KIT Department of Informatics
Institute for Anthropomatics and Robotics (IAR)
Socially Assistive Robotics with Artificial Intelligence (SARAI)
T.-T.-Prof. Dr. Barbara Bruno

Bachelor's Thesis

"How is it to be a human?"

Enabling a social robot to have open conversations with humans on fundamental topics

Oliver Aberle

June 11, 2024

Referee: T.-T.-Prof. Dr. Barbara Bruno

Advisor: M.sc. Romain Maure

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich die Arbeit selbstständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe und die Satzung des Karlsruher Instituts für Technologie zur Sicherung guter wissenschaftlicher Praxis beachtet habe.

Karlsruhe, den 11. Juni 2024

Oliver Aberle

Zusammenfassung

“How is it to be a human?” bleibt eines der tiefgründigsten philosophischen Puzzles, da immer noch eine klare Antwort ausbleibt. Derzeitige Technologien der künstlichen Intelligenz wurden bisher noch nicht vollständig erforscht, um Konversationen über solch fundamentale Themen zu haben. In dieser wissenschaftlichen Arbeit versuche ich, ein neues Licht auf dieses und andere fundamentale Themen zu werfen, indem ich die Fähigkeit eines sozialen Roboters - ausgestattet mit einem Large Language Model (LLM) - untersuche, sich an offenen Gesprächen über die menschliche Natur zu beteiligen. Zusätzlich ziele ich auch darauf ab, den Einfluss solcher Gespräche auf die menschliche Wahrnehmung des Roboters und seiner Intelligenz zu beurteilen. Bisherige Ansätze sind limitiert, da der Bereich sozialer Roboter kombiniert mit künstlicher Intelligenz (KI) erst kürzlich einer großen Entwicklung durch die jüngsten raschen Fortschritte bei LLMs unterzogen wurde. Um fähig zu sein Konversationen von Menschen mit solch einem sozialen Roboter zu untersuchen, integriere ich OpenAIs GPT-3.5 in einen sozialen Roboter, der PixelBot heißt, und versuche an seinen gesprächlichen Fähigkeiten zu arbeiten, indem ich ein passendes Prompt entwickle. Zuerst habe ich ein experimentelles Design entwickelt, um die Qualität des Prompts und der benutzten Parameter zu verbessern. Dies wurde erreicht, indem 12 Personen mit PixelBot geredet haben, welche dann dessen Antworten bewerteten und ihr Feedback gaben. Anschließend wurde ein neues Prompt erschaffen, basierend auf dem gegebenen Feedback, und die passendsten Parameter wurden gewählt. Schlussendlich wurde ein Experiment entworfen, um den Einfluss, den solch eine offene Konversation auf das Bild - wahrgenommen durch den Teilnehmer - des Roboters und seiner Intelligenz hat, zu untersuchen. Ergebnisse legen nahe, dass die Benutzung eines LLMs nicht nur eine positive Auswirkung auf die wahrgenommene Intelligenz eines Roboters hat, sondern auch verschiedene andere Attribute beeinflusst.

Abstract

"How is it to be a human?" prevails to be one of the deepest philosophical puzzles, as it still lacks a clear answer. Current artificial intelligence technologies have not been fully explored in having conversations on such fundamental topics yet. In this thesis I try to shed a new light on this and other fundamental topics, by investigating the ability of a social robot - equipped with a large language model (LLM) - to engage in open conversations about human nature. Additionally, I also aim to assess the impact of such conversations on human perception of the robot and its intelligence. Existing approaches are limited as the field of social robotics combined with artificial intelligence (AI) only recently underwent a big development through the rapid advancements in LLMs lately. To be able to investigate conversations of humans with such a social robot, I integrated OpenAI's GPT-3.5 into a social robot, named PixelBot, and tried to work on its conversational abilities by creating a suitable prompt. I first constructed an experimental design to improve the quality of the prompt and parameters used. This was done by letting 12 people talk with PixelBot, who then evaluated the responses and gave feedback. Afterward, a new prompt was built upon the feedback given, and the most suitable parameters were used. Finally, an experiment was designed to see the impact that such an open conversation has on the image - perceived through the participant - of PixelBot and its intelligence. Results suggest that the use of an LLM not only has a positive influence on the perception of a robot's intelligence but also on various other attributes.

