for increasing their health literacy and therefore their medication compliance. Even though this is an intriguing question, it will not be examined in this work but will only be answered after concluding the entire RealCo project. The focus of the current study lies in developing and testing the two chatbots. In the first phase of this current work, the suitable NLU platform of the bot is evaluated, the second phase consists of creating two bots (artifact) and lastly, in the evaluation phase the output of the bot is tested by healthcare professionals.

Due to the non-transparency of the NLU platforms, the choice of the most suitable platform is difficult. For this reason, instead of one, two chatbots are being converted. By developing two bots on two different NLU platforms, these two platforms will also be compared. In addition, a conclusion will be taken from this as to whether several chatbots combined lead to better results or only to additional effort. The scope of the paper is limited to the chatbot, which receives text inputs and produces text outputs and focuses on the verbal German language. Speech to text as well as nonverbal speech and interaction of the avatar with the patients are not part of this thesis and will be developed separately.

4.2 Research questions

The current thesis aims to answer the following research questions:

- Could the simultaneous use of two chatbots based on two different NLU platforms increase the number of correct responses provided by the chatbot?

- Which NLU platform is most suitable for the development of a Q&A bot for SGLT2 inhibitors?
- Can a Q&A bot provide correct answers to SGL2 inhibitors?
- Does a Q&A bot meet the acceptance criteria of healthcare professionals?

The findings of this thesis are expected to contribute valuable insights into the effectiveness and viability of chatbots as a tool for information provision and support in the domain of SGLT2 inhibitors. Furthermore, it contributes with the comparison of two NLU platforms to find the most suitable NLU platform for future chatbots. In addition, the thesis develops with two chatbot prototypes the knowledge foundation about SGLT2 inhibitors for the avatar, which is used in the RealCo project. In the following chapter the used methods as well as the development process will be presented.

5 Methods

In this master thesis, the research design was oriented towards the Design Science Research Methodology (Peffers et al., 2007). The model consists of six steps, whereby in this thesis the focus was on steps three and four. These included the development of an artifact in the form of a Q&A bot and its evaluation.

5.1 Evaluation of NLU platforms

In order to create an artefact, it was first necessary to determine the best platform for its development. One objective of this thesis was to compare the performance of two popular NLU platforms specifically focusing on their application in healthcare. To achieve this, a two-pronged methodology was employed: a literature review and a practical case study.

5.1.1 Literature Review

The first stage involved a review of existing academic papers focusing on NLU platform comparisons. Relevant literature was sought by performing a comprehensive search on key health databases, including PubMed and JMIR. The search strategy was designed to use specific terms such as, 'IBM', 'Watson Assistant', 'Google' and 'DialogFlow' and further terms which are presented later in chapter 6.1. While these database searches revealed the prevalence of NLU platforms in healthcare, a snowball search was conducted in IEEE Xplore and the ACM Digital Library for comparisons between NLU platforms in general. From this review, two NLU platforms were identified for further investigation.

5.1.2 Case Study

Following the literature review, a practical case study was designed to further assess and compare the performance of Watson Assistant and DialogFlow. Both platforms were trained as mentioned in chapter 5.2 and 5.3. with a dataset designed for this thesis, specifically related to SGLT2 inhibitors.

The training process involved feeding each platform with a range of potential questions as presented in **Table 1** in the next chapter and appropriate responses about SGLT2 inhibitors. Once the training was completed, both chatbots were tested with a set of 15 predetermined questions related to SGLT2 inhibitors as shown in **Table 5** in chapter 6.3. These questions were designed to mimic real-world queries and to test various aspects of the NLU performance.

The primary metrics used for comparison were the latency of responses and the accuracy of detected intents. Latency was measured as the time taken for each platform to respond to a query, while the accuracy of detected intents was assessed based on whether the chatbot's response correctly interpreted and addressed the question asked.

The collected data were then analyzed and compared to determine which platform delivered the best performance in terms of speed and accuracy. The outcomes of this comparative analysis, along with insights gleaned from the literature review, will be presented in the results.

5.2 Artefact development and design of the bot

This chapter presents the procedure for the development of the chatbot and the features of the chatbot.

5.2.1 Preparations for development

As a first step, the appropriate software solution implemented for a chatbot was evaluated. Furthermore, the required data was prepared for the use in the development of the chatbot. This included the most common questions from patients on the topic of SGLT2 inhibitors and the answers to them. The data was based on package inserts of medications as well as information about the medications from leaflets. Before starting programming a bot on the evaluated third-party platform, an excel spreadsheet with possible questions (intents) from patients and the answers (database for retrieval-based approach) was created. This served as database for the response of each intent. Since the healthcare industry is a specific area of research involving the importance of correct information for diagnosis and treatment (Lehto et al., 2012), the questions and answers for the bot were reviewed by the head of nephrology, Prof. Dr. med. Stephan Segerer, at the Cantonal Hospital Aarau to ensure the accuracy of the answers. **Table 1** shows examples of questions and intents.

Table 1

Possible patient questions and classified intents to the questions

Example question	Intents
Welche Nebenwirkungen können auftreten?	Nebenwirkungen
Wann muss ich einen Arzt aufsuchen aufgrund der Nebenwirkungen?	Nebenwirkungen
Wie sollte das Medikament eingenommen werden?	Einnahme
Wann wird das Medikament angewendet?	Anwendungsgründe
Wann darf ich das Medikament nicht einnehmen?	Einnahme

Wann muss bei der Einnahme Vorsicht geboten sein?	Vorsicht
Beeinträchtigt das Medikament meine Fahrfähigkeiten?	Einschränkungen
Darf ich Alkohol während der Behandlungszeit konsumieren?	Alkoholkonsum
Wann ist die Wirkung vom Medikament verringert?	Wirkung
Wie wirkt das Medikament?	Wirkung
Darf ich das Medikament einnehmen, wenn es abgelaufen ist?	Ablaufdatum
Was sind die Unterschiede zu früheren Medikamenten?	Alternativen
Darf ich das Medikament einnehmen, wenn ich schwanger bin?	Schwangerschaft

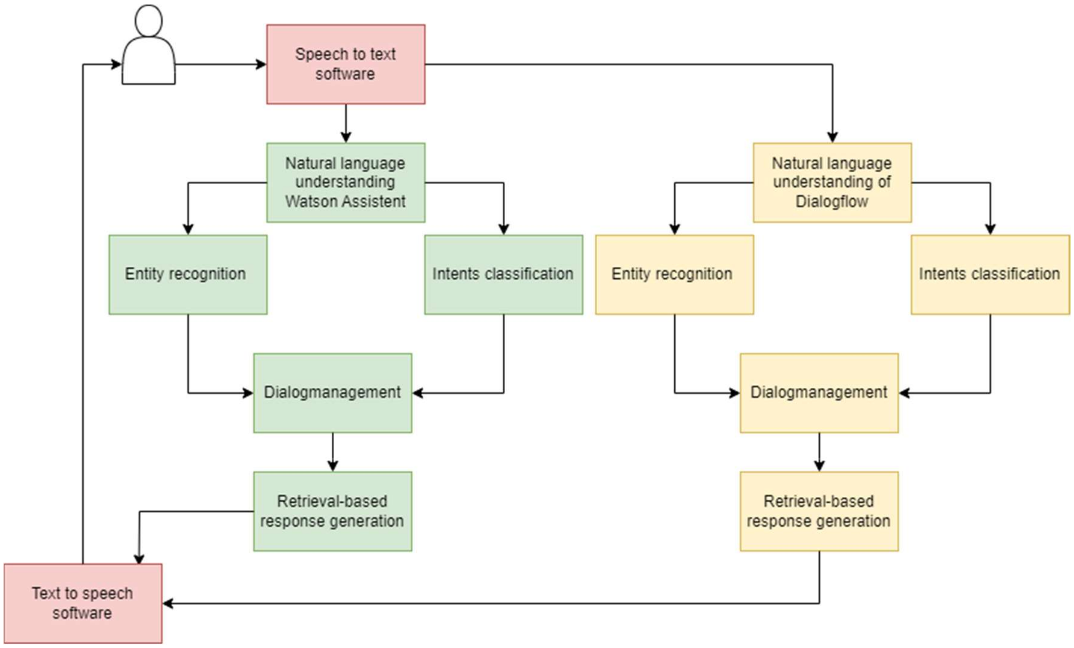
5.2.2 Chatbot design

Instead of developing one bot, two bots on two different NLU platforms were created. The added value of developing two bots should be less dependence on providers and their NLU platform. Whether this added value is worthwhile was investigated in the first research question. For this, latency and correct response were measured to compare the platforms.

To better classify what the chatbots can do, the design of the chatbots is presented. The goal of the bot was to answer questions about SGLT2 inhibitors. Due to this, the knowledge base is to be understood as a closed domain, as the bot only provides answers about SGLT2 inhibitors. Furthermore, it is a non-task-oriented bot, which returns information without triggering a task or similar. The answer generation was retrieval-based. Retrieval-based approaches offer the advantage over generative-based approaches that the answer takes on a predefined set of answers. The response which is the most accurate according to the algorithm is taken. This is significant for healthcare applications, as it allows control over the answers to be retained. If a different approach were taken on natural language generation (NLG), the answers are compiled by the algorithm itself and the risk of inaccurate responses may increase, as mentioned in chapter 3.3.5. While the interaction between user and avatar is voice, the chatbot is built for text interaction only. The development of the speech to text interaction underlies the responsibility of another part of the RealCo project. Hence, it was not considered while developing the chatbot. For a better understanding and differentiation of the chatbots from other software components in the RealCo project, a rough architecture is drawn in **Figure 3**. A bot consists of various components. Each component can contain different algorithms and metrics which are discussed in chapter 3.3.

Figure 3

High-level chatbot architecture


Speech to text software
Natural language understanding Watson Assistent
Entity recognition
Intents classification
Dialogmanagement
Retrieval-based response generation
Natural language understanding of Dialogflow
Entity recognition
Intents classification
Dialogmanagement
Retrieval-based response generation
Text to speech software

Note. Red boxes are not part of the scope of this thesis.

