Universiteit Leiden
The Netherlands

Opleiding Informatica

Integration of the EQ5D PROM questionnaire into a natural and unobtrusive conversation using a RASA-driven chatbot

Roderick Leito (s2927195)

Supervisors:
Marco Spruit & Armel Lefebvre

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

09/09/2024

Abstract

With the growing usage of conversational agents in healthcare, there are more ways than ever to improve patient involvement and data gathering. Especially when using Patient-Reported Outcome Measures (PROMs). This thesis investigates the development and evaluation of a RASA-driven chatbot for administering the EQ5D questionnaire, With a focus on improving user experience and ease of completion. The primary objective was to develop a system that could lead users through the questionnaire in a natural and unobtrusive way, accurately assessing their current health status while maintaining a supportive interaction.

The methodology involved designing a chatbot using the RASA framework. The chatbot incorporated custom actions and used log probability thresholds to manage categorization confidence and follow-up questioning. The system architecture was carefully developed to ensure that the chatbot could dynamically adapt to all type of user inputs, using OpenAI's GPT model for accurate response categorization. A small group of users was asked to engage with the chatbot and the evaluation was conducted using both quantitative and qualitative methods: usability was measured using the System Usability Scale (SUS) survey, while deeper insights into the user experience were obtained through semi-structured interviews.

The results showed that the chatbot effectively administered the EQ5D questionnaire. The evaluation demonstrated high usability of the chatbot with an average SUS score of 80.5. The chatbot was able to handle follow-up questions with an adaptive usage of log probability thresholds. However, certain limitations were noted, like truncated responses which were sometimes caused by parameter settings. The chatbot's overall usability and ease of use were appreciated by the participants, highlighting the chatbot's potential as an automated health assessment tool.

The results highlight the potential of RASA-driven chatbots in healthcare and emphasize the value of iterative testing and user-centered design. To further increase the chatbot's effectiveness, future study should look into growing the participant pool, improving system reliability, and improving categorization techniques. The field of AI-driven healthcare solutions is expanding, and this study adds to it by laying the groundwork for conversational agents to be developed further for the purpose of administering PROMs and improving patient care.

Contents

1 Introduction

Advancements in natural language processing (NLP) and artificial intelligence (AI) have revolutionized the way people engage with technology. One significant breakthrough is the rise of chatbots which are increasingly prevalent across sectors like education, healthcare and customer service. As noted by Ting, Carin, Dzau and Wong (2020) these chatbots leverage AI systems to engage in meaningful conversations with individuals while providing assistance, guidance and information.

The integration of chatbots presents an opportunity to enhance care and streamline administrative processes within the healthcare field. Leveraging chatbots for patient reported outcome measures (PROMs) administration stands out as an aspect of healthcare that holds potential for enhancement. These measures play a role in assessing patients well being and quality of life from their perspective. The EQ-5D questionnaire, created by the EuroQol Group, is a widely used PROM to evaluate health outcomes across different populations and diseases (Devlin & Brooks, 2017). It consists of five dimensions: mobility, self-care, activities, pain/discomfort and anxiety/depression with patients rating each on a five point scale.

While valuable, the traditional way of administering this survey can be time consuming and invasive requiring resources from healthcare providers for data collection and analysis. To overcome this challenge, integrating the EQ-5D questionnaire into a natural and unobtrusive conversation using a chatbot offers a solution. This approach not only simplifies data collection but also boosts engagement through an interactive and user-friendly experience(Milne-Ives, Lam, De Cock, Van Velthoven, & Meinert, 2020).

This thesis explores the integration of the EQ-5D questionnaire into a RASA-driven chatbot, focusing on creating a seamless and conversational user interface that can effectively gather patient-reported outcomes. RASA, a framework for creating conversational AI, provides powerful tools for developing advanced chatbots that can understand and respond to natural language inputs effectively (Bocklisch, Faulkner, Pawlowski, & Nichol, 2017). This study aims to use RASA's features to create a chatbot that can smoothly guide patients through the EQ-5D questionnaire in a way that feels non-intrusive (Schröder, Nielebock, Tinschert, & Kowatsch, 2021).

The main objectives of this thesis are to:

1. Develop a RASA-driven chatbot that can administer the EQ-5D questionnaire through natural conversation.

2. Evaluate how well the chatbot collects complete patient reported outcomes.

3. Examine the user acceptance when interacting with the chatbot-administered EQ-5D questionnaire.

4. Identify any obstacles and constraints in integrating Patient Reported Outcome Measures (PROMs) into a RASA driven chatbot.

This introduction paves the way for an in-depth exploration of how AI intersects with healthcare and patient-reported outcomes. Subsequent sections will delve into the background information and previous studies on chatbot technology within healthcare settings and the importance of the EQ-5D questionnaire as past efforts to incorporate PROMs into digital platforms. By focusing on these elements the goal of this thesis is to provide valuable insights on improving healthcare services using AI powered solutions.

2 Background and related work

2.1 Patient-Reported Outcome Measures (PROMs)

Patient Reported Outcome Measures (PROMs) serve as tools for collecting information on patients' health status from the patients themselves. These measures focus on aspects of health that patients have the best knowledge of, such as symptoms, daily functioning and overall well-being. PROMs play a big role in patient-centred care by offering insights into how diseases and treatments affect patients' lives from their perspective (Black, 2013; Devlin & Appleby, 2010).

The importance of utilizing PROMs lies in several factors. Firstly they promote shared decision-making between patients and healthcare providers by incorporating the patients viewpoint into practices. Secondly, they aid in tracking disease progression and evaluating the effectiveness of treatments over time. Finally, PROMs support research by supplying data, on outcomes that can enhance healthcare services and interventions (Kingsley & Patel, 2017).

The evolution of PROMs stretches back decades. PROMSs gained prominence in the 1980s and 1990s as healthcare systems began prioritizing quality of care and patient outcomes. Initially, PROMs were utilized in trials to assess the effectiveness of new treatments. PROMs have since expanded into everyday clinical practice and assessments of healthcare quality. This growth reflects an acknowledgment of the significance of understanding patients' viewpoints on their health and treatment results (Fitzpatrick et al., 1998). Both national and international healthcare policies have now incorporated PROMs into their frameworks. In one example the UKs National Health Service (NHS) introduced PROMs for surgeries in 2009 to enhance care quality and transparency. Similarly the U.S. Centers for Medicare and Medicaid Services (CMS) included PROMs in value based payment programs to promote quality patient-centred care (Appleby & Devlin, 2010; CMS, 2020).

Among various PROMs, the EQ-5D questionnaire from the EuroQol Group is one of the most widely recognized instruments for measuring health-related quality of life. The EQ-5D questionnaire focuses on five dimensions: mobility, self-care, activities, pain/discomfort and anxiety/depression. Patients rate each dimension on a five point scale to indicate the severity of issues they face. Additionally, there is an analog scale (VAS) where patients rate their overall health on a scale from 0 to 100. The simplicity, brevity and broad application of the EQ-5D questionnaires have made it popular in both practice and research (Devlin & Brooks 2017; Janssen et al., 2013).

The integration of tools like the EQ 5D has had a significant impact on healthcare research and practice. By gathering insights from patients regarding their well-being, Patient Reported Outcome Measures (PROMs) offer a perspective on the results of treatments that surpasses traditional clinical assessments. This patient focused data is crucial for customizing healthcare services to better address patients' requirements, thereby improving the quality and efficiency of healthcare provision. As the healthcare sector progresses towards patient-centred approaches, the significance of PROMs in guiding clinical and policy decisions is expected to expand further (Kingsley & Patel 2017; Black, 2013).

2.2 Chatbots in Healthcare

The rise of intelligence (AI) and natural language processing (NLP) has made it easier to develop chatbots, which are now widely embraced across various industries, including healthcare. These chatbots utilize AI to engage in dialogues with users, offering: assistance, information and support. In the healthcare sector chatbots present an opportunity to elevate patient care standards, streamline administrative tasks and enhance overall healthcare delivery efficiency.

2.2.1 Evolution and Current Applications

The landscape of healthcare chatbots has undergone some big transformation in the past decade. Initially, these programs were just simple rule based programs capable of answering simple questions. These systems have evolved with advancements in AI and NLP. Modern chatbots can now process language effectively, enabling them to handle more complex interactions. Leveraging machine learning algorithms, these advanced chatbots continuously learn from user engagements to enhance their performance over time (Ting, Carin, Dzau, & Wong, 2020).

Current applications of healthcare chatbots span over a range of purposes such as: mental health support services, appointment scheduling, symptom checks, prescription reminders and patient education. For example symptom checker chatbots can evaluate a patients symptoms inputted by the user and offer diagnoses or guidance on seeking medical help. Mental health chatbots provide therapy (CBT) and various therapeutic approaches to assist individuals, in managing issues like anxiety and depression (Fitzpatrick, Darcy, & Vierhile, 2017). Chatbots can also efficiently handle tasks such as scheduling appointments and sending medication reminders reducing the workload for healthcare staff and aiding patients in following their treatment plans (Laranjo et al., 2018).

2.2.2 Benefits and Challenges

The advantages of incorporating chatbots into healthcare are abundant. One key benefit is the improvement in engagement. Chatbots offer a platform for patients to easily express their health concerns and receive responses. This can especially be valuable for handling conditions that require monitoring and interaction (Laranjo et al., 2018). Moreover chatbots are available round the clock providing support to patients without the constraints of human staff availability.

Another advantage is the potential cost savings. By automating tasks like appointment scheduling and medication reminders healthcare providers can allocate resources better, allowing human efforts to focus on patient care needs. Additionally chatbots can aid in detection of health issues by encouraging patients to report symptoms and seek timely medical advice, potentially decreasing the necessity for more expensive treatments later on (Topol, 2019).

Despite the benefits these chatbots offer, incorporating chatbots into healthcare also brings a variety of challenges. One major issue revolves around the accuracy and dependability of the information given by these chatbots. It is crucial to ensure that chatbots provide safe guidance since any inaccuracies could result in serious health implications. To tackle this issue it is essential for chatbots to be equipped with strong algorithms and regularly updated with the latest medical insights (Laranjo et al., 2018).

Another obstacle is establishing trust and acceptance among users. Patients may feel reluctant to disclose health details to a chatbot due to privacy concerns. Thus strict data security measures must be implemented to protect data and build trust. Moreover the design of chatbot interactions should prioritize user friendliness and empathy to enhance user experiences and promote adoption (Bickmore & Picard, 2005).