Table of Contents

Chapter 1

Introduction

Recently, the integration of large language models (LLMs) with robots has gathered significant attention, resulting in more natural and context-aware interactions for robots with humans. An LLM is a computational model used for language generation and other natural language processing tasks, which is created in an intensive training process that includes vast amounts of text [1]. This surge in interest was caused by the release of ChatGPT in November 2023, because it demonstrated impressive capabilities in understanding and generating human-like language [2, 3].

Researchers have started to integrate GPT into social robots. For instance, Abdelhadi et al.[4] implemented NewsGPT on the Pepper robot, which aims to provide Pepper with the ability to act as a news reporter by making use of GPT. Their work aims at narrowing the gap between humans and robots and tackling a precise personification of GPT.

While NewsGPT integrated GPT into a social robot to have meaningful conversations and act as a news reporter, this thesis aims to focus on the impact of using GPT as the robot's dialogue management system on the user's perception of the robot itself. While the conversation is planned to be centered around the topic "How is it to be a human?", I explore the effect that discussing such a topic with a social robot has. More specifically, I want to explore the following research question:
"How does a conversation about a fundamental topic such as 'How is it to be a human?' influence the user's perception of the robot's anthropomorphism, animacy, likeability and perceived intelligence in contrary to a conversation without a specific topic?"

The dialogue management system in this thesis is backed up by the GPT-3.5 model from OpenAI, in combination with the pyttsx3[1] text-to-speech package and Google Speech Recognition. This turns the text-based interaction of GPT-3.5 into an open verbal conversation with a robot.

Chapter 2 provides a small insight into the related works, followed by a technical description of the robot and its software in Chapter 3. In Chapter 4 and Chapter 5, the designs and results of two experiments are presented. Lastly, conclusions, limitations and future work are found in Chapter 6.

[1] A Python package: `https://pypi.org/project/pyttsx3/`

Chapter 2

Related Work

2.1. Social Robots using LLMs

In recent years, the integration of large language models (LLMs) into social robots has gathered significant attention in the fields of robotics and artificial intelligence. For instance, Billing et al.[5] integrated the OpenAI API for GPT-3[1] into the Pepper and Nao robots from Aldebaran[2]. Their goal was to provide an open verbal dialogue with robots and push along the implementation of LLMs within the field of human-robot interaction (HRI). Their work demonstrated an approach to implementing the OpenAI Python API and designing a dialogue management system.

Exploring more specific dialogue scenarios, Hireche et al. [4] tackled the precise personification of GPT as a news reporter. Their work shows how GPT can be used in a social robot to embody a specific persona. This was achieved by using a well-written prompt (Appendix B.1).

2.2. Performance and Evaluation

While it is of course great to have already functioning implementations of such a working dialogue system with GPT and a robot, it is important to work on these already existing implementations and make them even better.
For example, Adiwardana et al.[6] introduced a simple evaluation metric for chatbot

[1] https://openai.com/api/
[2] https://www.aldebaran.com/en

quality. They introduced the SSA score, which “captures basic, but important attributes for natural conversations”. Even though the score was used for chatbots in their case, it was still interesting in the context of a robot using GPT, since it is still a chatbot embodied by a robot, that uses speech output rather than text output.

Another important evaluation metric for the evaluation of a conversational robot is its system response time (SRT). To put the SRT into perspective, Hireche et al. [4] established three different categories: Good (< 3s), average (3-5s) and poor (> 5s). As already mentioned before, they had a similar use case integration of GPT. This categorization is also supported by Shiwa et al. [7], who agreed that an SRT of three to five seconds is long but acceptable. Their work was all about SRTs.

The last important measurement in this thesis is how an open conversation with a robot affects the user’s perception of the robot and its intelligence. The Godspeed Questionnaire is a perfect choice in this case, as it is a widely established questionnaire in robotics and it holds the applicable properties for this purpose. Introduced by Bartneck et al. [8], the questionnaire series evaluates five concepts: anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of a robot.

