Large Language Models and RAG Approach for Conversational Coaching - Experiments for Enhancing e-VITA Virtual Coach

Kristiina Jokinen[1],*, Kristina Deryagina[2], Giulio Napolitano[3] and Abrar Hyder[3]

[1] AI Research Center, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto, Tokyo 135-0064, Japan
[2] University of Siegen, Adolf-Reichwein-Straße 2a, 57076 Siegen, Germany
[3] Institut für Angewandte Informatik (InfAI), Goerdelerring 9, 04109 Leipzig, Germany

Abstract

We give an overview of the dialogue modelling research conducted for the e-VITA Virtual Coach, developed in the EU-Japan collaboration project e-VITA with the aim of supporting personalised interaction with older adults, on topics related to active healthy living. The system integrates Rasa Conversational AI and RAG-based LLMs and addresses two main requirements for AI-based applications: how to provide reliable information, and how to maintain smooth personalised conversation. We focus on motivational coaching dialogues and qualitative user evaluation of the prototype system, which tailors dialogue interaction to develop user interest and to support their daily activities.

Keywords

e-VITA Virtual Coach, Dialogue system, Large Language Models, RAG, LangChain, reliable interaction

CEUR-WS.org/Vol-3906/paper2_ALTRUIST.pdf

1. Introduction

With the release of OpenAI's ChatGPT in November 2022, Large Language Models (LLMs) have made generative AI known to a large audience and excelled in several language-based tasks ranging from code writing to translation and from question-answering to chatty communication. LLMs also offer attractive solutions to build social robotics applications, which have so far suffered from clumsy and cumbersome interaction. However, much research has been conducted to alleviate the notorious "hallucinations" of the LLMs in order to improve their performance for various interactive tasks. This is especially important for practical social robotics applications which not only need smooth communication capabilities but also provide reliable information for users in a personalised manner.

In this article, we describe our work on one such practical application, the e-VITA Virtual Coach, and focus especially on the e-VITA Dialogue Manager. The system has been developed in the EU-Japan collaboration project e-VITA (www.e-vita.coach/homepage) to provide interactive coaching for active and healthy ageing. The project aims at improving subjective well-being of older adults in Europe and Japan, and supporting quality of life in daily activities through integrating advanced living technologies, robots, AI, and conversational interaction in the context of smart homes, personalised coaching, and companionship. The main characteristics of the project are described in [1, 2, 3].

The article is structured as follows. Section 2 describes the framework for our dialogue modelling, the e-VITA Coaching system, Section 3 describes the motivational coaching approach with personalised coaching cycle, and Section 4 discusses the three dialogue management approaches used in the e-VITA Dialogue Model. Section 5 provides an evaluation of the system, and Section 6 discusses ethical aspects related to reliable and sustainable coaching models. Finally, Section 7 concludes the paper.

ALTRUIST, BAILAR, SCRITA, WARN 2024: Workshop on sociAL roboTs for peRsonalized, continUous and adaptIve aSsisTance, Workshop on Behavior Adaptation and Learning for Assistive Robotics, Workshop on Trust, Acceptance and Social Cues in Human-Robot Interaction, and Workshop on Weighing the benefits of Autonomous Robot persoNalisation. August 26, 2024, Pasadena, USA

*Corresponding author.

kristiina.jokinen@aist.go.jp (K. Jokinen); Kristina.Deryagina@siegen.de (K. Deryagina); malicatu@googlemail.com (G. Napolitano); hyder@infai.org (A. Hyder)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. Overview of the e-VITA Coaching System

Previous work on coaching systems within healthcare and eldercare domains has focussed on technical aspects of the systems, or on behaviour change models and intervention methods using scripted dialogues. Only a few explicitly address dialogue aspects [4, 5]. Relational agents [6] engage users in a natural interaction by relating to their emotional state, while Zhou et al. [7] present a neurosymbolic goal summarizer to help health coaches keep track of the goals. An overview of virtual coaching for older adults is presented e.g. in [8].

2.1. Specification of the system

The main goal of the e-VITA Virtual Coach development has been *trustworthiness*, i.e. to design and develop a coaching system that can sustain older adults' well-being in everyday activities. For this, it is important that the coach can interact in natural language, and facilitate interaction through intuitive conversational coaching strategies. The Virtual Coach should also converse with the older adults in their native language, and the multilingual project setting of the EU and Japan is thus challenging: the Coach needs to manage five different languages (English, German, French, Italian, and Japanese) and also take into account intercultural differences such as presentation styles, politeness, and coaching strategies. Finally, the system is designed to be used in smart living environments so the integration of various devices and sensors is pertinent within the e-VITA platform, including their impact on the interaction. This is another challenge for the technical management of the e-VITA Coach which aims to support interoperability and standardization, and protect user privacy issues.

In this context, coaching dialogues have been designed to support and stimulate the daily activities of older adults, as well as provide information and advice in several domains, including exercising, sleeping, healthy eating, mental health, and spirituality. When interacting with a particular user, the Virtual Coach is required to take into account the personal profile of the user, including their cultural background, age, gender, health situation as well as their readiness level for motivational activity changes, which the content experts considered relevant features for personalising the Coach responses. The physical environment surrounding the user is exposed to the dialogue system from sensor readings, while emotion detection informs the system about the likely psychological status of the interlocutor. The challenges related to multilinguality are addressed by translation services and language-specific fine-tuning: user utterances and system responses are translated between the user's language and English, used as lingua-franca in system-internal processing and domain-specific databases, whereas language-specific aspects of greetings and database answers are manually fine-tuned and coded in the Dialogue Manager rules.