5.2.3 Development of the DialogFlow-Bot

First, a draft of the bot was created in DialogFlow. In the first development cycle the process went as follows:

1. Intent definition: The first step was to create new intents that describe different categories or types of user questions. This categorization allows for a structured understanding of user queries. For example, a relevant intent might be called "Vergessen", which addresses concerns related to missed medication. A special intent which was created, is the fallback intent. This intent is used if the chatbot does not find a suitable intent. It asks the user to rephrase the question.

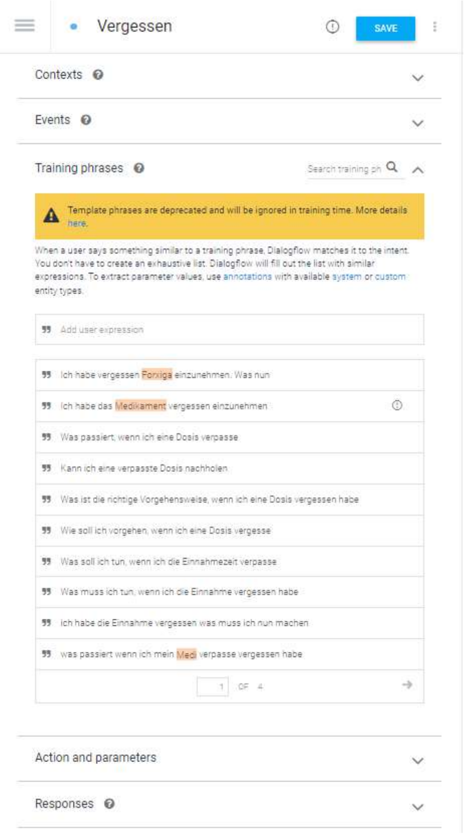
2. Intent naming: Each intent should be given a descriptive name that accurately reflects its purpose. This nomenclature helps to organize and distinguish between different intents. For example, the intent mentioned above could be called "Vergessen", thus aligning the name with its intended focus.

3. Selection of example questions: In order to train the conversational system to recognize and understand user questions, it is essential to provide representative sample questions within the defined intent category. These examples serve as training data, enabling the

system to recognize and understand similar questions in subsequent interactions. For example, for the “Vergessen”- intent, appropriate example questions as presented in **Figure 4** might be "Ich habe die Einnahme vergessen, was soll ich nun tun?"

Figure 4

Intent with training questions in DialogFlow



Template phrases are deprecated and will be ignored in training time. More details here.

When a user says something similar to a training phrase, Dialogflow matches it to the intent. You don't have to create an exhaustive list. Dialogflow will fill out the list with similar expressions. To extract parameter values, use annotations with available system or custom entity types.

- Ich habe vergessen Forxiga einzunehmen. Was nun
- Ich habe das Medikament vergessen einzunehmen
- Was passiert, wenn ich eine Dosis verpasse
- Kann ich eine verpasste Dosis nachholen
- Was ist die richtige Vorgehensweise, wenn ich eine Dosis vergessen habe
- Wie soll ich vorgehen, wenn ich eine Dosis vergesse
- Was soll ich tun, wenn ich die Einnahmezeit verpasse
- Was muss ich tun, wenn ich die Einnahme vergessen habe
- Ich habe die Einnahme vergessen was muss ich nun machen
- was passiert wenn ich mein Medi verpasse vergessen habe

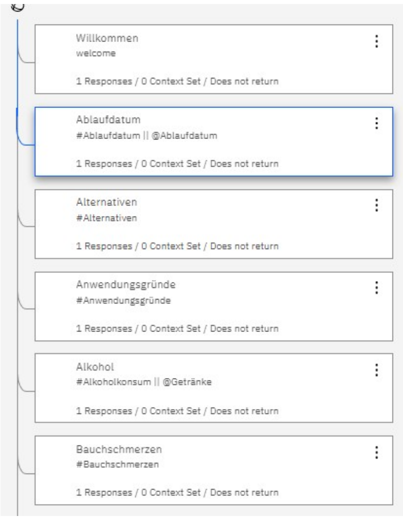
4. Pre-defined response inclusion: When a question is recognized as belonging to a specific intent, the conversational system can provide a pre-defined response tailored to that intent. In this context, the answer provided by Prof. Dr. med. Stephan Segerer can be used as an example response. This expert input serves as a reliable source of information for effectively responding to user queries.

5.2.4 Development of the Watson Assistant-Bot

The development in Watson Assistant was very similar to that described above. The difference between the two platforms is in the dialogue management. In DialogFlow, the answers were written to the respective intent. In Watson Assistant, an independent dialogue with dialogue nodes had to be created for each intent. This dialogue tree was processed for each user input in order to find the most suitable answer. **Figure 5** shows the dialogue tree created in Watson Assistant.

Figure 5

Dialog-Nodes in Watson Assistant



Willkommen
welcome
1 Responses / 0 Context Set / Does not return

Ablaufdatum
#Ablaufdatum || @Ablaufdatum
1 Responses / 0 Context Set / Does not return

Alternativen
#Alternativen
1 Responses / 0 Context Set / Does not return

Anwendungsgründe
#Anwendungsgründe
1 Responses / 0 Context Set / Does not return

Alkohol
#Alkoholkonsum || @Getränke
1 Responses / 0 Context Set / Does not return

Bauchschmerzen
#Bauchschmerzen
1 Responses / 0 Context Set / Does not return

Note. Only an extract of the whole dialog.

5.3 Testing the chatbots

Van Bussel et al. (2022) have explored which capabilities are relevant for the application and perception of bots for cancer patients. They highlight the following points, which are important in healthcare: performance expectancy, effort expectancy and trust. These factors can be achieved if physicians actively recommend bots, patients are involved in the development and receive training for the correct use of the bot. Therefore, development and testing in this thesis will be iterative in order to provide continuous enhancements.

The testing process for the chatbots in this thesis is structured into three stages. The initial stage involved conducting tests during the early development phase, with a primary focus on evaluating the functionality of the bots. This stage aimed to assess the basic capabilities and performance of the chatbots as they were being developed. These tests were performed with Botium, a chatbot testing software by Cyara. According to Motger et al. (2022), Botium is a reference in academic fields with respect to software-based solutions.

Botium could be connected via API with NLU platforms like DialogFlow and offer different functionalities like for example end-to-end user flow simulation for Chatbot-Testing (*Botium Box - The Chatbot Testing Tool*, o. J.). The second testing was executed by healthcare professionals from the Cantonal Hospital Aarau on site in Aarau. In addition to improved functionality, the focus here was on the content of the answers. In consultation with Prof. Dr. med. Stefan Segerer, further testing with patients will be carried out as soon as the interaction can take place in the metaverse. While the connection via API of the chatbots to the virtual reality software was successful in April 2023, there were still technical issues on the part of the VR software, which was the reason why it was not possible to test it within the scope of this thesis. Instead of the planned testing with patients, a third testing was carried out with the professional medical staff consisting of pharmacists and physicians specializing in nephrology. While the second testing was accompanied in order to explain the software, the third testing was done without supervision by the chatbot developer. Therefore, the second testing was more independent. After testing, all participants were instructed to fill out a questionnaire with nine questions as shown below in **Table 2** about their experience with the bot.

Table 2

Questionnaire for second testing with physicians and pharmacists

Nr.	Questions
1	Were any misstatements regarding medications identified in the responses?
2	What was your overall experience using the chatbot?
3	Did the chatbot provide you with the information you were looking for?
4	Did the chatbot respond promptly to your inquiries?
5	Were the chatbot's responses relevant to your questions?
6	Did you find the chatbot's personality engaging?
7	Were there any answer of the chatbot that you particularly disagree?
8	What, if anything, could be improved about the chatbot?
9	Do you have any additional comments or feedback about the chatbot?

Note. The questionnaire was translated into German.

6 Results

The following chapter presents the results from the comparison of the NLU platforms as well as the prototyping and testing procedure.

6.1 Evaluation of software

The first step for creating the chatbot was to choose a NLU platform from a third party. The existing research above indicated that the Watson Assistant by IBM performs best on average, which is why it was used in this thesis. In order to find the most suitable NLU platform for the use case of SGLT2 inhibitors, the second step was to investigate which NLU platform is frequently used in healthcare. Existing literature presented that Google's DialogFlow is commonly used (Kadariya et al., 2019; Nikitina et al., 2018; Rosruen & Samanchuen, 2018; van Heerden et al., 2017; Vasileiou & Maglogiannis, 2022). Watson Assistant (Fadhil et al., 2019) and Microsoft LUIS were also partially used. The reason for using DialogFlow so often in the research papers was not explained by the various authors. Explanation and justification for the use of DialogFlow would also be helpful for future papers in selecting the most appropriate platform.

In order to find the most suitable NLU platform, existing literature was reviewed. The most popular and most widely used NLU platforms were compared in various studies (Abdellatif et al., 2022b; Canonico & Russis, 2018; V. Shah & Shah, 2019; Thorat & Jadhav, 2020). As shown in chapter 3.2, it was determined that the two platforms DialogFlow and Watson Assistant appear in all comparisons and rank the highest. Furthermore, both were also used for health bots as mentioned above. To confirm the popularity of the platforms, a comprehensive search of databases such as PubMed and JMIR, which are primarily sourced from the healthcare sector, was deliberately conducted. Literature relevant to the topic was sought by utilizing specific search terms listed in **Table 3**. The findings reveal that Watson and DialogFlow were the platforms most frequently employed in healthcare research.