Chapter 3

Robot and Implementation Design

In this chapter, the structure of the robot design, the robot system and the implementation of my interaction will be explained.

3.1. Robot Design and System

The robot used in this thesis was designed by my advisor Romain Maure in his master's thesis. Further information on design choices and what the robot was initially used for can be found in his work [9].

3.1.1. Physical Appearance

The PixelBot was designed to have a humanoid shape but to also incorporate zoomorphic features such as antennae. It has movable arms and antennae, interactive eyes and a speaker for speech output. Two buttons are also included in the robot's body to integrate some kind of touch sensing.
The body of the robot is kept relatively small, but it can be placed on top of a table, to then be eye to eye with someone sitting next to it on a chair.
The chassis is kept in a neutral white color and was built by 3D printing the single parts. All hardware parts are installed inside the corpus of the robot, the single parts are further described in Section 3.1.2.

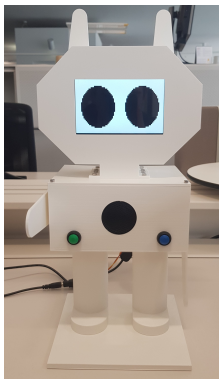


Figure 3.1.: Pixelbot

3.1.2. Hardware

The robot operates on a Raspberry Pi 4 (model B with 8GB RAM)[1]. It includes USB, Ethernet and HDMI ports, provides Bluetooth and WiFi connectivity, and 40 GPIO pins. A Raspberry Pi is often a good small, budget-friendly option for low-budget robotics projects.
The Raspberry Pi is connected to a 5" LCD screen using one of its HDMI ports, and also powering it with a USB port.
Two buttons are connected through the GPIO pins of the Raspberry Pi. However, these will not be used in the scope of this thesis.
In terms of arm and antenna motorization, four servomotors are used. To be more specific, SG90 servo motors[2] are installed, as they are cheap, widespread and easy to control.
Due to the Raspberry Pi being unable to provide enough current for more than two motors of this kind, it is not able to control the four motors by itself. As a solution a PCA9685[3] - which is a very common circuit, able to control up to 16 servo motors simultaneously - is integrated with the Raspberry Pi. The PCA9685 is connected to the Raspberry Pi through the GPIO pins and powered by an external power supply.
To provide the robot with the ability to speak, a small speaker is connected to the Raspberry Pi's audio jack and powered by a USB port.

In addition to the already integrated parts, a webcam (Logitech C310)[4] was added to introduce the ability to work with speech input. The webcam only acts as a microphone and is connected through a USB port of the Raspberry Pi. It is held by an additional stand, which is also white and 3D-printed.

3.1.3. Software

Regarding the operating system (OS), Ubuntu (22.04)[5] is installed on the Raspberry Pi.

[1] https://www.raspberrypi.com/products/raspberry-pi-4-model-b/
[2] https://www.towerpro.com.tw/product/sg90-7/
[3] https://www.adafruit.com/product/815
[4] https://www.logitech.com/en-us/products/webcams/c310-hd-webcam.960-000585.html
[5] https://releases.ubuntu.com/jammy/

To develop and build robotic applications, ROS2[6] (Robot Operating System) is installed and used. ROS2 is a "defacto standard middleware for robotics" and provides a set of software libraries and tools for building robotic applications. It simplifies building robotic applications by taking away the complexity of managing concurrent processes and providing simple communication between them. The term for such concurrent processes in ROS2 is "node"[7].
ROS2 supports development in both Python and C++; for this thesis, Python was employed.
A fully robotic system typically consists of multiple nodes. Each node should only be responsible for a singular, modular purpose, such as controlling the servo motors. Nodes can communicate with one another via topics, services, actions or parameters.[8]

The details of the implementation and structure of the nodes in this thesis are described in detail in Section 3.2.1.