2.2.3 Examples of Healthcare Chatbots in Use

Several healthcare chatbots have been effectively implemented and are currently operational. A nice example is Babylon Health's AI powered chatbot, which offers medical advice based on users' reported symptoms. This chatbot has been integrated into the UK's National Health Service (NHS) and is extensively used for consultations and triage purposes. Another great example is Woebot, a chatbot focused on mental health that uses cognitive-behavioural therapy (CBT) to help individuals cope with mental health issues. Research has shown that Woebot is effective in reducing symptoms of anxiety and depression (Fitzpatrick, Darcy, & Vierhile, 2017).

A more recent innovation is Welzijn.AI, a conversational AI system specifically designed to monitor mental well-being, currently under review for publication (Van Dijk, Lefebvre, & Spruit, under review). Welzijn.AI is significant for being the first chatbot to conduct the EQ-5D-5L questionnaire using a large language model (LLM). This system is developed to continuously track and assess users' well-being and is being explored as a tool to support vulnerable elderly individuals in maintaining their mental health through regular conversations. (Van Dijk, Lefebvre, & Spruit, under review).

2.3 Natural Language Processing (NLP) and RASA Framework

Natural Language Processing (NLP) is a subfield of artificial intelligence. NLP focuses on the interaction between computers and humans through natural language. It involves creating algorithms and models that give machines the possibility to comprehend interpret and generate human language. Within the healthcare sector, NLP plays a significant role in analysing unstructured data like clinical notes, patient reviews and medical records to extract valuable insights to enhance patient care.

2.3.1 Overview of NLP Techniques in Healthcare

The applications of NLP techniques in healthcare are broad and diverse. One common use is extracting information from electronic health records (EHRs). NLP algorithms can pinpoint relevant information such as diagnoses, treatments and patient outcomes from text data, facilitating better data analysis and decision making processes (Wang et al., 2018). Another significant application involves developing clinical decision support systems (CDSS). Powered by NLP technology that analyses patient data and clinical guidelines, CDSS provides evidence based recommendations to healthcare professionals. This improves the precision and quality of patient care (Mehta et al., 2019).

Sentiment analysis is another NLP application that is useful in understanding feedback and experiences. By analysing reviews from patients and social media posts, healthcare providers can gather insights on satisfaction, pinpoint areas needing improvement and effectively address patient concerns (Greaves et al., 2013). Furthermore, NLP techniques play a role in creating virtual health assistants and chatbots that engage with patients in natural language, offering: information, answering questions and aiding in various tasks regarding healthcare (Topol, 2019).

2.3.2 Introduction to the RASA Framework

RASA is an open-source framework specifically designed for developing conversational AI, like chatbots that understand and respond to natural language inputs. Co-founders Alex Weidauer and Alan Nichol initially released RASA in 2016, and thanks to its adaptability and customizability, RASA gained popularity fast, becoming one of the most popular platforms for creating advanced conversational agents (RASA, 2024). RASA has become an important tool in the conversational AI space, gaining contributions from a worldwide developer community and more than 18,000 stars on GitHub (GitHub, 2024).

RASA is adaptable for worldwide deployment because it supports a variety of languages. The most frequently used language is English, but RASA's adaptable pipeline makes it simple to utilize other languages as well, such as Dutch, Spanish, German, and many more. Customizable NLP components, which may be adapted to match the requirements of certain languages and dialects, are used to enable this wide language coverage (RASA, 2024).

The RASA framework consists of three components: RASA NLU (Natural Language Understanding), RASA Core and the newly introduced RASA CALM (Conversational AI with Language Models). RASA NLU focuses on interpreting the input of users by performing tasks such as classifying intents and recognizing entities. The classification of intents involves determining the users purpose based on their input, while entity recognition detects details, like dates, locations or symptoms, within the input. Leveraging machine learning models and NLP techniques enables RASA NLU to accurately understand and process natural language inputs (Bocklisch et al., 2017).

RASA NLU gives developers flexibility in selecting NLP techniques and machine learning models, enabling them to customize the pipeline to the requirements of individual projects. Traditional classifiers, such as Support Vector Machines (SVMs) for intent classification and Conditional Random Fields (CRFs) for entity extraction, are common techniques that can be used. However, the precise configuration will rely on how the developer implements them. RASA also facilitates integration with modern transformer models, such as BERT, which improves the system's comprehension of user input's context and semantic meaning (RASA Documentation, 2024). In order to improve text interpretation, the NLU pipeline also includes crucial NLP techniques including tokenization, word embeddings (such as Word2Vec, GloVe), and Named Entity Recognition (NER). However, the precise techniques employed can be modified to match the requirements of each deployment.

RASA Core manages the dialogue management of the chatbot by using machine learning to anticipate the action based on the conversation. It involves crafting replies, asking follow-up questions and executing tasks like accessing a database or setting up appointments. The system employs policies, which are either predefined rules or machine learning models that shape how the chatbot behaves. By combining these elements RASA empowers the development of chatbots that can adapt to contexts and carry out complex dialogues (Bocklisch et al., 2017).

RASA CALM brings in capabilities for understanding advanced dialogues, making chatbots better at handling contextually aware conversations. By utilizing cutting edge language models, RASA CALM enhances the chatbots comprehension of dialogue nuances and user intents. This feature is beneficial for managing more complex conversations by keeping context and offering more relevant responses. With transformer based models, like GPT or BERT, RASA CALM can pick up on details in user inputs that traditional models might overlook. This allows the chatbot to adeptly handle interruptions, shifts in context and multi-turn dialogues for a more natural and human-like interaction (Rasa, 2021).

The incorporation of RASA CALM into the RASA framework marks a big evolution in conversational AI. RASA CALM (Conversational AI with Language Models) is a new component that has been integrated into the RASA framework. RASA CALM is designed to improve task-oriented dialogue systems by using the in-context learning abilities of large language models (LLMs). In contrast to RASA's conventional intent-based NLU approach, which depends on predefined intents and entities to control conversations, RASA CALM interprets user inputs and dynamically generates commands that are relevant to a given domain, directing the chatbot's behavior.

By addressing some of the shortcomings of the intent-based method, RASA CALM enhances RASA NLU rather than replacing it, especially when handling complex conversations with context switching, interruptions, and nuanced user inputs. While RASA NLU converts human speech into structured dialogue acts (entities and intents), RASA CALM generates instructions on its own, taking into account the entirety of the conversation history as well as the system-defined business logic. This makes it possible for RASA CALM to handle dialogues in a scalable and flexible way, which lessens the requirement for detailed intent descriptions and facilitates the management of multiple tasks.

By using RASA CALM's methodology, the chatbot can continually comprehend and carry out business logic while adjusting to each user's distinct conversational patterns. In order to build sophisticated, context-aware dialogue systems that can navigate complex user interactions, which tends to be difficult for traditional NLU-based systems, Bocklisch, Werkmeister, Varshneya, & Nichol (2024) describe how RASA CALM combines the deterministic execution of business logic with the adaptive understanding capabilities of LLMs.

2.3.3 Use Cases of RASA in Healthcare and Other Industries

The RASA framework has found its way into a lot of different healthcare settings. For example, Tia, a start-up focused on women's health, uses a RASA based chatbot to allow users to ask health related questions and receive responses while ensuring compliance with HIPAA (Health Insurance Portability and Accountability Act) regulations, which are U.S. federal standards designed to protect the privacy and security of patient's health information (Rasa, 2024). Another instance is the Dialogue Virtual Clinic, which employs a RASA powered chatbot to streamline patient intake by asking relevant questions and collecting information prior to the patients appointment (Rasa, 2024).

In the field of mental health, RASA chatbots can play a role in offering cognitive behavioral therapy (CBT) techniques to assist users in coping with conditions like anxiety and depression. These conversational agents create an easily accessible platform for users to openly discuss their mental health issues and receive support (Fitzpatrick et al., 2017). Additionally, RASA is used for purposes such as scheduling appointments, following up with patients, which helps streamline operations and lessen the burden on healthcare personnel (Ting et al., 2020).

Apart from healthcare settings RASA is being utilized across customer service, finance and retail sectors. In customer service, RASA chatbots handle inquiries, give product information and resolve issues, leading to customer satisfaction and operational efficiency. In finance these chatbots assist with tasks like account management, fraud detection and financial planning. In the retail sector, chatbots assist customers by suggesting products, tracking orders and providing personalized shopping experiences (Mehta et al., 2019).

2.3.4 Using OpenAI API for NLP

The chatbot's conversational skills can be improved by integrating additional services like OpenAI in addition to RASA's built-in NLP capabilities. Advanced natural language understanding is provided by OpenAI's language models, such as GPT-3 and GPT-4. These models can be modified with key parameters like temperature and log probability.

The model's response randomness is controlled by the temperature parameter. More predictable and focused outputs are produced at lower temperatures (e.g., 0.1), which is perfect in situations when accuracy is crucial. On the other hand, the model responds more variably at higher temperatures (e.g., 0.7), which makes it better suited for handling creative or open-ended interactions. It is crucial to carefully choose the temperature setting depending on the context of the conversation, as studies like Holtzman et al. (2020) demonstrate that raising the temperature can result in more diverse but occasionally less coherent responses.

Another important parameter is log probability, which shows how confident the model is in producing particular tokens in a response. Greater confidence is indicated by a higher log probability, whereas uncertainty certainty by examining these values. In order to get more information, follow-up questions may be prompted when the model's confidence is low. This is a crucial part of fine-tuning conversational agents because, as Brown et al. (2020) showed, language models with higher log probability typically provide more accurate and contextually appropriate responses.

Chatbots that use OpenAI models are able to strike a compromise between response accuracy and natural conversation flow by utilizing temperature and log probability parameters. These settings help the chatbot deal with ambiguity more skillfully, especially in delicate use cases like healthcare where response quality and dependability are critical.

2.4 Integration of PROMs into Digital Platforms and Related Work

The incorporation of Patient-Reported Outcome Measures (PROM), into digital platforms has transformed how patient information is gathered, analysed and applied in healthcare. By using technology, healthcare providers can simplify the process of collecting patient-reported data, boost engagement and enhance the quality of care. This section delves into the options for collecting PROMs case studies involving chatbot operated surveys and comparisons between digital and traditional methods.

2.4.1 Existing Digital Solutions for PROMs Collection

Digital platforms for gathering PROMs consist of tools like surveys, mobile apps and systems integrated with electronic health records (EHR). These platforms offer benefits over paper-based approaches, such as capturing data in real time ensuring improved data precision and enabling continuous monitoring of patient progress (Rothman et al., 2015).

One notable example is the PROMIS (Patient Reported Outcomes Measurement Information System) program, that provides a set of measures for patient-reported health status for physical, mental and social well-being. PROMIS tools are accessible through digital interfaces to seamlessly integrate them into clinical practices (Cella et al., 2007).

