Universiteit Leiden
The Netherlands

Bachelor Computer Science & Datascience and Artificial Intelligence

LLM Personalization Using Summarization:

A Robotic Fitness Coach Case Study

Arie Klaver

Supervisors:
Joost Broekens & Peter van der Putten

BACHELOR THESIS

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

25/06/2024

Abstract

With the rapid advancement of large language models, virtual assistants as conversational agents could feasibly replace human coaching. This study aims to understand the level of contextual information needed by such virtual assistants to simulate personalized interactions. It evaluates the effect of summarization on personalizing human-robot interaction with a virtual assistant powered by an LLM, focusing on its impact on user experience. It also aims to identify the necessary architecture for a smooth interaction. A robotic fitness coach was developed, and experiments with personalized and unpersonalized versions were conducted. Though the sample size is too small for statistically significant conclusions, results show promise for summarization in personalizing interactions. While some metrics like anthropomorphism, attractiveness and efficiency were rated lower, the perceived perspicuity and novelty of the developed system saw substantial increases when using the personalized version. Users of the personalized version believed the assistant remembered their previous conversation, indicating the positive effect of summarization. The developed architecture, while improvable, was sufficiently smooth to provide a positive user experience.

Contents

1 Introduction

With the recent advancement of large language models (LLMs), conversational agents are not only becoming increasingly more difficult to distinguish from their human counterparts [JB23] but also have access to significantly more knowledge [ENT22]. These advancements have led to the increasing versatility and broad applicability of such agents across various domains, including business [BSR+20], health care [CDK+20], and personal fitness coaching [KtSM+20]. However, to what extent LLMs can achieve the personalization provided by human experts remains a critical question in the field of human-computer interaction.

In the field of human-robot interaction (HRI) in particular, personalization and adaptation to users will be necessary to maintain user engagement and to build rapport and trust between the user and the robot [IRS+19]. By being able to tailor responses based on individual user needs and preferences, the perceived interest and usefulness of the interaction with conversational agents can be enhanced significantly [CvK18]. This study aims to better understand the level of contextual information required to simulate personalized interactions. Specifically, it evaluates the effect of summarization as a technique for personalizing human-robot interaction with a virtual assistant powered by a Large Language Model (LLM) across longer histories than the current conversation, focusing on its impact on user experience and satisfaction. Additionally, it seeks to find out what architecture would have to be in place in order to facilitate a smooth human-robot interaction.

To investigate this, a robotic fitness coach, called FitBot, has been developed to provide personalized workout routines, fitness advice, and tips on maintaining a healthy lifestyle. Can FitBot effectively customize its advice with the minimal data input provided by the summaries, or does it need extensive user history to achieve a convincing level of personalization and provide a satisfactory user experience? We try to answer this question by running an experiment, where we split the participants into two groups. One where they will receive personalized advice based on the summarization of previous interactions, and one where they will interact without this personalization in place. The user experience of both groups will then be evaluated and compared, aiming to provide insights into the impact of summarization on perceived personalization and overall satisfaction with FitBot.

With this understanding we can gain a better grasp of the capabilities of LLM-powered conversational agents to simulate personalized interactions akin to human experts, while only having access to the limited amount of data provided by the summaries.

The remainder of this thesis is structured as follows. Section 2 lists the motivation and related work; Section 3 states the research question; Section 4 discusses the method used to perform the study; Section 5 describes the results; Section 6 discusses these results and Section 7 concludes.

2 Motivation and Related Work

First the topic of Large Language Models (LLMs) will be introduced, then conversational agents will be discussed and finally, virtual fitness coaches will be detailed.

2.1 LLMs

In this study a Large Language Model (LLM) will be used to interpret user input and generate responses as part of a system that functions as a robotic fitness coach, making use of summarization in order to personalize the coaching process. An LLM is an artificial intelligence system that can both comprehend and generate human-like text at a large scale. These models are built using deep learning techniques, particularly variants of neural networks known as transformers. They are trained on extensive datasets comprising text from various sources, such as website, books, articles and more. Because of the large amount of data used to train the models, they can have a great understanding of many topics and use this to perform a multitude of language-related tasks, including translation, question answering, summarization and conversation generation.

Using LLMs to augment human-robot interaction seems promising, as it introduces a new dimension of fluidity and intuition to human-robot conversations. With extraordinary reasoning and generation ability, large language models illustrate great potential to complete tasks meeting the requirement of users. [ZCL+23]

An example of this is the Pibot, short for "Humanoid Pilot Robot", which is a robot that can autonomously operate an airplane, using ChatGPT to understand the manuals of different airplanes. [McF23] Another noteworthy example is the Fruitcore robotic arm, which uses ChatGPT to allow the user to communicate with the robot using natural language, making them more adaptable and user-friendly. [Imp23]. Furthermore, at Princeton University, the TidyBot was created. This is a robot that tidies up your room by moving certain objects to specific locations. An LLM trained specifically for this task was used to interpret the rules that were to be followed, as inputted by the user. [WAK+23]

2.2 Conversational Agents

One of the main challenges of this study is creating a conversational agent and with this comes managing the dialogue. The conversation needs to be steered in the right direction and the right responses need to be given. There are multiple ways in which this can be handled.

First there are the handcrafted approaches to dialogue management. These rely on programs and/or models that are fully specified by developers or domain experts to track the dialogue state and define the policy. They can be distinguished between four kinds of handcrafted approaches. [BBB+22]

First there is a rule-based approach, where patterns are paired to specific responses. Although this approach is relatively easy to implement and does not require any training data, it lacks flexibility and requires considerable effort from developers to encode rules.

Second, there is a similar approach, which is the finite state-based approach. Here the dialogue transitions between different states based on the user responses. This method has the same shortcomings as the rule-based approach, along with other shortcomings such as versatility and robustness in situations where the user does not follow predefined sequences of states.

Third, there is the activity-based approach, which specifies a certain workflow the dialogue manager may go through during the specification, in terms of activities, rather than specific states.

The fourth approach is the frame-based approach. Here a set of frames is defined, each specifying the information that the conversational agent is required to acquire from the user in order to fulfill a dialogue task. This approach affords more flexibility, since it can efficiently process over-informative inputs from the user and the information can be provided in an arbitrary order, as long as the necessary information for the frame to be filled is eventually acquired.

Besides these handcrafted approaches, there are also data-driven approaches. These can be divided in supervised learning and reinforcement learning techniques. The supervised approach learns from a set of labeled data, whereas the reinforcement approach focuses on optimizing the learning by a trial-and-error process governed by a series of reinforcements. Hybrid approaches, combining multiple approaches, either handcrafted or data-driven, in order to capitalize on the benefits of each, also exist. [BBB+22] In this study a hybrid approach will be used, enabling us to first fill in a certain frame of information before initiating a LLM-driven conversation loop.

Using an LLM as a virtual assistant or conversational agent is a topic of great interest. OpenCHA, for instance, is an open-source LLM-powered framework made to empower conversational agents to generate a personalized response for users' healthcare queries. It does so by enabling developers to integrate external sources including data sources, knowledge bases, and analysis models, into their LLM-based solutions [AARJ23].

Conversational agents also have to deal with managing what information is to be stored for future interactions. Memory Sandbox, is a project that deals with memory management for conversational agents, and aims to provide the user with interaction affordances to manage how the agent should conceptualize the conversation, providing the user with more control [HGKM23]. The focus of the study discussed in this thesis, however, is not on user control, but on the effect of summarization as a memory technique for personalization on the perception of the user.

The RAISE framework was also introduced recently. This is an advanced architecture enhancing the integration of LLMs like GPT-4 into conversational agents, by incorporating a dual-component memory system, mirroring human short-term and long-term memory, to maintain context and continuity in conversations [LCT+24].

