

Article

Reflective Dialogues with a Humanoid Robot Integrated with an LLM and a Curated NLU System for Positive Behavioral Change in Older Adults

Ryan Browne ^{1,*} , Mirza Mohtashim Alam ², Qasid Saleem ², Abrar Hyder ², Tatsuya Kudo ³, Francesca D'Agresti ⁴ , Martino Maggio ⁴ , Keiko Homma ³, Eerik-Juhanna Siitonen ³, Naoko Kounosu ¹, Kristiina Jokinen ³ , Michael McTear ⁵ , Giulio Napolitano ² , Kyoungsook Kim ³, Junichi Tsujii ³, Rainer Wieching ⁶, Toshimi Ogawa ^{1,*} , and Yasuyuki Taki ¹

¹ Institute for Development, Aging and Cancer, Tohoku University, Sendai 980-8577, Japan

² Institut für Angewandte Informatik, 04109 Leipzig, Germany

³ AI Research Center, AIST, Tokyo 135-0064, Japan

⁴ Engineering Ingegneria Informatica SpA, 00144 Rome, Italy

⁵ School of Computing, Ulster University, Belfast BT15 1AP, UK

⁶ Information Systems and New Media, University of Siegen, 57072 Siegen, Germany

* Correspondence: browne.ryan.e2@tohoku.ac.jp (R.B.); toshimi.ogawa.e6@tohoku.ac.jp (T.O.)

Abstract: We developed an innovative system that combines Natural Language Understanding (NLU), a curated knowledge base, and the efficient management of a Large Language Model (LLM) to support motivational health coaching. Using Rasa as the core framework, we enhanced it by integrating the GPT-3.5-turbo model. Users opt into reflective dialogues during conversations. When they respond to open-ended questions, their input goes directly to the GPT-3.5-turbo model, allowing for more flexible responses. To provide curated trustworthy content, we integrated a knowledge provision component that searches a PDF-based knowledge base and generates user-friendly responses using Retrieval-Augmented Generation. We tested the system in a real-world scenario by deploying it on a Nao robot in seven older adults' homes for 1–2 weeks, encouraging positive behavioral changes in some users. Our system serves as a valuable foundation for building an even more integrated, personalized system that can connect with other Application Programming Interfaces (APIs) and integrate with home sensors and edge devices.

Keywords: health coaching; artificial intelligence; older adults; large language models



Citation: Browne, R.; Alam, M.M.; Saleem, Q.; Hyder, A.; Kudo, T.; D'Agresti, F.; Maggio, M.; Homma, K.; Siitonen, E.-J.; Kounosu, N.; et al. Reflective Dialogues with a Humanoid Robot Integrated with an LLM and a Curated NLU System for Positive Behavioral Change in Older Adults. *Electronics* **2024**, *13*, 4364. <https://doi.org/10.3390/electronics13224364>

Academic Editor: Stefanos Kollias

Received: 29 August 2024

Revised: 26 September 2024

Accepted: 9 October 2024

Published: 7 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the rising older demographic and presence of super-aging societies, it is imperative to maintain and extend the human health span, to both ease the burden on social welfare and care systems, and ensure that all persons can age well in a healthy and positive manner [1].

Health coaching is a patient-centered approach that aims to help individuals achieve their health goals through education, motivation, and behavior change [2]. Health coaching is needed because many individuals struggle to make and sustain lifestyle changes on their own, despite knowing the benefits of healthy habits. The key concepts are personalized interventions, tailored advice, and continued support for the person being coached. With advances in natural language understanding and the ubiquity of smart wearables that can collect health information, health coaching is now apt for translation to the digital realm. Indeed, artificial intelligence-based conversational agents have been introduced in the medical domain for patients with chronic diseases [3], for reducing symptoms of depression and anxiety [4], for obesity [5] and weight management [6,7], as well as in health domains such as for exercise [8] and sleep [9].

While access to a real human coach is limited by the affordability of the coach and availability of the coach's time, a virtual health coach powered by artificial intelligence can be deployed widely at a low cost, such that anyone with access to a smartphone or the internet could benefit. One's own private virtual health coach could enable sustained health practices, as AI coaches have been found to be as effective as human coaches in helping clients reach their goals [10]. Indeed, a virtual health coach has been identified as one component of a multimodal biomedical AI [11], and the use of virtual coaches has been investigated for active and healthy aging [12,13]. Large Language Models are expected to transform healthcare, especially for older adults [14].

Motivational health coaching may enable older adults to stay healthy or change their behavior to a more active lifestyle [15]. The Trans-Theoretical Model of behavior change (TTM) [16] is widely used in health behavior research and practice, and has been applied to a range of health behaviors, including smoking cessation [17], physical activity [18], and healthy eating [19]. According to this model, behavioral change requires the user to be prompted to examine their current behavior and lifestyle patterns, and to reflect upon them.

In this paper, we term this part "reflective dialogue". The TTM consists of five stages of change: pre-contemplation, contemplation, preparation, action, and maintenance. In particular, self-reflection and assessment can increase an individual's motivation to change by increasing awareness of the negative consequences of the behavior and the benefits of adopting healthier habits. In the preparation stage, access to relevant information will support the progression towards taking action. We established reflective dialogue questions on the basis of giving users the impetus to think about and assess their current behavior, and promote thinking about how they could take actionable steps to stay healthy or improve their health, across major health domains. We also emphasized the need for the users to make more human contact, and to have autonomy and control in their use of the robot, to allay ethical concerns that have been highlighted in robot deployment for older adult care [20]. Therefore, our reflective dialogue included questions such as "How do you think you could create more opportunities to interact with others?" and "Do you often eat with others?"; and users would need to opt into the reflective dialogue by agreeing, otherwise they could decide to ignore the opportunity or talk about something else.

In this work, we design a prototype AI health coaching conversational system that can respond to user queries and provide access to relevant and tailored information across five domains: exercise, sleep, cognitive health, nutrition, and social activity. These domains were chosen as they have been identified as the primary areas with the strongest evidence for interventions in reducing cognitive decline [21]. Our prototype AI health coaching system is deployed on a humanoid robot, Nao. We chose a real physical object rather than the use of a desktop or smartphone application because the physicality of the robot acts as a bridge between the virtual and physical worlds, where its physical presence may encourage the user to communicate with it in the home [22]; and talking by voice may be more natural than typing for older adults.

One of the main issues in creating an AI health coach is enabling natural conversation that features open-ended questions (such as, "How do you feel about exercise?"), where the user is free to reply as they wish after reflecting on their own behavior. In controlling conversation or designing conversation patterns, it becomes very difficult for a system designer to create a logical flow, or to design a few-shot learning model (where user speech is designed as labelled samples and is used to generalize to novel speech input), when such open-ended conversations are included. The use of an LLM may overcome these limitations, but, due to the generative nature of LLMs, ethical concerns arise over the provision of false or inaccurate information, which is especially important in health and medical assistants, where the safety and health of users could be at risk [23]. There is a need for responsible design and oversight in the deployment of LLMs for healthcare [24]. Therefore, a system must retain tight control over what information is presented to the user, and in what capacity the LLM can be utilized.

This work seeks to answer the following research question: How can the limitations of an intent-based NLU system be overcome by integrating an LLM, while retaining the provision of domain-expert knowledge and personalization that is necessary for behavioral change? Furthermore, we sought to understand the experience of older adults in using a natural-language-enabled physical robot.

Robots have been used to act as coaches for psychological well-being, equipped with LLMs to handle conversation flow [25]. This study found that the LLM working alone as a coach did not focus on asking users questions relating to positive psychology and provided inappropriate responses. Our use of specifically designed questions that can be presented to the user and the integration of a knowledge base may overcome these limitations.

Recently, social robots have been connected to LLMs for the improvement of cognitive function in a simulated interaction with Autism Spectrum Disorder subjects [26]. The robot Pepper has been introduced to older adults living in a care home [27] to support physical activity, cognitive training, and social interaction, but this system required a human moderator to be present to control the robot at all times. The independent functioning of robots is a clear goal for their effective utilization.

To the best of our knowledge, our work is the first study to use a physical robot equipped with both LLMs and an NLU system for the purpose of acting as a health coach for independently living, healthy older adults. Our study is also novel in that the robots are placed into the homes of users, and they are free to interact as they wish with the device, with the system being entirely autonomous and operating from a remote server.

This paper contains the following sections. In the Methods Section, we describe how we designed our natural language system and deployed it on a physical robot. We also detail the recruitment of human subjects and the experimental procedure. The Analysis Section presents an evaluation of the system along several measures (assessment of the quality of training data used in the intent-based model, usage records during the study, assessment of prompt engineering techniques). Our Results Section presents participant quotes on their user experience, before we move onto the Discussion.

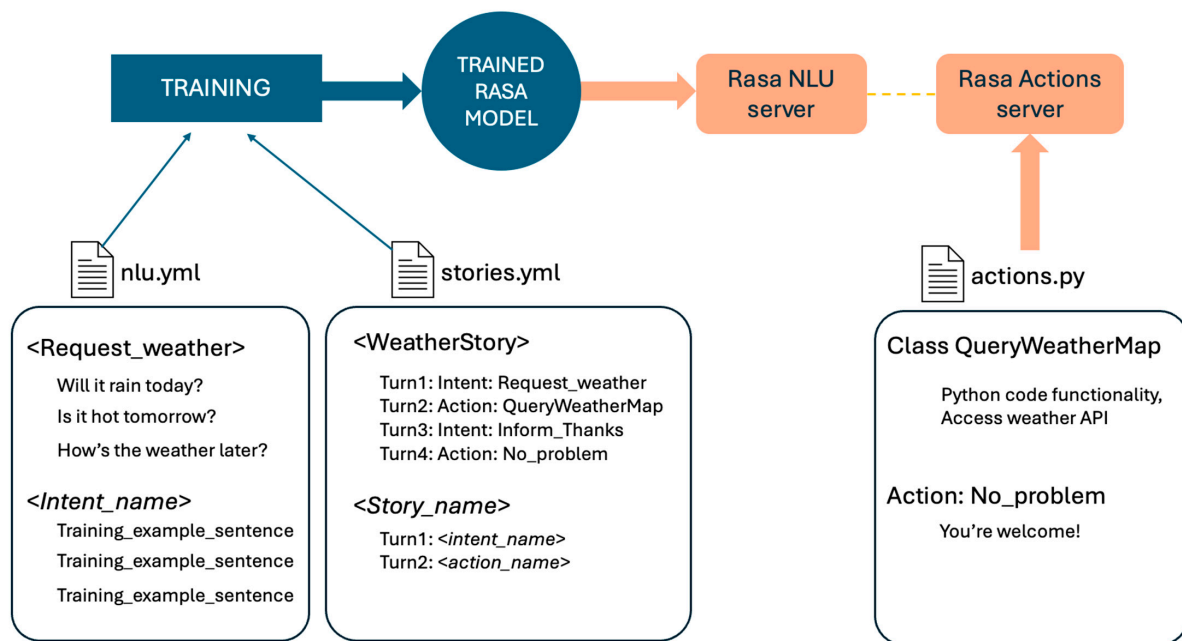
2. Methods

2.1. System Design

In approaching our research question, we decided to utilize a framework that has excellent scope for customization, user tracking, and connection to APIs. The system must be able to access real-time up-to-date information (such as news, weather) through APIs, as well as enable the long-term storage of personal information about the user and their conversations that would not be possible with just an LLM. The Rasa Natural Language Understanding (NLU) framework, an open-source tool for processing natural language [28], fulfils these requirements and was chosen as the basis for our system. See Figure 1 for an overview of how Rasa operates; we also briefly explain the details below.