The use of mobile health (mHealth) apps is also on the rise for collecting PROMs. These applications enable patients to fill out questionnaires, at their convenience leading to increased response rates and patient satisfaction. For example apps like myPROM and mHealth-PROMs have shown success in enhancing engagement and the quality of data in different clinical settings (Deering et al., 2017).

2.4.2 Case Studies on Chatbot-Administered Questionnaires

Utilizing chatbots for conducting PROMs is an innovative way to boost interaction and data gathering. Chatbots, powered by AI and NLP technologies, can engage patients in conversations making the process of completing PROMs more interactive and less burdensome.

A noteworthy case study, involves a chatbot named Florence, designed to support patients in managing their health. Florence can administer PROMs linked to chronic disease management, reminding patients about their assessments and offering feedback based on their responses. This chatbot has proven effective in increasing patient adherence to PROM schedules and enhancing the patient experience (Piao et al., 2020).

Another case is the Woebot chatbot that delivers behavioral therapy techniques to individuals experiencing symptoms of depression and anxiety. Woebot collects PROMs to monitor users mental health progress over time illustrating how chatbots can be effectively utilized for health monitoring and assistance (Fitzpatrick, Darcy, & Vierhile, 2017).

2.4.3 Comparative Studies: Digital vs. Traditional Methods

Studies comparing traditional methods for collecting patient reported outcome measures (PROMs) have indicated that digital techniques tend to outperform paper based methods in terms of efficiency, accuracy of data and patient satisfaction. According to a study conducted by Bennett et al. (2012), electronic PROMs collection reduced errors in data entry, Enhanced the completeness of data when compared to traditional paper surveys. Moreover digital platforms enable real time analysis and reporting of data, empowering healthcare providers to make more informed decisions. For instance, research by Lavallee et al. (2016) showcased that integrating PROMs into Electronic Health Records (EHRs) improved the tracking of patient outcomes and facilitated more personalized care planning.

Despite the benefits offered by methods, certain challenges remain with the implementation of PROMs. These challenges include ensuring the security and privacy of data, overcoming barriers related to patients proficiency with digital tools and integrating digital systems into existing healthcare infrastructures (Jongsma et al., 2020). Addressing these hurdles is crucial, for promoting acceptance and successful utilization of PROMs platforms.

3 Methodology

3.1 Research Design

The purpose of this research is to develop and optimize a RASA-driven chatbot designed to administer the EQ5D PROM questionnaire, with the aim of enhancing user experience and ease of completion. The focus is on incorporating Natural Language Processing (NLP) features into the chatbot system with an emphasis on improving the quality of interactions and the reliability of the survey process.

The research is guided by multiple key objectives. The initial goal is to design and implement a chatbot using the RASA framework that can effectively carry out the EQ5D PROM questionnaire.

The next objective involves refining the chatbot's performance to make conversations feel more natural and unobtrusive. This refinement will concentrate on improving the chatbot's ability to not only ask the questions of the questionnaire but to also respond empathetically, engage users in the conversation and ask relevant follow up questions. The final objective is to evaluate user satisfaction with the chatbot, regarding ease of use and overall experience, following completion of the optimization process.

The research is based on three hypotheses. Firstly it is predicted that a RASA driven chatbot can accurately conduct the EQ5D PROM survey with high reliability. Secondly, it is believed that users will perceive the chatbot led questionnaire as more user friendly and engaging compared to traditional methods. Thirdly, it is hypothesized that implementing optimization techniques such as improving intent recognition, improving dialogue management, implementing follow-up questions and incorporating user feedback, will significantly improve the chatbots performance and user experience.

The research methodology follows a mixed methods approach by combining qualitative and quantitative data collection and also using analysis methods.

Before data collection begins, the optimization phase of the chatbot will occur. The optimization phase focuses on improving the naturalness and unobtrusiveness of the interactions. This includes programming the chatbot to respond empathetically, engage users in conversations and ask follow-up questions. The aim is to ensure that the chatbot offers an interactive experience while accurately administering the EQ5D PROM questionnaire.

Following the optimization phase, quantitative data collection will be conducted using the System Usability Scale (SUS) to assess the chatbot's usability. After interacting with the optimized chatbot, users will complete the SUS to evaluate the usability, ease of use and overall user experience, of the chatbot with a standardized measure.

Qualitative data will be gathered by conducting in-depth interviews with a selected group of users after they have completed the SUS. These interviews will delve deeper into the user's interactions with the chatbot, offering a richer understanding of the challenges the users faced and their overall satisfaction levels. Additionally we will analyse user interactions with the chatbot to identify common issues and potential areas that can be improved.

The research focuses on optimizing the chatbot before collecting data to ensure it is user-friendly and efficient before evaluating the performance. By using both qualitative and quantitative methods we aim to evaluate the chatbots usability and its impact on the user experience. The insights gained from this study will guide future optimization strategies and enhance our understanding of chatbot usability in healthcare applications.

3.2 System Architecture

The architecture of the RASA-driven chatbot developed for conducting the EQ5D questionnaire has been designed to make users feel as if they are interacting with a human, providing a seamless, natural, and empathetic experience. Its system architecture incorporates components such as Rasa NLU, Rasa Core, the RASA CALM module and custom actions tailored for the EQ5D questionnaire. The following sections explain how these elements collaborate to achieve the chatbots goals.

The foundation of the chatbot is built on the RASA framework, which includes Rasa NLU for natural language understanding and Rasa Core for managing the dialogue flow. These components are enhanced by integrating the RASA CALM module, a crucial addition that enhances the naturalness and fluidity of the conversation. The Rasa NLU component is trained to understand user inputs in Dutch using a NLP pipeline that includes the 'LLMCommandGenerator' interfacing with the GPT-3.5-turbo model. This pipeline for extracting intents and entities from user messages is essential for accurate responses of the chatbot. For example, when a user begins the EQ5D questionnaire, the NLU component processes this to recognize intent and extract necessary entities, such as any current health conditions mentioned by the user.

Rasa Core is responsible for managing the conversation flow, ensuring it proceeds correctly. Rather than relying on the traditional stories and rules, it utilizes a more advanced flow feature introduced with RASA CALM. These flows provide a more adaptable approach to dialogue management that allows the bot to handle complex conversations effectively. These flows maintain context within a conversation, enabling the chatbot to respond appropriately even if the user deviates from the expected path.

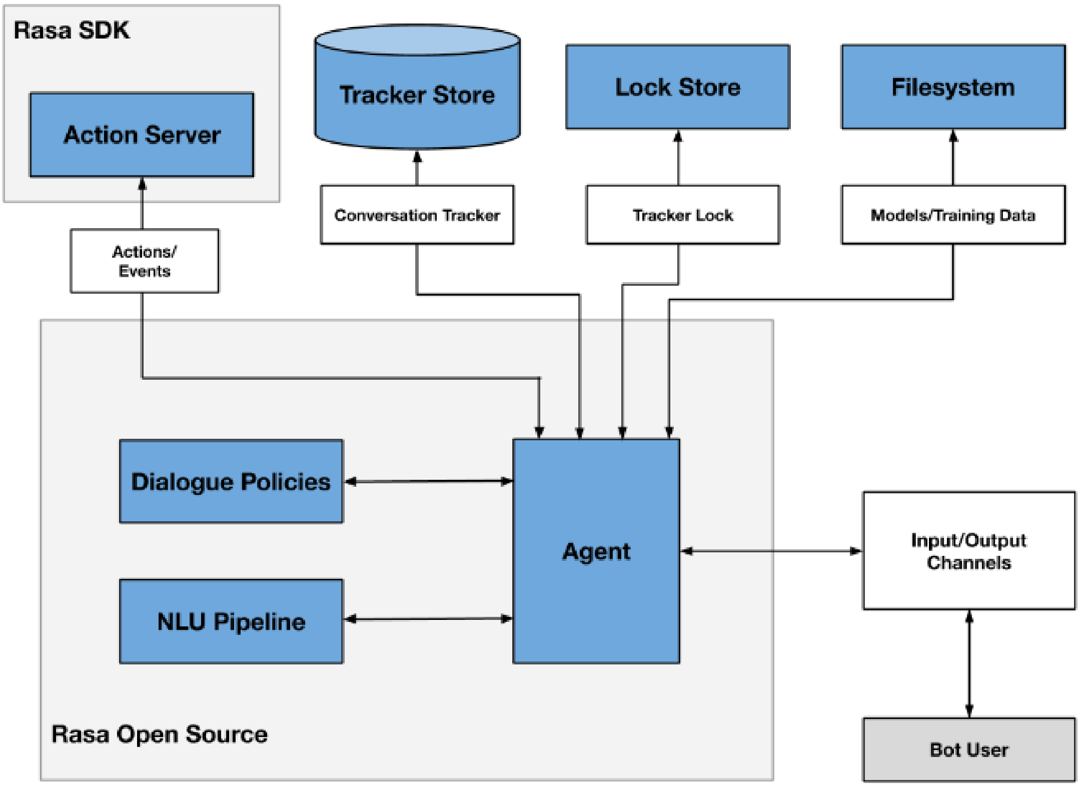


Figure 1: RASA Architecture Diagram (Source: Rasa, 2024)

Figure 1 illustrates the core components of the RASA architecture, including the NLU pipeline, Dialogue Policies, Tracker Store, and Action Server. These elements work together to process user inputs, manage conversation flow, and handle custom actions.

While the individual components of RASA CALM are not explicitly depicted in this figure, it plays a significant role in enhancing the chatbot's performance. Large language models (LLMs) are integrated by RASA CALM to enhance context preservation, enabling the chatbot to manage more complex, multi-turn conversations. RASA CALM offers the chatbot more human-like responses by enhancing the NLU pipeline and Dialogue Policies, especially when user inputs are unclear or incomplete. Although these improvements are part of the larger RASA framework, they serve as essential improvements to the existing architecture, resulting in more dynamic and responsive conversations.

The chatbot's structure, including its intents, entities, slots and actions is defined in the 'domain.yml' file. It is designed to identify user intents associated with the EQ5D questionnaire such as initiating the questionnaire or providing information about mobility or self-care. The data gathered from these interactions is stored in slots that keep track of the conversations progress. For example details regarding an user's mobility are saved in the 'mobility' slot while additional slots like 'mobility_confidence' are utilized to determine if follow-up questions are required. This framework ensures that the chatbot can engage in discussions when conversations get intricate or users provide incomplete details.

Custom actions, specified in the 'actions.py' file play a crucial role in enabling the chatbot to execute tasks related to the EQ5D questionnaire. These actions involve classifying user responses, formulating follow-up questions and ensuring that the chatbots responses are empathetic and relevant. For instance the 'ActionCategorizeMobility' custom action processes an user's reply concerning mobility related questions, updates conversation history and decides whether additional questions should be asked based on the log probability. The 'mobility_messages' slot keeps track of the series of interactions related to mobility in order to ensure that the chatbots responses remain consistent and contextually relevant.