2.3 Virtual Fitness Coach

The use of a virtual conversational agent as a potential replacement of a fitness coach is something that has been attempted before. Due to its wide availability, inexpensive, and ease of use, it makes for a powerful tool for behavioral change interventions. [KTM+17] [CLK10]

CoachAI, for example, is a messaging-based conversational agent built to support the development, classification and delivery of both individual- and group-based health interventions. Although some participants revealed preference to human support, it was demonstrated in the 1-month pilot performed in the study, that individuals have a generally positive reaction to a virtual agent providing them with health interventions and tracking their feedback. They felt the agent was interesting and easy to talk to. They also expressed high levels of trust in the agent and a desire to work with it again. Most participants felt that their conversation with the agent would help them improve their health and most were willing to reveal personal information about their daily life to the virtual agent. [FWR19]

In addition, the concept of multimodal Companions introduces an innovative method for

incorporating virtual beings into users' everyday life. Turunen et al. (2011) presented the Health and Fitness Companion, a system designed to build long-lasting relationships with users to support their everyday health and fitness activities. This Companion, in contrast to conventional task-based spoken dialogue systems, is designed to be a permanent part of the user's life, enabling interactions in home and mobile contexts. Users may be inspired to lead healthier lives by the conversational interaction and the physical presence, which promote social and emotional connections. The system employs a at the time novel interaction management model that distinguishes between dialogue management and cognitive modeling, enabling adaptable component interoperability. [THS+11]

In a review of conversational agents designed to aid in healthcare, performed by Laranjo et al. (2018), 17 studies involving 14 conversational agents were analyzed. Although most of the agents relied on either frame-based or finite-state methods of dialogue management, in contrary to the hybrid approach used for this particular study, they did all used some form of natural language as the user input, whether spoken or written. The review highlighted that while user satisfaction was generally high, common issues included spoken language understanding and dialogue management problems, leading to misunderstandings or the conversation being steered in the wrong direction. [LDT+18]

Research has also been done on the effect of a relational versus a non-relational robotic fitness companion on the intrinsic motivation of the user to engage in exercise tasks. In a particular study, the enjoyableness, usefulness and the companionship of a relational and non-relational version of a so-called socially assistive robot where compared, as evaluated by the users. The relational robot employed specific social interaction and personalization approaches, such as giving the user praise upon correct completion of a given exercise, providing reassurance in case of failure, displaying continuity behaviors, using humor and referring to the user by name. 85% of the participants rated the relational robot higher than the non-relational one in terms of enjoyment, and 77% rated it higher in terms of usefulness. Also in terms of companionship, the relational robot received higher scores. The users were also asked to directly compare both robots in terms of ten different evaluation categories (e.g. more intelligent, more useful). Here the participants again expressed a strong preference for the relational robot over the non-relational robot, with the relational one receiving 82% of the positive trait votes. [Fas24] Although the study focuses on intrinsic motivation, the measures that were evaluated are closely related to user experience, and continuity in particular is a trait the robot designed in this study, making use of the summarization technique, will also exhibit.

3 Research Question

One of the main obstacles of using a robot powered by an LLM as a fitness coach, or any virtual coach or assistant for that matter, is providing a personalized experience. A fitness coach often has many clients and for a real human being it would be trivial to differentiate these clients, remember their individual wants and needs, and account for these while answering their questions. For a robotic fitness coach this is not as self-explanatory. Systems have to be in place to identify the person standing in front of them, and data about previous interactions has to be stored and retrieved.

This study tries to investigate whether prompting the LLM to summarize the dialogues in order to store the key takeaways and retrieve these in future interactions, allows it to provide a personalized and improved user experience. The research question is: *what is the effect of summarization of past interactions as a technique for personalizing human-robot interaction with a virtual assistant, powered by an LLM, on the user experience in terms of overall satisfaction, perception of the robot and perceived personalization?*

Another important factor is ensuring a smooth interaction for the user. It is thus important to consider what architecture will have to be in place to facilitate this ease of use and low latency. Connecting the different components that need to be in place in a single pipeline could cause a delay. If this delay is too long, it could cause frustration among the users and decrease satisfaction [YD15][PMZ+20]. A second question is therefore: *what architecture needs to be in place to ensure a smooth user experience?*

3.1 Hypotheses

The hypothesis for this research is: using summarization as a technique for personalizing human-robot interaction with a virtual assistant, powered by an LLM, will have a positive effect on the user experience, measured as a difference on attractiveness, perspicuity, efficiency, dependability, stimulation and novelty using the User Experience Questionnaire, and on the perception of the robot, measured as a difference on anthropomorphism, animacy, likeability, perceived intelligence and perceived safety using the Godspeed Questionnaire Series and on the perceived personalization as expressed in the post-experiment evaluation.

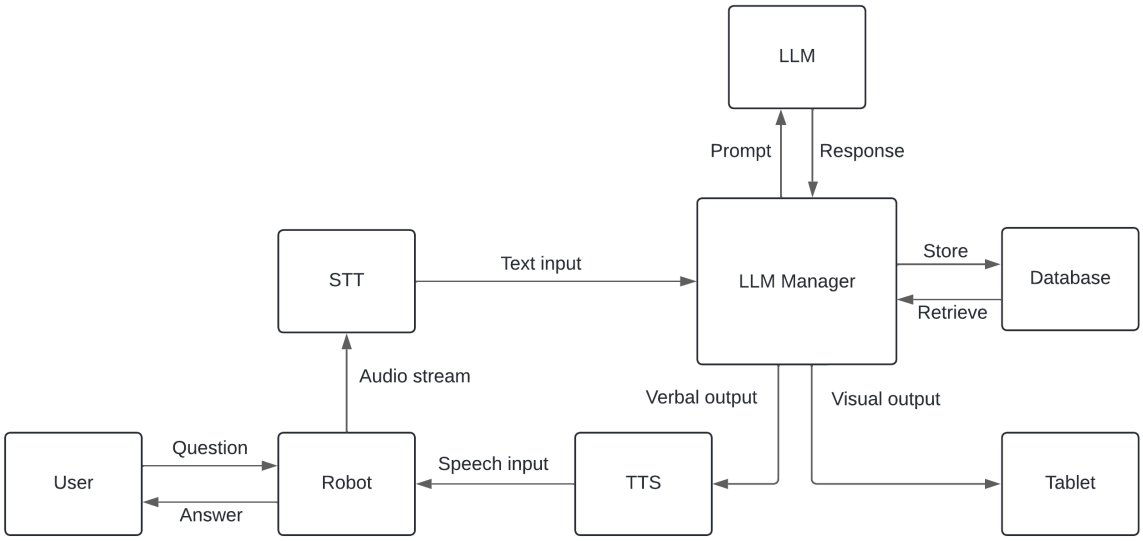

Figure 1: Overview of the architecture used for FitBot.

4 Method

The method used to perform this study will now be discussed. In Section 4.1, the developed system will be deconstructed. Then, the experimental setup will be explained in Section 4.2 and in Section 4.3, the measures used to evaluate the user experience will be detailed.

4.1 Overview

There are several requirements the developed system needs to address. It needs to take a voice-based input from the user through the sensors of the robot. It then needs to transcribe this input, so that it can be dealt with further. The dialogue of the conversation also needs to be managed. There should be a user profiling phase during which the details of the user are obtained and this should be followed by an LLM-driven question and answer loop. Thus, an LLM also needs to be connected; carefully refined prompts need to be sent and the generated responses need to be processed. The robots responses need to be dissected into a verbal and visual output and these need to be outputted accordingly, so either through the robots audio system or via the web page. When the conversation has concluded, it needs to be summarized and stored in a database, in order to be retrieved during a future interaction.

Having taken all of these requirements into account, the system developed to create the FitBot robotic virtual fitness coach consists of various interconnected components. Before discussing these components in further detail, let me first give a general overview of the system as a whole. A diagram of the implemented architecture can be seen in Figure 1.