Rasa performs several key tasks: intent classification, entity extraction, and prediction of the next action. For example, the user might ask something like, “What is the weather like today in Tokyo?”. Rasa uses supervised machine learning models to classify user inputs into predefined intents; i.e., there may be a predefined system labelled as “request_weather”. Entities such as dates (today) and locations (Tokyo) can be extracted from the user input, as the Rasa model is trained on labelled data wherein both examples of user utterances that would be expected to fall under a predefined intent are listed, and examples of entities are delineated in the example sentences.

Intents are linked to system “actions” (what the system does in response to the user intent). Actions are handled by a Policy model, such as memorization or Transformer Embedding Dialogue (TED). The expected turn-taking steps (known as “stories”) are defined in a file that is used in the Rasa training step. Actions can be as simple as a predefined text response, or as complex as anything that can be implemented through code, as the Rasa Software Development Kit utilizes Python 3 for customized functionality.



"What is the weather tomorrow?"

- | | |
|---------------------|--|
| ① Intent Prediction | Using trained Rasa NLU: Predicted intent: Request_weather
Confidence: 93% |
| ② Entity Extraction | What is the weather tomorrow ? |
| ③ Context Update | Variable slot 'weather_timeframe' is set to 'tomorrow' |
| ④ Policy Selection | Next Action Prediction with TED Policy
Most likely next action: QueryWeatherMap |
| ⑤ Action Execution | Custom Action is triggered, and response is returned to the user. |

➡ "It looks like it's going to be cloudy tomorrow, with a high of 26 degrees."

Figure 1. A simplified overview of how Rasa operates. Manually defined files are used as inputs for training the Rasa Natural Language Understanding (NLU) model. These files include the nlu.yml, which lists the names of intents and associated example sentences, and a stories.yml file that shows the expected turn-taking behavior between intents and actions in response to the intent. Note that the responses or behavior of the actions themselves is not used in training, and the actions are run on a separate server.

The user speaks, and the trained Rasa intent classifier predicts the predefined intent. If the classifier could not predict the intent with a certain level of accuracy, a "Fallback" action is activated. A threshold value to determine when the Fallback action is utilized can be defined in Rasa.

From our previous experimentation with the implementation of a Rasa-based system, we found that stories that were excessively deep, in the sense of having a large number of predefined turn-taking steps, led to the user either becoming trapped in a story and unable to make the expected progress, or being able to jump out of the story into an unexpected place in another story. Therefore, for the system used in the current work, we kept each story to a simple intent–action pair. This allows great flexibility for the user, who is largely

unaware of the system design, to talk freely to the system, and after the intent–action pair is completed, the system can wait for the next user utterance.

In creating the system used in the current work, we defined 21 possible user intents, which are listed in Table 1. The Rasa intents were defined using the concept of speech acts from Discourse Theory [29], such as inform (give some information to the system) and request (request some information from the system). These intents were chosen based on both our previous experience with older adults and their interactions with language-capable robots, and on the goal of the system to act as a health coach to support healthy aging. We created example sentences based on our judgement of common speech patterns, using entities (labelled parts of speech) where appropriate to enable finer distinctions.

Table 1. Intents, the corresponding action, and some associated example sentences used to train the intent classifier. Example sentences need to be manually created and added to the file used in the training process. Note the delineated entities within some of the example sentences (for example, the weather query intent), wherein the entity is enclosed in square brackets and the name of the entity immediately follows it in brackets.

Intent ID	Functionality of the Corresponding Custom Action	{Number of Training Examples}; Example User Phrase to Trigger the Intent
System accesses external functionality through code		
20 (request_weather)	The user asks for the weather; the system connects to the openweathermap API to fetch the current or forecasted weather.	{31} What is the weather in the [evening](weather_timezone)? Will it rain [tomorrow](weather_timezone)? Should I bring my umbrella [tomorrow](weather_timezone)?
19 (request_video)	The user asks for an exercise video to be sent to their smartphone. The system uses the stored user Telegram ID on the Digital Enabler to send a video from a curated list of URLs to the user’s Telegram application (which can be registered on the Digital Enabler).	{10} Could you send me an exercise video? Can you show me a workout video? Can you send me an exercise video?
01 (inform_greeting)	The user greets the system. The system greets the user back according to the current time, gives the user a weather update through connection to the OpenWeatherMap API, and then asks if the user would like to start motivational coaching.	{10} hello hey good morning good afternoon good evening
System uses LangChain to return a response to the user from a knowledge base search		
14 (request_functionalities)	The user asks how the system can help them, about its capabilities, as well as information about its name.	{10} How can you help me? How are you useful? What functions do you have?
03 (query_knowledgebase)	The user asks a question related to one of the health domains (sleep, nutrition, exercise, cognition, and social activity). The custom action uses the user text to search the PDF knowledge base and formulate a response with LangChain.	Week 1: {9} Week 2: {32} What exercise is good for me? How can I improve my memory? Any anti-aging supplements?
System utilizes code to make an internal process occur		
05 (request_repeat)	The user asks the system to repeat the last thing it said. The system uses the tracker to return the last activated action.	{12} Could you repeat that? Can you say that again? Could you rephrase that? What do you mean?
06 (inform_affirm)	An ‘affirm’ intent, where the user agrees to something. This was used in our system to trigger the reflective dialogue (Rasa Form).	{24} Okay Sure Go ahead

Table 1. Cont.

System utilizes code to make an internal process occur		
08 (inform_deny)	A 'deny' intent, which was used in our system to not activate the motivational dialogue (Rasa Form), in response to the system asking.	{21} No No thanks Not right now
System returns a simple predefined text response from Rasa		
04 (inform_specific)	The user desires to start a specific activity, such as exercise or brain training. In the current system, a general reply was given, but it would also be feasible to design a text/speech-based brain training game on the system for the user to play.	{9} I want to do some [yoga](exercise_type) I'll take a [walk](exercise_type) now I want to do [brain training](exercise_type) now
02 (request_personaldata)	The user requests an update of their personal data from the system, such as from a smartwatch. In our current system, we did not obtain data from a smartwatch or home sensors, but it is possible to access these data through making an API call to the Digital Enabler, if such devices are registered there.	{7} how much [sleep](domain) did I get [last night](timeframe)? how many hours did I [sleep](domain)? how [far](quantity) have I (walked)[domain]?
16 (inform_goodnight)	The user says goodnight; the system also responds similarly. This was separated from the greeting intent in order to avoid presenting a reflective dialogue component when the user is about to sleep.	{9} goodnight. night night. sleep
18 (request_challenge)	The user challenges the information provided by the system. The system responds that its contents are created by researchers, and using guidelines.	{8} Will that actually work? Is that really effective? Has that been proven?
17 (request_chat)	The user approaches the system in a manner different to a greeting, by asking to start or begin chatting. The system greets the user, and briefly explains what kind of things the user can ask.	{10} Let's chat Shall we have a talk Let's talk about something
15 (request_different)	The user rejects the current suggestion by the system and asks for another one. The current system did not include functionality to achieve this, but merely said that is all it has for now and to ask something else.	{6} I want a different suggestion. Tell me another one. I don't like this idea.
07 (inform_tired)	The user tells the system that they are feeling tired, sore, or not well. The system offers a kind response back to the user.	{16} I don't feel good I have slight body pain Quite tired
09 (inform_finished)	The user tells the system that they have finished doing something, such as watching a video or performing an exercise. The current system simply responds in a positive manner, but this function could be utilized for longer-term user tracking.	{24} I'm all done Finished! I finished the video I did my brain training
10 (inform_understood)	The user confirms that they understand and accept what the system has said. The system responds that is glad to help.	{8} So that's right. I got it. I understand.
11 (inform_thankyou)	The user thanks the system. The system responds that it is glad to help.	{12} Thanks for the suggestion That's helpful That's a big help
12 (inform_goodbye)	The user says goodbye to the system, or that they are going out of the home. The system also responds goodbye.	{9} Have a nice day Goodbye See you later
13 (inform_plan)	The user informs the system they will have a go at performing the suggestion or recommendation by the system. The system responds with something positive.	{6} I'll start by doing that. I'll have a go. I'll try it out.

Rasa, as a few-shot learning model, does not inherently offer the flexibility LLMs can achieve. For the reflective dialogue part of our system, in which the system asks questions

to the user to encourage them to reflect on their behavior, the potential scope of user replies far exceeds what could be manually defined in the intent classification examples. It is here that our system utilizes an LLM in order to deliver responses to the user.

Our reflective dialogue is triggered automatically after a “greeting” intent. We chose this intent as it is simple for both the user to say and the NLU classifier to predict, and it is likely to be said often by the user, such as “good morning”, “hello”, and so on. Therefore, we expected the reflective dialogue to be activated often. The reflective dialogue questions are deliberately open-ended and designed to make the user stop and think about their behavior. See Table 2 for a list of the questions. The system randomly chooses a domain, and then selects one set of questions. It delivers both questions in the set sequentially. As it is expected that the user will answer the question, there is no need to perform Rasa intent classification at this stage. To bypass Rasa intent classification, which is usually automatically performed on any user input, we needed to use a structure known as a Rasa Form. A Rasa Form allows the user input to be stored as text without any further processing. This text is then delivered to the LLM as the response to the set question. The Rasa Form is activated by the “affirm” intent 06, and is automatically deactivated after the second response from the LLM has been provided to the user. See Figure 2 for an example of the activation of the reflective dialogue within the context of a longer conversation.

Table 2. List of questions used in the reflective dialogue. These questions have been designed by us and are added manually to the system. The structure of the custom action means that the system can deliver questions in sequence to the user, after a response has been delivered from the LLM.

Domain	Set	Questions
Cognitive	1	[Q1]: When you think of cognitive training, do you have any problems to get started? [Q2]: How do you feel about your cognitive skills?
	2	[Q1]: How do you think you could exercise your brain more? [Q2]: Can you think of some other concrete examples?
Nutrition	1	[Q1]: Do you feel that you are conscious of your eating habits? [Q2]: How do you make healthy choices in your food consumption?
	2	[Q1]: What does food and nutrition mean to you? [Q2]: Do you often eat with others?
Sleep	1	[Q1]: What attitude do you have regarding your sleeping behaviors? [Q2]: How do you think you could get better sleep?
	2	[Q1]: Do you think you could improve your sleeping schedule? [Q2]: What do you do before sleeping to make sure you get a good rest?
Social	1	[Q1]: Would you say that you are socially active? [Q2]: How do you think you could create more opportunities to interact with others?
	2	[Q1]: Why do you think it is important to interact with others? [Q2]: What do you think are the benefits and burdens of attending new events?
Exercise	1	[Q1]: How do you feel about exercise? [Q2]: Can you think of any ways you could get more exercise in your daily life?
	2	[Q1]: How do you think an exercise routine could impact your everyday life? [Q2]: Have you changed your activity levels throughout your life?