In the chatbot, dialogue management is organized around the flow definitions outlined in the 'flows.yml' file. This file describes the progression of the EQ5D questionnaire, guiding users through each health dimension such, as mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The flow structure is designed to be adaptable allowing the chatbot to handle user inputs and scenarios while maintaining the core structure of the questionnaire. For instance if an user mentions difficulties with mobility the chatbot utilizes the 'ActionCategorizeMobility' custom action to categorize this response and determine the follow-up question if needed. This method ensures data collection provides support for users throughout their interaction with the chatbot.

The configuration for the NLP pipeline and dialogue policies is defined in the config.yml file. Within the pipeline there are components like the 'LLMCommandGenerator' that uses GPT 3.5 turbo for the language understanding. This setup is made to cater to the needs of the EQ5D questionnaire and Dutch language, guaranteeing accurate interpretation and responses by the chatbot. The dialogue policies, including the 'FlowPolicy' manage the flow of the conversation to ensure that the chatbot follows a predetermined sequence for the EQ5D questionnaire while also being adaptable to user inputs that are not expected.

The dataflow within the chatbot is designed to process user inputs and maintain context relevance throughout conversations. Whenever a user provides input it goes through the NLP pipeline for intent and entity extraction, which are then linked to slots. These slots, such as 'mobility' and 'pain messages' monitor conversation progress so that previous responses can be referenced, ensuring continuity. Custom actions are activated based on conversation status to handle data processing, update conversation logs and determine actions. This method guarantees that the chatbot can generate appropriate responses and keep the conversation going.

Upon completing the EQ5D questionnaire the chatbot has categorized all of the user's answers. This information can be stored for further analysis or integrated with other systems for comprehensive health monitoring. The integration of RASA CALM ensures that the chatbot's responses are contextually aware, contributing to a user experience that is more engaging.

3.3 Development Process

An organized and iterative procedure was used in the creation of the RASA-driven chatbot that will administer the EQ5D PROM questionnaire. The method included multiple crucial stages, beginning with the preliminary setup and configuration, followed by domain design, conversation flow development, and custom action creation, and ending with testing, optimization, and deployment.

3.3.1 Initial Setup

Setting up the RASA environment and making sure all required dependencies were installed and configured appropriately marked the beginning of the development process. This involved starting a new RASA project and putting up a virtual environment for efficient dependency management.

Setting up the ”config.yml” file, which contained the definition of the natural language processing (NLP) pipeline, was the first step. In order to improve language comprehension, this pipeline was designed to process inputs in the Dutch language and integrated the ”LLMCommandGenerator” with GPT-3.5-turbo. This configuration was necessary to make sure the chatbot could appropriately handle user input, especially in the context of the EQ5D PROM questionnaire.

3.3.2 Domain Design

The next crucial stage after the initial setup was creating and setting up the chatbot's domain. Since it establishes the core structure for handling the data during the interaction process, the domain configuration is essential.

- **Slots**: The slots required for capturing and storing user inputs connected to the EQ5D PROM questionnaire were defined in the ”domain.yml” file configuration. These slots included elements like ”mobility level,” ”self-care level,” and ”pain level,” all of which were essential for monitoring the user's reactions on various health-related areas. To evaluate whether more questions should be asked, other slots such as ”mobility confidence” were added. This helped to guarantee that user inputs were fully comprehended and precisely documented.

- **Responses**: The questions from the EQ5D PROM questionnaire made up the majority of the responses defined in the domain. The format of these answers was designed to lead respondents through every aspect of the survey, including mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. In order to provide a thorough and personalized interaction, follow-up questions were also included in the response framework. These questions were made to delve more into the user's first responses when the confidence level of the categorization was low.

- **Custom Actions**: Custom actions are defined within the domain to handle more complex tasks that go beyond predefined responses in the 'domain.yml' file. These custom actions are defined in the 'domain.yml' file, which enables the chatbot to run specific Python functions during conversations. For instance, the chatbot uses the OpenAI API to classify a user's response to a mobility question by executing the relevant custom action such as

'ActionCategorizeMobility'. Additionally, custom actions use log probability to determine the degree of confidence in this categorization and, if needed, raise a follow-up question A more intelligent and personalized interaction is ensured by the chatbot's ability to dynamically modify its responses depending on real-time analysis provided by the integration of custom actions within the domain.

3.3.3 Development of Flows

After the domain structure was established, the advanced flow-based dialogue management system from RASA CALM was used to construct the conversation flows. The sequences of interactions that the chatbot would follow were defined in the 'flows.yml' file.

The series of actions and choices the chatbot makes while interacting with the user is referred to as the "conversation path". This path consists of the main questions asked to the user, any follow-up questions that come up due to the user's answers, and any clarifications needed throughout the conversation. Depending on the user's input, the conversation path might take various routes, which makes the flow interactive and adaptive rather than rigid. Every user input has the potential to take the conversation in a slightly different path, allowing the chatbot to respond appropriately and making the exchange seem dynamic and natural. Figure 2, located on page 18, offers an additional illustration of this concept by showing the discussion flow visually.

The main questions from the EQ5D questionnaire as well as the contextually triggered follow-up questions depending on user replies were handled by the flows inside the implementation. During this phase, the main goal was to make sure the flows could handle different conversational situations, such as handling interruptions, offering clarifications, and smoothly continuing with the questionnaire.

The chatbot had the ability to make the interaction feel less scripted and more natural by utilizing RASA CALM's flow system to dynamically modify the conversation path based on the user's inputs. This adaptability was essential to preserving a supportive and engaging user experience, particularly as the chatbot led users through health evaluation offered by the EQ5D PROM questionnaire.

3.3.4 Custom Actions

The chatbot's ability to carry out particular tasks that went beyond the scope of typical dialogue flows was made possible by its custom actions. Specifically created to manage the classification of user responses across all five aspects of the EQ5D questionnaire: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. These actions were implemented in the 'actions.py' file. Every unique action was created to accurately categorize user responses and handle follow-up questions in cases where the initial classification confidence was low.

A specific custom action, such as 'ActionCategorizeMobility', 'ActionCategorizeSelfCare', and others, was made for every EQ5D dimension and question. Through real-time analysis of user input, these actions performed the categorization using the OpenAI API. For every action, a thoughtfully designed prompt was created, telling the GPT model to read the user-chatbot dialogue and categorize the response into one of the predetermined levels relevant to the specific EQ5D dimension:

- **Mobility**: Categorizes responses from ”Ik heb geen problemen met lopen” to ”Ik ben niet in staat om te lopen”
- **Self-care**: Categorizes from ”Ik heb geen problemen mezelf te wassen en aan te kleden” to ”Ik ben niet in staat om mezelf te wassen en aan te kleden”
- **Usual activities**: Ranges from ”Ik heb geen problemen met het uitvoeren van dagelijkse activiteiten” to ”Ik ben niet in staat om dagelijkse activiteiten uit te voeren”
- **Pain/discomfort**: Ranges from ”Ik heb geen pijn of ongemak” to ”Ik heb extreme pijn of ongemak”
- **Anxiety/depression**: Ranges from ”Ik ben niet angstig of somber” to ”Ik ben extreem angstig of somber”

The GPT model was specifically told by the prompts to only provide the number that corresponded to the chosen category in order to produce a normalized output from which log probabilities could be extracted.

Below are the exact prompts used for each EQ5D dimension:

Mobility Prompt:

```
Lees het volgende gesprek tussen gebruiker en chatbot:
{chat_history}

Categorizeer de mobiliteit van de gebruiker in 1 van deze categorieën:
1. Ik heb geen problemen met lopen
2. Ik heb een beetje problemen met lopen
3. Ik heb matige problemen met lopen
4. Ik heb ernstige problemen met lopen
5. Ik ben niet in staat om te lopen

Antwoord alleen met het getal van de categorie en niks anders.
```

Self-Care Prompt:

```
Lees het volgende gesprek tussen tussen gebruiker en chatbot:
{chat_history}

Categorizeer de zelfzorg van de gebruiker in 1 van deze categorieën:
1. Ik heb geen problemen mezelf te wassen en aan te kleden
2. Ik heb een beetje problemen mezelf te wassen en aan te kleden
3. Ik heb matige problemen met mezelf te wassen en aan te kleden
4. Ik heb ernstige problemen met mezelf te wassen en aan te kleden
5. Ik ben niet in staat om mezelf te wassen en aan te kleden

Antwoord alleen met het getal van de categorie en niks anders.
```

Usual Activities Prompt:

```
Lees het volgende gesprek tussen tussen gebruiker en chatbot:
{chat_history}

Categorizeer aan de hand van het gesprek de gebruiker in 1 van deze categorieën:
1. Ik heb geen problemen met het uitvoeren van dagelijkse activiteiten
2. Ik heb een beetje problemen met het uitvoeren van dagelijkse activiteiten
3. Ik heb matige problemen met het uitvoeren van dagelijkse activiteiten
4. Ik heb ernstige problemen met het uitvoeren van dagelijkse activiteiten
5. Ik ben niet in staat om dagelijkse activiteiten uit te voeren

Antwoord alleen met het getal van de categorie en niks anders.
```

Pain/Discomfort Prompt:

```
Lees het volgende gesprek tussen tussen gebruiker en chatbot:
{chat_history}

Categorizeer aan de hand van het gesprek de gebruiker in 1 van deze categorieën:
1. Ik heb geen pijn of ongemak
2. Ik heb een beetje pijn of ongemak
3. Ik heb matige pijn of ongemak
4. Ik heb ernstige pijn of ongemak
5. Ik heb extreme pijn of ongemak

Antwoord alleen met het getal van de categorie en niks anders.
```

Anxiety/Depression Prompt:

```
Lees het volgende gesprek tussen tussen gebruiker en chatbot:
{chat_history}

Categorizeer aan de hand van het gesprek de gebruiker in 1 van deze categorieën:
1. Ik ben niet angstig of somber
2. Ik ben een beetje angstig of somber
3. Ik ben matig angstig of somber
4. Ik ben ernstig angstig of somber
5. Ik ben extreem angstig of somber

Antwoord alleen met het getal van de categorie en niks anders.
```

The log probability provided a measure of the model's confidence in its categorization decision. If the log probability exceeded a predefined threshold, indicating high confidence, the chatbot confirmed the categorization and continued without further questioning.

The chatbot dynamically produced a follow-up question customized to the particular EQ5D dimension when the logprobability dropped below the threshold, signifying uncertainty. This follow-up question was created via an additional OpenAI call, in which a prompt was created to tell the model to react sympathetically, recognizing the user’s prior input and asking an appropriate question intended to provide clarity on the classification. For example, if a user reported having trouble walking but did not clearly fit into one of the predefined categories, the follow-up question would gently ask them to provide further information in order to improve the categorization.