The interaction starts with the user. They can interact with the user using voice commands. The robot thus has to record a continuous audio stream. This audio stream will then be transcribed, so that the LLM Manager can receive a text input. The LLM Manager handles the flow of the

conversation and is also the bridge between the robot and the large language model. In addition, the LLM Manager also operates the database, where the user details are stored. The LLM Manager has two outputs; a verbal output and a visual output. As FitBot is a fitness coach, it will occasionally provide long lists of exercises in the form of routines. Since such routines are tedious for the robot to vocalize, they will be outputted visually using a simple web page. This web page also shows the current state of FitBot (e.g. listening or processing) and its latest response. The rest of the output, the verbal output, will be converted back to audio format using speech-to-text and the audio will then be played by the robot, reaching the user, which means the process has come full circle.

We will now discuss the individual components in more detail. Note that the public GitHub repository can be found among the references [**?**].

4.1.1 Programming Environment

The programming language that has been used is Python. The connection with the robot is made with WAMP [WAM22] using the Autobahn Twisted library [Aut22]. This allows for communication with the robot using remote procedure calls. It also enables us to easily create our own remote procedure calls that are hosted within the same session, which comes in handy when updating the web page. Also making use of the Autobahn Twisted library, `inlineCallbacks` and `defer` are used to handle the asynchronous nature of the system.

4.1.2 Robot

The robot used to represent the virtual assistant in this study is the NAO robot made by SoftBank Robotics. [SRA24] This robot is a programmable personal teaching assistant developed to aid in making learning about programming and robotics more fun and more concrete for students. It is connected to the Robotsindeklas Portal [Rob23], created by Interactive Robotics [Int22], where its realm ID can be obtained. Through this ID, connection to the WAMP server can be made and the remote procedure calls can be accessed.

The NAO has a human-like shape, motorized joints and advanced algorithms for control, allowing it to mimic the movement of humans to a remarkable degree. It also has built-in speech recognition and touch, audio, and camera sensors, enabling a responsive interaction. All of these features combined make it the perfect candidate for a personal fitness coach. An image of the NAO robot can be found in Figure 2.

4.1.3 LLM

When selecting the large language model that will be used, there are a few things to consider. First of all, the performance of the LLM is obviously of great importance. The model needs to be trained using a large dataset, ensuring a vast knowledge of the fitness domain. It also needs to have sufficient language understanding, granting it the ability to deconstruct as well as generate coherent and contextually appropriate responses and also enabling it to properly mark routines, i.e. lists of exercises which FitBot should not be vocalizing but will instead display on the web page, for them to be extracted. In addition, the cost needs to be minimal, the API should be easily accessible and the latency should be minimal. When considering all of these factors, GPT-3.5 Turbo 0125 by OpenAI [GPT24] was selected as best meeting these requirements. It has a 16K context window, is optimized for dialog and, as the name suggests, is relatively fast. At a cost of just $\$0,50/1\text{M}$

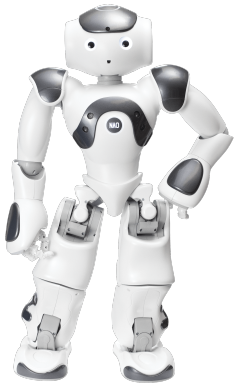

Figure 2: NAO, the humanoid and programmable robot made by SoftBank Robotics and the robot used to represent FitBot in this study.

input tokens and $\$1,50/1M$ output tokens it is one of the most affordable large language models out there. Furthermore, the API is straightforward to implement and it performs reasonably well at marking the routines within the output text.

OpenAI also provides Assistants API [Ass24], which is an API designed more specifically for the purpose of using their models for creating assistants. However, since this API has its own built-in memory management, it is unsuitable to be used for this study.

4.1.4 Audio Transcription

As mentioned earlier, the NAO has built-in speech recognition. However, in practice this seemed too inconsistent to be used for this study. Although GPT-3.5 Turbo is able to correct for a certain degree of misinterpretation, when using NAO's speech-to-text capabilities, the transcriptions were too different from the actual words spoken to the point where the user experience was severely compromised. Because of this, it was decided to make use of the Google Cloud Speech-to-text API [Goo24]. This API uses advanced speech AI to recognize and transcribe speech and outperforms NAO's built-in speech-to-text. The quality of the transcriptions increases even further when making use of the "video" model.

Since the audio stream needs to be processed continuously, multi-threading was used to handle the asynchronous nature of the program. After subscribing to the hearing stream event of the NAO and yielding the result to a callback function, the incoming data-frames are perpetually added to a queue to be dealt with further. Now a third thread is instantiated, in which the configuration for the speech-to-text API is initialized. In this same thread, the streaming requests are gathered. It takes the queue of audio data and yields a new chunk of this data to the streaming recognition process, ensuring continuous and seamless transcription of incoming audio in real-time. The responses are then transcribed individually and when the API marks the `is_final` flag of a generated result as true, the transcription is sent to the LLM Manager to be processed.

The API can also take a speech context. This is a list of phrases that the transcription model will look out for, allowing for them to be transcribed properly more often. In this list we included various terms relating to fitness, which it seemed to struggle with during testing. Additionally, the list of user names is added to this speech context. This way, the names returning users will be

recognized with a little more ease.

4.1.5 LLM Manager

The LLM Manager could be considered the brain of FitBot. It handles the conversation flow, processes the incoming transcriptions and is the bridge between the robot and the large language model. Let's start with addressing the flow of the conversation. This differs for the two versions of the program, i.e. the personalized and the unpersonalized version. Although the unpersonalized version is not entirely unpersonalized, as the context the LLM is naturally provided with throughout the conversation will be incorporated into the generated responses, the version of the program that does not take into account any of the previous interactions with a specific user will from now on be referred to as such, for the sake of simplicity.

Before discussing the flow of the conversation for both versions, let's first break down some of the developed methods that were used.

4.1.5.1 End of Sentence Detection

As noted earlier, once a transcription is considered final by the speech-to-text API, it is propagated to the LLM Manager to be processed. The problem however, is that the sentence is not always actually complete. Sometimes the API thinks the user is finished speaking while they might have just had a bit of a longer pause in the middle of their sentence than usual. To solve this, there needs to be a way to predict whether the sentence thus far is complete, or the user is likely to add something to it in the next transcribed phrase.

In order to do this, the LLM has been utilized. The transcribed phrase that is received by the LLM Manager is fed to the LLM, combined with a prompt that instructs it to predict whether it thinks the user is finished speaking or not. Within this prompt, the number of phrases that have yet to be marked as complete is also provided. If this number is higher, the LLM is more likely to indicate the sentence to be concluded. If the sentence is most likely not complete yet, the LLM Manager will wait for a new phrase to be transcribed and stitch the two together, forming a new potentially complete sentence to be analyzed by the LLM. This process repeats until the LLM marks the sentence as complete, or until there is a silence of 1.5 seconds. The silence detection is in place to account for cases where the user is finished speaking but the LLM mistakenly thinks it is not.

4.1.5.2 End of Conversation Detection

In order to conclude the conversation, rather than using a specific key phrase to detect the end of interaction, a more dynamic approach has been opted for. Within the system prompt the LLM is initially provided with, the LLM is instructed to mark the end of the conversation using the `<stop>` terminator. By utilizing the natural language understanding of the LLM to detect the end of the conversation, the interaction can be concluded in a much more natural and fluent matter, compared to using a key phrase, pressing a button or using silence detection.

4.1.5.3 Check Whether Response Affirmative

During the introductory profiling phase of the program, the user is often asked for confirmation. They are, however, provided with the freedom to answer as they please. Accounting for all possible

answers manually would be a tedious and error-prone approach, so once again the LLM is carefully prompted to handle every potential scenario. It is provided with some positive and negative examples and based on the input text it receives, it will either respond with `True` or `False` for a affirmative or negative case respectively.

4.1.5.4 Check Whether Response Includes Name and PIN

Also during this profiling phase, the user is asked to provide their name and a 4-digit PIN code. To smooth out and simplify this process, two other LLM-driven functions are used: `IncludesName()` and `IncludesPin()`. These functions, in order, determine whether the obtained user input includes a name and a PIN code and if so, returns them. This way, the user won't have to list them one by one and they do not have to worry about uttering exactly the right phrase in order for the name or PIN code to be registered correctly.