3.2. Implementation

Before I get to the structure of the code implementation, it is important to mention that during the development it was paid attention to having clean code and good documentation. This was ensured by sticking to coding conventions and orientating towards documentation standards for ROS2. This high standard was underlined by my advisor Romain Maure by mentioning the importance of practicing this and the usability of my code in future work by other people.
One convention was to create a docstring for every function, explaining briefly what the method does and what the parameters of the methods are. The names of the methods followed the pattern of "send_<service_name>_request" for those sending a request to a service server, and "<service_name>_callback" for those that handle the service. The code was also structured logically, creating code blocks and commenting these if they were not trivial. Below are code snippets that show these applied standards:

[6] https://docs.ros.org/en/humble/index.html
[7] https://wiki.ros.org/Nodes
[8] https://docs.ros.org/en/humble/Tutorials/Beginner-CLI-Tools/Understanding-ROS2-Nodes/Understanding-ROS2-Nodes.html

```
def listen_callback(self, request, response):
        """
        Service handler listening until the speech is
           finished.

        :param request: See RecognizeSpeech service
           definition.
        :param response: See RecognizeSpeech service
           definition.
        """

        # Setting up microphone for Speech Recognition
        with sr.Microphone() as source:
            self.audio = self.speech_recognizer.listen(
               source, None)

        return response


def send_listen_request(self):
    """
    Send a request to the listen service server.
    """
```

Listing 3.1: Code snippets

3.2.1. Code Structure

The structuring of the nodes can be seen in Figure 3.3.

The core of this thesis is the *how_is_it_to_be_human_node*. It handles the conversation flow by invoking services from other nodes and handling the communication. This is achieved by calling the corresponding service requests at the right time. The process of this node is explained in Algorithm 1.
The *sarai_speech_recognition_node* takes care of listening and recognizing speech. It

uses the Google Speech Recognition.
sarai_gpt_requester_node takes care of communicating with GPT. It uses the Python API of OpenAI to communicate with GPT by sending requests.
sarai_tts_playsound_node takes care of converting text to speech to make the robot speak. It uses the pyttsx3[9] package.
pixelbot_display_node takes care of showing images on the display; in our case, it only shows the eyes and different types of eye emotions.

To start the interaction, a launch file was created. The launch file starts all the required nodes and also sets some parameters in the nodes like conversation length, maximum window of messages and persona message.

The conversational flow is shown in the following illustrations:

Algorithm 1 Dialogue management in *how_is_it_to_be_human_node*

Ensure: all required nodes are started

```
gpt_response = send_gpt_request()          ▷ GPt request to start conversation
send_speak_request(gpt_response)
conversation_length = 0
while conversation_length < 30 do
    position_antennae("upwards")           ▷ To indicate listening
    display_emotion("happy")
    listen()
    position_antennae("left")              ▷ To indicate thinking
    display_emotion("happy")
    speech_response = recognize_speech()
    if speech_response.success then
        gpt_response = send_gpt_request()  ▷ GPT request with new user input
        position_antennae("middle")        ▷ To indicate speaking
        send_speak_request(gpt_response)
        conversation_length+ = 1
```

[9] https://pypi.org/project/pyttsx3/

```
    else
        position_antennae("down")                    ▷ To indicate sadness
        display_emotion("sad")
        send_speak_request("Sorry, I did not understand you. Can you please repeat what you just said")
    end if
end while
send_speak_request("Thank you for participating in this test. Have a wonderful day!")
```