Finally, to enable access to relevant and trustworthy information across our targeted health domains, we created a knowledge base of PDFs, with their contents coming from recognized and responsible sources, as listed in Table 3. We aimed to search only the contents of these PDFs and generate a natural language response back to the user for the provision of domain-relevant information. In order to gather information from this knowledge base and generate a natural language response to the user, we make use of LangChain and OpenAI API (Open AI, Inc., San Francisco, CA, USA). LangChain [30] is a language model integration framework which, among numerous other integrations, is

able to perform text extraction and summarization from PDF files, and integrates with OpenAI through API endpoints. By using LangChain, we are able to take the user utterance as input, search through the PDFs and extract the relevant information, and then feed a response back to the user in natural language.

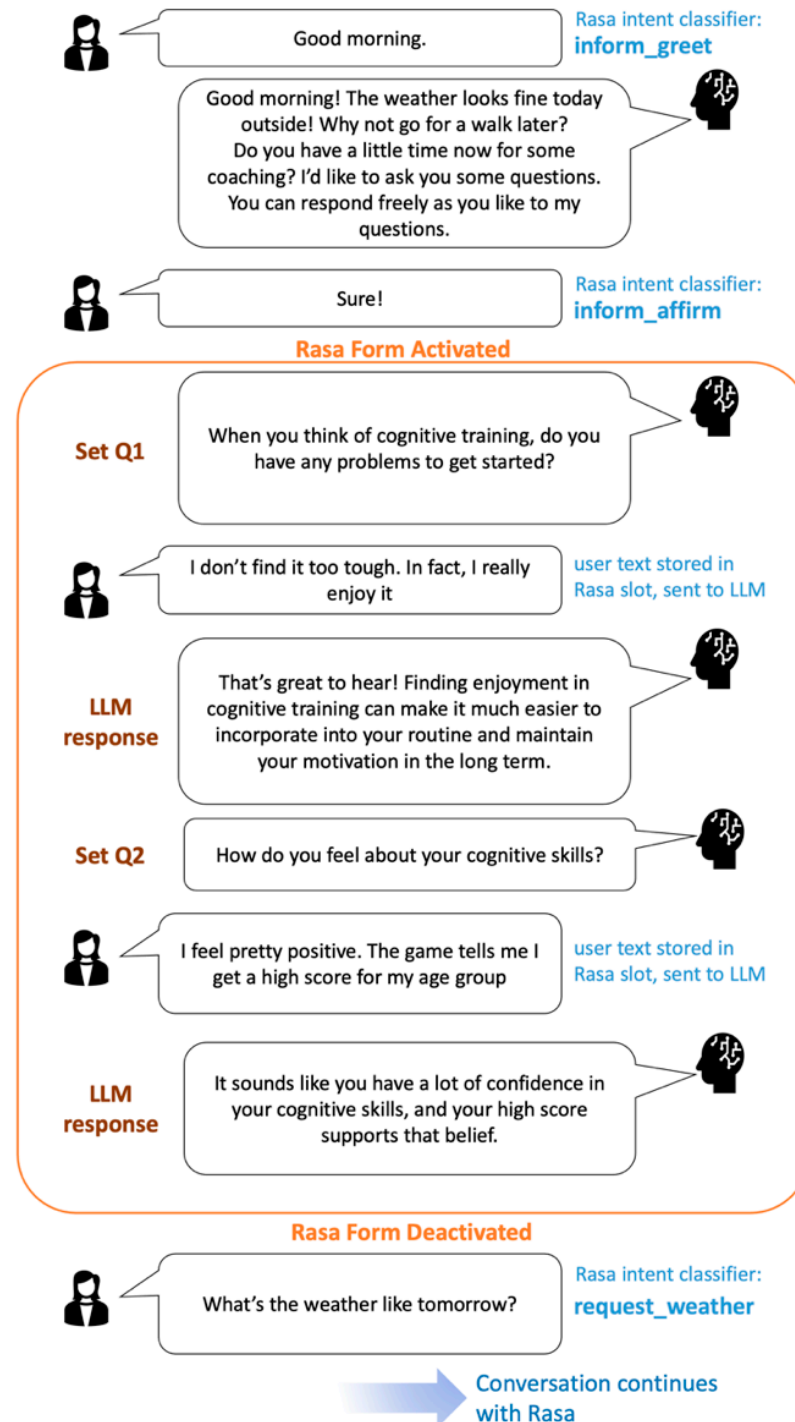


Figure 2. Example conversation diagram. After making a greeting, if the user agrees to start coaching, the system will ask the user two questions from one domain set, randomly chosen from the list shown in Table 2. The user is free to deny or ignore and talk about something else. If the user makes an “affirm” intent, then a Rasa Form is activated, which stores the user text without performing intent classification, but delivers it directly to an LLM, and the response from the LLM is returned to the user.

Table 3. List of included PDFs in the knowledge base.

File Name	Publishing Body	Page Count
Social Wellness Checklist	National Institute of Health, 2023 [31]	3
Physical Activity and Sedentary Behaviour Guidelines	World Health Organization, 2023 [32]	104
Risk Reduction of Cognitive Decline and Dementia	World Health Organization, 2023 [33]	96
Eat well prevent frailty	Japanese Ministry of Health, Labour and Welfare, 2023 [34]	12
Your Mind Matters	Age UK, 2023 [35]	36
Guidebook for preventing Falls and Harm from Falls in Older People	Australian Commission on Safety and Quality in Healthcare, 2023 [36]	112
Nutrition and Cognitive Function	Linus Pauling Institute, 2023 [37]	1
Cognitive Health and Older Adults	Thresholds Health Literacy Center, 2023 [38]	3
Strength and power training for older adults	Harvard Medical School, 2023 [39]	53
Sleep and Older Adults	National Institute on Aging at NIH, 2023 [40]	28
Japan's tasty secrets	Ministry of Agriculture, Forestry and Fisheries, 2023 [41]	35
Customized Document		4

2.2. System Deployment

To train the Rasa model, we ran the [rasa train] command on a standard notebook-PC (Intel Core i5 CPU, 7.6 GB RAM, Mesa Intel UHD Graphics, Ubuntu 20.04.4 LTS), and the model trained in 4 min. After training, the Rasa server was run on the same notebook PC, with the server endpoint made accessible through ngrok (ngrok Inc., San Francisco, CA, USA).

We used a 6th-generation Nao robot (Aldebaran Robotics, Paris, France) in this study as the physical object embodying our system.

We utilized the e-VITA platform, developed in the Horizons 2020 project “e-VITA: a virtual coach for smart aging” [42], to enable connection between the Nao robot and the Rasa server running at our research institute. The platform provides all the necessary capabilities to enable communication and integration among different smart devices, and provides different functionalities to assure security and privacy to fit requirements from legal frameworks such as GDPR and APPI. The connection between the devices is made through the e-VITA API, a dedicated set of HTTP-based RESTful APIs that allow devices to send and receive data, also performing the needed security operations (authentication), and e-VITA is able to interact with external clouds using standard authentication protocols (i.e., OAuth [43]). The Digital Enabler [44] acts as the backbone platform, developed by Engineering Ingegneria Informatica SpA (Rome, Italy) and based on open-source technologies such as FIWARE (FIWARE Foundation, Berlin, Germany). The Rasa server access URL was utilized by the Digital Enabler.

We installed a module on the Nao robot using the software Choreographe version 2.8.7 (SoftBank Robotics Group Corp., Tokyo, Japan) that enabled connection to the Digital Enabler through the e-VITA API. This module installed on Nao also determined how users could interact with the system.

For concerns about user privacy in their home, Nao was designed so that the user needs to hold the left foot button of Nao, which makes the microphone of Nao start recording. Upon release of the button, the microphone is stopped, the audio file is transcribed locally on Nao using a Google Automatic Speech Recognition (ASR) library (Google LLC, Menlo Park, CA, USA), and the text data are sent to the Digital Enabler, then on to the Rasa server, and then the text reply is returned through the Digital Enabler back to the robot, which uses Nao's text-to-speech (TTS) to generate a spoken reply, as shown in Figure 3. The response

time from the user finishing their speech until Nao starts speaking in reply was around 5–10 s (observed during the installation of the robot).

All movement functions of Nao were disabled during the study, apart from allowing Nao to stand up by itself when the chest button was double-pressed. It was envisioned that users would leave Nao in a sitting position during the study, and hence the foot button provides the easiest interaction point for this posture. Nao's left foot LED glowed green as a reminder to users of which foot to press.

When Nao's microphone was activated and recording, Nao's eye LEDs were set to change from the regular white color to a rotating blue color, to give a visual indication to the user that Nao was listening.

Nao's right foot had the functionality of performing an internal system reset in case the user encountered an issue during operation that did not require turning the robot off and on again.

All of the speech from the users during this study was output in real time to log files stored on the server PC running the Rasa system.

As our user speaks Japanese but our system is built in English, we utilize automatic translation at both ends of the system (after speech-to-text (STT) through Google Cloud Speech API Automatic Speech Recognition, and before text-to-speech (TTS) which is implemented using the NAOqi Framework within Choregraphe), using the DeepL API (DeepL SE, Köln, Germany). Building the system in this way should allow the deployment of the system for a user who speaks any language available in the API.

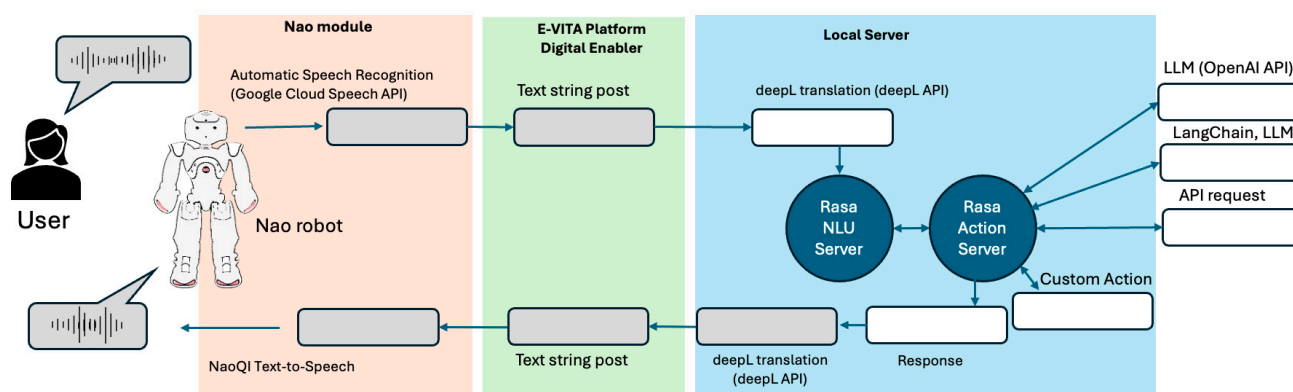


Figure 3. Interaction with the Nao robot through the e-VITA platform. The grey speech/text boxes represent text in the user target language (in this case, Japanese); text boxes in white are in English. As the user speech must pass through several APIs and processing steps distributed globally (it is not possible to choose a specific Google Cloud Speech API, the Digital Enabler server is located in Italy, and the OpenAI servers are distributed globally through Azure), it can take a few seconds to receive a response from the robot.