4.1.5.5 Load Conversation

If the user has interacted with FitBot before, it will retrieve the last conversation from the database. Or rather, it will retrieve the summary of the last conversation from the database. The LLM is then prompted to come up with a response in which it quickly highlights the most important parts of the last conversation in order to create the assumption that FitBot has remembered the entire conversation, and thereafter ask about how they liked the advice or whether they have tried out the routine specified during the previous interaction.

4.1.5.6 Save Conversation

Once the conversation is deemed complete by the LLM and subsequently marked using the same `<stop>` terminator as mentioned before, the second difference compared to the unpersonalized version comes into play; the conversation is summarized by the LLM. It is specifically prompted to only include the parts it considers important, so it can retrieve these in the future for a personalized interaction. The exact prompt used to summarize the conversation is as follows:

```
"Summarize the entire conversation up to this point, including the summary you may
have been provided with at the start of the conversation, using short sententences
or keywords. Do this in a way so you could retrieve the most important
conversation topics if you were to read this summary later, in order to give more
personalized feedback in future conversation. Especially focus on personal details
of the client, not so much on the flow of the conversation. Be sure to include the
last conversation in your summary. You do not have to mention every conversational
detail that occurred, just the most important things that were discussed."
```

This summary is then inserted into the database. The old summary is overwritten after every conversation. A different approach could be to add to the existing summary every time, but in order to keep the summaries short, this is not the approach that has been opted for.

In the prompt that instructs the LLM to summarize the conversation, however, the LLM is asked to also include the summary of the previous conversation that may or may not have been provided at the start of the conversation into the context of their newest one. So even though the summary is completely overwritten, details that have been deemed important by the LLM are still incorporated into the newly generated summary.

This approach also ensures that the old information becomes less and less important as it is repeatedly summarized, resulting in some natural form of information decay.

Additionally, it makes sure the chat summaries will not keep increasing in size over time, slowing down the time it takes to process them and increasing the amount of tokens that need to be processed by the LLM.

4.1.5.7 Unpersonalized Conversation Flow

As previously stated, we make use of a hybrid approach to handle the conversation flow. Both versions of the program start by listening for a specific key phrase. They continuously transcribe incoming audio and wait until they hear the phrase "activate coaching". Once the user has uttered this phrase, FitBot will output an introductory statement. This statement differs for the two versions and beyond this point, the resemblance between the two version diminishes further.

The unpersonalized version directly enters the main conversation loop, driven by the LLM. So after the short handcrafted introductory phase, it enters the data-driven central part of the conversation. This is a simple loop where the input from the user is obtained and a response is generated by the LLM. If the aforementioned terminator is detected within the generated response, it is extracted and the response is outputted one final time. After this, the conversation is over and FitBot no longer listens to the user. This marks the end of the unpersonalized loop.

4.1.5.8 Personalized Conversation Flow

In the personalized version of the program, the introductory statement made by FitBot ends with the question whether it and the user have crossed paths before.

Based on the response of the user, the program splits into two segments. One for returning and one for new users. From this point on, a handcrafted approach, most closely resembling a frame-based approach, is used to obtain some information about the user until eventually the LLM-driven conversation loop starts. FitBot already knows whether the user has interacted with it before from the first question it asked, but during this next user profiling phase, there are several other details it needs to obtain. Firstly, it needs to know whether the user is okay with sharing their information. If the user does not grant permission, the program will continue in similar fashion to the unpersonalized version, directly entering the LLM-driven conversation loop. If they are okay with sharing their information, their name and a 4-digit PIN code will subsequently be acquired. FitBot also asks the user for confirmation of both their name and PIN code in order to assure it will be correctly inserted into the database later on.

Once both the name and PIN code have been confirmed, a LLM-driven conversation loop, similar to that of the unpersonalized version will commence, but with two key differences. If the user is new, the conversation continues by FitBot asking how it can assist the user. However, if the user has interacted with FitBot before, it will provide the more personal response that has been generated, as mentioned in Section 4.1.5.5. Now the loop continues by continually obtaining a user input and generating a response and outputting this. Once the conversation is deemed complete by the LLM and subsequently marked using the same `<stop>` terminator as mentioned before, the second difference compared to the unpersonalized version comes into play; the conversation is now summarized and stored in the database. After this, the interaction is over and the personalized loop has been completed.

4.1.6 Web Page

FitBot is accompanied by a web page. Ideally this web page would be displayed on some sort of tablet that is placed at about the same height has the NAO, but during the experiments the web page was shown on a laptop sitting next to it. This web page shows various things.

To begin with, it shows the current status of FitBot. This can be either of four different states; listening, processing, speaking or offline. By showing this visual cue to the user, they will hopefully feel less confused when FitBot happens to take a bit longer then expected to generate a response, further increasing the smoothness of the interaction.

In addition to showing the status, the latest response is also shown. This way the user is able to read the response if they mishear something and it also allows for better confirmation of details such as their name and PIN code.

Finally, routines provided by FitBot will be displayed on the screen. Due to the repetitive nature of routines, vocalizing them entirely quickly becomes a tedious process and listening to them can be tiresome and even frustrating to the user. By extracting these routines from the response and displaying them on the screen, instead of outputting them verbally, we aim to solve this issue.

Moreover, it allows for the user to carefully read through the routine and by displaying the routine in an organized table, the routines will hopefully also be easier to grasp for the user. An example of the web page in action can be found in Figure 3.

4.2 Experimental Setup

In order to find out what the effect of summarization as a technique for personalization of the human-robot between the user and FitBot on the user experience is, a version of FitBot with and without personalization have been created.

In an effort to see what version achieves the best user experience, an experiment has been conducted. This experiment consisted of ten participants. These participants have been split into two samples of five. The first group have tested the unpersonalized version and the second the personalized version.

Since FitBot does not have any conversation history to refer to during the first conversation, the personalization really only starts to take shape in subsequent conversations. Because of this, all participants have had two conversations with FitBot. These conversation can be as short or as long as they like, depending on how much information they intended to obtain from FitBot.

The conversation starts when the user utters the phrase “activate coaching” in order to wake FitBot and it ends when the LLM thinks the conversation is over and has inserted the terminator into its response. FitBot then states that the session is over and the status on the screen changes to “offline”.

The ten participants vary from ages between 21 and 62 years old and have been distributed over the two samples as fairly as possible, ensuring about the same amount of participants of the same age group are in each sample. By keeping the differences between the groups as small as possible, the combined user experience should be affected as little as possible by individual differences between participants.

During the interactions FitBot was standing on a desk in front of the participant and the web page was displayed on a laptop standing next to FitBot. The experiments took place at the homes of the participants, often in their living room or in some sort of study room. Either way, there