2.2. The e-VITA Dialogue Manager

The e-VITA Dialogue Manager implements dialogue management of the e-VITA Coach, taking care of the interpretation of the user's intent and the generation of an appropriate system response. It uses the Rasa Open-Source Framework (https://rasa.com) to build the conversational agent. To capture a wider array of user intents and to make system responses more fluent, the e-VITA Dialogue Manager takes advantage of the GPT models, and incorporates ChatGPT in the system via an extension of the Retrieval Augmented Generation (RAG) approach.

Figure 1 shows the integration of the Rasa dialogue model in the e-VITA coach framework. The coach platform, called the Digital Enabler, coordinates the information flow and also contains an API hub which includes information from the sensors, the emotion recognition system, motion trackers, and environmental sensors. It also contains the user's private information (such as BMI, exercise, fitness, hobbies, eating habits), which is used to drive the dialogues through different individualized pathways.

When a user starts a conversation with the e-VITA Coach, the audio signal is captured by the microphone of the interface agent (Nao, Gatebox, Android, Daruma) and sent to the Google speech recognizer (https://cloud.google.com/speech-to-text), except for the Gatebox interface, which relies on

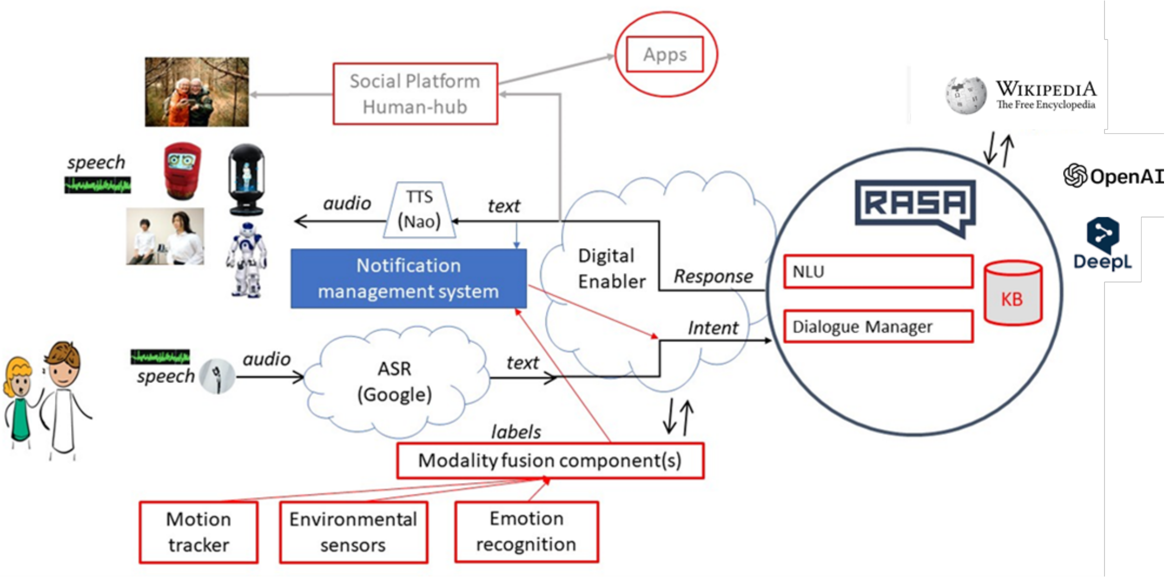


Figure 1: Integration of the Rasa dialogue model in the e-VITA coach framework.

the Microsoft Azure speech engines (https://portal.azure.com/). The transcribed text is subsequently managed by Digital Enabler, which uses DeepL translator (https://www.deepl.com/en/translator) to translate the text into English (if not originally in English), and sends the text to the Rasa dialogue service. The NLU component of the Rasa system processes the input and uses the DIET (Dual Intent and Entity Transformer) to recognize dialogue intents and entities from the user input. Rasa's Dialogue Manager component produces the next system response with the help of stories (model dialogues) and exploiting the TED (Transformer Embedding Dialogue) policy. The response text is returned to Digital Enabler, which translates it into the user's native language and sends this further to the selected coaching device. The robot agent then speaks the text to the user using its built-in TTS technology.

The system also has access to external services (Wikipedia, News, and Weather services), and it includes domain-specific knowledge bases (KB), i.e. json-type databases and collections of relevant documents for the RAG approach, compiled by the project's domain experts. Digital Enabler also provides a notification management system which enables external triggers to start dialogues based on sensor status or timing. For instance, the user can be prompted at 8 am every morning to exercise outside, if the weather is not too hot.

An example of a coaching dialogue is given in Table 1. It shows the recognized user intent and the AI agent's response intent (to be executed as a system action like knowledge-base search, request to external services, request to the robot utterancece). As the dialogues are tailored according to the user's specific characteristics, Table 1 also shows some user features and their values that will guide the dialogue. The features are provided by the content experts and include interests, family relations, age, etc., and can also include more personal data depending on the privacy constraints.