If the log probability dropped below the threshold, the follow-up question prompt was used to ask for more information from the user to improve categorization. Below is the follow-up prompt used for the mobility dimension:

Follow-up Question Prompt (Mobility):

```
Je bent een vriendelijk en behulpzame assistent. Lees het volgende gesprek tussen
gebruiker en chatbot:
{chat_history}

Erken de laatste reactie van de gebruiker en reageer op een natuurlijke en
empatische manier. Stel aan de hand van het gevoerde gesprek
een gerichte vervolgvraag om meerinformatie over de mobiliteit
van de gebruiker te krijgen, zodat het betergecategoriseerd kan
worden in 1 van de volgende categorieën:

1. Ik heb geen problemen met lopen
2. Ik heb een beetje problemen met lopen
3. Ik heb matige problemen met lopen
4. Ik heb ernstige problemen met lopen
5. Ik ben niet in staat om te lopen

Geef antwoord zonder "Bot:" ervoor te zetten.
```

For the other EQ5D dimensions (self-care, regular activities, pain/discomfort, anxiety/depression), similar follow-up prompts were used, with the corresponding categories modified to correspond with the dimension that was being assessed. These follow-up questions enabled the chatbot to seek clarification when needed, maintaining a natural and user-friendly dialogue flow.

This strategy improved the accuracy of the data gathered while enabling the chatbot to keep an adaptive and user-centered interaction flow. It also guaranteed that the dialogue remained friendly and natural. Across all EQ5D questionnaire dimensions, the custom actions developed in ’actions.py’, successfully balanced thoroughness with user comfort by dynamically regulating when and how to seek clarification based on log probability data.

3.3.5 Personal Testing and Optimization

Once the core functionalities were in place, the development process moved to an iterative testing and refinement phase. I tested the chatbot myself throughout this step to find any performance problems and improve its responses and behavior.

- **Personal Testing**: I performed a thorough user interaction simulation to verify the chatbot's ability to manage slots, carry out customized actions, and follow the predefined flow. Through testing, it was possible to determine where the chatbot's accuracy and responsiveness needed to be improved.

During this testing phase, the effectiveness of the follow-up questions in eliciting more detailed responses was a key focus. Based on my observations, several iterations of development were conducted, leading to adjustments in the NLP pipeline configuration, custom actions, and flow structures. These refinements enhanced the chatbot's overall accuracy, responsiveness, and user experience.

To encourage reproducibility and offer complete openness of the development process, the complete implementation of the chatbot, including all code files, is available in a public GitHub repository. Please refer to Appendix B for the repository link and further instructions on accessing the source code.

The main goals of the last stage of development were optimizing the chatbot's performance and making sure it was prepared for deployment. This required streamlining the flow management to provide a seamless and coherent user experience. The chatbot's performance under real-world conditions was given particular consideration, including its capacity to manage a variety of user inputs without disrupting the flow of the conversation. The end result would be a highly functional chatbot that could efficiently administer the EQ5D PROM questionnaire while interacting with people in a kind and understanding manner. The deployment process also included setting up necessary monitoring and logging tools to track the chatbot's performance in production, allowing for ongoing improvements based on user interactions.

3.4 Testing and Evaluation

The last stage of this proof-of-concept research was the testing and evaluation phase, which was meant to evaluate the performance of the RASA-driven chatbot created for administering the EQ5D PROM questionnaire. This phase consisted of two parts: tuning the system's parameters (threshold tuning) and user evaluation.

3.4.1 Threshold Tuning for Follow-Up Questions

Before conducting the user evaluation, a testing phase was carried out to determine the optimal log probability threshold and temperature setting for the chatbot's categorization process. Inputs with varying Ambiguity Levels were used to provide a range of test cases that helped evaluate the model's performance.

- **Temperature** Various temperature settings were tested, ranging from 0.2 to 1.0, to observe how randomness in the model's responses impacted its classification performance. Lower temperatures were expected to yield more deterministic outputs, while higher temperatures introduced variability, which might affect categorization accuracy, especially when user inputs were unclear. The goal was to identify the temperature that provided the best balance between consistent classification and flexibility in uncertain situations.

- **Log Probability** Log probability was used to measure the chatbot's confidence in its classifications. The testing phase was aimed at identifying a log probability threshold below which the chatbot would trigger follow-up questions to clarify uncertain responses. By testing different temperatures and levels of input ambiguity, the system was calibrated to determine at which log probability value the follow-up question should be triggered.

- **Ambiguity Level** Inputs with varying ambiguity levels (from 1 to 5) were used to test the effects of temperature and log probability. These levels served as a way to create inputs that were progressively less clear, from unambiguous statements to highly ambiguous ones. The intention was to observe how well the chatbot handled increasingly vague or unclear inputs and to refine the log probability threshold accordingly.

The testing phase allowed for the identification of the most effective temperature setting and log probability threshold, ensuring that the system could categorize user responses accurately while triggering follow-up questions only when necessary.

3.4.2 User Evaluation

After the system was optimized through threshold tuning, a group of five users was invited to interact with the chatbot. This allowed for a hands-on evaluation of how the system handled user inputs, controlled the flow of the discussion and executed the customized actions meant to deliver sympathetic and relevant responses. The conversations were set up to mimic real-life scenarios, providing a clear view of the chatbot's functionality.

The participants provided a variety of perspectives to evaluate the system because they differed in terms of their ages, jobs, level of IT experience, and physical health. A brief description of each participant's background can be found below:

- **Participant 1**: A 25-year-old employee, who has a bachelor's degree in finance and control and is employed as a financial controller at the moment. Her limited IT experience combined with her mild knee difficulties gave insight into how the system works for non-technical users.

- **Participant 2**: A 23-year-old student at Leiden University, studying Computer Science. His input was helpful in evaluating the technical performance of the system because he was an IT-savvy participant. He did not have any health problems, and his comments focused on the chatbot's usability and dependability.

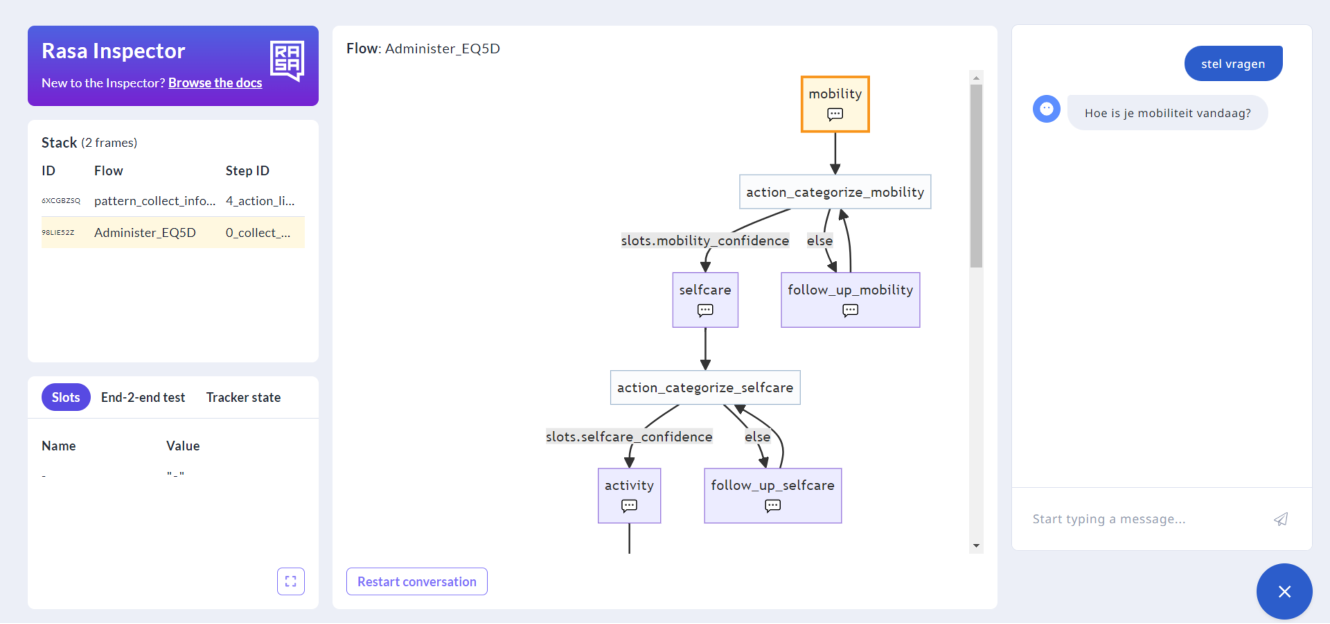
- **Participant 3**: A 48-year-old software tester with a wealth of IT knowledge. His mild back issues were pertinent to the EQ5D questionnaire's health-related inquiries. His observations were on technical correctness and the chatbot's capacity to manage inputs related to health.

- **Participant 4**: A 23-year-old student, enrolled in Hogeschool Rotterdam's second year of Industrial Engineering and Management. He had a little IT knowledge, no physical health concerns, and was able to concentrate just on the chatbot's conversation flow and user interface.

- **Participant 5**: A 23-year-old student of Chemistry at Hogeschool Rotterdam, sustained a meniscus injury while participating in football. His experience was especially relevant for evaluating how successfully the chatbot handled health-related questions, particularly those related to mobility, as a participant with a mobility impairment and less IT knowledge.

This diverse group provided valuable feedback, ranging from technical expertise to health-related concerns, ensuring that the system was tested from multiple perspectives.

RASA Inspect was a crucial debugging tool used during development to help this review, offering comprehensive visual insights into the chatbot's decision-making procedures. I was able to observe the steps that the chatbot took, including how it handled slot filling, evaluated user inputs, and initiated custom actions like follow-up questions based on log probability thresholds, thanks to RASA Inspect. This tool was quite helpful in pinpointing areas that needed to be adjusted so that the chatbot's responses matched the intended conversational design.


Rasa Inspector
New to the Inspector? Browse the docs
Slots
End-2-end test
Tracker state
Flow: Administer_EQ5D
mobility
action_categorize_mobility
slots.mobility_confidence
else
selfcare
follow_up_mobility
action_categorize_selfcare
slots.selfcare_confidence
else
activity
follow_up_selfcare
Restart conversation
Start typing a message...

Stack (2 frames)

ID	Flow	Step ID
6XCGBZSQ	pattern_collect_info...	4_action_li...
98LIE52Z	Administer_EQ5D	0_collect_...

Name	Value
-	"-"

stel vragen

Hoe is je mobiliteit vandaag?