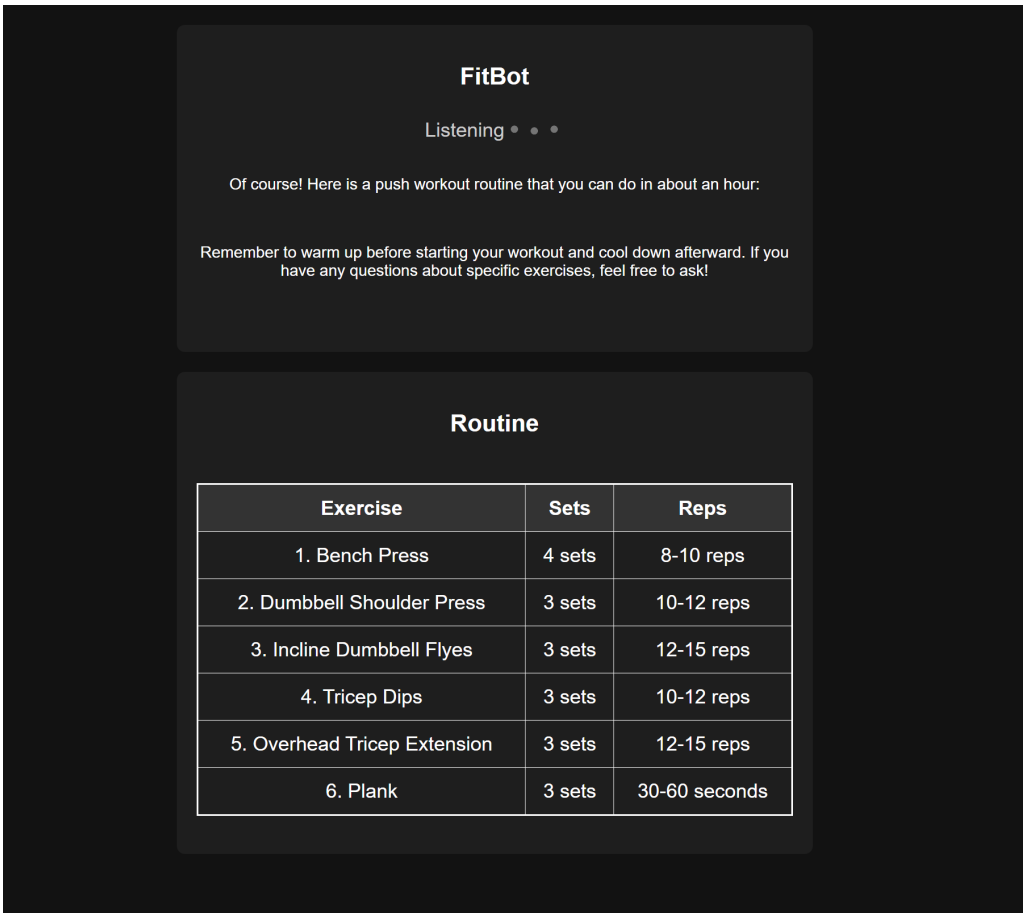

Figure 3: Example of the web page during an interaction with FitBot. As can be seen, the current status of FitBot is "Listening" and it has just provided a push workout routine that can be performed in about an hour.

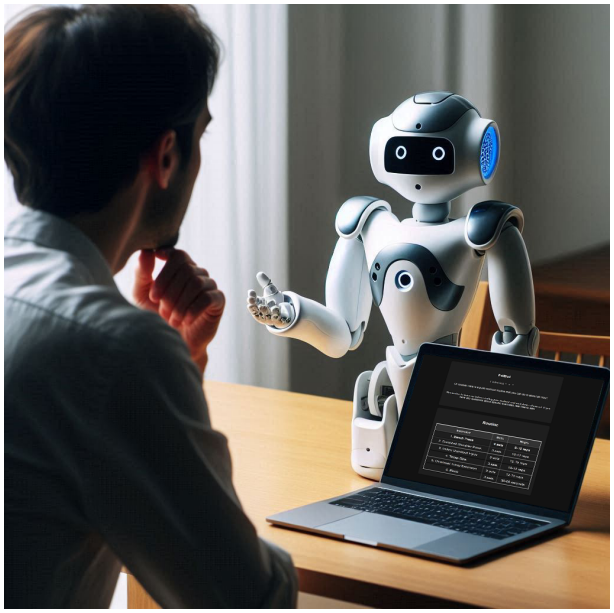

Figure 4: Recreation of the experimental setup. Note that this is not a real image, hence the exact scale and look varied slightly in the actual experiment.

was never anyone else in the room and the background noise was minimized, in an effort to ensure proper silence detection. A recreation of the experimental setup can be found in Figure 4.

Before starting the conversation, the participants were asked to read a short introductory text, explaining what is expected of them. This text can be found in Appendix A. In a realistic use case there would be some time between the two conversations, however due to time constraints both conversations were held within a relatively short time frame of about 20 minutes. After the participants had completed both interactions with FitBot, they were interviewed in order to assess their perception of the robot, the user experience and the perceived personalization.

4.3 Measures

After the participants have had two conversations with FitBot, they were asked to answer questions.

Firstly, they were asked to fill in the Godspeed Questionnaire Series (GQS) [Bar23]. It aims to assess the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of the robot the user is interacting with. It does so, by listing a set of items represented by two terms with opposite meaning, such as fake and natural. The user will then have to select which term more accurately describes FitBot by picking a number on five-stage scale. The exact questionnaire used can be found in Appendix B.

After filling in the GQS, the participant's user experience has been assessed using the User Experience Questionnaire (UEQ). This questionnaire focuses on measuring the perceived attractiveness, efficiency, perspicuity, dependability, stimulation and novelty of the product. It does so in similar fashion as the GQS, but instead uses a seven-stage scale. The exact questionnaire used can be found in Appendix C.

After filling in these questionnaires, the participants will answer a couple of additional questions,

aiming to assess the level of personalization perceived. These are simple questions such as: "Did you feel like FitBot remembered your previous interaction?". The exact questions asked can be found in Appendix D.

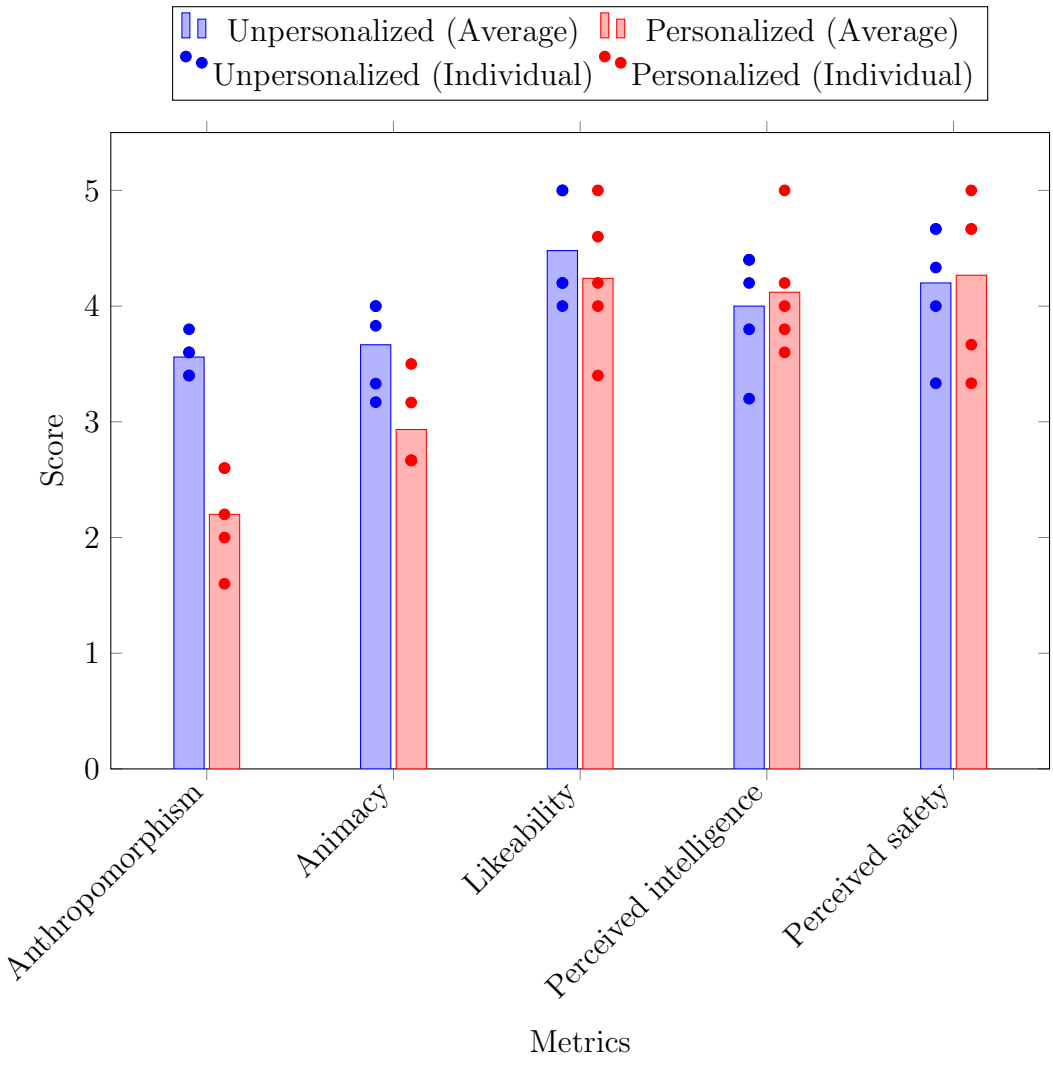

Figure 5: Results from the Godspeed Questionnaire Series.