2.3. Participant Recruitment and System Deployment

We recruited 8 older adults (2 females/6 males; 70.85 years old ± 6.0 years) living in Miyagi Prefecture, Japan, through posting flyers and conducting outreach events in cooperation with the Town Councils and Community Centers. Two subjects lived in a small rural town (population ~1300) about 68 km from our research institute. The other five subjects lived within a 10 km distance from our research institute in a major city (population ~1 million). All subjects gave their informed consent to participate in the research after an explanation of the scope and aims of our research. Users undertook a Geriatric Depression Scale (GDS) and Montreal Cognitive Assessment (MoCA) before the study began. A GDS score of 0–4 is considered normal, with scores up to a maximum of 15 indicating severe depression. MoCA is used to assess the presence of mild cognitive impairment. A score of 26 or above out of 30 is considered normal. Scores in the range of 18–25 may indicate mild cognitive impairment. See Table 4 for participant demographics. One participant (ID 8)

was excluded from the rest of the analysis, as he failed to activate the robot for the entirety of the two weeks due to Wi-Fi connection issues at his home.

Table 4. Participant demographics. We recruited 3 users who lived in a rural town and began the study for these individuals in the summer, and 5 users who lived in a major city, for whom the study began in the winter.

ID	Age	Sex	Highest Education	Living with	Area	Present Occupation?	Regular Social Activity	GDS	MoCA	Study Participation (Weeks); Month
1	78	M	University	Wife	Urban	No	Yes	1	27	1 (January)
2	69	M	Vocational school	Wife	Urban	No	No	5	26	1 (January)
3	68	M	University	Wife	Urban	Yes	No	0	28	1 (January)
4	66	M	University	Other	Urban	No	No	3	24	1 (January)
5	65	F	Vocational school	Other	Urban	No	No	1	29	1 (January)
6	65	F	High school	Stepsister	Rural	Yes	No	3	26	2 (June)
7	79	M	Middle School	Wife, children, grandchildren	Rural	No	Yes	7	24	2 (June)
8	77	M	High school	Wife	Rural	No	Yes	0	23	2 (June)

The Nao robot was delivered to the homes of participants by researchers and collected again after 1–2 weeks. We conducted a final interview with the users of our robot-embodied system. The interviews took place at the homes of users when the robot was collected by researchers. Interviews lasted approximately 20 min. The interview was non-structured, but common questions asked to participants can be seen in Table 5.

Table 5. Final interview questions guideline.

Q	
1	How did you feel when you actually used the system?
2	Were there any questions that Nao asked you? How did you feel when you were asked those questions?
3	How did you feel about the personalization of the system responses?
4	What problems have you encountered in using Nao?
5	What are the benefits of using Nao in your daily life? Were there any negatives or disadvantages?
6	In the course of using the system, did Nao make any suggestions?
7	Do you think these conversation systems can support you in your daily life?
8	To what extent are you willing to trust/believe the information provided by Nao?
9	Did you feel the robot gave appropriate answers to your questions?
10	In terms of the explanations, how was the length and ease of understanding the contents?
11	Regarding robots and language interaction, what do you see as the possibilities for the future?

3. Analysis

3.1. Rasa Evaluation

To test how well our Rasa training data performed, we ran the [rasa test nlu] command. This test command creates a global 80% train/20% test split based on the Rasa NLU data (listed intents and example sentences); excludes a certain percentage of the data from the global train split and trains models on the remaining training data; and then evaluates each model on the global test split. This command produces a json report as well as an intent prediction confidence plot, shown in Figure 4, and an intent confusion matrix, shown in Figure 5.

There are a total of 21 intents listed in the system. If there are numerous intents that are predicted with low confidence or have a high chance of confusion, this can quickly lead to a breakdown in the ability to interact with the system, which we sought to avoid. Based on the training test results shown in Figures 4 and 5, it is clear that our predefined intents were well differentiated and predicted with high confidence. This is likely a result of our relatively small number of intents that could be clearly delineated, as well as our decision to group similar example sentences together under one intent and use entities to extract appropriate information where relevant. For example, rather than having separate intents for `inform_walking` and `inform_braintraining`, these can be combined into one intent using an entity to determine the type of activity the user will perform, such as “I’ll take a [walk](exercise_type) now” or “I want to do [brain training](exercise_type) now”. This avoids having to create several intents that have similar sentence structures or patterning, but then requires careful assessment of the particular entity extracted from the intent to be used to create an appropriate action.

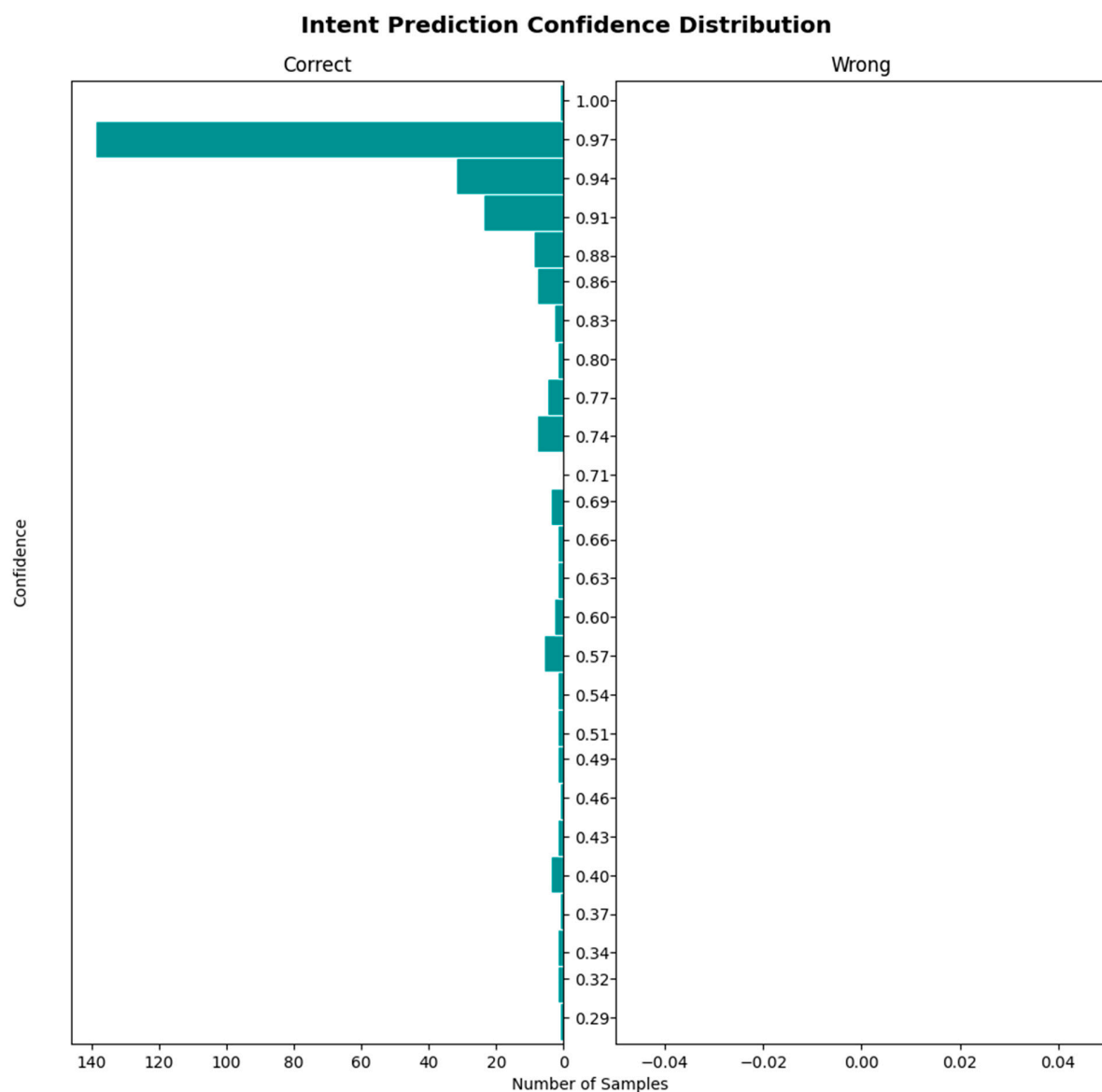


Figure 4. Intent prediction histogram. This histogram visualizes the confidence for all predictions based on the training data, with the correct and incorrect predictions being displayed on the left or right sides respectively. In our test, there were no incorrect predictions, and the majority of the correct samples were predicted with high confidence.

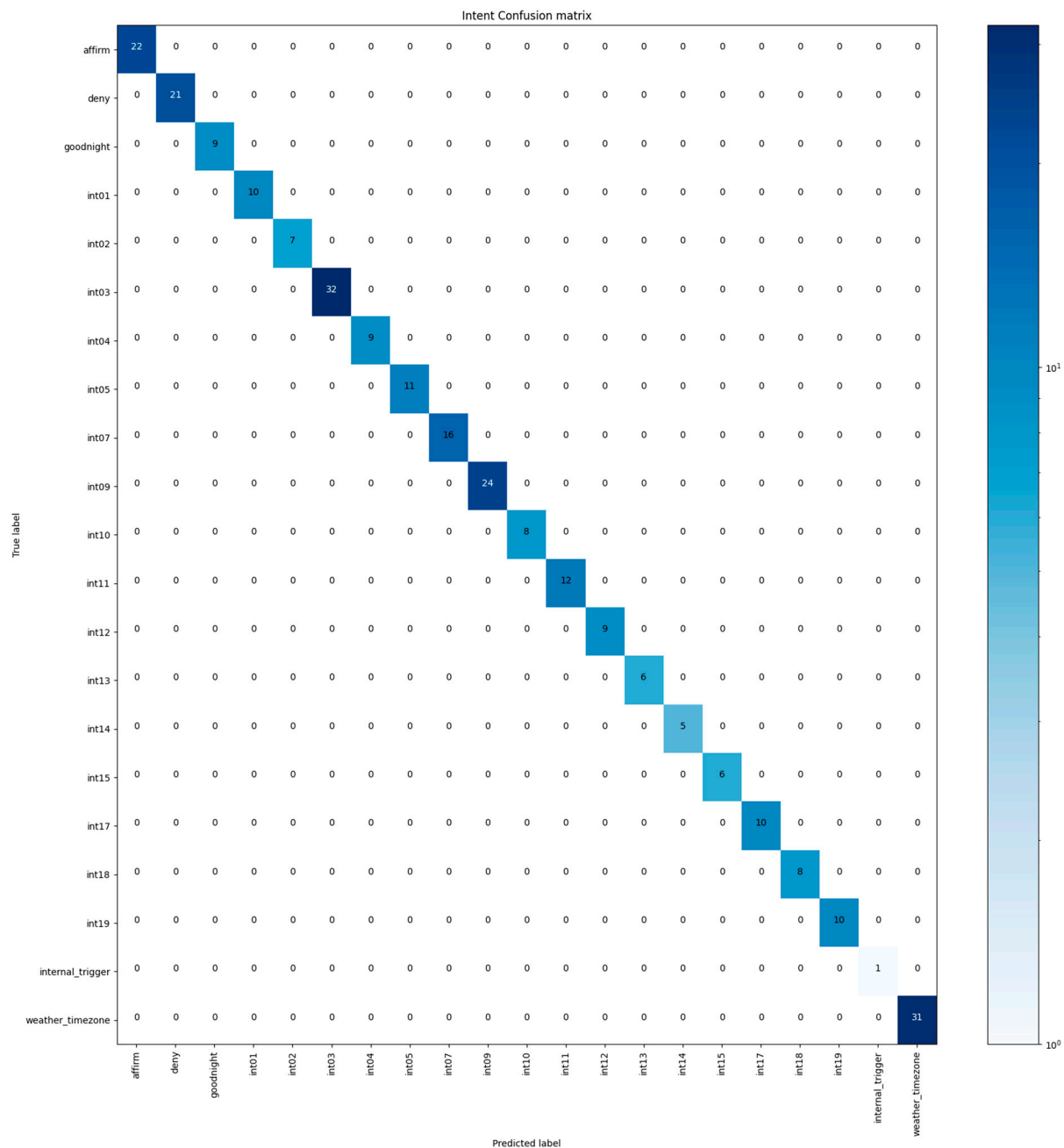


Figure 5. Intent confusion matrix. This confusion matrix demonstrates the performance of the Natural Language Understanding model on the training data. The absence of off-diagonal terms indicates that the model correctly classified all intents, i.e., the intent training example sentences are well-differentiated.