Figure 2: Interface of RASA Inspect Command Used During Chatbot Development

Figure 2 illustrates the RASA Inspect interface used to monitor the EQ5D flow. Real-time evaluation of each step was made possible by a visual representation of the conversation flow, which highlighted the execution of custom actions, slot checks, and decision points. This degree of detail was essential for system optimization, especially when it came to adjusting the follow-up question prompts to improve the chatbot's overall effectiveness.

Participants were requested to complete the System Usability Scale (SUS) survey after interacting with the chatbot. The SUS is a well-respected instrument for evaluating system usability because it offers a consistent way to measure how users feel about a system's functionality and ease of use (Brooke, 1996). This quantitative evaluation was essential for determining how users felt about the chatbot's usability and for determining the proof-of-concept's overall success.

To obtain qualitative input, semi-structured interviews were carried out with every participant in addition to the SUS survey. The purpose of the interviews was to investigate various aspects of the user experience, including the EQ5D questions' clarity, the conversation's natural flow, the chatbot's empathy, and any difficulties that arose during the conversation. A more comprehensive assessment of the chatbot's performance was made possible by this qualitative method, which also provided insights that were not possible to obtain by quantitative measurements alone. The specific questions used in the interviews are provided in Appendix A.

The evaluation offered a thorough summary of the usability and efficiency of the chatbot by combining the quantitative data from the SUS survey with the qualitative insights from the interviews. The chatbot's response to follow-up questions, which were set off by log probability thresholds in the custom actions, received greater attention. The system's ability to fulfill the research objectives was assessed by looking at how well the chatbot handled these conversations. This observation also yielded insightful input for possible future improvements.

The testing and evaluation phase, which was the last stage of this research, was crucial to verifying the created proof of concept. The insights gained from this phase not only demonstrated the chatbot's capabilities but also offered a foundation for understanding its strengths and areas for improvement, informing any further developments or applications of the system.

4 Results

This chapter presents the results obtained from the evaluation of the RASA-driven chatbot developed for administering the EQ5D PROM questionnaire. The results are divided into two main sections: the determination of the log probability threshold for categorization decisions, and the findings from the System Usability Scale (SUS) survey and user interviews.

4.1 Log Probability Threshold for Categorization

During the development and testing of the chatbot, a subquestion emerged regarding the confidence level of the categorization made by the OpenAI model used in the chatbot's custom actions. The research specifically aimed to identify a suitable log probability threshold that, in the event that the initial categorization's degree of confidence was judged to be too low, would prompt a follow-up question. In order to address this, a number of experiments were carried out using various user inputs, looking at how modifications to the temperature parameter impacted the log probability and classification of answers related to the mobility question.

The results of these tests are summarized in Table 1, which also illustrates how various user inputs, corresponding ambiguity levels and temperature settings affected the predicted category and the log probability.

As the table shows, lower log probabilities were generally the result of higher levels of ambiguity in user inputs, suggesting a loss of confidence in the categorization. A temperature setting of 1.0 yielded a log probability of -1.434428 for the input "Lopen gaat meestal wel goed" with an ambiguity level of 4, which led to a categorization of level 3, which was more uncertain than lower temperatures and ambiguity levels. Similar to this, the system demonstrated high confidence at lower ambiguity levels for inputs aimed at Category 5 ("I am unable to walk"), such as "Ik ben niet in staat om te lopen." and "Ik kan helemaal niet meer lopen, ik gebruik een rolstoel." These inputs frequently produced high log probabilities, indicating a high degree of confidence in the Category 5 classification.

However, as ambiguity increased for Category 5, log probabilities decreased. At a temperature of 1.0, for instance, "Het is vrijwel onmogelijk voor mij om te lopen, ook met hulp" had a log

Table 1: Log Probability Results for Mobility Question at Different Temperatures

User Input	**Temperature**	**Ambiguity Level (1-5)**	**Log Probability**	**Predicted Category**
”Ik heb geen problemen met lopen”	0.2	1	-0.000093	1
”Ik loop prima zonder problemen”	0.2	2	-0.000216	1
”Geen problemen met lopen, denk ik”	0.2	3	-0.003571	1
”Lopen gaat meestal wel goed”	0.2	4	-0.168717	2
”Soms een beetje moeite met lopen”	0.2	5	-0.041283	2
”Ik heb geen problemen met lopen”	0.5	1	-0.000122	1
”Ik loop prima zonder problemen”	0.5	2	-0.000259	1
”Geen problemen met lopen, denk ik”	0.5	3	-0.001349	1
”Lopen gaat meestal wel goed”	0.5	4	-0.168717	2
”Soms een beetje moeite met lopen”	0.5	5	-0.040976	2
”Ik heb geen problemen met lopen”	0.7	1	-0.000065	1
”Ik loop prima zonder problemen”	0.7	2	-0.000441	1
”Geen problemen met lopen, denk ik”	0.7	3	-0.004559	1
”Lopen gaat meestal wel goed”	0.7	4	-0.276574	2
”Soms een beetje moeite met lopen”	0.7	5	-0.061726	2
”Ik heb geen problemen met lopen”	1.0	1	-0.000114	1
”Ik loop prima zonder problemen”	1.0	2	-0.000170	1
”Geen problemen met lopen, denk ik”	1.0	3	-0.002744	1
”Lopen gaat meestal wel goed”	1.0	4	-1.434428	3
”Soms een beetje moeite met lopen”	1.0	5	-0.071238	2
”Ik ben niet in staat om te lopen.”	0.2	1	-0.000536	5
”Ik kan helemaal niet meer lopen, ik gebruik een rolstoel.”	0.2	2	-0.003342	5
”Het is vrijwel onmogelijk voor mij om te lopen, ook met hulp.”	0.2	3	-0.967409	5
”Lopen is erg moeilijk voor mij, ik moet altijd hulp krijgen.”	0.2	4	-0.001263	4
”Ik kan soms helemaal niet lopen, maar het hangt af van de dag.”	0.2	5	-0.697157	5
”Ik ben niet in staat om te lopen.”	0.5	1	-0.001900	5
”Ik kan helemaal niet meer lopen, ik gebruik een rolstoel.”	0.5	2	-0.001014	5
”Het is vrijwel onmogelijk voor mij om te lopen, ook met hulp.”	0.5	3	-0.373736	4
”Lopen is erg moeilijk voor mij, ik moet altijd hulp krijgen.”	0.5	4	-0.004297	4
”Ik kan soms helemaal niet lopen, maar het hangt af van de dag.”	0.5	5	-0.258531	4
”Ik ben niet in staat om te lopen.”	0.7	1	-0.001107	5
”Ik kan helemaal niet meer lopen, ik gebruik een rolstoel.”	0.7	2	-0.002438	5
”Het is vrijwel onmogelijk voor mij om te lopen, ook met hulp.”	0.7	3	-0.785506	5
”Lopen is erg moeilijk voor mij, ik moet altijd hulp krijgen.”	0.7	4	-0.001698	4
”Ik kan soms helemaal niet lopen, maar het hangt af van de dag.”	0.7	5	-0.331540	4
”Ik ben niet in staat om te lopen.”	1.0	1	-0.000734	5
”Ik kan helemaal niet meer lopen, ik gebruik een rolstoel.”	1.0	2	-0.001014	5
”Het is vrijwel onmogelijk voor mij om te lopen, ook met hulp.”	1.0	3	-1.173162	5
”Lopen is erg moeilijk voor mij, ik moet altijd hulp krijgen.”	1.0	4	-0.001698	4
”Ik kan soms helemaal niet lopen, maar het hangt af van de dag.”	1.0	5	-0.385581	4

likelihood of -1.173162, indicating a lower level of confidence in the system’s classification. Lower log probabilities were also the outcome of ambiguous statements such as ”Ik kan soms helemaal niet lopen, maar het hangt af van de dag,” which indicated the system’s uncertainty when processing less clear inputs.

The results of these tests helped establish a log probability threshold that served as a cutoff point for asking follow-up questions. Through the analysis of log probabilities under various scenarios, especially those with greater ambiguity, a threshold that struck a compromise between accuracy and the need for additional clarification could be determined. For instance, a threshold was selected to guarantee that further questions would be asked of the user to get more information when the system’s confidence decreased (e.g., log probability below -0.01).

4.2 Categorization Results

The chatbot's effectiveness in distributing the EQ5D PROM questionnaire was evaluated with five participants once the ideal log probability threshold was determined. Answering the basic questions about mobility, self-care, usual activities, pain/discomfort and anxiety/depression, each participant interacted with the chatbot.

The interactions were analyzed to see if the follow-up questions were prompted at the right times and how well the chatbot classified the responses according to the predetermined thresholds. For instance, there were a few slight inconsistencies in Participant 1's responses, especially when it came to classifying mobility-related responses when follow-up questions were triggered because of ambiguous log probabilities. This pattern has been observed in several participants, demonstrating the usefulness of the threshold while also pointing up potential areas for improvement in categorization accuracy through more modification.

Based on the interactions, it can be seen that the chatbot managed when to ask follow-up questions by using the log probability threshold of -0.01. If chatbot's confidence in the category was low, more questions were asked to clarify their response. On occasion, it was noticed that the max tokens parameter set in the OpenAI API request caused some of the follow-up questions created by the AI to be cut short. These incidents, nevertheless, did not seriously interrupt the conversation because most of the follow-up questions had already been generated and the intended meaning was still clear even with the final words missing.

Overall, the interactions demonstrated that the chatbot could effectively guide participants through the questionnaire, but some challenges with the consistency of triggering follow-up questions were noted. This feedback was crucial for understanding how the chatbot performed under real-world conditions and provided insights into areas for further adjustment.

4.3 System Usability Scale (SUS) and User Feedback

After interacting with the chatbot, each of the five participants completed the System Usability Scale (SUS) survey. A quantitative assessment of the chatbot's usability was provided by the SUS survey, which evaluated user opinions on a range of topics including system complexity, ease of use, and user confidence. The survey scores were calculated on a scale from 0 to 100. Higher ratings denoted better usability.