5 Results

Before listing the results of the experiment, it is important to note that due to the small sample size, no statistically significant conclusions can be drawn. We can, however, still analyze the results and speculate about what may have caused them. We will first go over the quantitative results of the questionnaires in Section 5.1 and then lay out the qualitative results in Section 5.2.

5.1 Quantitative Results

Let us start by laying out the quantitative results from the experiment. We will start by reviewing the Godspeed Questionnaire Series. As can be seen in Figure 5, the unpersonalized version of FitBot scores higher in terms of anthropomorphism animacy. How these differences could be explained will be discussed in Section 6. Both versions score about the same in the categories likeability, perceived

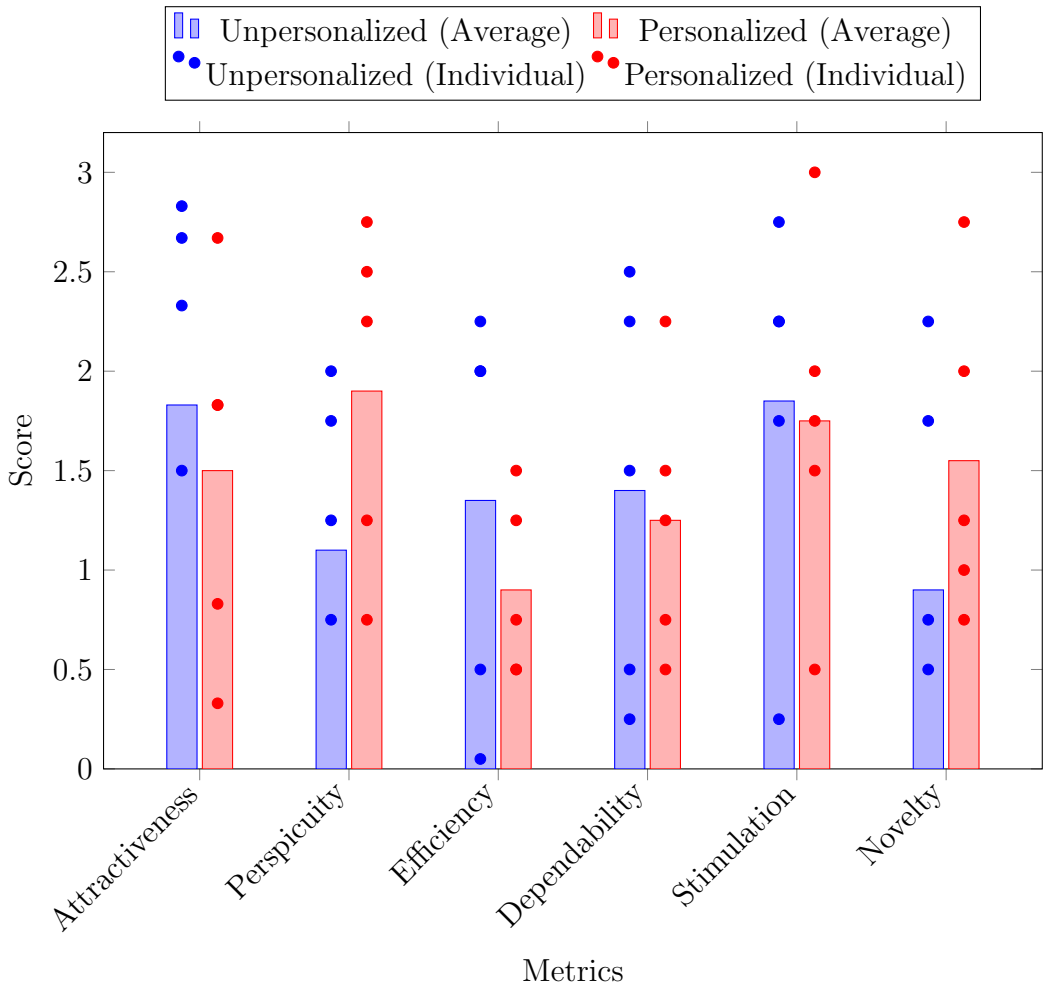

Figure 6: Results from the User Experience Questionnaire.

intelligence and perceived safety.

Now, for the user experience as measured by the User Experience Questionnaire (UEQ), the results can be seen in Figure 6. Here the results are a little more varied. The unpersonalized version scores higher in terms of attractiveness, efficiency, dependability and stimulation, but the personalized version scores substantially higher in terms of perspicuity and novelty.

5.2 Qualitative Results

Finally, we will take a look at the perceived personalization as indicated in the series of more open-ended questions that the participants were asked. The results have been interpreted and a general overview can be found in Table 1.

Noteworthy, is that all of the participants of the personalized version were not hesitant about the fact that FitBot remembered their previous conversation, even though it had only stored a summary of the interaction.

Interestingly, one of the participants of the unpersonalized group also felt like FitBot remembered

User felt like FitBot...	Unpersonalized		Personalized	
	No	Yes	No	Yes
...provided tailored recommendations.	0	5	0	5
...remembered previous interaction.	4	1	0	5
...adapted responses based on previous interaction.	4	1	1	4

Table 1: User perceptions of FitBot as interpreted from the evaluatory questions asked after the experiment.

their previous interaction, because it coincidentally ended up talking about roughly the same topic as before but a bit more in depth. The other participants in the unpersonalized group did not feel like FitBot had remembered their previous conversation.

With the unpersonalized version, there were some participants that indicated that it would have been nice if FitBot had referred to their previous interaction in some way. One participant also said that if the robot needed to be more personal, it could have been useful to ask some questions about the user at the beginning of the conversation in order to create a sort of profile of the user. These are both features the personalized version does have.

With the personalized version, all participants felt like FitBot did not only remember their previous conversation, but also adapted its responses based on what had been discussed. One participant did mention that, although FitBot asked how the workout went and whether the user liked it, it would have been nice if it asked specifically what the user had done in the meantime. So how many sets and reps they had performed, what exercises they had and hadn't done, etc. Another participant mentioned that they found it a nice touch that FitBot mentioned their name every now and then, adding to the perceived personalization.

Participants from both groups indicated that they enjoyed the experience and found FitBot to be friendly to them. In some cases, detecting whether the conversation had ended and providing the terminal keyword in their response took FitBot a bit long, leading to a rather superfluous final greeting and confirmation that the user was indeed out of questions. Also, especially during the profiling phase at the beginning of the personalized conversation, the 1.5 seconds of silence that was needed for FitBot to acknowledge that the user was finished speaking after it had said a phrase that could potentially have been added upon, felt too long in some cases, leading to somewhat of an awkward silence. In other cases, particularly when the user had a slower rate of speech, the silence detection threshold felt perfectly fine.

5.3 Reflecting on Research Questions

As to the effectiveness of the summarization technique in convincing the user of the fact that FitBot remembered their previous conversation, the participants felt confident that it had retained the entirety of the former interaction. Of course it's unknown to what extent this would be the case when interacting with FitBot in a more frequent or more in-depth manner, but the results from the experiment are promising.

We should also discuss the smoothness of the interaction. As mentioned in Section 4.1, many considerations where taken into account when designing the architecture in order to ensure a smooth human-robot interaction. Delays were minimized and systems were put in place to interpret

the user's voice input as efficiently and accurately as possible.

Although the user experience was generally positive for both the unpersonalized and personalized version, limiting factors became apparent during the experiment, especially during the introductory phase of the conversation. If the user mentioned only their name or PIN code in their response, FitBot sometimes had trouble recognizing the transcribed input as complete and continued listening for a potential extension.