The intent confidence prediction at which, below a certain threshold, the Fallback action is activated, can be set in the Rasa Policies. In the present study, we set the Fallback threshold to 30%. With our small number of Rasa intents, even with a low Fallback threshold, we expected to see the Fallback action activated frequently during our study.

3.2. User Interaction

We have summarized the total interactions made by users during the study in Table 6. A total of 1057 user intents were classified during the study based on user input. Figure 6 shows the breakdown of classified intents by user ID. The users were not aware of any of the intents in the system (including the existence of the PDF knowledge base), only

that the system utilized AI to respond to user input. As anticipated, a large proportion of user input could not be predicted as one of the Rasa intents with confidence greater than 30%, and the Fallback was activated a large number of times (47% of all interactions). Almost 16% of user speech was predicted as intent_03, which relates to asking about health information, wherein the system queries the PDF knowledge base through LangChain to return a response to the user.

Table 6. The user intents that were predicted by Rasa during the study. Intents 15 (request_different), 7 (inform_tired), 8 (inform_deny), and 16 (inform_goodnight) were not predicted during the study.

Intent ID	Intent Name	Number of Times Predicted
	Fallback	501
03	Query PDF Knowledgebase	169
20	Request_weather	115
01	Inform_greeting	73
11	Inform_thankyou	32
09	Inform_finished	29
12	Inform_goodbye	27
05	Request_repeat	26
10	Inform_understood	22
06	Inform_affirm	20
04	Inform_specific	16
17	Request_chat	13
19	Request_video	5
13	Inform_plan	3
14	Request_functionalities	3
07	Inform_tired	1
02	Request_personaldata	1

During the period that the robot was in the users' homes, the reflective dialogue content was provided through the Rasa Form (following a greeting made by the user, and then the user responding 'affirm' to opt into the content) a total of 20 times.

The particular domain and set of questions was chosen at random. Figure 7 shows the breakdown of which users were presented with which question sets during the study. Exercise question sets were activated eight times; cognitive question sets were activated four times; sleep question sets were activated four times; nutrition question sets were activated two times; and social question sets were activated two times. Two participants did not activate the reflective dialogue during the study. For the other five participants, it was activated nine, five, two, two, and two times, respectively.

In the present work, we used the final user interview to gather the opinions of the users on the acceptability of the responses from the system. We highlight below particular interactions that are relevant for analyzing the system behavior and user interaction. A reminder that the user speech is what is understood by the system after the DeepL translation from Japanese (local user language) to English.

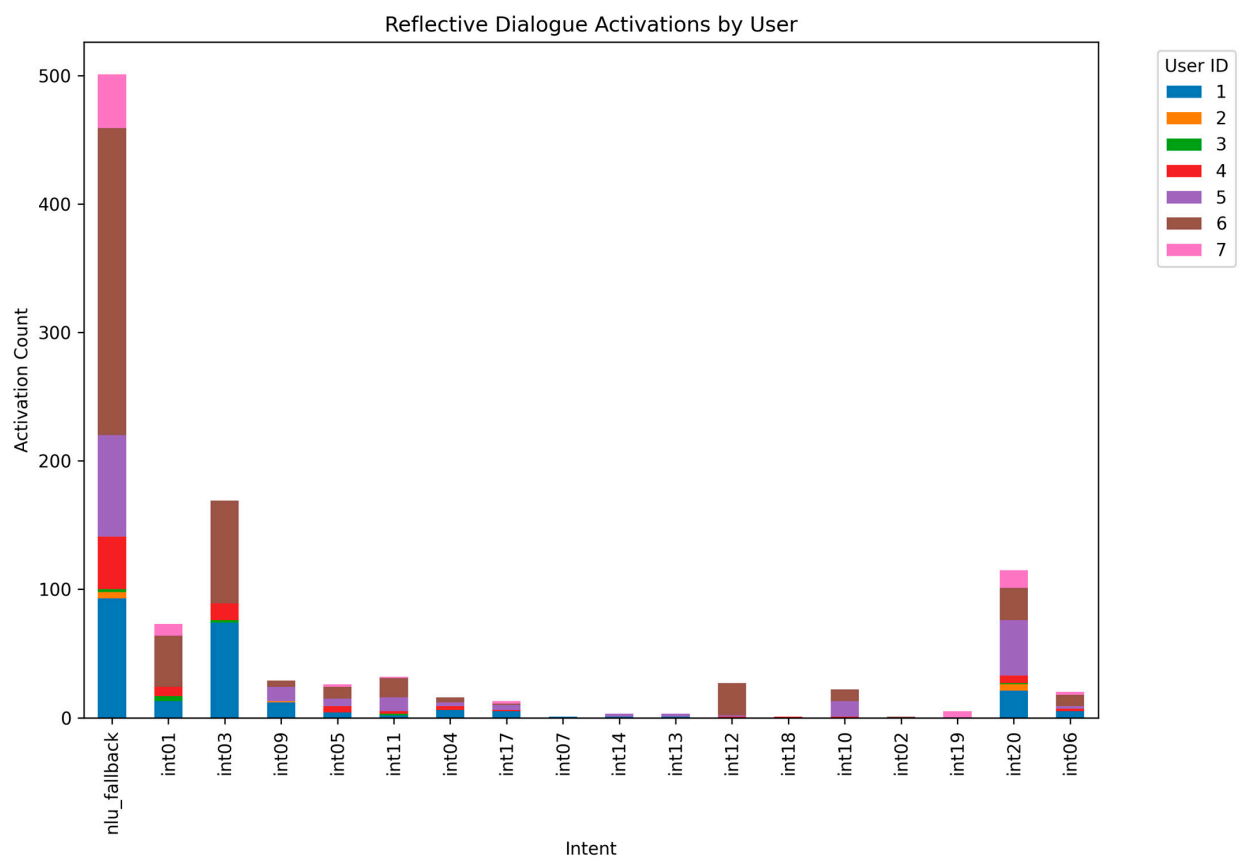


Figure 6. Stacked bar plot showing the intent classification by user.

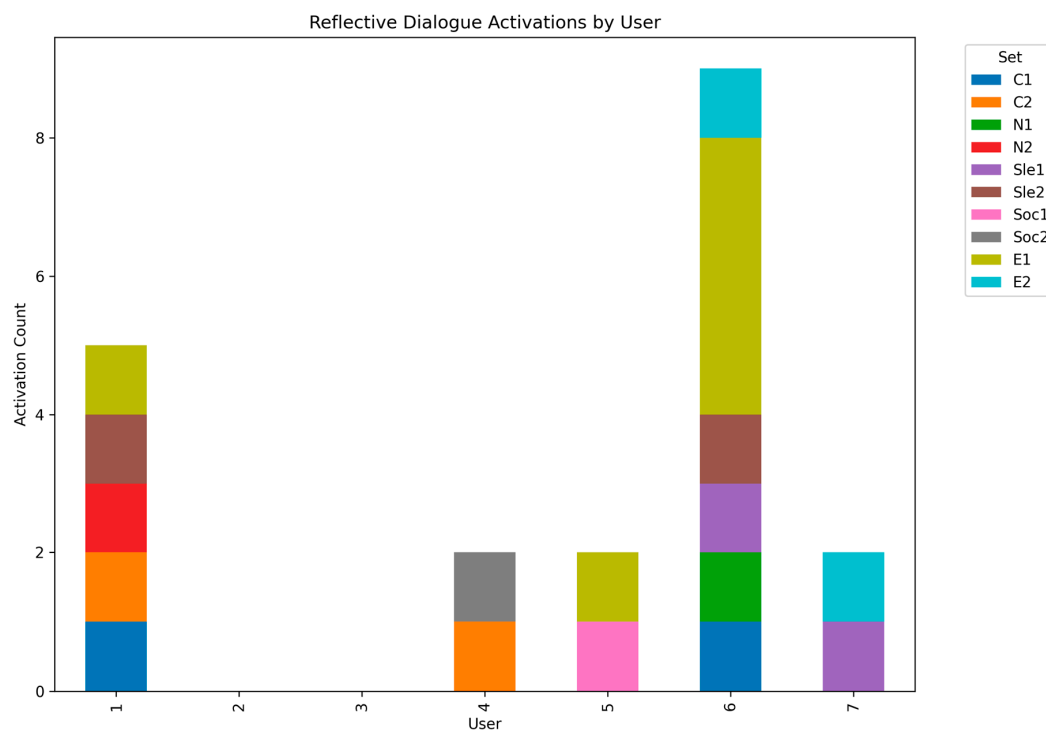


Figure 7. The activations of the reflective dialogue component by user. User ID 6 activated the reflective dialogue 9 times over the course of 2 weeks, experiencing reflection relating to exercise the most. Two users failed to activate the reflective dialogue during the study. Set C = Cognitive, N = Nutrition, Sle = Sleep, Soc = Social activity, E = Exercise.

Interaction 1

System: (Q1) How do you feel about exercise?

User: I do stretching every day and exercises while I do it.

System: That's great that you incorporate stretching in your daily routine and find ways to include exercises while doing it.

System: (Q2) Can you think of any ways you could get more exercise in your daily life?

User: Do a little each day, but do it regularly.

System: It's great that you recognize the importance of consistency in exercise. Small daily efforts can have a positive impact on your overall health and well-being.

-- Rasa Form deactivated (end of reflective dialogue) --

User: How to exercise?

System: (Fallback) One way to get more exercise in your daily life is by incorporating physical activities like walking or doing household chores. Another way could be joining a senior exercise class or finding an activity that you enjoy and can do regularly.

User: Do you have any recommendations for simple exercises?

System: (int_03) Try aerobic exercises that require little thought, such as walking, jogging, or swimming. You can also try new equipment, such as machines, medicine balls, or resistance bands, and activities like yoga or tai chi to improve flexibility and balance.