The SUS scores for each participant were as follows:

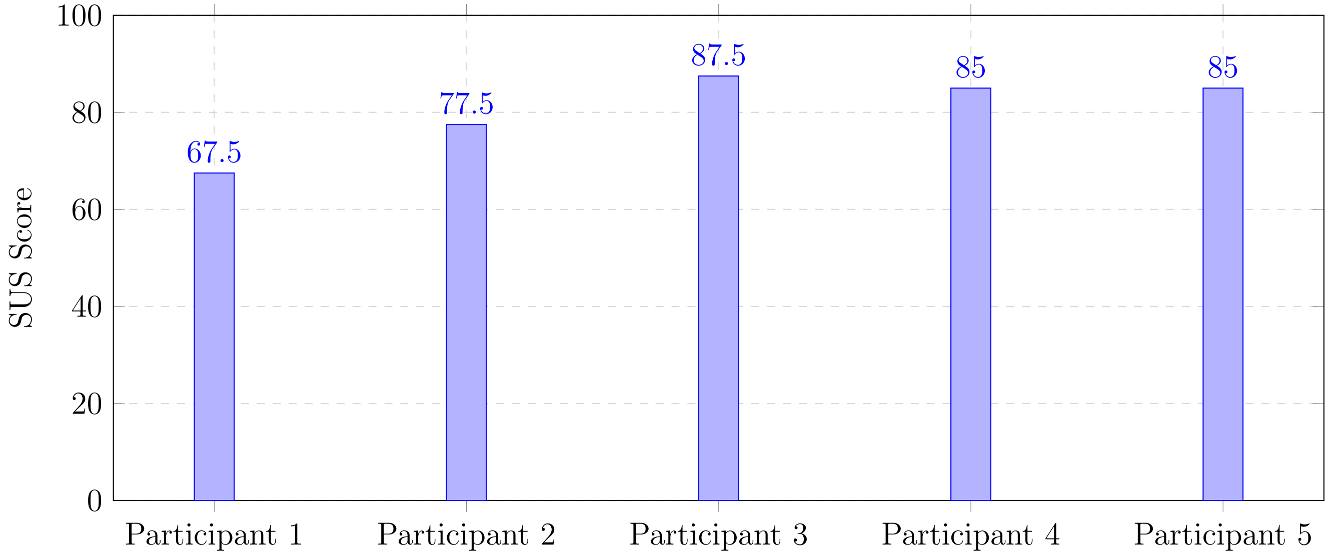


Figure 3: System Usability Scale (SUS) Scores by Participant

- **Participant 1 (Age: 25)**: SUS Score: 67.5

  During the interaction with the chatbot, this participant's chatbot crashed, which probably influenced the overall usability of the experience. Looking at the score of 67.5, this suggests that the participant found the system somewhat usable but identified challenges, possibly due to the disruption caused by the technical issue.

  To better understand Participant 1's experience, her individual responses to the SUS survey are analyzed below:

Question	Response
I think I would like to use this tool frequently.	3
I found the tool unnecessarily complex.	2
I thought the tool was easy to use.	3
I think that I would need the support of a technical person to be able to use this system.	2
I found the various functions in this tool were well integrated.	3
I thought there was too much inconsistency in this tool.	2
I would imagine that most people would learn to use this tool very quickly.	4
I found the tool very cumbersome to use.	2
I felt very confident using the tool.	4
I needed to learn a lot of things before I could get going with this tool.	2

Table 2: Participant 1's Responses to the SUS Questionnaire

Participant 1's responses imply a mixed experience with the chatbot. The responses were rated on a scale from 1 to 5, with 1 indicating strong disagreement or dissatisfaction, and 5 representing strong agreement or satisfaction. Although the participant gave it a score of 3 for ease of use, she did have several issues that affected her opinion overall. While giving the tool a score of 2 for complexity, the participant gave the system's integration and ease of use a similar rating of 3, meaning that although the tool is accessible, it could still benefit from further refinement.

The participant expressed confidence in using the tool, scoring it a 4, but noted inconsistencies in the user experience, giving it a 2, likely due to the technical crash they experienced. The tool's perceived cumbersomeness also contributed to the lower score, with another 2 given for ease of use. Despite their faith in the system, Participant 1's lower SUS score appears to be mostly impacted by perceived inconsistencies and technical problems.

- **Participant 2 (Age: 23)**: SUS Score: 77.5

  This score indicates that the participant had a positive usability experience, finding the product to be well-integrated yet simple to use. The high score indicates that the user felt comfortable utilizing the chatbot and that there was little perceived complexity.

- **Participant 3 (Age: 48)**: SUS Score: 87.5

  The highest SUS score among the participants, 87.5, shows us that this participant found the chatbot very user-friendly and intuitive. Comments emphasized how well-thought-out the questions were and how the conversation flowed, which all added to an extremely positive experience.

- **Participant 4 (Age: 23)**: SUS Score: 85.0

  Like Participant 3, this score indicates that the participant had a strong satisfaction with the usability of the chatbot. According to the participant's feedback, the tool was simple to use, and most users could pick up the skills necessary to utilize it efficiently very quickly.

- **Participant 5 (Age: 23, Meniscus Injury)**: SUS Score: 85.0

  This participant, who suffered a meniscus injury while playing football, offered insightful feedback on how the chatbot handled mobility-related health issues. The individual's high score implies that the chatbot effectively assisted them in completing the EQ5D PROM survey, accurately capturing their answers concerning mobility challenges. This particular instance demonstrates the chatbot's ability to handle actual health issues by offering appropriate classifications that were in line with the participant's condition.

The average SUS score was 80.5, meaning that the chatbot was generally well-liked and thought to be user-friendly. But the wide range of results, especially Participant 1's lower score because of the system crash, showed areas in which technical reliability needed to be improved.

These scores were further evaluated by an analysis of the individual SUS survey replies. The majority of participants agreed that the tool was simple to use and that its features were effectively integrated. However, certain users' experiences were hindered by technical issues and inconsistent behavior. The qualitative feedback obtained from the interviews provided additional context for these findings. It was found that although the majority of users had no trouble interacting with the chatbot, there were times when the engagement felt interrupted or repetitive, because of redundancy in the follow-up questioning.

4.4 Participant Interviews

To obtain qualitative input from participants, informal interviews were carried out in addition to the SUS survey. The qualitative information from these interviews enhanced the quantitative data from the SUS by revealing more about the user experience.

The majority of participants reported having a good experience with the chatbot, praising its general usability and the clarity of its questions. One participant wrote, "Het was makkelijk om de vragen te beantwoorden, want de chatbot was wel duidelijk en simpel om te gebruiken." However, some comments indicated that the chatbot's handling of follow-up inquiries needed to be improved. Participants said that when there was already a high level of confidence in the categorization, the engagement could feel repetitive.

The chatbot's sympathetic attitude was also noted by participants, who liked its conversational tone but thought it might be improved with more tailored feedback at the end of the interaction. Additionally, recommendations were made to improve how the chatbot handles unclear answers in order to improve conversation flow and lessen the number of pointless follow-up questions.

These qualitative observations gave the SUS scores important context and made it easier to pinpoint certain places where the chatbot's interaction design needed to be changed in order to better satisfy user expectations.

5 Discussion

This chapter discusses the findings from the development, implementation, and evaluation of the RASA-driven chatbot designed to administer the EQ5D PROM questionnaire. The discussion focuses on interpreting the results in relation to the research objectives, exploring the implications of these findings, identifying limitations, and suggesting directions for future research.

5.1 Key findings

The main objective of this research was to develop a chatbot powered by RASA that could administer the EQ5D PROM questionnaire in a natural and unobtrusive way, enhancing user experience and ease of completion. The outcomes showed that the chatbot could successfully lead users through the EQ5D questionnaire. It did this by using log probability thresholds to determine if follow-up questions were needed, using customized actions. This controlled the flow of the conversation.

According to the results of the classifications, the chatbot correctly recognized user responses that were below the log probability threshold, asking follow-up questions when needed. This mechanism made it possible for the chatbot to consistently classify the responses with high accuracy, even in cases where the user inputs were unclear. Cases such as Participant 5, who suffered from a meniscus injury, demonstrated how the chatbot could adjust to certain mobility-related health issues, offering relevant classification and an interaction that seemed contextually appropriate and personalized.

Using a log probability barrier of -0.01 was a crucial step in controlling the chatbot's questioning strategy. This threshold was chosen to achieve a balance between maintaining a smooth and natural flow of interaction and the requirement for an accurate categorization. According to the findings, this threshold successfully reduced the number of pointless follow-up questions by only posing new ones when the categorization confidence was low.

However, occasionally, the OpenAI API call's use of the "max_tokens" argument resulted in the prompts being somewhat incomplete because the follow-up questions were cut short. Fortunately, the majority of the questions were generated and comprehensible, so these cutoffs did not really disrupt the conversation. However, this problem brought to light a response management shortcoming that might be improved in future versions. The effectiveness and clarity of the engagement would be improved by adjusting the parameters to better manage the length of responses and guarantee that all follow-up questions are given completely.

The chatbot was also well-received overall, according to the System Usability Scale (SUS) statistics, which showed an average score of 80.5, above-average usability. The majority of participants reported that the system was user-friendly and intuitive, but technical problems, like the crash that Participant 1 encountered, emphasized how crucial dependability is to preserving user pleasure.

These results were supported by the qualitative comments from the interviews, which provided more in-depth understanding of the user experiences. Participants praised the chatbot's helpful tone and the EQ5D questions' clarity, but they also pointed out areas that needed work, such as reducing repetitive follow-up questions and improving the chatbot's ability to respond to unclear inputs.

5.2 Practical Implications

The results of this study offer several possible uses for utilizing chatbots in healthcare, especially when administering PROMs like the EQ5D questionnaire. The effective use of log probability thresholds to manage follow-up questioning, illustrates a valuable approach for maintaining the balance between accuracy and user engagement. Chatbots can better handle ambiguous responses without overburdening the user by adjusting the confidence levels that trigger additional questions. This will lead to an enhanced overall interaction experience.

Furthermore, the positive usability ratings of the participants imply that chatbots can function as a productive and user-friendly substitute for traditional survey methods. The high acceptability among the users suggests that people are becoming more comfortable with AI-driven interactions, particularly when the system is made to appear sympathetic and helpful. This demonstrates how such technologies may be adopted more widely in a variety of healthcare situations where patient participation and efficient data gathering are essential.

5.3 Limitations

Although the results are promising, there are a few things to be aware of. The study's first goal was to test the chatbot on a group of medical patients who were at least 50 years old. But because of practical problems, like possible access issues, ethical considerations and the logistical difficulties of dealing with elderly patients, this proposal was changed. Because of this only a small number of younger participants were used in the study, which restricts the applicability of the results to the intended target group.

The chatbot's dependency on slot filling to get user answers is another major flaw. Slots were useful for collecting structured data, but they also presented problems when respondents provided information that was applicable to more than one aspect of health. When a user describes pain while walking, for instance, the chatbot may mistakenly allocate the response to the pain slot rather than the mobility slot. This misclassification highlights a crucial area for development in the chatbot's answer processing capabilities and may have an impact on the overall health assessment's accuracy.

Another limitation that surfaced was technical reliability, which was demonstrated by Participant 1's system crash, which had an impact on their usability score and overall interaction experience. These technological hiccups highlight how important it is to have stable systems and strong error handling in order to guarantee reliable functioning.