Table 3

Result from literature search to NLU platforms

Search key (1.1.2020 – 1.12.2022)	Journal of Medical Internet Research	PubMed
IBM AND Watson AND Assistant	16	51
Google AND DialogFlow	8	3
Microsoft AND LUIS	1	1
Amazon AND Lex	2	0

Canonico & Russis (2018) developed a taxonomy (presented in the chapter 3.2) in their comparison, whereby 13 factors are considered. After the review of the existing literature in chapter 3.2 of this thesis of bot comparisons, the taxonomy of Canonico & Russis (2018) was updated in **Table 4** and was modified by one parameter. The reason for the update and modification was, that since 2018 the platforms have evolved and this thesis aims to reflect the latest state of the art. Instead of the supported programming languages, the parameter transparency of AI was included. In this point, both platforms were equally lacking in transparency, since they are proprietary platforms. Thus, IBM and Google only communicate very superficially about the methods and techniques used. Furthermore, it can be observed that the two most widespread NLU platforms provide almost identical features (Dialogflow | Google Cloud, o. J.; Watson Assistant - Einführung, o. J.). From this chapter it emerged that instead of one bot, two bots will be developed, since both platforms evaluated demonstrate many similarities. Whether the algorithms of one of the two bots were superior to the other platform will be discussed in a later chapter. It may also be possible to determine improved outcome by using the two bots, which would justify the additional effort of two developments.

Table 4

Updated taxonomy according to Canonico & Russis (2018)

Platform	Usability	Languages	Transparency about the AI	Pre-build entities	Pre-build intents	Default fallback intent	Automatic context	Composition mode	Online-integration	Webhook/SDK Availability	All-in Platform	Linkable intents	Price
Dialog-Flow Essentials	High	Over 30 languages	low	available	available	available	yes	Form-based bot builder	many built in integrations	yes	yes	yes	Free trial edition
Watson Assistant	High	13 languages	low	available	available	Fallback available as dialog node	yes	Form-based bot builder	many built in integrations	yes	yes	yes	Free lite plan

6.2 Developing and testing the artefacts

The development of the artifact was conducted in several iterative phases. The procedure and the results are presented in the next chapters.

6.2.1 Developing the prototype

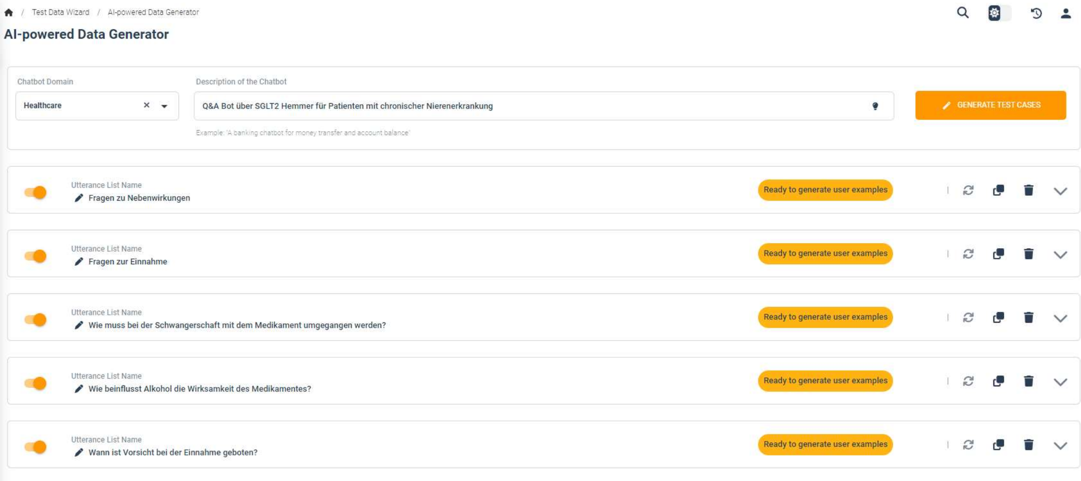
In the first development cycle as presented in chapter 5.2, around 400 questions distributed over 23 intents were enriched. Entities were created for the most frequently used expressions such as "Medikament" or "SGLT2 Hemmer".
The building of the same solution with the same content using IBM Watson Assistant was carried out after the completion of the fourth development cycle in order to keep the development effort to a minimum.

6.2.2 Testing the prototype with testing-software

The first testing was performed with Botium. This is a software from Cyara that is specialized in the testing of chatbots. Botium offers a free and a paid version of their software. The paid version provides more features, such as the generation of questions using AI which is used in this thesis. A free 30-day trial version was given to use the advanced tools. Hence, manual testing was dispensed and testing was carried out with natural language generative questions. For each intent, it was possible to briefly explain what the intent was about. Based on that, the system generated with AI a total of 500-600 questions which formed the test set. These questions were then entered into the bot via API. This increased the number of questions contained in the bot and improved its robustness.

Figure 6

AI-powered data generator by Botium



/ Test Data Wizard / AI-powered Data Generator
GENERATE TEST CASES
Ready to generate user examples

AI-powered Data Generator

Chatbot Domain: Healthcare

Description of the Chatbot: Q&A Bot über SGLT2 Hemmer für Patienten mit chronischer Nierenerkrankung

Example: 'A banking chatbot for money transfer and account balance'

Utterance List Name: Fragen zu Nebenwirkungen

Utterance List Name: Fragen zur Einnahme

Utterance List Name: Wie muss bei der Schwangerschaft mit dem Medikament umgegangen werden?

Utterance List Name: Wie beinflusst Alkohol die Wirksamkeit des Medikamentes?

Utterance List Name: Wann ist Vorsicht bei der Einnahme geboten?

Figure 6 provides an insight into Botium's tool and the possibility to generate questions via NLG. By testing with the test set, the newly generated questions could be assigned to the existing intents in the DialogFlow or, if completely new questions had arisen, they could be assigned to new intents. With this, the second development phase was completed.

6.2.3 Testing the prototype with physicians

In order to ensure that the chatbot would later answer patients' questions correctly from a professional's point of view, the chatbot was tested with four physicians from the department of nephrology and one pharmacist from the Aarau Cantonal Hospital on site. This testing was the first time the bot was tested by professionals, whereas within the whole testing phase, this was already the second testing step. Based on their experience with patients' questions regarding SGLT2 inhibitors, test persons were able to verbally ask the bot their own questions. For this, the integrated speech to text component of DialogFlow was used. This ensured that the spoken language in the chatbot was trained which was beneficial for the further development of the interaction between patient and bot in the metaverse.

After consulting with the specialists, there were four main findings. First, new questions could be generated by the test participants. These new formulations were incorporated into the existing intents. Second, the general feedback was that the bot should communicate with the patients in the German polite form of speaking (in German: "Sie"). Hence, this was adjusted in the chatbot. Third, the physicians wished the bot to address more symptoms such as nausea, symptoms of fever, abdominal pain and eczema. Based on this feedback, new intents were created and trained. Lastly, it was also concluded that the differentiation of SGLT2 inhibitors from other medications and/or themes cannot always be fully ensured. To avoid false statements, the bot was programmed to refer to consult a physician or a pharmacist if the question is not within the trained scope.

After the adjustments mentioned above, the chatbot was once again revised by Prof. Dr. med. Stephan Segerer. Due to the fact that Forxiga as a drug is now reimbursed by Swiss health insurance companies for chronic kidney disease, the answers in the chatbot are based on this medicine. Slight differences with other SGLT2 inhibitors such as Jardiance, Vokanament or Qtern are possible, but according to Prof. Dr. med. Stefan Segerer do not have to be considered in the chatbot. Thereby, the medication within the SGLT2 inhibitors can be treated equally.

6.2.4 Second testing of the prototype with physicians

As a third testing step, the chatbot was tested with professionals once more. This was their second testing and it was carried out with the medical staff consisting of pharmacists and physicians specialized in nephrology. While the first testing was accompanied on site by the chatbot developer in order to explain the software, the second testing was done without supervision. Therefore, the second testing was more independent and the risk of a bias due to presence of the chatbot developer could be reduced. Hence, in the second testing, the professionals once again received the task to ask the bot their questions. After testing, all participants were instructed to fill out a questionnaire with nine questions about their experience with the bot as presented in **Table 2** in chapter 5.2. The questionnaire and access to the chatbot for testing was sent by e-mail to two pharmacists and four physicians. The chatbot was tested by two pharmacists and two physicians which all returned the questionnaire by e-mail. Despite reminders for the testing by e-mail, no further testing was carried out by the remaining two physicians contacted. The results from the questionnaires were as following:

Regarding any misstatements concerning medications in the chatbot's responses, in general, all test participants agreed with the answers regarding correctness. However, one participant suggested a slight wording adjustment which was then implemented. One respondent was repeatedly asked by the chatbot to rephrase the question in order to obtain the answer she was looking for. The overall experience with the chatbot was perceived as good by all test participants. An improvement over the first testing was also noted by one person. Furthermore, the information that the users were looking for was provided for the most part. This means that the chatbot was mostly able to deliver the correct answers but that there were still subjects that the chatbot did not yet cover. One participant even pointed out that there were more questions to be added to the chatbot's scope. Certain issues were subsequently included as intent in the chatbot or deliberately delimited, such as medication prices. There was feedback that different SGLT2 inhibitors vary in their costs and that the chatbot did not consider this in its answers. Since different SGLT2 inhibitors can be treated equally within the chatbots according to consultation with Prof. Dr. med. Stephan Segerer and the primary focus of this thesis was on Forxiga, the bot only provides prices for this specific pharmaceutical. Regarding latency, all users were satisfied with the response time between the input from the user and the output from DialogFlow. The answers provided were relevant to the user's questions. In some cases, it was noted that the answers were rather general, even if the question asked about something more specific. Furthermore, no answer was provided for one question, which was concerning the ingredients of SGLT2 inhibitors. This point was also resolved afterwards and programmed as an intent. The persona of the chatbot was changed to the German polite form after the first tests and the formulations were formal. All participants found the persona of the chatbot appealing. Additionally, the speech tone, which was delivered via text-to-speech by DialogFlow, was described as pleasant or very friendly by two people. One person noted a slight hesitation within a word. In the STT area, grammar and spelling are of central importance, and avoiding the use of ä, ö, ü can help to reduce this stuttering. Any errors are immediately noticed in the audio and can disturb the interaction When asked if test participants disagreed with certain content, one case was reported where DialogFlow detected the wrong intent. Furthermore, when a follow-up question was asked, the chatbot in the current version could not handle such follow-up questions except for the pharmaceutical ingredients. Instead of referring to physicians when the chatbot cannot assist, one person recommended also considering pharmacists, given that contact is available without an appointment and can also provide answers to questions about the medication. In the eighth question, more variation in the training questions was requested by the participants. Under other comments, no further input was provided. Overall, the project was assessed by the test persons as an interesting project and their feedback highlight current advantages and disadvantages of the chatbot. This will be further discussed in later chapters.