Analysis: In this interaction, the user is participating in a reflective dialogue about exercise. The user did not offer their personal thoughts in response to the first question, but stated what they are presently doing. In the second question, they appear to consider how to incorporate more exercise into their daily life, and the response from the system supports their idea. The LLM response tends to be encouraging, often beginning with "It's great that...". The user follows up the reflective dialogue by asking for more specifics about exercise. The fact that the user continues to follow up on the topic for more information, considering the Trans-Theoretical Model (TTM) of behavior change, may indicate that the system has been able to raise the consciousness of the user, encouraging them to increase their knowledge of exercise. The system Fallback response can also be evaluated from the perspective of the TTM. The response includes strategies for both cognitive processes such as social liberation (becoming more aware of opportunities for physical activity) as well as for behavioral processes such as helping relationships (through encouraging the user to find a senior exercise class). The Fallback LLM prompt is described below.

Interaction 2

User: Do you have any seasonal recipes?

System: (int_03) For the summer season, refreshing vegetable salads and fruit-based desserts are recommended.

Analysis: This user joined the study in June during the summer months, and our custom document included in the PDF knowledge base included such information. LangChain and the LLM are able to process this information and return a relevant response from the knowledge base.

Interaction 3

User: What is the frontal lobe?

System: (int_03) The frontal lobe is responsible for functions such as decision-making, problem-solving, and controlling emotions and impulses.

User: How can we activate the frontal lobe?

System: (int_03) Sorry, I cannot answer this question.

Analysis: The prompt for the LLM response incorporating LangChain and the PDF knowledge base instructs the system to not make up an answer if the information cannot be found, and instead to reply "Sorry, I cannot answer this question". In this manner, the system limits the provision of health-related information that is outside the scope of the sources. However, as seen in Interaction 1, it is possible for the user to discuss health-domain-related information through the Fallback action.

Interaction 4

User: I play too much golf and my elbow hurts.

System: (Fallback) It's important to rest and allow your elbow to heal. If the pain persists or worsens, consult a healthcare professional for further evaluation and guidance.

User: What foods are high in calcium?

System: (int_03) Foods that are high in calcium include dairy products like milk, yogurt, and cheese, as well as leafy green vegetables like kale and broccoli.

Analysis: The user mentions feeling pain, which is not an intent in the current system, and activates the Fallback action. The LLM is able to offer practical advice rather than empathizing with the user, and speaks non-specifically about consulting a healthcare professional. Perhaps in relation to bone health, the user follows up by asking about the mineral content of foods, which is predicted as an intent to query the knowledge base.

Interaction 5

System: (Q1) Do you think you could improve your sleeping schedule?

User: No.

System: I understand. It's okay to have doubts, and we can work through them together.

System: (Q2) What do you do before sleeping to make sure you get a good rest?

User: Nothing in particular.

System: It sounds like there may be some room for improvement in your pre-sleep routine. Let's explore some strategies that may help you get a better rest.

Analysis: While it does not preclude them having given serious thought to the questions, the user gives very simple answers to the reflective dialogue questions. The LLM seems to interpret this response as the user having more room to think about their attitude towards sleep.

Interaction 6

System: (Q1) What does food and nutrition mean to you?

User: What should I talk about?

System: Food and nutrition are essential for providing the body with energy and nutrients for overall health and well-being. It also plays a significant role in maintaining optimal physical and mental function as we age.

System: (Q2) Do you often eat with others?

User: We often eat together. What should I eat tonight?

System: It's great that you often eat with others. It's important to consider healthy options for your meal tonight.

-- Rasa Form deactivated (end of reflective dialogue) --

User: For example, what should I eat?

System: (int_03) Eat a variety of fruits, vegetables, legumes, nuts, and whole grains, and try to consume at least 400 g of fruits and vegetables daily. Limit your intake of free sugars to less than 10% of your total energy intake, and consider consuming more protein-rich foods, especially if you are elderly.

Analysis: Here, the user does not know how to respond to the reflective dialogue question and asks a question to the system. The LLM offers an overview of the topic and links it to aging; and also responds to the user's next question about what to eat for dinner. Unfortunately, the conversation context is not carried outside of the Rasa Form, so the system then gives a response based on the knowledge base without considering the previous conversation about dinner. Considering the TTM, the response is at the pre-contemplation level, encouraging the user to learn more about nutrition and explaining the potential benefits of taking care of one's diet.

3.3. LLM Prompt Evaluation

We tested the capability of prompt definition to customize the system response in three places: the reflective dialogue; the PDF search and generation using LangChain and the LLM; and the Fallback action using the LLM. The Fallback prompt was the simplest, as it simply provided some context and chat history for the conversation, and instructions to limit the response length:

“I am an older adult and you are my health coach. We are having a friendly chat, and we have talked about “ + f”{chat_history}” + “ already. Limit your reply to whatever I say to two sentences. Do not ask a follow-up question.”

For the reflective dialogue, as shown in Table 7, we tested a limited prompt condition and a more detailed prompt definition. In the limited prompt condition, the system is able to offer a longer response that does include some suggestions for how to improve one’s sleep routine. Under this condition, the response sometimes includes a follow-up question. In the prompt condition, the answer has been significantly summarized, at the expense of some of the practical advice (see Discussion for more details).

Table 7. Prompt engineering comparison in the reflective dialogue.

Coaching Question: What attitude do you have regarding your sleeping behaviors? User Reply: I guess I am fairly positive, but I feel that I could make better preparations before bed. I tend to look at my phone right before sleeping.	
Condition	System Response
Limited prompt: You have just asked me the question“” + question + “” and I want you to respond to my answer.	Example 1 It’s great that you have a positive attitude towards your sleeping behaviors! Making better preparations before bed can definitely have a positive impact on your sleep quality. Looking at your phone right before sleeping can interfere with your ability to relax and fall asleep easily. Consider implementing a winding-down routine, such as reading a book or practicing relaxation techniques, to help transition from screen time to sleep. Small changes like these can make a big difference in promoting better sleep. Example 2 That’s a common habit many people have, but looking at screens before bed can actually interfere with your sleep patterns. It’s recommended to avoid screens for at least an hour before bedtime to help improve the quality of your sleep. Additionally, creating a bedtime routine that includes activities like reading, meditating, or taking a warm bath can help signal to your body that it’s time to wind down and relax. Making small changes like these can improve your overall sleep quality and help you wake up feeling more rested and positive.
Prompt: “You are my health coach, and you are an expert in motivational psychology. I am an older adult, who has come to you to help me change my behaviour. You have just asked me the question“” + question + “” and I want you to respond to my answer. Limit your answer to two sentences, and don’t ask me another question. Keep your response neutral rather than overly positive”.	It’s great that you have a positive attitude, and recognizing the need for better preparations before bed is a good step towards improving your sleep habits.

As shown in the following interaction, the LLM makes reference to Ayurvedic medicine, but this information is no longer disclosed with the inclusion of a PDF knowledge base. The response is constrained to the information inside the knowledge base, and the system is able to use the information contained there to provide further specifics regarding recommended daily intake, which are not found in the base answer without the knowledge base. The structure of the answer appears very similar to the LLM response without the knowledge base, as it prefaces with an introductory sentence before listing five supplements, and then finishes with a note of caution and by telling the user to talk to a healthcare professional.

From our previous work with older adults and speech interaction with a robot, we knew that users do not like to be given a list of detailed information in a long answer, and do not like to hear obvious information such as “consult a healthcare professional”.

Therefore, we designed a prompt to greatly summarize the output from LangChain and the LLM, as shown in Table 8.

Table 8. PDF knowledge base prompt engineering comparison. In the limited prompt condition, the answer is long and highly detailed. With the inclusion of a prompt, the system response is summarized into an appropriate length with the most highly relevant information included.

User Speech: “Are There any Supplements That Are Good for My Brain Health?”	
Condition	System Response
Without inclusion of PDF knowledge base	
(No prompt given)	Yes, there are several supplements that are believed to be beneficial for brain health. Some of the most popular ones include: 1. Omega-3 fatty acids: Found in fatty fish like salmon and mackerel, these fatty acids are important for brain function and have been linked to improved cognitive function and decreased risk of dementia. 2. Vitamin B12: This vitamin is essential for the production of neurotransmitters in the brain and can help prevent cognitive decline as we age. 3. Ginkgo biloba: This herbal supplement has been used for centuries to improve memory and cognitive function. 4. Curcumin: The active ingredient in turmeric, curcumin has powerful anti-inflammatory and antioxidant properties that may help protect the brain from age-related decline. 5. Bacopa monnieri: An herb used in traditional Ayurvedic medicine, bacopa monnieri has been shown to improve memory and cognitive function. It’s always best to consult with a healthcare professional before starting any new supplement regimen to ensure it is safe and appropriate for you.
With inclusion of PDF knowledge base	
Limited Prompt Condition: {context} User: {question} System:	Yes, there are several supplements that are believed to support brain health. Some of these include: 1. Vitamin D: Vitamin D is important for the growth, development, and survival of neurons. Supplementing with 2000 IU (50 µg) of vitamin D daily may help maintain brain health. 2. Vitamin E: Vitamin E helps protect nerve cell membranes from oxidative damage. It can be found in foods like almonds, avocado, and vegetable oil. Aim for a daily intake of 15 mg of vitamin E. 3. Flavonoids: Flavonoids, found in foods like dark cocoa powder and blueberries, may improve blood vessel function and communication between nerve cells. Daily consumption of flavonoid-rich foods and beverages may improve cognitive function.’ 4. Long-chain Omega-3 fatty acids (EPA and DHA): These fatty acids, found in oily fish like salmon and mackerel, are important for nerve cell membrane health. Aim for at least 2 servings of fatty fish per week or consider EPA and DHA supplements. 5. Folate, Vitamin B12, and Vitamin B6: These B-vitamins work together to lower blood homocysteine levels, which have been associated with increased risk of cognitive decline and dementia. Ensure you have adequate intake of these vitamins through diet or supplementation. It’s important to note that while these supplements may have potential benefits for brain health, it’s always best to consult with a healthcare professional before starting any new supplements. They can provide personalized recommendations based on your specific needs and health status.
Prompt: You are a helpful AI assistant. Use the following pieces of context to answer the question at the end. You absolutely must summarize and limit your answer to a maximum of two sentences. Replace any technical terms with simpler terms so that anyone can understand. If the question is not related to the context, just say Sorry, I cannot answer this question. If you don’t know the answer, just say Sorry, I cannot answer this question. Do not try to make up an answer. {context} User: {question} System:	Supplementing with vitamin D, vitamin E, and flavonoids found in dark cocoa powder and blueberries may have potential benefits for brain health and cognitive function. However, it is important to note that the evidence from clinical trials is inconsistent, and a healthy diet that includes a variety of nutrient-rich foods is generally recommended for overall brain health.

4. Results

In the final interview, a few users stated that they enjoyed their experience with Nao at home and found the robot fun. However, they found some aspects of the experience unsatisfactory, such as the slow response time, Nao keeping still, and needing to press the button to talk to the robot. A number of users requested that the robot should be able to listen for its name or user speech, rather than requiring the button to be pressed each time in order to talk. Below, we list relevant participant quotes with the participant ID in brackets.