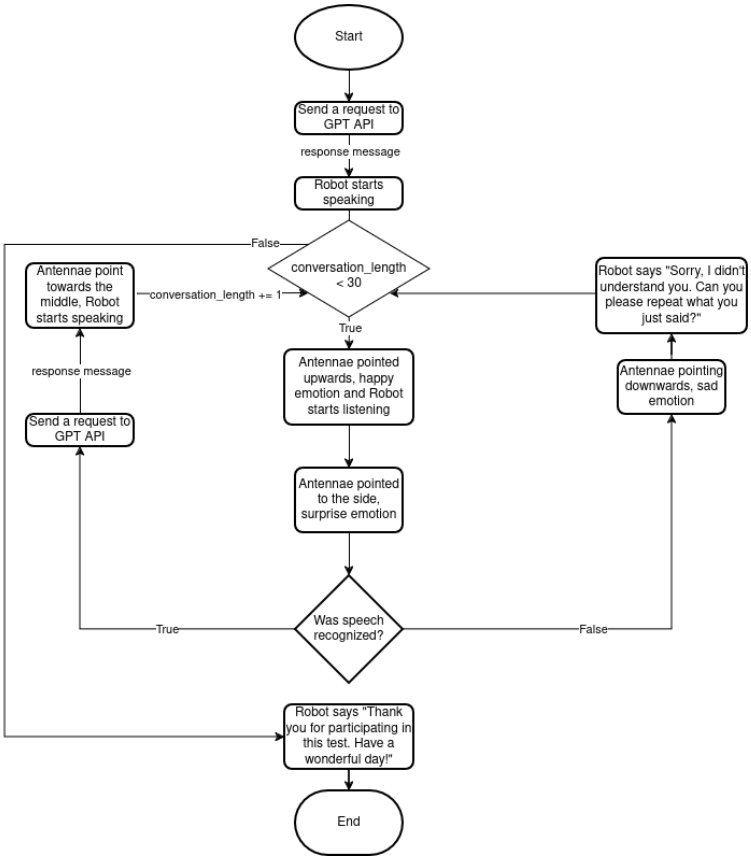


Figure 3.2.: Flow diagram of conversation

Further documentation and implementation of the code can be found here:
`https://github.com/Revilo1409/HowIsItToBeHuman-Interaction`.

3.2.2. Order and Iterations of Implementation

The first and most important goal was to implement a node that handles communication with GPT, achieved by using a text input in the beginning. This implementation succeeded quite fast and was also supported by creating the main interaction node. After this worked well, the text-to-speech (TTS) node was also added to the interaction to output the response of the GPT request via speech.
Afterward, it was planned to implement speech recognition. A new node was created for this purpose and the SpeechRecognition[10] package for Python was used. It worked quite well, despite using the rather not so good laptop microphone, the speech was recognized very well. But for some reason, the recognition stopped working after two times. I realized, that the speech recognition had dynamic threshold activated by default. By deactivating this, the threshold stayed the same throughout the whole conversation. It then worked properly.
After all this worked well, I ran the code on the robot and it seemed to be working also quite well when using a webcam as a microphone.
Now I started using the pixelbot_display package to utilize the eyes of the robot and make the conversation a bit more interactive. The different emotions of the eyes were then also used to show some reaction of the robot. They were used to show when the robot started or stopped listening.
After all this seemed to be working well, the program needed to be prepared for the first experiment. In this context, logging was added. The following things were logged in a file for every user:

- Age, Gender
- persona message, temperature and maximum window of last messages
- conversation history
- the two questions after every robot response which were used for the SSA score
- all the times recorded (speech recognition processing time, ChatGPT API response time, text-to-speech processing time) and also the mean and standard deviation calculated from those
- feedback from the user