While the implemented silence detection helped ensure that conversations would eventually resume without significant disruption, these interruptions did retract from the overall smoothness of the interaction.

Also, if FitBot has a hard time identifying the user's name, the user will have to explicitly say only their name in their response, without including any opening statement, such as "My name is". This led to some confusion in one of the conversations.

Additionally, the inability of FitBot to listen while speaking, could have contributed to a less natural flow of the conversation, further impacting its smoothness. Overall, the developed architecture was sufficiently smooth for the sake of the experiment and in order to provide a satisfactory user experience according to the UEQ, and the conversations were conducted with a reasonable level of fluidity, but the system could definitively be improved upon in further iterations.

6 Discussion

We will first reflect on the findings of this study, then address some of its limitations and lastly discuss potential future work.

6.1 Reflecting on Results

Firstly, we will discuss the results of the GQS. The decrease in anthropomorphism could be because of the somewhat unnatural profiling phase at the beginning of the personalized interaction. For instance, a real fitness coach would not request a PIN code, nor would they likely confirm the client's name in the manner that FitBot does. For FitBot it is crucial to obtain this information and to do so accurately, however, it is fathomable that such interrogation does not feel particularly natural, leading to the users attributing fewer human-like characteristics to FitBot when interacting with the personalized version. A similar thing can be said for the perceived animacy.

Now let us move on to the findings resulting from the UEQ. The most noteworthy drawbacks of the personalized version are the decrease in attractiveness and efficiency. This could also be explained by the perhaps rather cumbersome introductory section, where the user's details are acquired. With the unpersonalized version, the user can immediately start asking questions, seemingly leading to a smoother user experience.

After this initial profiling phase, though, the user does receive a more personal treatment. This enhanced personalization seemed to have the greatest impact on the perspicuity and novelty, as rated by the participants.

The increase in perspicuity, which is defined as the quality of being clear and easy to understand, could be explained by the continuity across the two conversations. FitBot's contextual understanding of the interaction could have led to less repetition in explanations and general responses, as well as maintained coherence in conversations.

The higher rated novelty experienced by the participants while interacting with the personalized version could also be explained by FitBot referring to the topics discussed in the previous interaction. This provided personalization may have made the experience feel fresh and different compared to the generic responses generated in the unpersonalized version. The personal touch could have made the interaction feel more novel.

As to the results of the final handful of questions that were asked, it is promising that all of the participants interacting with the personalized version were under the assumption that FitBot had fully remembered their previous interaction. This indicates that not the entire conversation needs to be stored in order for the virtual assistant to provide a personalized experience and that perhaps just a summary is sufficient. Of course, it should be mentioned that with more intense use the limitations of this approach could become more apparent, especially when referring to specific exercises mentioned in previous conversations, as these were often not included in the generated summary.

6.2 Limitations

As mentioned before, due to the low sample size used during the experiment, the results of this study are not statistically significant.

Additionally, the experiments did not fully reflect a real-world use case of FitBot. The time between two conversations was only short. If there was more time in between the interactions this could perhaps alter the results of the experiment.

Moreover, the participants only had two relatively short interactions with FitBot. In a realistic scenario FitBot would be used more often and more intensively. This more vigorous use could potentially reveal some of the limitations of using summarization as a technique for personalizing human-robot interaction compared to other techniques.

In addition to this, it could be argued that using summarization might not be necessary at all. The context size of current LLMs is already rather large and will only increase over time, so why bother summarizing the conversations at all? Although the LLM might be able to make its way through a large history of conversations, these conversations still need to be stored somewhere. Especially when a system has a lot of users and when conversations are long or frequent, storing all of these conversations could require a lot of available data storage.

Additionally, shrinking the conversations down to mere summaries could greatly reduce the time it takes the LLM to process the contents of the conversation, reducing delay and improving user experience.

Moreover, by forcing the LLM to summarize the conversations, it narrows its focus to what it thinks are the key points of the conversation. As interactions accumulate over time, the sheer volume of information can become overwhelming, even for large context window LLMs. Summarization helps manage this information overload by presenting only the most crucial elements.

And finally, summarization could be beneficial when human oversight or intervention is required. Humans can process the summaries much faster than the entire conversation history, so they provide a quicker way for human operators to understand user contexts and preferences.

Besides, we are in a transitional phase in terms of AI. It could very well be the case that human assistants will be replaced by virtual assistants in the future. These human assistants are likely to have important details and events from the conversations with their clients documented somewhere. If AI was to take over their clients, it would be beneficial to know that they are able to provide a personalized experience using only these details or summaries.

6.3 Future Work

It would be interesting to see a similar study being done at a larger scale. By having more participants and by allowing for more time in between and possibly more conversations in general, the resulting findings would hold a lot more weight.

There are several other complementary studies that could be done of which the outcome would be intriguing. A more long-term study, for instance, could be performed where participants interact with FitBot or a similar virtual assistant over the span of several weeks or months.

Or different summarization techniques could be experimented with and compared, such as extractive versus abstractive summarization techniques.

Another interesting study could be to somehow include the users non-textual information, such as their physical activity data, voice tone or visual cues into the summaries to provide an even more personalized experience and see what effect this has.

Besides carrying out these follow-up studies, it could be constructive to improve upon the developed system. This could be done by allowing the user to interrupt FitBot if desired so, as well as by further perfecting the flow of the profiling phase. For instance by providing the user with

more freedom in the manner in which they answer. Furthermore, although the LLM-driven method used to determine whether the user was finished speaking worked well in most cases, there were instances were it hindered the smoothness of the interaction. This could potentially be mitigated by using an AI that is trained specifically for this purpose.

Also, a designated screen for the FitBot would be beneficial compared to using a laptop to display the web page. This would make for a more integrated system and possibly a better user experience.

7 Conclusions

To summarize our findings, we can state that summarization as a technique for personalizing human-robot interaction with a virtual assistant, powered by an LLM, seems to have a varying effect on the user experience. Although the anthropomorphism and animacy of the robot, as perceived by the users, along with the attractiveness and efficiency of the general user experience seemed to have decreased, the rated perspicuity and novelty of the product were rated substantially higher. Of course, due to the low sample size, these are mere suggestions and definitive conclusions can only be drawn if further research is to be done.

Additionally, the summarization technique seems to be effective in convincing the user that FitBot had remembered the previous interaction in it's entirety, although, again, more intensive testing would have to be done before conclusions can be drawn.

In terms of smoothness, the developed architecture was sufficiently smooth in order to provide a satisfactory user experience according to the UEQ and the conversations were conducted rather intuitively and fluently. The delay between the user being finished speaking and FitBot providing a response was generally acceptable and did not hinder the smoothness of the user experience too much, especially considering the addition of the status indicator on the web page was able to provide the user with a better sense of what FitBot was currently occupied with.

References

[AARJ23] Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, and Ramesh Jain. Conversational Health Agents: A Personalized LLM-Powered Agent Framework. *arXiv*, October 2023.

[Ass24] OpenAI Platform, July 2024. [Online; accessed 10. Jul. 2024].

[Aut22] Autobahn|Python — autobahn 22.8.1.dev1 documentation, October 2022. [Online; accessed 10. Jul. 2024].

[Bar23] Christoph Bartneck. *Godspeed Questionnaire Series: Translations and Usage*, pages 1–35. 02 2023.

[BBB+22] Hayet Brabra, Marcos Báez, Boualem Benatallah, Walid Gaaloul, Sara Bouguelia, and Shayan Zamanirad. Dialogue management in conversational systems: A review of approaches, challenges, and opportunities. *IEEE Transactions on Cognitive and Developmental Systems*, 14(3):783–798, 2022.

[BSR+20] R. Bavaresco, D. Silveira, E. S. Reis, J. Barbosa, R. Righi, C. M. Costa, Rodolfo Stoffel Antunes, M. M. Gomes, Clauter Gatti, Mariângela Vanzin, Saint Clair Junior, Elton Silva, and Carlos Moreira. Conversational agents in business: A systematic literature review and future research directions. *Comput. Sci. Rev.*, 36:100239, 2020.