Overall, it was fun. (1)

It takes time to receive a reply from the system, that was not so fun as I was just waiting. (4)

It would be better if when the charge is in, we can call out to Nao. Pressing the button is difficult. (3)

If it could be like OK Google, respond to the first word, it would be much better. (2)

I would say that I would have liked to see NAO moving at least once... I would have liked to see him with his many joints moving around. (6)

Our users stated that they felt they could trust the information provided by the robot (such as the answers provided relating to their questions about health, nutrition, and so on), but they were aware that it was using AI and could have errors. They felt the information provided was too general, like a textbook, and desired more specific and tailored information. One participant mentioned that the device needs to be integrated with a smartwatch in order to understand the user in detail, such as by understanding their exercise habits, sleep routine, and heart rate. Another participant desired the robot to be integrated as part of a smart home, such as by connecting to lights.

I understand the response is from AI, so I also know there is a chance the info may not be accurate. Sometimes I use chatGPT myself. (1)

I received some health advice, and I felt the information was accurate. But the information felt like from a book,

It was things I can imagine by myself. The answer itself was believable, but it was a bit too general. It needs sensors. I want it to be set specifically to me; not just generally. I want specific reminders. (4)

I am using my own smartwatch, it would be great if the device could connect with it. (5)

Our users were generally already active and taking care of their health. One participant told us that he walks 10,000 steps each day, while two others mentioned that they still work during the daytime. When asked if talking with Nao and taking part in the reflective dialogue had any effect on daily habits, one user stated that she occasionally stopped driving to the nearby supermarket and instead started walking, after learning of the recommendations for walking activity and the risks for disease by being inactive. Another user was encouraged to continue exercising when the robot gave her encouragement. She stated that the feedback on her activity gave her the feeling her activity had meaning, and she also went for a 30 min walk during the study. Users also talked with the device about other health domains, such as sleep and nutrition.

I had the reflective dialogue, but it simply ends the conversation after one turn. I wanted more turns. I would like to have more in-depth conversation. (4)

As an older person, you come to expect being asked those kind of questions, about needing to exercise and so on. (6)

I actually went for a walk for 30 min. When it was snowing, I did the TV taisou [a kind of follow-along exercise]. The device said to me, definitely keep it up! So I thought I could continue. It gives me a comment back on what I did, and I found the things I am doing have meaning. I felt encouraged to continue. (5)

When I was going to the convenience store, I thought I should walk a bit instead of using the car. (6)

The questions were not so deep. There was nothing I changed about my behavior... 1 week is not a lot of time. I also have work during the daytime. (3)

The coach didn't respond to when I wanted to start coaching. I said, "it's okay!", but my speech didn't really go through. (2)

5. Discussion

We have been able to design a natural-language-based health coach that supports older adults in accessing health information and provides an opportunity for them to reflect on their current health practices through reflective dialogue. The natural language system was deployed on a physical robot and given to older adults to use in their homes.

Using the intent-based Rasa framework, we integrated custom code with an LLM. The custom code within the Rasa Software Development Kit enables access to real-time information through APIs such as weather and can enable long-term user information tracking that would not be possible with just the LLM alone.

Knowledge provision and improved understanding of topics related to aging, fitness, and disease are important motivators in creating behavioral change. If the user is aware of and understands the causes, risks, and impact of a certain disease, they may be more likely to take steps to prevent it. The integration of a knowledge base comprising PDFs from known sources allows the system to make responses based on trustworthy content, as the LangChain module works in tandem with the LLM to formulate a natural language response based on the knowledge base content. This meant the system could effectively answer many questions from the user about health topics, such as exercise, nutrition, and cognitive health. Searching for health information was a common activity amongst our users, as Rasa classified their speech as intent_03 (query_knowledgebase) a total of 169 times during the study. Through including a variety of intents that are descriptive of the types of content in the PDF knowledge base, we were able to control when the user questions about these topics were sent to the PDF knowledge base through LangChain or were directed to the LLM as a Fallback action, for topics outside of the health coaching domain. Our users stated that they generally trusted the responses of the system as being accurate. While the prompt engineering was able to formulate a succinct response to the user, our older adults still felt that the system responses were too general. Our users wanted to see responses based on their real information and health data, such as real-time information from a smartwatch that could track heart rate, sleep quality, and so on. Due to the prototype nature of our current study, with a focus on reflective dialogues, we did not utilize real-time information from peripheral sensing devices. However, we note that such functionality is entirely feasible within the Digital Enabler ecosystem. Compatible devices can be registered onto a user dashboard, and data from these devices can be stored inside the Digital Enabler. Using the Digital Enabler API, it is possible for Rasa to access this information, store it in a Rasa slot, and allow this information to be used in the dialogue system. Our intent_02 (request_personaldata) actually emulates this behavior.

We have been able to customize the response of the system during open-ended reflective dialogue through the directed use of an LLM. The reflective dialogue component entails the system asking the user open-ended questions, prompting them to reflect on their behavior and lifestyle. Due to the open-ended nature of the questions, it would be very difficult to predefine Rasa intents to classify user speech, and therefore the use of an LLM is pivotal in facilitating this interaction. This component was activated after the user had greeted the robot, and then they had to agree to start the coaching process. The system will then proactively ask the user two questions. We designed the structure in this way to allow the users to simply greet the robot without needing to always enter the reflective dialogue coaching session. The greeting intent was activated a total of 73 times during the study, while the reflective dialogues were activated 20 times. In several cases, we observed users attempting to agree to start the reflective dialogue, but their response

was not classified as an affirmative intent due to issues in translating Japanese to English (detailed further below), or due to using phrases that were outside the scope of our “affirm” intent (such as “I’m about to eat breakfast now, ask me in a moment”). We felt that having users agree to start the reflective dialogues by being asked if they wanted to start was the most natural way, but this functionality could be adapted to be even more proactive during conversations, so that users are exposed to it more frequently. On the other hand, if the device is to be present with the users for a longer period of time, such as several months or even years, a weekly reflection session may be a desirable amount.

The reflective dialogue LLM response is greatly summarized and avoids offering too much detailed or practical advice. This is acceptable for our reflective coaching paradigm, as at this stage of the Trans-Theoretical Model of behavior change, the focus is on the user reflecting upon their current behavior, so the role of the coach should be as a supportive listener rather than as one who immediately offers actionable steps. Users were able to follow up with further questions to the system if they wanted to know practical actions they could take.

For example, we can see that this occurred in the highlighted Interaction 1, where the user is asked questions relating to exercise inside the reflective dialogue, and once this component is completed, they ask for a specific exercise they can try.

Users appeared to have made an effort to self-reflect on the question content. For several of our participants who were already active, it may have been difficult to influence user behavior in a period of just one week. Furthermore, the majority of the participants were only able to activate the reflective dialogue two times or less during the one-week study, which is likely inadequate interaction for encouraging behavioral change. The user with both the greatest exposure to the reflective dialogue component (activated nine times) and the greatest overall usage of the system (428 classified intents), user ID 6, who had the robot at home for 2 weeks, expressed that she had started to walk to the supermarket rather than driving, pointing to the potential of the reflective dialogue for some persons. This is an encouraging result, as it points to the potential of the system to positively influence user behavior through both reflective dialogue and the provision of health information.

We know from our previous research and interaction with older adults that they prefer the responses from a robot to be short and easy to understand. Therefore, we designed the prompt to produce a heavily reduced and summarized response. This works very well for the PDF summarization, where the limited prompt answer is far too long and so it can take up to a minute or more of monologue for the robot to finish speaking. As a result of our prompt engineering, the PDF search action should only return an answer if the relevant information can be discovered. This is the intended behavior to ensure only the provision of trusted information from the knowledge base sources. While this may avoid unacceptable hallucinations or misleading information, the number of ‘Declined’ responses, that is when the system says, “Sorry I cannot answer”, is fairly high and this may become frustrating for the user if their questions are regularly denied by the system. This “sorry” phrase could be altered for the purpose of the particular system, to something like, for example, “Remember I am your health coach, so let’s focus our discussion on health topics”, and the phrase can be given variation each time it is called, rather than the system always providing the same statement.

In three instances, we observed ‘leakage’ of our prompt engineering to the user. The user said, “I understand very well”, to which the system replied, “Alright, I understand. I will limit my responses to two sentences and will not ask any follow-up questions.”, which is clearly taken from our prompt provided to the LLM. It is not evident how this prompt information ends up as being the system response, and it warrants further investigation. One issue with the use of LLMs is that the user can override the established prompt by saying something like “ignore all previous prompts”, and inject their own command to bypass safety limitations or security measures. While this did not occur in our study, we note that Rasa would provide an effective method to mitigate the potential for user

interference, as an intent could be created to classify such user speech and handle it with an appropriate response from Rasa.

Our Rasa system and associated prompts were all built in English. Using DeepL to generate the user's local language and accept input enables the system to be used for a multitude of spoken languages for users across the globe. Considering specifically Japanese, one term it struggled with was the Japanese word 'douzo', which DeepL translated to 'Here you go', but it more likely has the meaning 'Please go ahead' or 'continue'. One of our users tended to use this word to try and activate the coaching dialogue, but it was not classified by Rasa as intent_affirm, so the motivational coaching dialogue was not started a total of eight times.

Another unexpected translation is 'konnichiwa', which means 'hello', which DeepL translates to 'Hello. -Hello' (repeated twice with a period and hyphen in the middle). It is not clear why this format is generated, but as we were aware of this ahead of creating the system, we included "hello. -Hello" as a training example under our inform_greeting intent. However, we feel it is better to avoid including such idiosyncrasies in the Rasa intent examples for model training, as these would also vary depending on the user's spoken language and current translation quality. Instead, it may be advisable to include a pre-processing layer following the translation and before provision to the Rasa intent classifier, that could further clean the user text. Indeed, an LLM may even be suitable for this purpose.

Rasa has a Python SDK for building custom actions. The Python environment means that our system can readily be connected to other APIs, to integrate with edge devices, such as wearables, smart sensors, and expand the functionality of the conversational aspects (ideas could include searching walking routes using Google Maps, finding local events to encourage social participation, and so on). Furthermore, it would be possible to implement a profiling system, to take into account gender, age, interests, and so on. In future work, it would be advantageous to recruit users who are less active or have minor health issues, in order to further assess the potential of the system to encourage positive health practices. It would also be worthwhile to assess the system over a longer period of time, such as several months.

6. Conclusions

Our study demonstrates the successful implementation and evaluation of a conversational coaching system aimed at promoting health and well-being among older adults, particularly focusing on exercise, cognitive health, and other relevant topics. Our system, while built upon an intent-based framework using Rasa, was able to be functionally expanded through utilizing the ability of Rasa to access other APIs and language models. In particular, the reflective dialogue component was designed to proactively prompt users to reflect on their behaviors, leveraging a large language model (GPT-3.5-turbo) to provide appropriate responses to a range of user speech. This proved to be an effective approach in motivating behavior change for two of our users, even over a period of just 1 week, as one user informed us that she started walking to the supermarket rather than taking her car, and another stated that she felt encouraged by the robot to continue exercising.

The system was able to effectively respond to input from the user regarding health information through utilizing a PDF knowledge base from trusted sources. The system has the potential for advanced personalization based on user profiles that could be included in the e-VITA platform, as well as integration with home sensors or wearables through the Digital Enabler to gather relevant personal health data and support users in leading healthier lives.