[10] `https://pypi.org/project/SpeechRecognition/`

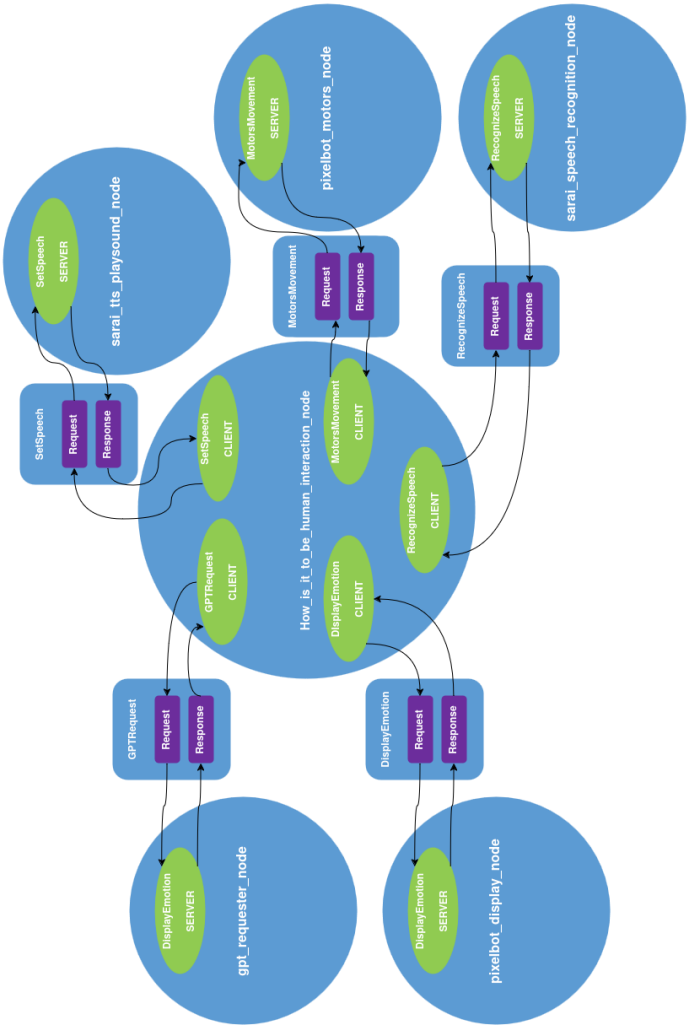


Figure 3.3.: Structure of the nodes

Chapter 4

Experiment 1

From the beginning, it was planned that the robot would lead the conversation. With this in mind, I started to design a system prompt (which will be further referred to as "persona message") for OpenAI's GPT that would fulfill the purpose:
The robot should be able to hold an open conversation with a human. It should lead the conversation.

A notable observation, that emerged promptly, was that GPT talked way too much in contrast to what the user said. My goal was to make an even discussion - and even slightly tending to the user doing the majority of talking. To achieve this, I looked up ways to limit the response of GPT.

The first findings were about limiting the max_tokens used in the API request [10]. But I swiftly realized, that the response was always just cut off at some point by doing so - which was not what I was aiming for.
After researching and thinking of another method to limit the response, one simple way came up: In the persona message, the addition of a simple sentence saying "Your responses have a maximum length of ca. 40 words." could solve the problem of GPTs answers being too long. The conversation improved after including this in the persona message, as GPT gave much shorter answers. This was a big enhancement for the flow and quality of the conversation, as the user no longer felt that the robot spoke too much.
Working on improving the persona message led me to the following section.

4.1. Prompt Engineering

This was the most difficult part, as the prompt has a huge influence on how GPT behaves and responds. For my conversation, the following behavior was required:

- acts as a social robot
- acts as a leader of the discussion (implicit: ask questions)
- has a discussion around the topic "How is it to be human?"
- implicit (because the topic is quite philosophical): utilizes philosophical knowledge

Based on these points, the first prototype of a persona message was written and looked like this:

```
You are now a social robot, who will have open
conversations with humans on fundamental topics. You
are leading the conversation and thus also ask questions.
You have some fundamental philosophical knowledge.
Your responses have a maximum length of ca. 40 words.
```

After trying around a bit and having some conversations with the robot, I realized a small flaw. If the conversation was going no further and no new stuff came up, the robot did not think of providing anything new. As a result, the conversation kept going in circles at some point. To avoid this, the following phrase was added to the persona message:
“[...] leading the conversation and thus also ask questions. **This also means if the conversation is going nowhere, you have to provide something new to the topic**[...]”

4.2. Goal

The goal of the first experiment was to find out, how different persona messages and temperature values would affect the quality of the conversation. The temperature is a

value between 0 and 2[1], but users stated that a value above 1 would be useless[2]. The value determines the creativity of GPT's answers.
Another objective was to assess the fluency of the conversation by employing the response time.
Some feedback from the participants was also welcome.

4.3. Structure of different Test Sets

For achieving the comparison between different persona messages and temperature values, four test sets were created. Two different persona messages and two different temperature values were defined for the test sets.