6.2.5 Development of the second prototype in Watson Assistant

Once the testing was completed, the chatbot included over 800 training questions and over 40 intents on the topic of SGLT2 inhibitors. These training questions and intents were then transferred from DialogFlow to Watson Assistant in order to have two identical chatbots for the comparison presented in the next chapter.

6.3 Experimental comparison of NLU platforms

Since NLU platforms are relatively non-transparent with their deployed algorithms, the exact same chatbot was developed twice. This is intended to identify which platform applies the superior algorithms in the background. Once this was carried out in Google's DialogFlow and once on IBM's NLU platform Watson Assistant. Both platforms offered different versions of their product, that goes from free versions to paid version depending on the amount of user or messages etc. DialogFlow ES, which was applied in this thesis, is recommended for medium and small bot which was the case in this research. Watson

Assistant also has different offerings whereby the free version was used for Watson Assistant and DialogFlow which was intended to ensure the comparability of the platforms. The comparison in **Table 5** is intended to answer the research questions defined in chapter 4.2. The first research question was whether the development of multiple chatbots adds value, meaning that the number of correct detected intents is increased, and the second question was which NLU platform is most suitable for the use case of SGLT2 inhibitors. For answering these questions, 15 questions were collected manually about SGLT2 inhibitors that were not trained in the intents before. These questions were entered into the different chatbots. **Table 5** and **6** presents the results of this comparison which consisted of two parameters for evaluation. First: Is the chatbot delivering the correct answer? Second: How quickly does the chatbot provide the answer? The second point is essential for the use case of the RealCo project, since a high latency would make the avatar's speech appear unrealistic.

The first comparison parameter is the correct intent detection. In **Table 5**, columns 2 and 4 show the detected intents of the corresponding NLU platform, and with which intent the platform answers the question posed in column 1. In addition to the detected intent, the confidence score, which lies between 0 and 1, was recorded.

It can be stated that for Watson Assistant all reported intents matched the question submitted. In DialogFlow, five intents were reported incorrectly and are marked in red in **Table 5**. Furthermore, in **Table 5** the latency per question is presented, which was measured in milliseconds. As with the first testing in chapter 6.2 the latency was measured with the software from Cyara called Botium. Both bots were connected via API with Botium and the test set with the 15 questions below were executed once. While executing the test cases, Botium can record the time it takes for the chatbot to respond to each message. This can be done using performance testing features or custom scripts. The response time is calculated as the difference between the timestamp when Botium sent a message and the timestamp when it received a response from the chatbot. Text to speech (TTS) and speech to text (STT) was turned off for both platforms. This was because in the project RealCo TTS and STT are not a part of the bot system, but of a separate software.

Table 5

Comparison of SGLT2-Bots in DialogFlow and Watson Assistant

Sentences	DialogFlow Intent detection	Latency	Watson Assistant Intent detection	Latency
Darf ich das Medi beim Stillen trotzdem nehmen?	0.56 (#Schwangerschaft)	382ms	0.85 (#Schwangerschaft)	167ms
Kann ich während der Einnahme des SGLT2-Hemmers Alkohol trinken?	0.80 (#Alkoholkonsum)	554ms	0.93 (#Alkoholkonsum)	166ms
Wie beeinflusst Forxiga meinen Blutdruck?	0.71 (#Blutdruckauswirkungen)	349ms	0.91 (#Blutdurckauswirkungen)	167ms
Was sind die häufigsten Nebenwirkungen des SGLT2-Hemmers?	0.67 (#Nebenwirkungen)	429ms	0.84 (#Nebenwirkungen)	172ms
Ich musste mich nach der Einnahme übergeben. Soll ich nochmals eine Tablette einnehmen?	0.84 (#Einnahme)	399ms	0.79 (#Erbrechen)	172ms
Ist es in Ordnung, wenn ich auch mal zwei Tabletten am Tag nehme?	0.55 (#Einnahme)	482ms	0.57 (#Einnahme)	172ms
Was löst das Medikament in meinem Körper aus	0.61 (#Wirkung)	444ms	0.41 (#Wirkung)	158ms
Sind SGLT2 Hemmer giftig bei gesunden Menschen?	0.79 (#Toxizität)	369ms	0.97 (#Toxizität)	158ms
Was soll ich tun, wenn ich eine Infektion im Genitalbereich erhalte?	0.65 (#Infektion)	528ms	0.45 (#Nebenwirkungen)	169ms
Ich hatte gestern das Medikament vergessen. Soll ich heute zwei Tabletten nehmen?	0.75 (#Vergessen)	290ms	0.87 (#Vergessen)	160ms
Wie lange muss ich nun SGLT2 Hemmer nehmen?	0.66 (#DauerderEinnahme)	401ms	0.83 (#DauerderEinnahme)	144ms
Was kann ich mit abgelaufener Packung von SGLT2 Hemmern tun?	0.52 (#Überdosis)	643ms	0.62 (#Ablaufdatum)	186ms
Ich kriege einen Hautauschlag seit ich SGLT2 Hemmer nehme. Was soll ich tun	0.55 (#Anwendungsgründe)	472ms	0.95 (#Hautauschlag)	162ms

Ich nehme Forxiga ein. Wie teuer ist dieses Medikament und zahlt meine Krankenkasse?	0.81 (#Preis/Vergütung)	392ms	0.91 (#PreisVergütung)	168ms
Ich nehme Metformin ein. Nun muss ich auch noch SGLT2 Hemer einnehmen. Ist das sinnvoll?	0.43 (#Anwendungsgründe)	683ms	0.54 (#Unwissen)	190ms

Note. The intents in red are wrong classified.

The average latency as well as the number of successful test cases of all questions were presented in **Table 6**. The average latency was calculated by summing the latency per question and dividing by the total number of questions. The same calculation was applied for the average intent confidence instead of latency the intent confidence per question was used, not considering misrecognized intents. The comparison of the latency times indicates that Watson Assistant provided faster responses across all 15 questions. On average, Watson Assistant gave the answer to the asked question within 167.40 milliseconds. On the other hand, DialogFlow needed 448.47 milliseconds on average for the same questions. This is 2.6 time longer than Watson Assistant. In summary, DialogFlow performed worse than Watson Assistant in terms of successful recognized intents as well as latency.

Table 6

Summary of the test results

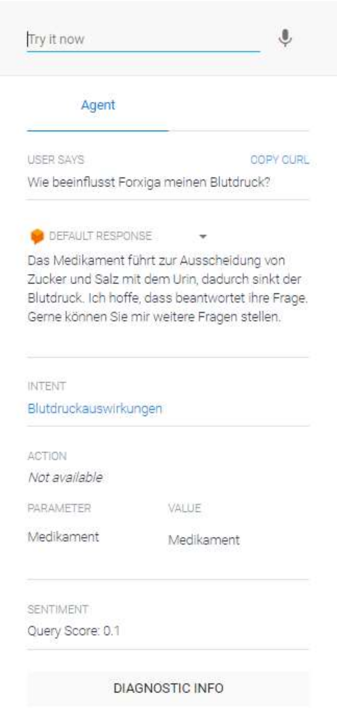
	Successfull test cases	Average latency	Average intent confidence
DialogFlow	10/15	448.47ms	0.691
Watson	15/15	167.40ms	0.763

Note. Only correctly detected intents considered in the average intent confidence calculation.

Another question which arose was, how confident the chatbot was in delivering the answer? In **Table 6** the average intent confidence for each NLU platform is shown. While the average score for DialogFlow was 0.691, Watson Assistant reached a score of 0.763. Intent detection confidence refers to the probability that the algorithm has correctly detected a user's intent. The intent detection confidence is usually expressed as a value between 0 and 1, where 1 means the highest confidence that the detected intent is correct. At a high confidence, the system is more likely to provide a correct and helpful response. At a low certainty, the system is more likely to have misunderstood the user's intent, which may result in an incorrect response. Although both provided confidence scores for the identified intents, the scores were not directly comparable as discussed later in this thesis in chapter 7.1 because they were generated using different algorithms and models which were not publicly available. **Figures 7** and **8** illustrate the developed chatbots in DialogFlow and Watson Assistant.

Figure 7

Chatbot in DialogFlow



USER SAYS
Wie beeinflusst Forxiga meinen Blutdruck?

DEFAULT RESPONSE
Das Medikament führt zur Ausscheidung von Zucker und Salz mit dem Urin, dadurch sinkt der Blutdruck. Ich hoffe, dass beantwortet ihre Frage. Gerne können Sie mir weitere Fragen stellen.

INTENT
Blutdruckauswirkungen

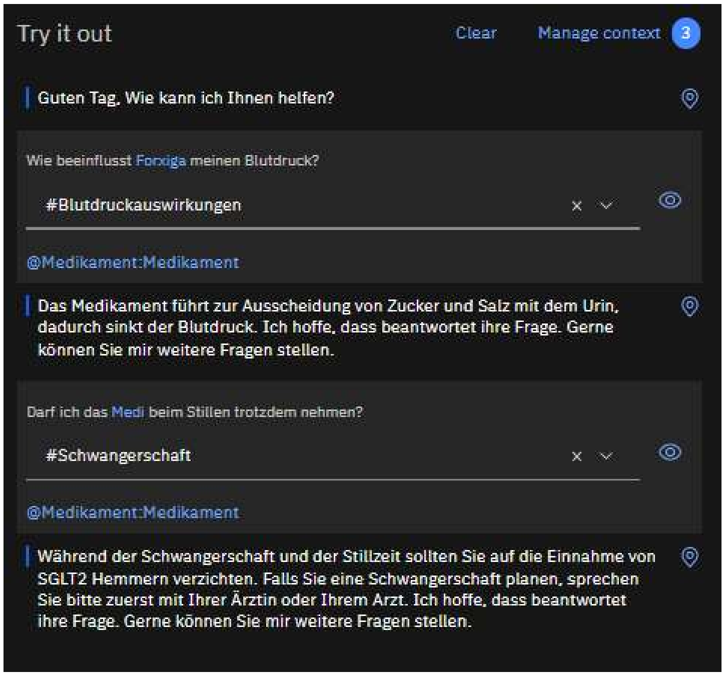
ACTION
Not available

PARAMETER	VALUE
Medikament	Medikament

SENTIMENT
Query Score: 0.1

Figure 8

Chatbot in Watson Assistant



Guten Tag. Wie kann ich Ihnen helfen?