3. Coaching Model

3.1. Motivational Coaching

The e-VITA coaching model is based on Motivational Coaching Theory [9, 10], which incorporates an understanding of the user's preparedness to change their life style. The user's preparedness is categorized into four *readiness levels*: non-interested, interested, ready (prepared), and maintenance, and the e-VITA Coaching Model interprets the first three levels as different *user types*, for whom the Virtual Dialogue Coach should tailor the coaching dialogues accordingly, in order to motivate behavioural change to the next level. These are:

Table 1
Sample dialogue

Speaker	Dialogue	User Intent or Robot Action	User Features

User	I am interested in improving my life by regular exercise. Can you help?	request_coaching-exercise	interest-level=0,
			hobby=gardening,
			gender=male

Robot	Absolutely, I'd be happy to help. Do you have specific goals in mind, like improving strength, flexibility, or overall fitness? Let's work together to create a plan that suits your interests and lifestyle.	agree_coaching-exercise	

User	I have problems with motivation. I cannot make myself do even a little bit of sport every day. Can you help me?	request_finding-motivation	interest-level=0,
			hobby=cooking,
			gender=female

Robot	Yes, setting small realistic goals and tracking your progress can help with motivation. You can also try doing activities you enjoy or joining a group for accountability and camaraderie.	agree_finding-motivation	

- Not interested person: Mainly "working on the idea". The goal is to *make the user aware* of the need for behavior change. Increase the user's knowledge of health behaviors and help them understand the benefits of behavior change and the risks of not doing so. Also ask users to express their feelings.
- Interested person: The main focus is on "giving the idea". The goal is to *motivate the user* and give the user more confidence in their ability to change their behavior. Identify obstacles to behavior change. Continue to increase knowledge of health behaviors.
- Prepared person: Main focus is on "working on behavior". The goal is to *clarify the action plan*, i.e. have the user make a concrete and achievable plan and be determined to implement the behavior. Once the user has started even a little, follow up with them so that their determination does not waver. To encourage the use of self-monitoring, rewards such as points, and social support.

For the current implementation of the coaching dialogues, particular attention is given to the first two stages, non-interested and interested, operating under the assumption that the user intends to modify their behavior. As for the "maintenance" level users, it was assumed that they would benefit more from a system that monitors their performance and accurately tells what is the current situation, encouraging them to continue or notifying them of any anomalies. Since the project focussed on behavioural change and personalised coaching, dialogue management for the "maintenance" level users was not included in the system development.

The coach provides customized advice for older adults on domains such as active healthy living, nutrition, and exercises. Motivational coaching is delivered through a structured process known as the coaching cycle, a conceptual framework which guides users through steps designed to foster self-awareness, encourage self-reflection, and facilitate the adoption of healthier lifestyle choices. For each of the three focussed user types, different "stories", i.e. sample dialogues, are manually defined by the content experts. The stories represent prototypical dialogues that a user with a given profile and readiness level can conduct with the coach, and they are included in the document pool from which data is retrieved in order to generate the system response. For instance, if the user is "interested", then the sample stories for an interested user are selected and processed together with the content documents. A sample story with different options for the user utterances is given in Table 2.

The interested/non-interested user types were extensively studied in the experiments exploring usability of ChatGPT and the RAG approach in the project-specific tasks such as generating personal fitness and exercise plans [11]. While ChatGPT is capable of generating exercise plans, its use as a

Table 2
Sample stories related to lighting conditions.

Speaker	Intent/Action	Content
User	request-lighting-condition-practices	I am worried about the lighting conditions inside my home, could you tell me about the best practices for lighting for older adults?
		I am concerned about the illumination levels inside my house, could you inform me what are the best guidelines for lighting for seniors?
		I am anxious about the brightness quality inside my home, could you explain to me what are the best standards for lighting for older people?
		I am nervous about the light intensity inside my house, could you advise me what are the best tips for lighting for elder adults?
		I am troubled by the lighting situation inside my home, could you teach me what are the best methods for lighting for senior citizens?
Robot	action-suggest-proper-lighting	Absolutely. Good lighting is important for older adults to see clearly and avoid falls. Here are a few things you can do to improve lighting —
User	satisfied-from-bot-answer	That's helpful, thank you.
		Thank you, that's very helpful.
		That's useful, thanks.
		That's informative, many thanks.
		That's beneficial, thank you.
		That's valuable information, thank you very much.
Robot	action-ask-again	You're welcome. If you have any other questions or concerns about indoor lighting or any other topic related to indoor safety, I'd be happy to help.

direct generator of the plans is not recommended for the e-VITA Coach, since the produced plans are general and not necessarily suitable for older adults. Moreover, the paper discusses the caveats and ethical considerations of LLMs in general, which are mportant to be taken into account when developing reliable and trustworthy coaching systems. Instead, to achieve personalisation and reliability on the coaching advice, the e-VITA Dialogue Manager opted for the RAG approach where ChatGPT is embedded in the processing pipeline and its input is guided by the separate data retrieved from reliable documents (see Section 4).