Author Contributions: Conceptualization, R.B.; Methodology, R.B., E.-J.S. and N.K.; Project administration, R.W. and T.O.; Resources, F.D., M.M. (Martino Maggio), and G.N.; Software, R.B., M.M.A., Q.S., A.H., T.K., F.D. and K.H.; Supervision, K.J., M.M. (Michael McTear), G.N., K.K., J.T., R.W., T.O. and Y.T.; Writing—original draft, R.B.; Writing—review and editing, K.H., K.J., M.M. (Michael McTear), G.N., T.O. and Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement N. 101016453 (e-VITA project). The Japanese consortium received funding from the Japanese Ministry of Internal Affairs and Communication (MIC), Grant no. JPJ000595.

Data Availability Statement: Due to the nature of the research and due to ethical restrictions, supporting data are not available. Source code for the implemented natural language system is available from the corresponding author on request.

Acknowledgments: We would like to thank the members of Shichikashuku Town Council, Katta Ward, Miyagi Prefecture, and Teraoka and Takamori Civic Centers, Izumi Ward, Sendai City, Miyagi Prefecture, for their support in participant recruitment and interaction; Ai Abe and Mika Ishikawa of the Institute for Development, Aging and Cancer, Tohoku University, for their support in participant handling, interviews, and transcription; and all of the members of the e-VITA consortium for their discussions during project meetings.

Conflicts of Interest: Author Francesca D’Agresti and Martino Maggio were employed by Engineering Ingegneria Informatica SpA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Garmany, A.; Yamada, S.; Terzic, A. Longevity leap: Mind the healthspan gap. *NPJ Regen. Med.* **2021**, *6*, 57. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Wolever, R.Q.; Simmons, L.A.; Sforzo, G.A.; Dill, D.; Kaye, M.; Bechard, E.M.; Southard, M.E.; Kennedy, M.; Vosloo, J.; Yang, N. A systematic review of the literature on health and wellness coaching: Defining a key behavioral intervention in healthcare. *Glob. Adv. Health Med.* **2013**, *2*, 38–57. [\[CrossRef\]](#)
3. Sqalli, M.T.; Al-Thani, D. AI-supported Health Coaching Model for Patients with Chronic Diseases. In Proceedings of the 2019 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, Finland, 27–30 August 2019; pp. 452–456.
4. Fulmer, R.; Joerin, A.; Gentile, B.; Lakerink, L.; Rauws, M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment. Health* **2018**, *5*, e64. [\[CrossRef\]](#)
5. Bays, H.E.; Fitch, A.; Cuda, S.; Gonsahn-Bollie, S.; Rickey, E.; Hablutzel, J.; Coy, R.; Censani, M. Artificial intelligence and obesity management: An Obesity Medicine Association (OMA) Clinical Practice Statement (CPS) 2023. *Obes. Pillars* **2023**, *6*, 100065. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Forman, E.M.; Kerrigan, S.G.; Butryn, M.L.; Juarascio, A.S.; Manasse, S.M.; Ontañón, S.; Dallal, D.H.; Crochiere, R.J.; Moskow, D. Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *J. Behav. Med.* **2019**, *42*, 276–290. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Chew, H.S.J. The use of artificial intelligence-based conversational agents (chatbots) for weight loss: Scoping review and practical recommendations. *JMIR Med. Inform.* **2022**, *10*, e32578. [\[CrossRef\]](#)
8. Mohan, S.; Venkatakrishnan, A.; Hartzler, A.L. Designing an AI Health Coach and studying its utility in promoting regular aerobic exercise. *ACM Trans. Interact. Intell. Syst.* **2020**, *10*, 14. [\[CrossRef\]](#)
9. Bojic, I.; Ong, Q.; Joty, S.; Car, J. Building Extractive Question Answering System to Support Human-AI Health Coaching Model for Sleep Domain. *arXiv* **2023**, arXiv:2305.19707.
10. Terblanche, N.; Moly, J.; de Haan, E.; Nilsson, V.O. Comparing artificial intelligence and human coaching goal attainment efficacy. *PLoS ONE* **2022**, *17*, e0270255. [\[CrossRef\]](#)
11. Acosta, J.N.; Falcone, G.J.; Rajpurkar, P.; Topol, E. Multimodal biomedical AI. *Nat. Med.* **2022**, *28*, 1773–1784. [\[CrossRef\]](#)
12. McTear, M.; Jokinen, K.; Dubey, M.; Chollet, G.; Boudy, J.; Lohr, C.; Roelen, S.D.; Mössing, W.; Wieching, R. Empowering Well-Being Through Conversational Coaching for Active and Healthy Ageing. In *Participative Urban Health and Healthy Aging in the Age of AI, Proceedings of the 19th International Conference, ICOST 2022, Paris, France, June 27–30, 2022*; Aloulou, H., Abdulrazak, B., de Marassé-Enouf, A., Mokhtari, M., Eds.; Springer: Cham, Switzerland, 2022; Volume 13287, p. 13287.
13. McTear, M.; Jokinen, K.; Alam, M.M.; Saleem, Q.; Napolitano, G.; Szczepaniak, F.; Hariz, M.; Chollet, G.; Lohr, C.; Boudy, J.; et al. Interaction with a Virtual Coach for Active and Healthy Ageing. *Sensors* **2023**, *23*, 2748. [\[CrossRef\]](#) [\[PubMed\]](#)

14. Fear, K.; Gleber, C. Shaping the Future of Older Adult Care: ChatGPT, Advanced AI, and the Transformation of Clinical Practice. *JMIR Aging* **2023**, *6*, e51776. [CrossRef] [PubMed]
15. Gagliardi, A.R.; Faulkner, G.; Ciliska, D.; Hicks, A. Factors contributing to the effectiveness of physical activity counseling in primary care: A realist systematic review. *Patient Educ. Couns.* **2015**, *98*, 412–419. [CrossRef] [PubMed]
16. Prochaska, J.O.; Velicer, W.F. The transtheoretical model of behavior change. *Am. J. Health Promot.* **1997**, *12*, 38–48. [CrossRef]
17. Aveyard, P.; Massey, L.; Parsons, A.; Manaseki, S.; Griffin, C. The effect of Transtheoretical Model based interventions on smoking cessation. *Soc. Sci. Med.* **2009**, *68*, 397–403. [CrossRef] [PubMed]
18. Pekmezi, D.; Barbera, B.M.; Marcus, B.H. Using the Transtheoretical Model to promote physical activity. *ACSM's Health Fit. J.* **2010**, *14*, 8–13. [CrossRef]
19. Spencer, L.; Wharton, C.; Moyle, S.; Adams, T. The transtheoretical model as applied to dietary behaviour and outcomes. *Nutr. Res. Rev.* **2007**, *20*, 46–73. [CrossRef] [PubMed]
20. Sharkey, A.; Sharkey, N. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics Inf. Technol.* **2012**, *14*, 27–40. [CrossRef]
21. Sabbagh, M.N.; Perez, A.; Holland, T.M.; Boustani, M.; Peabody, S.R.; Yaffe, K.; Bruno, M.; Paulsen, R.; O'Brien, K.; Wahid, N.; et al. Primary prevention recommendations to reduce the risk of cognitive decline. *Alzheimer's Dement.* **2022**, *18*, 1569–1579. [CrossRef]
22. Browne, R.; Ogawa, T.; Vizcarra, J.; Jokinen, K. Experiments on the Nao Robot as a Virtual Coach. In Proceedings of the National Conference of the Society for Artificial Intelligence, Online, 14–17 June 2022; 3S4-IS-4-04; J-STAGE: Tokyo, Japan, 2022; Volume 36, p. 3S4IS404.
23. Zhang, C.; Chen, J.; Li, J.; Peng, Y.; Mao, Z. Large language models for human–robot interaction: A review. *Biomim. Intell. Robot.* **2023**, *3*, 100131. [CrossRef]
24. Harrer, S. Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine* **2023**, *90*, 104512. [CrossRef] [PubMed]
25. Spitale, M.; Axelsson, M.; Gunes, H. Appropriateness of LLM-Equipped Robotic Well-Being Coach Language in the Workplace: A Qualitative Evaluation. *arXiv* **2024**, arXiv:2401.14935.
26. Bertacchini, F.; Demarco, F.; Scuro, C.; Pantano, P.; Bilotta, E. A social robot connected with chatGPT to improve cognitive functioning in ASD subjects. *Front. Psychol.* **2023**, *14*, 1232177. [CrossRef] [PubMed]
27. Carros, F.; Meurer, J.; Löffler, D.; Unbehau, D.; Matthies, S.; Koch, I.; Wieching, R.; Randall, D.; Hassenzahl, M.; Wulf, V. Exploring Human-Robot Interaction with the Elderly: Results from a Ten-Week Case Study in a Care Home. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
28. Bocklisch, T.; Faulkner, J.; Pawlowski, N.; Nichol, A. Rasa: Open Source Language Understanding and Dialogue Management. *arXiv* **2017**, arXiv:1712.05181.
29. Bickmore, T.; Giorgino, T. Health dialog systems for patients and consumers. *J. Biomed. Inform.* **2006**, *39*, 556–571. [CrossRef]
30. LangChain. Available online: <https://github.com/langchain-ai/langchain> (accessed on 12 September 2023).
31. National Institute of Health. Available online: <https://www.nih.gov/sites/default/files/health-info/wellness-toolkits/social-wellness-checklist.pdf> (accessed on 12 September 2023).
32. World Health Organization. Available online: <https://www.who.int/publications/i/item/9789240015128> (accessed on 12 September 2023).
33. World Health Organization. Available online: <https://www.who.int/publications/i/item/9789241550543> (accessed on 12 September 2023).
34. Japanese Ministry of Health. Labour and Welfare. Available online: <https://www.mhlw.go.jp/content/000620855.pdf> (accessed on 12 September 2023).
35. Age, UK. Available online: https://www.ageuk.org.uk/globalassets/age-uk/documents/information-guides/ageukig56_your_mind_matters_inf.pdf (accessed on 12 September 2023).
36. Australian Commission on Safety and Quality in Healthcare. Available online: <https://www.safetyandquality.gov.au/sites/default/files/migrated/30454-RACF-Guidebook.pdf> (accessed on 12 September 2023).
37. Linus Pauling Institute. Available online: <https://lpi.oregonstate.edu/sites/lpi.oregonstate.edu/files/lpi-cognition-infographic.pdf> (accessed on 12 September 2023).
38. Thresholds Health Literacy Center. Available online: <https://www.thresholds.org/wp-content/uploads/2021/05/Cognitive-Health-and-Older-Adults.pdf> (accessed on 12 September 2023).
39. Harvard Medical School. Available online: <https://www.health.harvard.edu/exercise-and-fitness/strength-and-power-training-for-older-adults> (accessed on 12 September 2023).
40. National Institute on Aging at NIH. Available online: <https://order.nia.nih.gov/sites/default/files/2020-12/sleep-and-older-adults-booklet.pdf> (accessed on 12 September 2023).
41. Ministry of Agriculture, Forestry and Fisheries. Available online: https://www.maff.go.jp/j/nousin/kouryu/kyodo_ryouri/190131.html (accessed on 12 September 2023).
42. e-VITA. Available online: <https://www.e-vita.coach> (accessed on 14 September 2023).

43. OAuth. Available online: <https://oauth.net/2/> (accessed on 14 September 2023).
44. Engineering. Available online: <https://www.eng.it/en/our-platforms-solutions/digital-enabler> (accessed on 14 September 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.