Wie beeinflusst Forxiga meinen Blutdruck?

#Blutdruckauswirkungen

@Medikament:Medikament

Das Medikament führt zur Ausscheidung von Zucker und Salz mit dem Urin, dadurch sinkt der Blutdruck. Ich hoffe, dass beantwortet ihre Frage. Gerne können Sie mir weitere Fragen stellen.

Darf ich das Medi beim Stillen trotzdem nehmen?

#Schwangerschaft

@Medikament:Medikament

Während der Schwangerschaft und der Stillzeit sollten Sie auf die Einnahme von SGLT2 Hemmern verzichten. Falls Sie eine Schwangerschaft planen, sprechen Sie bitte zuerst mit Ihrer Ärztin oder Ihrem Arzt. Ich hoffe, dass beantwortet ihre Frage. Gerne können Sie mir weitere Fragen stellen.

7 Discussion

This chapter presents a comprehensive analysis and synthesis of the findings regarding the development and comparison of two chatbots about SGLT2 inhibitors on Watson Assistant and DialogFlow. It contextualizes these findings within the broader literature on chatbot development and NLP platforms, offering insights into the implications and limitations of the research and identifying potential future directions for further research in this area.

7.1 Experimental comparison of NLU platforms

The aim of this thesis was to provide a chatbot for CKD patients. Before this objective could be achieved, the appropriate platform for implementing the chatbot had to be evaluated. For this purpose, in chapter 3.2, different studies were presented that compared NLU platforms. Unlike Abdellatif et al. (2022a) and Canonico & Russis (2018) which compared confidence scores delivered by the NLU platform, this thesis compared the latency between question-and-answer delivery and the correctness of the intent detection. This was because the method behind confidence scores at Watson Assistant and DialogFlow were unexplained and unavailable to the public. This is the main difference between this thesis and all mentioned comparisons about NLU platforms (Abdellatif et al., 2022a; Canonico & Russis, 2018; V. Shah & Shah, 2019). Although this paper used different parameters for the comparison than the existing research, the results were very similar. The results in this thesis demonstrated that Watson Assistant achieved the best results regarding latency and intent detection. Abdellatif et al. (2022a) and Canonico & Russis (2018) also rated Watson Assistant as the best platform for chatbot development.

For the research question about the most suitable platform, the numbers of correctly assigned intents can be analyzed. **Table 5** in the results chapter showed that with Watson Assistant, intent detection worked very well and a suitable answer was found for all questions. With DialogFlow, this was only the case for 10 out of 15 questions. This means that inadequate answers were provided for five questions. Although these wrong responses are not wrong in terms of content, the answers do not add value to the patient in response to the question asked. This can reduce acceptance as well as trust among patients. To avoid this, a threshold can be set. For example, a threshold of 0.6 means, that only if an intent with a higher confidence than 0.6 is detected, the chatbot will provide an answer. If all intents detected lower than 0.6, the fallback intent steps in and asks the user to rephrase the question. However, this thesis showed in **Table 5** that DialogFlow made false statements with a relatively high confidence so that the set threshold did not add any value. This behavior had already been observed by Abdellatif et al. (2022a) and can therefore be confirmed by the results of this thesis. Watson Assistant also performed better than DialogFlow in terms of latency. While milliseconds are less noticeable in text-based chatbots, delivering the answer in a visual environment like virtual reality (VR) is one of the challenges, as time lags are more perceived and the user expect real-time interactions (Hu et al., 2020). Due to these results, the second research question about the most appropriate NLU platform for the use case can also be answered. From the comparison shown in **Table 6** between Watson Assistant and DialogFlow, conclusions can be made to answer the first research question. The effort for the development of two bots can be classified as higher, whereby no added value can be generated for the use case of SGLT2 inhibitors. In all compared parameters, Watson Assistant performs better. Hence, DialogFlow would add no additional value for the presented use case. Although, the additional effort experienced in development in this thesis is not considerable, it is recommended to develop only one chatbot in the future. The reason for the minimal effort is that all intents and entities could be copied once. In operating mode, this would have to be ensured on a recurring cycle, which is the reason why additional time and effort, hence the costs, would be generated for the maintenance of two chatbots.

With current disruptive innovation within the chatbot area, it remains necessary to observe the developments around Med-PaLM2, the currently most appraised chatbot in the healthcare area. As the first LLM, Med-PaLM2 was able to perform an "expert" test-taker level performance on the MedQA dataset of US Medical Licensing Examination style questions, reaching 85%+ accuracy. Currently, MedPaLM2 is in a testing phase for selected Google Cloud customers (Gupta & Waldron, 2023). Also, since this LLM comes from Google, a later connection to DialogFlow could be much simpler than via IBM's Watson Assistant. Other large language models such as ChatGPT have currently received worldwide attention, the developed chatbots in DialogFlow and Watson Assistant do not include an LLM for response generation. The reason for this is the objective of the chatbot. As mentioned above, this thesis is about the development of two bots that are supposed to provide answers to SGLT2 inhibitors. This would also be the case with an LLM, however, there is a loss of control over the answers. Since the developed bots make recommendations to the patient and thus, influence the medical therapy, it must be ensured that the content of the answer is correct. This can be ensured in DialogFlow and Watson Assistant after checking with a physician as it was done in this thesis. Since the chatbot could be classified as medical product, this is also a criterion for approval by swissmedic which is responsible for the approval of medical products in Switzerland. The approval for medical products may change in the future with further development in industry-specific large language models like Med-PaLM2 by Google. The developments regarding LLM must be observed for the use case with patients with CKD as well as for general information on medications. The integration of an LLM could contribute to a better overall user experience in terms of scope, response variety and individuality. A reaction of NLU vendors to this latest development is necessary and will be interesting to observe in the future. Today, the individual NLU platforms are rather non-transparent regarding what is occurring in the background. It remains necessary to observe whether the strong momentum and more competition will also create more transparency. It may be suspected that companies which still do not publicly share their algorithms, might have troubles to keep up with the market leaders in the field of AI.

7.2 Quality of the artefacts

In this thesis, a great focus was laid on testing the created chatbot. As mentioned above, this was conducted in three consecutive steps. The results reported in detail about the development and testing of the artefacts. In this chapter, the aim is to assess the quality of the artefact.

The third research question was about, whether a Q&A bot could provide correct answers about SGLT2 inhibitors. Watson Assistant showed in a comparison test with DialogFlow that this was indeed the case. More than 800 training questions have been collected in this area. It was possible to cover a wide range of patient questions including topics such as application of the medication as well as side effects. In the future, extensions by testing with patients or even extensions after a possible roll-out would further increase the quality of the bot. As already explained above, this was not possible for this thesis. Nonetheless it would be crucial to further inspect the use of a chatbot with CKD patients.

Based on the questionnaires and testing with physicians, it could also be concluded that longer questions with explanations can be more difficult task for the chatbot. While longer questions were also trained, it remains difficult to reproduce answers with a high confidence for longer explanations. Since it is a question-and-answer bot in this case, it can be assumed that in the use with real CKD patients almost only short questions with little explanation will be entered. Hence, this limitation of the bot may not necessarily present a fault. According to one test person, there was some learning with the bot about how to enter questions in a way that gets the desired answer. This goes along with Zuccon & Koopman (2023) that prompt knowledge is important in receiving correct answers from chatbots. In other words, this means that patients should be instructed in how to use the chatbot in order to gain correct and reliable answers. While the acceptance of the patients is crucial for the usage in the real-life, the acceptance on the part of physicians and pharmacists was surveyed in this thesis. To measure whether the chatbot meets the acceptance criteria of healthcare professionals, a questionnaire was completed by them. With the restriction of only being able to answer questions about SGLT2 inhibitors, the chatbot was perceived as friendly and the questions were answered correctly from a professional point of view. The answers were predominantly perceived as relevant to the question asked. The answers can influence the medical therapy of patients, which is why the professional correctness of the answers is mandatory. After reviewing all answers by Prof. Dr. med. Stephan Segerer, together with the test results, the correctness of the answers can preliminarily be considered as given and the third research question was positively answered.

Another important statement made during the testing with physicians and pharmacists was that they found the project interesting and were curious about the further developments. This can be essential for the later use of such chatbots. These chatbots may only be implemented in everyday practice if physicians and pharmacists recommend them to their patients. Therefore, it is important to closely involve medical personnel for further testing. This is also why their feedback, as shown above, was of great value and included in the final development phase of the bots.

7.3 Limitations

The results of this thesis show that the use of a chatbot with CKD patients can be successful. However, it has several limitations with are as follows.

First, the thesis used two popular closed-source NLU platforms. A limitation of the used platforms is about their lack of transparency. Without notification of the users, there is the possibility of changes in the implemented algorithms. Due to this, replicating this thesis might lead to different results. This creates a challenge in verifying the thesis findings and comparing them with future research or similar studies.

A second limitation might have been the possibility of a bias in the testing set. The applied test set with 15 questions about SGLT2 inhibitors is manually labeled which can lead to human bias. The testing by medical personnel served to check the answers and to ask further questions (intents) to the bot. Since the testing only took place with DialogFlow, the Watson Assistant was not directly reviewed by the participants. From the authors viewpoint, this might not have impacted the results because lastly, both chatbots contained the same intents. Nonetheless it should be mentioned as a possible limitation and can be pursuit in future works. Based on the observed results regarding correct answers in chapter 6.3, testing in the RealCo project of Watson Assistant is suggested.