Moreover, the limited sample size and controlled testing setup restrict the applicability of these findings to larger, real-world contexts. Even though the input from friends and acquaintances is enlightening, it might not accurately reflect the range of viewpoints from a larger patient group, especially in the case of older persons who could have different expectations and needs when it comes to engagement.

5.4 Reflection of Methodology

This study's mixed-methods approach, which combined prototyping with qualitative interview feedback and quantitative SUS scores, worked well to get a full picture of the chatbot's performance. Nonetheless, the dependence on a limited participant pool and regulated testing environments implies that larger-scale, real-world experiments need to be taken into account in future studies in order in order to verify the results and enhance the chatbot's usefulness even more.

Though innovative, the use of log probability thresholds to control follow-up questions brought attention to the necessity of constant tuning and adjusting to satisfy user expectations and conversational flow. This study's finding highlights how crucial user input and iterative testing are to the creation of conversational AI agents.

6 Conclusions and Further Research

The main objective of this study was to create a chatbot powered by RASA that could automatically and naturally administer the EQ5D PROM questionnaire, improving user experience and ease of completion. The study's objective was to develop a system that, by using customized actions and log probability levels, not only reliably captured user responses but also retained a friendly and engaging interaction.

The results show that these goals were mainly achieved by the chatbot. It effectively guided participants through the EQ5D questionnaire, which demonstrated its ability to adapt to different user inputs and provide contextually appropriate responses. The system's use of log probability thresholds to control follow-up questions was very successful in striking a balance between user engagement and accuracy, making sure that users were only asked for more information when it was absolutely required.

The chatbot obtained a good rating according to the System Usability Scale (SUS), with an average score of 80.5, indicating a generally positive user experience. The participants liked how the questions were clear and how simple it was to use the chatbot. Technical problems, such as random system crashes and difficulties with the slot-filling procedure, however, pointed out areas that needed more improvement. In particular, the inaccurate assignment of answers to the wrong slots, for example, labeling a mobility-related reaction as pain—emphasized the necessity for more sophisticated natural language comprehension tools.

Overall, this research successfully showed that a RASA-driven chatbot can efficiently deliver the EQ5D PROM questionnaire while offering a user-friendly and empathetic interaction. The research offers insightful information about the possibilities of AI-powered health tools, especially with regard to improving patient-reported outcomes and streamlining data collecting in healthcare settings.

Although the research met most of its objectives, it also identified areas in which more investigation is required to improve and enhance the chatbot's functionality. The chatbot's method of filling slots is one major area that needs work. Currently, when user inputs cross multiple health dimensions, the chatbot may incorrectly assign user responses to slots. A significant improvement would be the ability to dynamically manage slot filling. That is to turn off specific slots at particular points throughout the conversation. This would decrease misassignments and increase the overall accuracy of the categorization process by enabling more precise management of which slots can be filled based on the context.

Since the EQ5D is a standardized tool that is usually not modified, Further research should examine the reliability of using a chatbot to deliver the questionnaire. It is crucial to look at whether using this digital administration approach has an impact on the validity or reliability of the replies when compared to traditional formats. Because any modification to the questionnaire's delivery could have an impact on the outcomes. This is particularly important because maintaining the EQ5D's standardized application requires sticking to its original format and language.

Increasing the number of participants to include a wider range of demographics, especially older persons and patients with different medical conditions, would provide a greater understanding of the chatbot's usability in various settings. The previously mentioned expansion could aid in determining whether the existing design of the system effectively caters to the requirements of different types of users and identify any necessary modifications.

Enhancing the chatbot's technical reliability is also crucial. More effective error-handling and recovery techniques could reduce user interaction interruptions and guarantee that the system is stable and consistent even in challenging real-world situations. Investigating real-time learning and adaptable algorithms could help to better personalize the user experience by enabling the chatbot to dynamically modify responses in response to specific needs.

Future studies may also examine the application of the GPT-4 model to improve the chatbot's conversational abilities. Initial observations indicate that the improvements in contextual comprehension and response quality provided by GPT-4 may result in a more effective and engaging user experience. Researchers can evaluate usability and user satisfaction gains by comparing GPT-4's performance with the current system. This will help to design more advanced AI-driven health products.

Lastly, looking into multi-modal input techniques like speech recognition could increase the chatbot's accessibility and inclusivity for users of all skill levels. This would be in line with the objective of developing a conversational interface that is more natural and engaging, especially for users who might have trouble reading or entering text.

Appendix A

The following questions were used in semi-structured interviews to gather feedback on the user experience of the RASA-driven chatbot:

1. Hoe je jouw algehele ervaring met de chatbot beschrijven?
2. Vond je de chatbot makkelijk te gebruiken? Waarom wel of niet?
3. Ervaarde je het gesprek als natuurlijk en onopvallend?
4. Heeft de chatbot je effectief door de EQ5D vragenlijst geleid?
5. Als je één ding aan de chatbot mocht verbeteren, wat zou dit dan zijn?
6. Zou je deze chatbot opnieuw gebruiken voor vergelijkbare taken? Waarom wel of niet?
7. Als je naar jouw eindresultaat kijkt, vind je dan dat de chatbot jou goed heeft gecategoriseerd bij elke vraag?

Appendix B: Access to the Chatbot Code

The complete implementation of the RASA-driven chatbot is available in a public GitHub repository. This repository contains all the the source code.

- **GitHub Repository Link**: `https://github.com/RoderickLei/RASAchatbot.git`

References

[1] Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181.*

[2] Devlin, N. J., & Brooks, R. (2017). EQ-5D and the EuroQol Group: Past, present and future. *Applied Health Economics and Health Policy, 15*(2), 127-137. https://doi.org/10.1007/s40258-017-0310-5

[3] Milne-Ives, M., Lam, C., De Cock, C., Van Velthoven, M. H., & Meinert, E. (2020). The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of Medical Internet Research, 22*(10), e20346. https://doi.org/10.2196/20346

[4] Schröder, J., Nielebock, C., Tinschert, P., & Kowatsch, T. (2021). A digital intervention for a clinical trial: Monitoring and improving adherence to a web-based pre-treatment in routine psychotherapy. *Internet Interventions, 26*, 100445. https://doi.org/10.1016/j.invent.2021.100445

[5] Ting, D. S. W., Carin, L., Dzau, V., & Wong, T. Y. (2020). Digital technology and COVID-19. *Nature Medicine, 26*(4), 459-461. https://doi.org/10.1038/s41591-020-0824-5

[6] Appleby, J., & Devlin, N. (2010). Measuring success in the NHS: Using patient-assessed health outcomes to manage the performance of healthcare providers. *London: King's Fund.*

[7] Black, N. (2013). Patient reported outcome measures could help transform healthcare. *BMJ, 346*, f167.

[8] Centers for Medicare & Medicaid Services (CMS). (2020). Value-Based Programs. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs

[9] Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment, 2*(14), 1-74.

[10] Van Dijk, B., Lefebvre, A., & Spruit, M. (under review). Welzijn.AI: A Conversational AI System for Monitoring Mental Well-being and a Use Case for Responsible AI Development. http://welzijn.ai

[11] EuroQol Group. (1990). EuroQol–a new facility for the measurement of health-related quality of life. *Health Policy, 16*(3), 199-208.

[12] Janssen, M. F., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., ... & Swinburn, P. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Quality of Life Research, 22*(7), 1717-1727.

[13] Kingsley, C., & Patel, S. (2017). Patient-reported outcome measures and patient-reported experience measures. *BJA Education, 17*(4), 137-144.

[14] Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2), 293-327.

[15] Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavioral 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.

[16] Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... & Coiera, E. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248-1258.

[17] Richards, M., Caldwell, P. H., & Go, H. (2020). Impact of the Babylon Health Chatbot on Health Outcomes. *Frontiers in Digital Health*, 2, 12.

[18] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research, 15*(11), e239.

[19] Mehta, N., Pandit, A., & Shukla, S. (2019). Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. *Journal of Biomedical Informatics, 100*, 103311.

[20] RASA. (2024). RASA Documentation and Overview. https://rasa.com

[21] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

[22] Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. *International Conference on Learning Representations (ICLR).*

[23] Bocklisch, T., Werkmeister, T., Varshneya, D., & Nichol, A. (2024). Task-Oriented Dialogue with In-Context Learning. *arXiv preprint arXiv:2402.12234*. https://doi.org/10.48550/arXiv.2402.12234

[24] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44-56.

[25] Wang, Y., Wang, L., Rastegar-Mojarad, M., Shen, F., Liu, S., Afzal, N., ... & Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics, 77*, 34-49.

[26] Bennett, A. V., Jensen, R. E., Basch, E., et al. (2012). Electronic patient-reported outcome systems in oncology clinical practice. *CA: A Cancer Journal for Clinicians, 62*(5), 337-347. https://doi.org/10.3322/caac.21150

[27] Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., & PROMIS Cooperative Group. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3-S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55

[28] Deering, S., Rowland, P., Voelkel, S., & Butz, K. (2017). A demonstration of the impact of an innovative mHealth model for the collection of patient-reported outcomes. *Journal of Patient-Reported Outcomes, 1*(1), 1-7. https://doi.org/10.1186/s41687-017-0004-1

[29] Jongsma, K. R., Janssens, A. C., Kraft, S. A., & Wouters, H. (2020). The implementation of PROMs in cancer care: exploring physicians' perceptions of their impact on the doctor-patient interaction. *BMC Health Services Research, 20*(1), 1-9. https://doi.org/10.1186/s12913-020-05628-w

[30] Lavallee, D. C., Chenok, K. E., Love, R. M., Petersen, C., Holve, E., Segal, C. D., & Franklin, P. D. (2016). Incorporating patient-reported outcomes into health care to engage patients and enhance care. *Health Affairs, 35*(4), 575-582. https://doi.org/10.1377/hlthaff.2015.1362

[31] Piao, M., Ryu, H., Lee, H., Kim, J., & Lee, J. Y. (2020). Development and user research of a chatbot for chronic disease management. *JMIR mHealth and uHealth, 8*(12), e21735. https://doi.org/10.2196/21735

[32] Rothman, M., Burke, L., Erickson, P., Leidy, N. K., Patrick, D. L., & Petrie, C. D. (2015). Use of existing patient-reported outcome (PRO) instruments and their modification: The ISPOR good research practices for the use of patient-reported outcomes in clinical trials task force report. *Value in Health, 12*(8), 1075-1083. https://doi.org/10.1111/j.1524-4733.2009.00609.x

[33] Brooke, J. (1996). SUS: A quick and dirty usability scale. In *Usability Evaluation in Industry* (pp. 189-194). Taylor & Francis.