3.2. Personalised Coaching Mode

The primary distinction between a personalised coaching mode and a general, non-personalised coaching mode lies in the level of customization and specificity of interactions based on the user data. In the personalised coaching mode, the system adapts its responses using detailed information about the user, while in the general mode, the system has no access to the user's private information and thus cannot tailor its responses according to the individual's unique context and needs.

The user's private information is categorized into several subdomains in the e-VITA Coach as follows:

- Personal Information: the user's login ID, name, and location
- Medical Information: the user's height, weight, age, and medical history (e.g., any illnesses or allergies)
- Daily Routine: details about the user preference for activities like grocery shopping (morning, evening, or afternoon) and whether they are independent or dependent in doing these tasks
- Eating Habits: the number of meals the user consumes per day and daily calorie intake

- Exercise: the user's preferred exercises and their readiness level for exercising
- Social Habits: details about the user living alone or with family, their social preferences (e.g., preference for socializing or solitude), pet ownership, religious affiliation.

Users can update their private information via the Evita dashboard (an interface part of the Dialogue Enabler framework). In the beginning of a dialogue session between a user and the Virtual Coach, the user has to login to the system using their unique ID, which is detected and stored in the Dialogue Manager. The user ID is then used to collect user-specific information from the Digital Enabler, thus enabling e-VITA Dialogue Manager to access the user's private information in order to personalise its responses appropriately.

By default, the system is designed to run fixed, tailor-made dialogues based on RASA's stories. It is assumed that the personalized coaching cycle enhances user engagement by leveraging detailed personal information, whereas the normal coaching cycle provides broader, less tailored responses due to the lack of user-specific data. Tables 3 and 4 illustrate the differences in responses between personalized and normal coaching cycles, highlighting the impact of user-specific data on the quality and relevance of interactions.

Table 3
Response generated without user information

Speaker	Dialogue
User	What do you think about my living condition?
Robot	test_user99, I do not know the answer.

Table 4
Personalised response generated using user information

Speaker	Dialogue
User	What do you think about my living condition?
Robot	Hello, test_user99. Living alone and feeling lonely can be challenging, especially when you don't have frequent contact with friends and relatives. It's important to find ways to stay connected and engaged with others. You might consider joining community activities or groups that interest you, which can help reduce feelings of loneliness. Additionally, maintaining a routine that includes physical activity and hobbies can improve your overall well-being. How do you feel about exploring some new activities or groups in your area?

4. RASA and RAG Dialogue Management

As discussed in Section 2, dialogue management for the e-VITA Coaching system is operationalised with the help of the Rasa Conversational AI framework, but it also utilises LLMs to produce smooth system responses, and experiments with ChatGPT in a special coaching mode. Dialogue management in the e-VITA Coach thus deals with three types of dialogues:

- Story-driven Dialogues, using typical stories designed in RASA
- LLM-assisted Dialogues, based on the RAG approach and motivational Coaching cycles
- Open-ended Conversations with OpenAI's ChatGPT.

The conversation flow is given in Figure 2. After recognition of the user intent, i.e. processing the utterance through the NLP module and classifying it into one of the available user intents, a relevant story is triggered when possible. If the confidence for triggering a relevant story is below the threshold, the system will start the LLM-assisted dialogue management, i.e. it utilises the RAG-based approach to generate an appropriate response with the reliable data source.

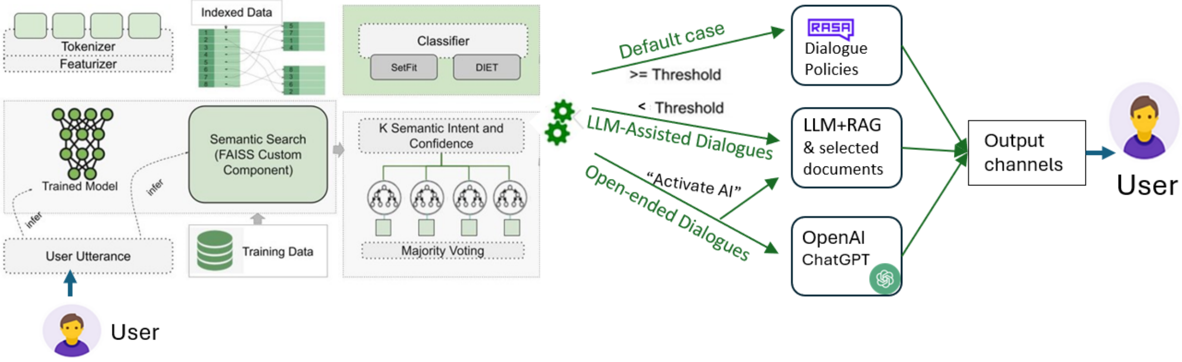


Figure 2: Dialogue management flow.