A further limitation is the unbalanced dataset. In order to have a balanced dataset, all intents should have roughly the same number of training queries. Further, NLU platforms recommend a minimum number of training queries. For example, Watson Assistant recommends at least 5 questions per intent. While this requirement is met in the SGLT 2 bot, the dataset is not considered very balanced. One reason for this is due to the fact that certain intents were addressed more frequently than others during the training with the physicians. Furthermore, there are certain intents that have a higher importance, since they refer to a physician, for example, or have longer answers with a lot of general information about SGLT2 inhibitors.

7.4 Next steps and future work

Having successfully developed and trained two chatbots with an extensive dataset of over 800 queries in this thesis, this chapter explains where future work can start to improve the existing chatbots.

The chatbot developed in this study was meant to answer questions regarding SGLT2 inhibitors only but still contained 40 intents. This is still an easily manageable number of intents. The larger the chatbot becomes, the more intents and entities have to be managed manually. Also, the administration of the responses must be kept up to date manually. With rapidly changing conditions, a fast response time is required so that the bot remains usable. Future research could investigate how to ensure maintenance and automatic expansion of the chatbot based on user input. Neither Watson Assistant nor DialogFlow offer automatic self-learning of the chatbot during operation. The extension of training queries in the intents as well as the creation of new intents has to be done manually. However, both platforms facilitate manual enhancements by storing all user inputs and suggesting intents. These suggestions can then be accepted or, if misclassified, easily assigned to the correct intent. An automation of the self-learning would have to be done by a separate customized solution. Due to the importance of maintaining the control of the answers, it is recommended to use the available tools of the NLU platforms for the next development steps.

Due to the circumstance that the testing with DialogFlow was performed, but the comparison afterwards revealed Watson Assistant as the preferred solution, testing with Watson Assistant should be aimed for in the next phase. Such a testing should also be conducted using virtual reality glasses in order to use the separate speech-to-text software and also evaluate it.

As mentioned in the beginning of this thesis, chatbots may present an opportunity to enhance health literacy and medication compliance. Since this was not part of this study, this could be addressed in future research. Doing this, future studies should also measure the health literacy of patients with CKD before and after the use of the chatbot in order to find out whether a real increase in health literacy and medication compliance can be observed. Depending on the results, this can have a great impact in the further use of chatbots in the healthcare industry.

The issues of privacy and ethical considerations are not elaborated in this research. In future research, the subject of data protection in Switzerland and the handling of large language models in medicine must be investigated. In this context, the laws regarding the approval of medical products will have to be considered as well.

8 Conclusion

In conclusion, a first prototype of two chatbots were developed using NLU platforms DialogFlow and Watson Assistant to reply to questions about SGLT2 inhibitors. The two chatbots were also verified and tested by the medical staff at the cantonal hospital in Aarau. The answers of the chatbots are professionally accurate and could contribute to the improvement health literacy for patients with chronic kidney disease in the future. Furthermore, the comparison of the two most popular NLU platforms reveals differences in response latency and accuracy, despite the same training queries as well as intents and entities. This helps to better compare NLU platforms, as there is a lack of transparency regarding the algorithms applied. However, it should also be noted that the design of chatbots and the goal pursued have an impact.

Due to newly available API for large language models, future research should investigate how industry-specific large language models influence chatbots and to what extent control over the responses is necessary or control is handed over to the AI. Further, testing with patients combined with VR-glasses are crucial to succeed.

9 References

This reference list was created with Zotero.

Abdellatif, A., Badran, K., Costa, D. E., & Shihab, E. (2022a). A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering. *IEEE Transactions on Software Engineering*, *48*(8), 3087–3102. https://doi.org/10.1109/TSE.2021.3078384

Abdellatif, A., Badran, K., Costa, D. E., & Shihab, E. (2022b). A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering. *IEEE Transactions on Software Engineering*, *48*(8), 3087–3102. https://doi.org/10.1109/TSE.2021.3078384

Adamopoulou, E., & Moussiades, L. (2020a). An Overview of Chatbot Technology. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Hrsg.), *Artificial Intelligence Applications and Innovations* (S. 373–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-49186-4_31

Adamopoulou, E., & Moussiades, L. (2020b). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. https://doi.org/10.18653/v1/N19-4010

Anastasiadou, M., Alexiadis, A., Polychronidou, E., Votis, K., & Tzovaras, D. (2020). A prototype educational virtual assistant for diabetes management. *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 999–1004. https://doi.org/10.1109/BIBE50027.2020.00169

Asahara, M., & Matsumoto, Y. (2003). Japanese Named Entity Extraction with Redundant Morphological Analysis. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 8–15. https://aclanthology.org/N03-1002

Assayed, S., Shaalan, K., & Alkhatib, M. (2022). *A Chatbot Intent Classifier for Supporting High School Students* (SSRN Scholarly Paper Nr. 4397536). https://papers.ssrn.com/abstract=4397536

Aula, A., & Käki, M. (2005). Less is more in Web search interfaces for older adults. *First Monday*. https://doi.org/10.5210/fm.v10i7.1254

Bibault, J.-E., Chaix, B., Guillemassé, A., Cousin, S., Escande, A., Perrin, M., Pienkowski, A., Delamon, G., Nectoux, P., & Brouard, B. (2019). A Chatbot Versus Physicians to Provide Information for Patients With Breast Cancer: Blind, Randomized Controlled Noninferiority Trial. *Journal of Medical Internet Research*, *21*(11), e15787. https://doi.org/10.2196/15787

Bickmore, T. W., Utami, D., Matsuyama, R., & Paasche-Orlow, M. K. (2016). Improving Access to Online Health Information With Conversational Agents: A Randomized Controlled Experiment. *Journal of Medical Internet Research*, *18*(1), e5239. https://doi.org/10.2196/jmir.5239

Bikel, D., Schwartz, R., & Weischedel, R. (1999). An Algorithm that Learns What's in a Name. *Machine Learning*, *34*. https://doi.org/10.1023/A:1007558221122

Biro, J., Linder, C., & Neyens, D. (2023). The Effects of a Health Care Chatbot's Complexity and Persona on User Trust, Perceived Usability, and Effectiveness: Mixed Methods Study. *JMIR Human Factors*, *10*(1), e41017. https://doi.org/10.2196/41017

Boren, S. A. (2009). A Review of Health Literacy and Diabetes: Opportunities for Technology. *Journal of Diabetes Science and Technology*, *3*(1), 202–209. https://doi.org/10.1177/193229680900300124

Boroghina, G., Corlatescu, D. G., & Dascalu, M. (2022). Multi-Microworld Conversational Agent with RDF Knowledge Graph Integration. *Information*, *13*(11), Article 11. https://doi.org/10.3390/info13110539

Botium Box—The chatbot testing tool. (o. J.). Botium. Abgerufen 18. Dezember 2022, von https://www.botium.ai/

Braun, D., Hernandez Mendez, A., Matthes, F., & Langen, M. (2017). Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 174–185. https://doi.org/10.18653/v1/W17-5522

Brucker-Kley, E., Kleinberger, U., Keller, T., Christen, J., Keller-Senn, A., & Koppitz, A. (2021). Identifying research gaps: A review of virtual patient education and self-management. *Technology and Health Care*, *29*(6), 1057–1069. https://doi.org/10.3233/THC-202665

Buck, C., Doctor, E., Hennrich, J., Jöhnk, J., & Eymann, T. (2022). General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study. *Journal of Medical Internet Research*, *24*(1), e28916. https://doi.org/10.2196/28916

Buehrig, K., Fienemann, J., & Schlickau, S. (2020). *On Certain Characteristics of 'Diabetes Consultations'* (S. 65–82). https://doi.org/10.1007/978-3-658-27120-6_4

Canonico, M., & Russis, L. D. (2018). A Comparison and Critique of Natural Language Understanding Tools. *CLOUD COMPUTING*, 7.

Carstensen, K.-U. (2010). Anwendungen. In K.-U. Carstensen, C. Ebert, C. Ebert, S. J. Jekat, R. Klabunde, & H. Langer (Hrsg.), *Computerlinguistik und*

Sprachtechnologie: Eine Einführung (S. 553–658). Spektrum Akademischer Verlag. https://doi.org/10.1007/978-3-8274-2224-8_5

Casas, J., Tricot, M.-O., Abou Khaled, O., Mugellini, E., & Cudré-Mauroux, P. (2020). Trends & Methods in Chatbot Evaluation. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 280–286. https://doi.org/10.1145/3395035.3425319

Chen, X., Hay, J. L., Waters, E. A., Kiviniemi, M. T., Biddle, C., Schofield, E., Li, Y., Kaphingst, K., & Orom, H. (2018). Health Literacy and Use and Trust in Health Information. *Journal of Health Communication*, *23*(8), 724–734. https://doi.org/10.1080/10810730.2018.1511658

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, *9*(1), 10. https://doi.org/10.1186/s40537-022-00561-y

Crutzen, R., Peters, G.-J. Y., Portugal, S. D., Fisser, E. M., & Grolleman, J. J. (2011). An Artificially Intelligent Chat Agent That Answers Adolescents' Questions Related to Sex, Drugs, and Alcohol: An Exploratory Study. *Journal of Adolescent Health*, *48*(5), 514–519. https://doi.org/10.1016/j.jadohealth.2010.09.002

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A Practical Part-of-Speech Tagger. *Third Conference on Applied Natural Language Processing*, 133–140. https://doi.org/10.3115/974499.974523

De Spindler, A. (2022, März 8). *Applied data science—Graphen*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dialogflow | Google Cloud. (o. J.). Abgerufen 17. Dezember 2022, von https://cloud.google.com/dialogflow?hl=de