[CDK+20] Lorainne Tudor Car, D. Dhinagaran, B. M. Kyaw, T. Kowatsch, Shafiq R. Joty, Y. Theng, and R. Atun. Conversational agents in health care: Scoping review and conceptual analysis. *Journal of Medical Internet Research*, 22, 2020.

[CLK10] Heather Cole-Lewis and Trace Kershaw. Text Messaging as a Tool for Behavior Change in Disease Prevention and Management. *Epidemiologic Reviews*, 32(1):56–69, 03 2010.

[CvK18] I. Cornelisz and Chris van Klaveren. Student engagement with computerized practising: Ability, task value, and difficulty perceptions. *J. Comput. Assist. Learn.*, 34:828–842, 2018.

[ENT22] A. Egri-Nagy and Antti Törmänen. Advancing human understanding with deep learning go ai engines. *IS4SI 2021*, 2022.

[Fas24] Using Socially Assistive Human–Robot Interaction to Motivate Physical Exercise for Older Adults, May 2024. [Online; accessed 23. May 2024].

[FWR19] Ahmed Fadhil, Yunlong Wang, and Harald Reiterer. Assistive Conversational Agent for Health Coaching: A Validation Study. *Methods Inf. Med.*, 58(01):009–023, June 2019.

[Goo24] Speech-to-Text AI: speech recognition and transcription, July 2024. [Online; accessed 2. Jul. 2024].

[GPT24] GPT-3.5 Turbo LLM by OpenAI, July 2024. [Online; accessed 2. Jul. 2024].

[HGKM23] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen Macneil. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. In *UIST '23 Adjunct: Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–3. Association for Computing Machinery, New York, NY, USA, October 2023.

[Imp23] Mvs Import. Fruitcore Robotics zeigt Roboter Horst mit integriertem Sprachmodell und neue Solution Kits. *Automationspraxis*, September 2023.

[Int22] Innovatieve en interactieve robots in het onderwijs en de zorg, December 2022. [Online; accessed 10. Jul. 2024].

[IRS+19] Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Dylan F. Glas, Iolanda Leite, and K. Koay. Personalization in long-term human-robot interaction. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 685–686, 2019.

[JB23] Cameron Jones and Benjamin Bergen. Does gpt-4 pass the turing test? *ArXiv*, abs/2310.20216, 2023.

[KTM+17] Spyros Kitsiou, Manu Thomas, G. Elisabeta Marai, Nicos Maglaveras, George Kondos, Ross Arena, and Ben Gerber. Development of an innovative mhealth platform for remote physical activity monitoring and health coaching of cardiac rehabilitation patients. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 133–136, 2017.

[KtSM+20] L. L. Kramer, Silke ter Stal, B. Mulder, E. de Vet, and L. van Velsen. Developing embodied conversational agents for coaching people in a healthy lifestyle: Scoping review. *Journal of Medical Internet Research*, 22, 2020.

[LCT+24] Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. From LLM to Conversational Agent: A Memory Enhanced Architecture with Fine-Tuning of Large Language Models. *arXiv*, January 2024.

[LDT+18] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 07 2018.

[McF23] Christopher McFadden. Meet Pibot: Korea's LLM-powered smart robotic pilot. *Interesting Engineering*, August 2023.

[PMZ+20] Zhenhui Peng, Kaixiang Mo, Xiaogang Zhu, Junlin Chen, Zhijun Chen, Qian Xu, and Xiaojuan Ma. Understanding user perceptions of robot's delay, voice quality-speed trade-off and gui during conversation. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[Rob23] Innovatieve en interactieve robots in het onderwijs - Robotsindeklas, October 2023. [Online; accessed 1. Jul. 2024].

[SRA24] Inc SoftBank Robotics America. NAO, June 2024. [Online; accessed 1. Jul. 2024].

[THS+11] Markku Turunen, Jaakko Hakulinen, Olov Ståhl, Björn Gambäck, Preben Hansen, Mari C. Rodríguez Gancedo, Raúl Santos de la Cámara, Cameron Smith, Daniel Charlton, and Marc Cavazza. Multimodal and mobile conversational health and fitness companions. *Computer Speech Language*, 25(2):192–209, 2011. Language and speech issues in the engineering of companionable dialogue systems.

[WAK+23] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. TidyBot: Personalized Robot Assistance with Large Language Models. *arXiv*, May 2023.

[WAM22] WAMP Programming — autobahn 22.8.1.dev1 documentation, October 2022. [Online; accessed 10. Jul. 2024].

[YD15] Euijung Yang and M. Dorneich. The effect of time delay on emotion, arousal, and satisfaction in human-robot interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59:443 – 447, 2015.

[ZCL+23] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics*, 3(4):100131, December 2023.

Appendix A: Introductory Text for Experiment

FitBot Experiment

First of all, thank you for participating in this experiment. You will be testing the FitBot, a robotic fitness coach designed to answer all your fitness-related questions. It can create personalized workout routines, offer tailored fitness advice, and provide tips on maintaining a healthy lifestyle. For the sake of the experiment, please think of a workout routine or fitness-related question you would like to ask before initiating the conversation. Also, please wait until FitBot is done speaking until you respond.

You will be having two separate conversations with FitBot. In a realistic scenario, these conversations would take place at different points in time, allowing you to implement FitBot's advice or try out the recommended workout routine. However, due to time constraints, both conversations will take place today, with only a short break in between.

To help simulate a more realistic experience, please imagine that a week has passed between the two conversations. During this imagined week, assume you have had the time to incorporate FitBot's advice into your routine and prepare a new conversation topic accordingly. This will help us understand how FitBot's guidance might impact your fitness journey over time.

After completing both conversations with FitBot, you will be asked to fill out a user experience questionnaire. Your feedback is crucial for evaluating the effectiveness and user satisfaction with FitBot.

Thank you again for your participation.

Appendix B: Godspeed Questionnaire Series

English

Translated by: Christoph Bartneck
Publication: https://doi.org/10.1007/s12369-008-0001-3
Instructions: Please rate your impression of the robot on these scales:

Anthropomorphism

Fake	1	2	3	4	5	Natural
Machinelike	1	2	3	4	5	Humanlike
Unconscious	1	2	3	4	5	Conscious
Artificial	1	2	3	4	5	Lifelike
Moving rigidly	1	2	3	4	5	Moving elegantly

Animacy

Dead	1	2	3	4	5	Alive
Stagnant	1	2	3	4	5	Lively
Mechanical	1	2	3	4	5	Organic
Artificial	1	2	3	4	5	Lifelike
Inert	1	2	3	4	5	Interactive
Apathetic	1	2	3	4	5	Responsive

Likeability

Dislike	1	2	3	4	5	Like
Unfriendly	1	2	3	4	5	Friendly
Unkind	1	2	3	4	5	Kind
Unpleasant	1	2	3	4	5	Pleasant
Awful	1	2	3	4	5	Nice

Perceived Intelligence

Incompetent	1	2	3	4	5	Competent
Ignorant	1	2	3	4	5	Knowledgeable
Irresponsible	1	2	3	4	5	Responsible
Unintelligent	1	2	3	4	5	Intelligent
Foolish	1	2	3	4	5	Sensible

Perceived Safety

Anxious	1	2	3	4	5	Relaxed
Calm	1	2	3	4	5	Agitated
Still	1	2	3	4	5	Surprised

Appendix C: User Experience Questionnaire

Please assess the product now by ticking one circle per line.

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

Appendix D: Questions to Assess Perceived Personalization

1. **Did FitBot provide recommendations that felt tailored to your specific needs?**

   .....

   .....

   .....

2. **Did you feel like FitBot remembered your previous interaction?**

   .....

   .....

   .....

3. **To what extent did FitBot adapt its responses based on your previous interaction?**

   .....

   .....

   .....

4. **Did you feel that FitBot took into account your personal fitness history and progress during the conversations?**

   .....

   .....

   .....

5. **Were there any instances where you felt FitBot could have better personalized its responses? If so, please describe.**

   .....

   .....

   .....