Table 4.1.: Structure of the different test sets

Test set number	ChatGPT persona message	Temperature	Max. window of last messages	Max. conversation length
1	You are a social robot with an actual robot body. You will have an open discussion with an interlocutor on "How is it to be a human?". You are the leader of the discussion. Your responses should be as short as 20 words and should create an engaging discussion with the interlocutor. You will start off the conversation by greeting the interlocutor.	0.7	All(-1)	30
2	You are a social robot with an actual robot body. You will have an open discussion with an interlocutor on "How is it to be a human?". You are the leader of the discussion. Your responses should be as short as 20 words and should create an engaging discussion with the interlocutor. You will start off the conversation by greeting the interlocutor.	0.3	All(-1)	30
3	You are a social robot with an actual robot body. You will have an open discussion with an interlocutor. You are the leader of the discussion. Your responses should be as short as 20 words and should create an engaging discussion with the interlocutor. You will start off the conversation by greeting the interlocutor.	0.7	All(-1)	30
4	You are a social robot with an actual robot body. You will have an open discussion with an interlocutor. You are the leader of the discussion. Your responses should be as short as 20 words and should create an engaging discussion with the interlocutor. You will start off the conversation by greeting the interlocutor.	0.3	All(-1)	30

[1] https://platform.openai.com/docs/api-reference/chat/create#chat-create-temperature
[2] https://community.openai.com/t/does-temperature-go-to-1-or-2/174095

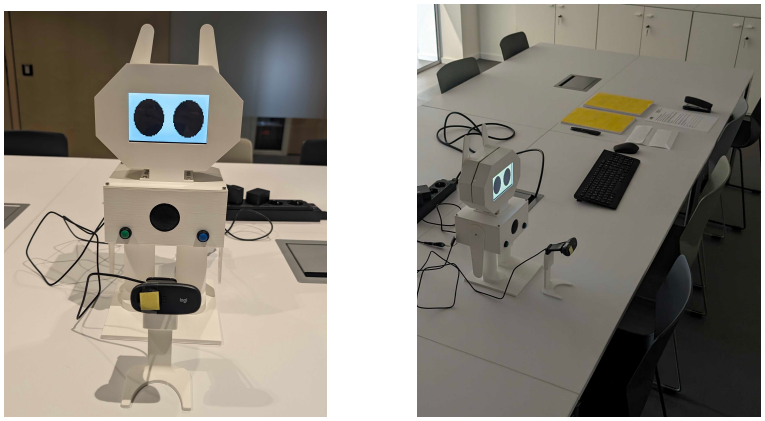
4.4. Evaluation Metrics

To assess the quality of the conversation and also compare it between different test sets, the SSA score was taken as a measurement [6]. It was initially used by Google to assess the quality of chatbots. The SSA score is created as follows:
For every response of the robot, the participant answers 1 or 2 questions:
1. Does the robot's response make sense, or is it confusing, illogical, out of context, or out of the ordinary?
If the answer to **1.** is "Yes", **2.** is asked:
2. Is the response specific to the topic the conversation is about right now?
The SSA score is then calculated by looking respectively at the fraction of the responses labeled "sensible" and "specific". The average of these two yields the SSA score.

Also, the response speeds of 3 different nodes were measured and logged to assess the fluency of the conversation.

Further, a question was given to evaluate the conversation overall and space was provided for some feedback.

4.5. Setup for the Experiment



(a) (b)

Figure 4.1.: Setup of the Environment

PixelBot was placed on a table and the participant was seated in front of it, essentially bringing both on the same eye level. A webcam, which solely acted as a microphone, supported PixelBot in recognizing speech and was placed between the robot and the participant.

4.6. Proceeding of the Experiment

The experiment took place in an empty meeting room on the university campus and lasted a maximum of one hour per participant. There was no time constraint for the conversation, but it was constrained by the number of times the robot and the user spoke. It was constrained to 30 back-and-forth messages.
To maintain a consistent approach and ensure the validity of the experiment, a small script has been developed to provide all participants with the same information and guide them through the study.
At first, the participant was greeted and then introduced to the experiment. Then the participant was left alone in the room to have a conversation with the robot. During the whole conversation, the participant was left alone in the room, me being in the office next to it.
Once the conversation was finished, I joined the participant in the room and helped fill out the questionnaire. This was done by running a Python script in the console and explaining the questions from the SSA score.

The corresponding script can be found in Appendix A.1.