Drozdal, J., Chang, A., Fahey, W., Murthy, N., Mogilisetty, L., Sunray, J., Powell, C., & Su, H. (2021). The Design and Evaluation of a Chatbot for Human Resources. In C. Stephanidis, M. Antona, & S. Ntoa (Hrsg.), *HCI International 2021—Late Breaking Posters* (S. 239–248). Springer International Publishing. https://doi.org/10.1007/978-3-030-90176-9_32

Eiselen, R., & Bukula, A. (2022). IsiXhosa Named Entity Recognition Resources. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(2), 35:1-35:19. https://doi.org/10.1145/3531478

Fadhil, A., AbuRa'ed, A., & Information & Communication Technologies, Universitat Pompeu Fabra Barcelona, Spain. (2019). OlloBot - Towards A Text-Based Arabic Health Conversational Agent: Evaluation and Results. *Proceedings - Natural Language Processing in a Deep Learning World*, 295–303. https://doi.org/10.26615/978-954-452-056-4_034

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, *4*(2), e19. https://doi.org/10.2196/mental.7785

Forni Ogna, V., Ogna, A., Ponte, B., Gabutti, L., Binet, I., Conen, D., Erne, P., Gallino, A., Guessous, I., Hayoz, D., Muggli, F., Paccaud, F., Péchère-Bertchi, A., Suter, P. M., Bochud, M., & Burnier, M. (2016). Prevalence and determinants of chronic kidney disease in the Swiss population. *Swiss Medical Weekly*, *146*(1718), Article 1718. https://doi.org/10.4414/smw.2016.14313

Gavagnin, E. (2022, April 5). *Applied data science—Natural language processing*.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical

Education and Knowledge Assessment. *JMIR Medical Education*, *9*, e45312. https://doi.org/10.2196/45312

Gossen, T., Nitsche, M., & Nürnberger, A. (2012). Knowledge journey: A web search interface for young users. *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, 1–10. https://doi.org/10.1145/2391224.2391225

Great Learning Team. (2022, Oktober 24). An Introduction to Bag of Words (BoW) | What is Bag of Words? *Great Learning Blog: Free Resources What Matters to Shape Your Career!* https://www.mygreatlearning.com/blog/bag-of-words/

Gregori, E. (2017). *Evaluation of Modern Tools for an OMSCS Advisor Chatbot*. https://smartech.gatech.edu/handle/1853/58516

Guerino, G. C., & Valentim, N. M. C. (2020). Usability and User eXperience Evaluation of Conversational Systems: A Systematic Mapping Study. *Proceedings of the XXXIV Brazilian Symposium on Software Engineering*, 427–436. https://doi.org/10.1145/3422392.3422421

Gupta, A., & Waldron, A. (2023, April 13). Sharing Google's Med-PaLM 2 medical large language model, or LLM. *Google Cloud Blog*. https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model

Heerspink, H. J. L., Stefánsson, B. V., Correa-Rotter, R., Chertow, G. M., Greene, T., Hou, F.-F., Mann, J. F. E., McMurray, J. J. V., Lindberg, M., Rossing, P., Sjöström, C. D., Toto, R. D., Langkilde, A.-M., & Wheeler, D. C. (2020). Dapagliflozin in Patients with Chronic Kidney Disease. *New England Journal of Medicine*, *383*(15), 1436–1446. https://doi.org/10.1056/NEJMoa2024816

Helmi Setyawan, M. Y., Awangga, R. M., & Efendi, S. R. (2018). Comparison Of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent

Classification In Chatbot. *2018 International Conference on Applied Engineering (ICAE)*, 1–5. https://doi.org/10.1109/INCAE.2018.8579372

Hien, H. T., Cuong, P.-N., Nam, L. N. H., Nhung, H. L. T. K., & Thang, L. D. (2018). Intelligent Assistants in Higher-Education Environments: The FIT-EBot, a Chatbot for Administrative and Learning Support. *Proceedings of the Ninth International Symposium on Information and Communication Technology - SoICT 2018*, 69–76. https://doi.org/10.1145/3287921.3287937

Horne, R., Weinman, J., Barber, N., Elliott, R., Morgan, M., Cribb, A., & Kellar, I. (2005). *Concordance, Adherence and Compliance in Medicine Taking*.

Hu, F., Deng, Y., Saad, W., Bennis, M., & Aghvami, A. H. (2020). Cellular-Connected Wireless Virtual Reality: Requirements, Challenges, and Solutions. *IEEE Communications Magazine*, *58*(5), 105–111. https://doi.org/10.1109/MCOM.001.1900511

Kadariya, D., Venkataramanan, R., Yip, H. Y., Kalra, M., Thirunarayanan, K., & Sheth, A. (2019). kBot: Knowledge-Enabled Personalized Chatbot for Asthma Self-Management. *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, 138–143. https://doi.org/10.1109/SMARTCOMP.2019.00043

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13428-4

Khyani, D., Siddhartha, Niveditha, & Divya. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology*, *22*, 350–357.

Kim, J., Lee, H.-G., Kim, H., Lee, Y., & Kim, Y.-G. (2018). Two-Step Training and Mixed Encoding-Decoding for Implementing a Generative Chatbot with a Small

Dialogue Corpus. *Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*, 31–35. https://doi.org/10.18653/v1/W18-6707

Koetter, F., Blohm, M., Kochanowski, M., Goetzer, J., Graziotin, D., & Wagner, S. (2019). Motivations, Classification and Model Trial of Conversational Agents for Insurance Companies. *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, 19–30. https://doi.org/10.5220/0007252100190030

Koman, J., Fauvelle, K., Schuck, S., Texier, N., & Mebarki, A. (2020). Physicians' Perceptions of the Use of a Chatbot for Information Seeking: Qualitative Study. *Journal of Medical Internet Research*, *22*(11), e15185. https://doi.org/10.2196/15185

Lang, N. (2022, Oktober 24). *Stemming vs. Lemmatization in NLP*. Medium. https://towardsdatascience.com/stemming-vs-lemmatization-in-nlp-dea008600a0

Lee, S.-W., Kim, N., Kwon, J.-H., Choi, H. D., Lee, S.-B., & Kim, E.-J. (2022). Comparative Study of Word Embeddings for Classification of Scientific Article on Human Health Risk of Electromagnetic Fields. *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)*, 391–392. https://doi.org/10.1109/GCCE56475.2022.10014242

Lehto, T., Oinas-Kukkonen, H., Pätiälä, T., & Saarelma, O. (2012). CONSUMERS' PERCEPTIONS OF A VIRTUAL HEALTH CHECK: AN EMPIRICAL INVESTIGATION. *ECIS 2012 Proceedings*. https://aisel.aisnet.org/ecis2012/154

Maddox, T., Chmielewski, C., & Fitzpatrick, T. (2022). Virtual Reality in Chronic Kidney Disease Education and Training. *Nephrology Nursing Journal: Journal of the American Nephrology Nurses' Association*, *49*(4), 329–381.

Maher, S., Kayte, S., & Nimbhore, S. (2020). Chatbots & Its Techniques using AI: A Review. *International Journal for Research in Applied Science and Engineering Technology*, *8*(12), 503–508. https://doi.org/10.22214/ijraset.2020.32537

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019). A Survey on Evaluation Methods for Chatbots. *Proceedings of the 2019 7th International Conference on Information and Education Technology*, 111–119. https://doi.org/10.1145/3323771.3323824

Martin, P.-Y. (2017). Fortschritte in der Nephrologie: Vorteilhaft für Patienten und Kosten. *Schweizerische Ärztezeitung*, *98*(45), 1484–1486. https://doi.org/10.4414/saez.2017.06160

Mash, R., Schouw, D., & Fischer, A. E. (2022). Evaluating the Implementation of the GREAT4Diabetes WhatsApp Chatbot to Educate People With Type 2 Diabetes During the COVID-19 Pandemic: Convergent Mixed Methods Study. *JMIR Diabetes*, *7*(2), e37882. https://doi.org/10.2196/37882

McCallum, A., & Li, W. (2003). Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 188–191. https://aclanthology.org/W03-0430

Mokmin, N. A. M., & Ibrahim, N. A. (2021). The evaluation of chatbot as a tool for health literacy education among undergraduate students. *Education and Information Technologies*, *26*(5), 6033–6049. https://doi.org/10.1007/s10639-021-10542-y

Motger, Q., Franch, X., & Marco, J. (2022). Software-Based Dialogue Systems: Survey, Taxonomy, and Challenges. *ACM Computing Surveys*, *55*(5), 91:1-91:42. https://doi.org/10.1145/3527450

Mu, X., Shen, X., & Kirby, J. (2017). Support vector machine classifier based on approximate entropy metric for chatbot text-based communication. *International Journal of Artificial Intelligence*, *15*, 1–16.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists* (First edition). O'Reilly Media, Inc.