Due to the popularity of OpenAI's ChatGPT, the e-VITA system also gives the user an opportunity to experiment with it. If the user wishes to engage in direct ChatGPT conversations through the e-VITA Coach, they can activate the AI-mode by a particular keyword phrase "Activate artificial intelligence". Similarly, to end the interaction with ChatGPT, the user must say "Deactivate artificial intelligence". This is to ensure that the users are fully aware of starting an interaction with Open AI's ChatGPT, and that the responses may be inaccurate or false so they need to assess responses carefully. Since the LLM-assisted dialogues with the e-VITA Coach use a curated document set for response generation, the responses are assumed to be reliable given the expert knowledge available in the project. However, the direct conversations with ChatGPT were only available in the Japanese version of the e-VITA Coach. This was due to differing views between the ethical committees in the EU and Japan of the acceptability of the direct use of ChatGPT in user experiments. In the EU system, the command "Activate artificial intelligence" always uses the path for LLM-assisted dialogue management, i.e. the RAG approach, whereas in the Japanese system, it evokes direct conversation with ChatGPT. In the EU-system, the LLM-assisted Dialogues and the Open-ended dialogues thus end up using the same RAG-based approach, as seen in Figure 2 where the this path branches out to LLM-assisted dialogue management.

Retrieval Augmented Generation (RAG) [12] is utilised in the e-VITA Coach to address the issue of reliable and truthful response generation. Custom-selected content is used as a resource to guide the truthfulness of the responses and enable the dialogue manager to answer queries based on the material which is pertinent to the application domain. The dialogue manager thus adopts a form of *guided text generation* shown in Figure 3. The retrieved text, together with the original user query, is input to the LLM, which generates a suitable response to the user.

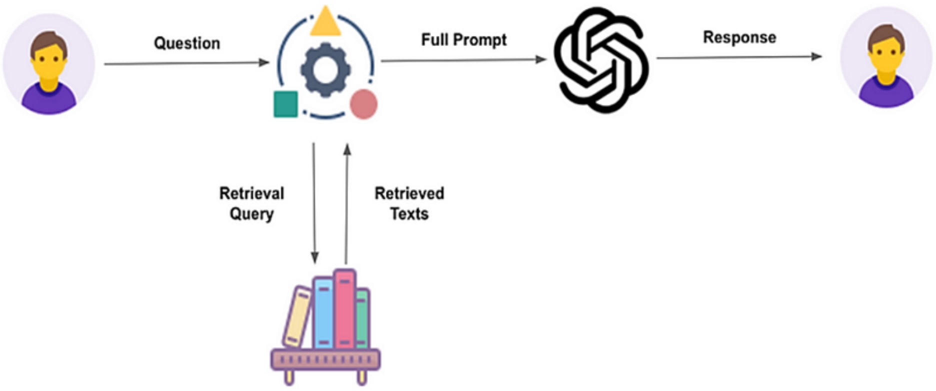
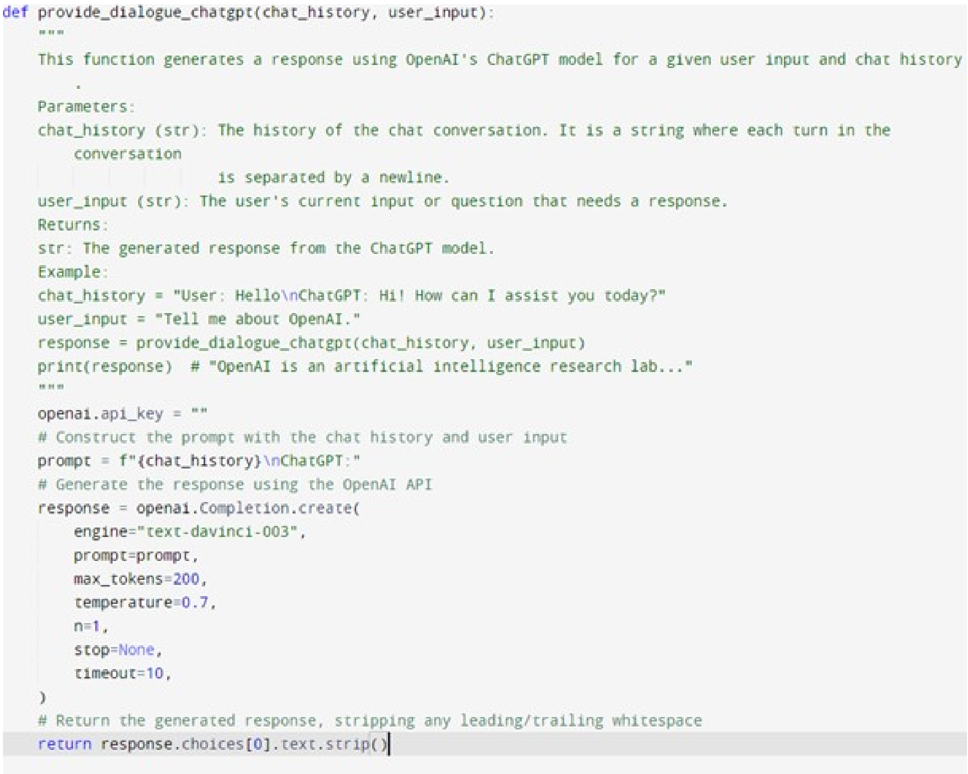


Figure 3: RAG approach. Image from [13] (CCA licence).

The documents are selected by the content experts. The texts are encoded in a vector format using LangChain (https://www.langchain.com/) and the FAISS (Facebook AI Similarity Search) vector store.

Since there is a limit for tokens in QA tasks, token size is reduced by extracting information to create chunks of information.