Muñoz, S., Araque, O., Llamas, A. F., & Iglesias, C. A. (2018). A Cognitive Agent for Mining Bugs Reports, Feature Suggestions and Sentiment in a Mobile Application Store. *2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data)*, 17–24. https://doi.org/10.1109/Innovate-Data.2018.00010

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *DIGITAL HEALTH*, *5*, 2055207619871808. https://doi.org/10.1177/2055207619871808

Nakhleh, A., Spitzer, S., & Shehadeh, N. (2023). ChatGPT's Response to the Diabetes Knowledge Questionnaire: Implications for Diabetes Education. *Diabetes Technology & Therapeutics*. https://doi.org/10.1089/dia.2023.0134

Nikitina, S., Callaioli, S., & Baez, M. (2018). Smart conversational agents for reminiscence. *Proceedings of the 1st International Workshop on Software Engineering for Cognitive Services*, 52–57. https://doi.org/10.1145/3195555.3195567

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on Medical Challenge Problems* (arXiv:2303.13375). arXiv. https://doi.org/10.48550/arXiv.2303.13375

Nov, O., Singh, N., & Mann, D. M. (2023). *Putting ChatGPT's Medical Advice to the (Turing) Test* (S. 2023.01.23.23284735). medRxiv. https://doi.org/10.1101/2023.01.23.23284735

Pai, A. (2020, Mai 25). What is Tokenization in NLP? Here's All You Need To Know. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/

Perkovic, V., Jardine, M. J., Neal, B., Bompoint, S., Heerspink, H. J. L., Charytan, D. M., Edwards, R., Agarwal, R., Bakris, G., Bull, S., Cannon, C. P., Capuano, G., Chu, P.-L., de Zeeuw, D., Greene, T., Levin, A., Pollock, C., Wheeler, D. C., Yavin, Y., … Mahaffey, K. W. (2019). Canagliflozin and Renal Outcomes in Type 2 Diabetes and Nephropathy. *New England Journal of Medicine*, *380*(24), 2295–2306. https://doi.org/10.1056/NEJMoa1811744

Ramesh, K., Ravishankaran, S., Joshi, A., & Chandrasekaran, K. (2017). A Survey of Design Techniques for Conversational Agents. In S. Kaushik, D. Gupta, L. Kharb, & D. Chahal (Hrsg.), *Information, Communication and Computing Technology* (S. 336–350). Springer. https://doi.org/10.1007/978-981-10-6544-6_31

Ramos, R. (2017, Februar 3). Screw the Turing test—Chatbots don't need to act human. *VentureBeat*. https://venturebeat.com/ai/screw-the-turing-test-chatbots-dont-need-to-act-human/

Reed, T. R., & Dubuf, J. M. H. (1993). A Review of Recent Texture Segmentation and Feature Extraction Techniques. *CVGIP: Image Understanding*, *57*(3), 359–372. https://doi.org/10.1006/ciun.1993.1024

Reis, L., Maier, C., Mattke, J., & Weitzel, T. (2020). Chatbots in healthcare: Status quo, application scenarios for physicians and patients and future directions. *ECIS 2020 Research Papers*. https://aisel.aisnet.org/ecis2020_rp/163

Rodríguez-Cantelar, M., D'Haro, L. F., & Matía, F. (2021). Automatic Evaluation of Non-task Oriented Dialog Systems by Using Sentence Embeddings Projections and Their Dynamics. In L. F. D'Haro, Z. Callejas, & S. Nakamura (Hrsg.), *Conversational Dialogue Systems for the Next Decade* (S. 71–84). Springer. https://doi.org/10.1007/978-981-15-8395-7_6

Rosruen, N., & Samanchuen, T. (2018). Chatbot Utilization for Medical Consultant System. *2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, 1–5. https://doi.org/10.1109/TIMES-iCON.2018.8621678

Shah, P., Shah, S., & Joshi, S. (2022). A Study of Various Word Embeddings in Deep Learning. *2022 3rd International Conference for Emerging Technology (INCET)*, 1–5. https://doi.org/10.1109/INCET54531.2022.9824963

Shah, V., & Shah, D. S. (2019). *A Comparison of Various Chatbot Frameworks*. *6*(0975), 9.

Song, Y., Yan, R., Li, X., Zhao, D., & Zhang, M. (2016). *Two are Better than One: An Ensemble of Retrieval- and Generation-Based Dialog Systems* (arXiv:1610.07149). arXiv. https://doi.org/10.48550/arXiv.1610.07149

Tesnière, L. (1959). *Élements de syntaxe structurale*. C. Klincksieck.

Thirunavukarasu, A. J., Hassan, R., Mahmood, S., Sanghera, R., Barzangi, K., Mukashfi, M. E., & Shah, S. (2023). Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care. *JMIR Medical Education*, *9*(1), e46599. https://doi.org/10.2196/46599

Thorat, S. A., & Jadhav, V. (2020). *A Review on Implementation Issues of Rule-based Chatbot Systems* (SSRN Scholarly Paper Nr. 3567047). https://doi.org/10.2139/ssrn.3567047

Toxtli, C., Monroy-Hernández, A., & Cranshaw, J. (2018). Understanding Chatbot-mediated Task Management. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3173574.3173632

Tun, T. N., & Soe, K. M. (2020). Intent Classification on Myanmar Social Media Data in Telecommunication Domain Using Convolutional Neural Network and Word2Vec. *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 1–5. https://doi.org/10.1109/O-COCOSDA50338.2020.9295031

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

van Bussel, M. J. P., Odekerken–Schröder, G. J., Ou, C., Swart, R. R., & Jacobs, M. J. G. (2022). Analyzing the determinants to accept a virtual assistant and use cases among cancer patients: A mixed methods study. *BMC Health Services Research*, *22*(1), 890. https://doi.org/10.1186/s12913-022-08189-7

van Heerden, A., Ntinga, X., & Vilakazi, K. (2017). The potential of conversational agents to provide a rapid HIV counseling and testing services. *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, 80–85. https://doi.org/10.1109/FADS.2017.8253198

Vasileiou, M. V., & Maglogiannis, I. G. (2022). The Health ChatBots in Telemedicine: Intelligent Dialog System for Remote Support. *Journal of Healthcare Engineering*, *2022*, e4876512. https://doi.org/10.1155/2022/4876512

Vernon, J., Trujillo, A., Rosenbaum, S., & DeBuono, B. (2007). Low Health Literacy: Implications for National Health Policy. *Health Policy and Management Faculty Publications*. https://hsrc.himmelfarb.gwu.edu/sphhs_policy_facpubs/172

Watson Assistant—Einführung. (o. J.). Abgerufen 17. Dezember 2022, von https://cloud.ibm.com/docs/cloud.ibm.com/docs/assistant

Weizenbaum, J. (1966). *ELIZA—a computer program for the study of natural language communication between man and machine*. 10.

WHO. (2020, Dezember 9). *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Wilson, H. J., Daugherty, P. R., & Morini-Bianzino, N. (2017). The Jobs That Artificial Intelligence Will Create. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/will-ai-create-as-many-jobs-as-it-eliminates/

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2017). Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 496–505. https://doi.org/10.18653/v1/P17-1046

Zuccon, G., & Koopman, B. (2023). *Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness* (arXiv:2302.13793). arXiv. http://arxiv.org/abs/2302.13793

10 Appendix

10.1 Exports from chatbots

DialogFlow_CKD-Bot.zip

Watson_Assistant_CKD-Bot.json

10.2 Questionnaire and answers from third testing

1. **Wurden in den Antworten falsche Angaben zu Medikamenten gemacht?**
   P1: Nicht bei den Fragen, die ich gestellt habe, sofern sie richtig erkannt wurden. Teilweise musste ich die Frage nochmals umformulieren, um die gesuchte Antwort zu bekommen.

   P2: Wirkung dadurch sinkt der Blutzucker und es gehen auch Kalorien verloren (das Gewicht kann leicht abnehmen)

   Aufklärung wünschen

   P3: nein

   P4: Soweit ich das sehe, sind keine falschen Angaben enthalten. Bei spezifischerer Antwort steht drin welches Medikament von der Gruppe gemeint ist.
2. **Wie war Ihre allgemeine Erfahrung mit dem Chatbot?**
   P1&3: Gut

   P2: Besser.

   P4: Gut, die Bedienung ist einfach und die Antwort kommt schnell
3. **Hat der Chatbot Ihnen die Informationen geliefert, nach denen Sie gesucht haben?**
   P1: Grösstenteils.

   P2: Ja

   P3: Ja

   P4: Zum grössten Teil ja, es gibt Fragen, die mir der Bot nicht beantworten konnte

- Inhaltsstoffe (zB Laktose)
- Preis (zu anderen Medis als Forxiga)
- Unterschied zwischen den einzelnen SGLT-2 Hemmer
- Gibt es ein Generikum?
- Sind die Medikamente Insulinabhängig

4. **Hat der Chatbot prompt auf Ihre Anfragen geantwortet?**
   P1 & P3: ja

   P2: Ja, fand ich schon

   P4: Ja, innert Sekunden

   Auch das Mikrophon funktioniert super!

5. **Waren die Antworten des Chatbots für Ihre Fragen relevant?**

   P1: Ja, sofern die Frage richtig erkannt wurde. Bei konkreten Fragen (z.B. ich nehme jetzt ein Schmerzmedikament ein..) halt nur generell

   P2: Ja aber ich habe ja auch die Fragen formuliert gehabt.

   P3: Überwiegend

   P4: Ja, die Antworten waren relevant. Einzig die Frage mit den Inhaltstoffen führte zu keiner Antwort

6. **Fanden Sie die Persönlichkeit des Chatbots ansprechend?**

   P1: Sprachton ist sympathisch

   P2: Durchaus, manchmal stockt es innerhalb eines Wortes

   P3: Ja

   P4: Ja, er ist sehr freundlich.

7. **Gab es Antworten des Chatbots, mit denen Sie besonders nicht einverstanden waren?**

   P1: Grundsätzlich finde ich, sollte nicht nur exklusiv auf den Arzt / die Ärztin verwiesen werden. Viele Fragen kann auch der Apotheker / die Apothekerin beantworten. Diese Anlaufstelle ist niederschwelligier und braucht keinen Termin!

   P2: Auf die Frage was ist wenn ich nicht Essen kann, kommt eine Antwort zur Diät und nicht der Hinweis zum pausieren des Medikamentes

   P3: Nein

   P4: «Was wenn es nicht hilft?»

8. **Was könnte an dem Chatbot verbessert werden?**

   P1: Die Frage nach dem Einnahmezeitpunkt sollte noch mit mehr Varianten erkannt werden. Sowohl ich als auch mein Partner (wollte wissen, was ich da mache) mussten diese Frage umformulieren.

   P2: Bin Gespannt auf die 3d Variante

   P3: Das verarbeitbare Fragenspektrum erscheint weiterhin relativ klein.

   P4: Antworten zum Teil spezifisch für Forxiga und nicht SGLT-2 Hemmer insgesamt.

   (Beispiel Krankenkasse / Bezahlung)

9. **Haben Sie weitere Kommentare oder Rückmeldungen zum Chatbot?**

   P1: Interessantes Projekt

   P2: -

   P3: Nein

   P4: -