Figure 4 shows a prompt designed to produce an answer with 3-4 sentences, in a given context. The token limit is set to 256 and the temperature parameter to 0, favoring tokens with the highest probability in the output string. The prompt template strictly instructs the system to generate responses from the given documents and avoid any made-up information. This helps to ensure accuracy and eliminate the risk of hallucinations or fabricated answers.



```
def provide_dialogue_chatgpt(chat_history, user_input):
    """
    This function generates a response using OpenAI's ChatGPT model for a given user input and chat history
        .
    Parameters:
    chat_history (str): The history of the chat conversation. It is a string where each turn in the
        conversation
                    is separated by a newline.
    user_input (str): The user's current input or question that needs a response.
    Returns:
    str: The generated response from the ChatGPT model.
    Example:
    chat_history = "User: Hello\nChatGPT: Hi! How can I assist you today?"
    user_input = "Tell me about OpenAI."
    response = provide_dialogue_chatgpt(chat_history, user_input)
    print(response)  # "OpenAI is an artificial intelligence research lab..."
    """
    openai.api_key = ""
    # Construct the prompt with the chat history and user input
    prompt = f"{chat_history}\nChatGPT:"
    # Generate the response using the OpenAI API
    response = openai.Completion.create(
        engine="text-davinci-003",
        prompt=prompt,
        max_tokens=200,
        temperature=0.7,
        n=1,
        stop=None,
        timeout=10,
    )
    # Return the generated response, stripping any leading/trailing whitespace
    return response.choices[0].text.strip()
```

Figure 4: A prompt with dialogue history.

The input variables in the prompt template include the context and the user question, along with the additional parameter: "user info". This parameter contains a comprehensive list of the user's private data, collected by the Dialogue Manager across the subdomains of user information (see Section 3.2). It is used to direct the system to produce highly personalised responses. When generating the response, the information in all three parameters in the prompt template is taken into account, thus enabling the coach to adapt responses according to the user data. On the other hand, the prompt for the general coaching cycle lacks the "user info" parameter, and thus the system does not have access to the specific user data. Without knowledge of the user's personal details, the system cannot tailor its responses to the individual's needs, and response generation results in more generalized utterances.

5. Evaluation

The present article details the user evaluation results of the earlier versions of the dialogue system [14, 2], and provides a comparison with the version incorporating the coaching cycle. At the time of writing, the results of user tests for the latest version of the system were still under analysis and will be reported in forthcoming papers and project documentation. On the technical side, the e-Vita platform evaluation is described in [3], while the initial evaluations of ChatGPT and the RAG approach are reported in [11].

5.1. Evaluation methodology

The evaluation methodology comprised mostly qualitative approaches, ensuring user satisfaction and a thorough understanding of the system's performance. The project included a long evaluation period through series of user studies when the coaching system was placed in an elder person's home and the user was asked to test and evaluate the system usage at their convenient time and technical assessment. Each session was designed to simulate real-life interactions and included tasks such as daily exercise recommendations, nutritional advice, and mental well-being support.

Participants were recruited from various demographics, focusing on older adults who would benefit from the e-VITA coaching system. The recruitment process involved collaboration with community centers and healthcare providers to ensure a diverse participant pool.

Data were collected through questionnaires, interviews, and system logs. Pre-study and post-study questionnaires measured changes in participants' health behavior and attitudes. Interviews provided qualitative insights into user experiences, while system logs recorded interaction patterns and system performance.

5.2. Evaluation results

In one of the studies, primarily aimed at assessing participants' interaction with the robot and their acceptance of it, participants were engaged in various activities, such as motivational talks, memory training, cognitive challenges, and physical exercises. In interviews, participants mentioned they appreciated the opportunity to engage in conversations with the robot on topics such as health and family, which provided a sense of interaction and companionship. Users expressed that having the robot equipped with a dialogue system at home could motivate them to engage in exercises, even outside of formal exercise classes. Some participants found the conversations with the robot to be intellectually stimulating. Thus, it can be concluded that one of the practical values of the dialogue system inside the robot for the participants was facilitating interactions and providing a sense of connection and support. However, the restricted scope of dialogue topics and technical constraints such as limitations in the robot's ability to engage in spontaneous and natural conversations were noted as drawbacks.

In the final version of the system, OpenAI's text generation models were utilized. The advanced capabilities of Large Language Models enable more dynamic and contextually relevant responses by leveraging a vast amount of pre-trained knowledge, allowing the robot to engage in more spontaneous and natural conversations. This method reduces the occurrence of repetitive or irrelevant responses and enhances the system's ability to handle a broader range of topics and user intents.

Another study aimed to enhance user experience by offering a means of communication with the robot for daily living tasks. Participants especially appreciated the system's capability to engage in conversations about everyday routine. Users found the system useful for tasks such as checking the weather forecast, expressing emotions, and planning activities. The functional advantage of the dialogue system inside the robot for the participants was to provide them with a conversational interface that could assist them in various daily living activities. However, issues with timing and the speed of responses were mentioned, with some feeling that conversations were rushed or one-sided, and delays between conversations. Certain users found it challenging to maintain a smooth conversation with the robot, requiring them to repeat themselves or speak loudly.

These issues were addressed by optimizing the response generation process. The integration of the Retrieval-Augmented Generation model and improvements in the dialogue management flow reduced response times and enhanced the fluidity of conversations. Additionally, the use of advanced Automatic Speech Recognition and Text-to-Speech technologies improved the system's ability to understand and respond promptly and accurately, thereby minimizing delays and ensuring smoother interactions. In [11], RAG's promising start with increased reliability of the responses ia confirmed, but the need to validate the initial data to ground responses in real-world representation and to thoroughly evaluate the assistant's responses are also pointed out.

6. Trustworthy Coaching

The basic principles of Trustworthy AI underpin the research and are an important part of the standardization work [15], which continues after the project end. There are working groups and organizations like MPAI (https://mpai.community/standards/mpai-mmc/) which aim to unify AI tools and interfaces to give recommendations on use cases and functional requirements. As for dialogue modeling research, there are standardization efforts related to ISO-standards of dialogue acts [16]. We also draw on risk management research to mitigate epistemic (the process of producing knowledge) as well as interactive (usability) risks, especially in the context of healthy aging in the digital world (AAL guidelines).

Ethical aspects of speech-based AI agents also need to be considered carefully, and the main points are discussed in [17]. Advances in LLM technology enable chatbots to generate and analyze the content of conversations and store personal data on the web. As generative AI agents can predict responses and produce coherent sounding but inaccurate or fabricated content, this can cause both mistrust in the service (even in cases where the information is true) and over-trust (even in cases where the information is clearly wrong). In coaching scenarios, mistrust and over-trust are not desired, so the design of the coaching dialogues needs to pay attention to the accuracy of the information as well as how to support the facts and explain their validity in context. The AI agents do not 'know' the meaning of the utterances they generate, so technically their understanding and generation capability needs to be extended with such information and be grounded in the true facts of the context and the multimodal environment, besides in the dialogue history and context of the conversation (see discussion in [18]).

Moreover, other issues appear as important aspects in the quest for balanced trustworthy interaction. As is briefly noted in [11], interactive learning and adversarial learning may be useful to adapt the system to the user's behavior and to control the LLM agent's behavior, but also have a downside in the vulnerability of the development which can include malicious intents. It is thus necessary to support ethical and sustainable practices in designing and using practical AI agents such as virtual coaches and friendly assistants.

7. Conclusion

This article has provided an overview of the Dialogue Manager in the e-VITA Virtual Coach, focussing on trustworthy and reliable interaction for older adults in order to support their well-being and active healthy life. The work comprises of the necessary tools to acquire, represent, integrate, and analyze a variety of data generated in the interactions between the human user and the robotic interface devices, and the implementation of the e-VITA Dialogue Manager, which integrates advanced AI technology and various natural language processing components into an interactive system aiming to provide useful and flexible coaching for older users.

The Dialogue Manager is capable of personalised coaching conversations with the users, and its design emphasises reliable and responsible development of LLM-based systems. With the three different dialogue management models, i.e. regular Story-driven, LLM-assisted, and Open-ended dialogues, the e-VITA Dialogue Manager exemplifies a hybrid combination of rule-based and AI-based dialogue management approaches, which together support various functionalities of the Virtual Coach, and contribute to the active research area of LLMs and interactive AI systems. Furthermore, the experience gained through co-creating requirements for technical development together with the end-users will be valuable when designing and developing future applications in the context of smart home techology, generative AI, and robot interfaces, with the purpose of supporting well-being and quality of life in a responsible and sustainable manner.

The combination of the dialogue management approaches is to be evaluated further in different real-world situations and application scenarios, so as to reach full understanding of its (combined) benefits and drawbacks for practical applications. For instance, integration of data from sensors and smart living environments can be extended and elaborated further in order to supply the coaching system with a model of how to include dynamic data in the triggering system, to enable system-initiated interaction management. The current system deals with environmental sensors and time-scheduling, and interesting aspects concern extensions to multimodal data from smart home scenarios. An important extension is also integration of external knowledge (environment, world, situation, user) into dialogue modelling, and in particular, how to best use knowledge graphs in LLM-based interaction management. Much research is currently ongoing and published in this respect, here we only refer to two events where these challenges have been discussed: the e-VITA final workshop (https://www.e-vita.coach/homepage/finalconference/) and the SIGDial Special Session (https://2024.sigdial.org/call-for-papers-ss/).

Finally, the technical aspects related to the deployment of the components of the platform in the e-VITA project are fundamental issues connected to key technical and privacy-related issues. It may not be possible to define one single methodology for the deployment and provisioning of coaching and NLP services because this is strictly related to the technical requirements of the components (e.g. base development technology, sensor specifications), typology of data processing (e.g. algorithms), and use cases (e.g. performance and privacy restrictions). In order to choose the best way to realize a concrete components deployment, it is necessary to adopt multiple strategies that cover both cloud and edge scenarios to fulfill specific requirements related to these features. In the hybrid approach, a single application or service can use components that are on the edge or cloud side, or a combination of the two depending on the unique requirements of the application.

Acknowledgments

We thank all the project partners for useful and fruitful discussions and enjoyable cooperation. Special thanks go to the colleagues in technical workpackages cooperating in the integration of the dialogue module in the e-VITA platform and to the content experts on the example dialogues for coaching.

References

[1] K. Jokinen, K. Homma, Y. Matsumoto, K. Fukuda, Integration and interaction of trustworthy AI in a virtual coach: An overview of EU-Japan collaboration on eldercare, in: Y. Takama, N. Matsumura, K. Yada, M. Matsushita, D. Katagami, A. Abe, H. Kashima, T. Hiraoka, T. Uchiya, R. Rzepka (Eds.), Advances in Artificial Intelligence: Selected papers from the Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2021), Springer, 2022, pp. 190–200. doi:`https://doi.org/10.1007/978-3-030-96451-1_17`.

[2] M. McTear, K. Jokinen, M. M. Alam, Q. Saleem, G. Napolitano, F. Szczepaniak, M. Hariz, G. Chollet, C. Lohr, J. Boudy, Z. Azimi, S. D. Roelen, R. Wieching, Interaction with a virtual coach for active and healthy ageing, Sensors 23 (2023). URL: https://www.mdpi.com/1424-8220/23/5/2748. doi:`10.3390/s23052748`.

[3] R. Naccarelli, F. D'Agresti, S. D. Roelen, K. Jokinen, S. Casaccia, G. M. Revel, M. Maggio, Z. Azimi, M. M. Alam, Q. Saleem, A. H. Mohammed, G. Napolitano, F. Szczepaniak, M. Hariz, G. Chollet, C. Lohr, J. Boudy, R. Wieching, T. Ogawa, Empowering smart aging: Insights into the technical architecture of the e-VITA virtual coaching system for older adults, Sensors 24 (2024). URL: https://www.mdpi.com/1424-8220/24/2/638. doi:`10.3390/s24020638`.

[4] M. I. Torres, J. Tenorio, G. Cordasco, E. Gonzalez-Fraile, A. Esposito, M. S. Kornes, G. Chollet, B. Fernandez-Ruinova, Ageing in the 21st century in europe: social challenges and innovation opportunities to support elderly independency and wellbeing, 2021. URL: https://addi.ehu.es/bitstream/handle/10810/50944/EMPATHIC_book_Chapter.pdf, technical Report EHU-ADDI.

[5] Z. Callejas, D. Griol, K. Benghazi, M. Noguera, G. Chollet, M. I. Torres, A. Esposito, Measuring and fostering engagement with mental health e-coaches, in: Companion Publication of the 2020 International Conference on Multimodal Interaction, ICMI '20 Companion, Association for Computing Machinery, New York, NY, USA, 2021, p. 275–279. URL: https://doi.org/10.1145/3395035.3425316. doi:`10.1145/3395035.3425316`.

[6] T. W. Bickmore, Relational Agents in Health Applications: Leveraging Affective Computing to Promote Healing and Wellness, in: The Oxford Handbook of Affective Computing, Oxford University Press, 2015. URL: https://doi.org/10.1093/oxfordhb/9780199942237.013.010. doi:`10.1093/oxfordhb/9780199942237.013.010`.

[7] Y. Zhou, B. D. Eugenio, B. Ziebart, L. Sharp, B. Liu, N. Agadakos, Modeling low-resource health coaching dialogues via neuro-symbolic goal summarization and text-units-text generation, 2024. URL: https://arxiv.org/abs/2404.10268. `arXiv:2404.10268`.

[8] N. Lete, A. Beristain, A. García-Alonso, Survey on virtual coaching for older adults, Health Informatics Journal 26 (2020) 3231–3249. URL: https://doi.org/10.1177/1460458220926894. doi:`10.1177/1460458220926894`, pMID: 32744137.

[9] W. R. Miller, S. Rollnick, Motivational interviewing: Helping people change, Guilford Press, 2012.

[10] M. Droppa, H. Lee, Motivational interviewing: A journey to improve health, Nursing 44 (2014). doi:`10.1097/01.NURSE.0000443312.58360.82`.

[11] K. Jokinen, Is the plan ready yet? - Exploring LLMs when talking about well-being and health, in: 14th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2024), Sapporo, Japan, 2024.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, 2020, pp. 9459–9474.

[13] Welcome to World of LLMs - Session 2 - Knowing RAGS, Youtube video https://www.youtube.com/watch?v=Q6shHtNFiuQ, 2024.

[14] R. Browne, T. Ogawa, J. Vizcarra, K. Jokinen, Experiments on the Nao robot as a virtual coach, in: Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2022), Kyoto, Japan, 2022.

[15] ITU, Definition of "Open Standards", https://www.itu.int/en/ITU-T/ipr/Pages/open.aspx, 2005.

[16] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, D. Traum, ISO 24617-2: A semantically-based standard for dialogue annotation, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 430–437. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/530_Paper.pdf.

[17] T. Williams, C. Matuszek, K. Jokinen, R. Korpan, J. Pustejovsky, B. Scassellati, Voice in the Machine: Ethical Considerations for Language-Capable Robots, Communications of the ACM 66 (2023) 20–23.

[18] G. Wilcock, K. Jokinen, To err is robotic; to earn trust, divine: Comparing ChatGPT and knowledge graphs for HRI, in: 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2023), Busan, Korea, 2023, pp. 1396–1401.