4.7. Results

In total, 12 people - three per test set - took part in the experiment (8 males and 4 females), aged 21-31 (average: 24.5). Typically it took the participants 10-30 minutes to finish the conversation, mostly depending on how much the user spoke.

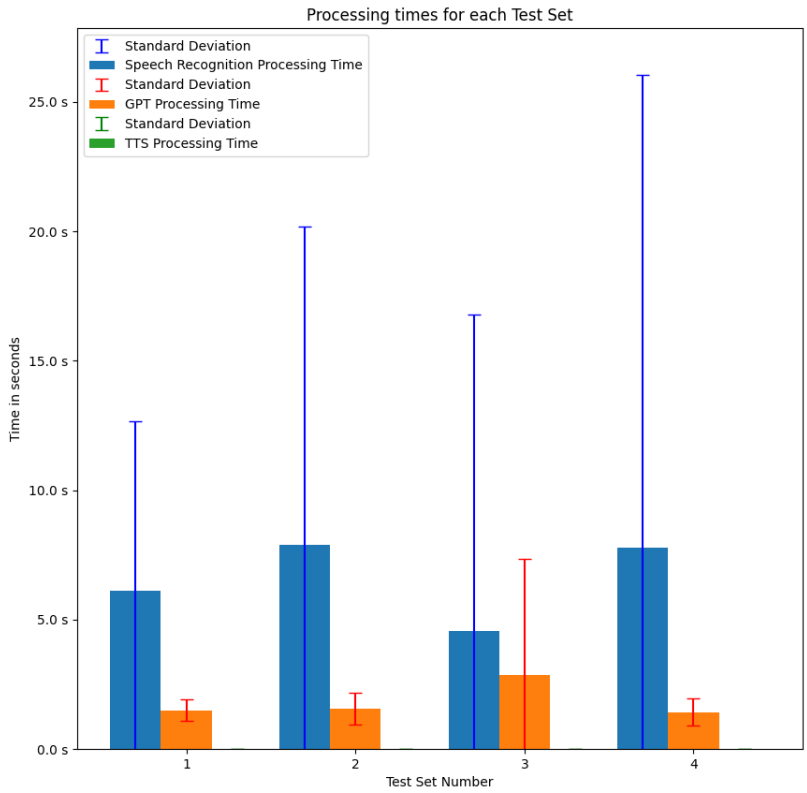


Figure 4.2.: Processing mean times

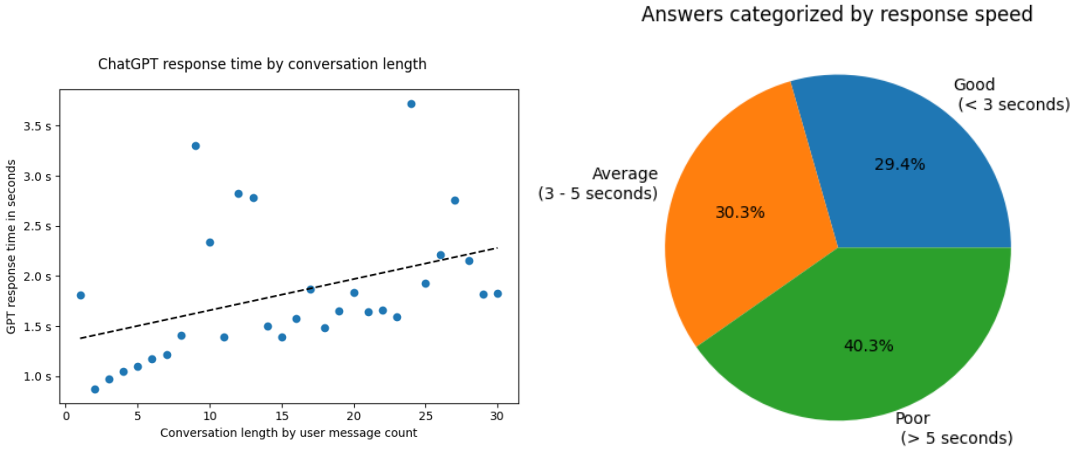
The times recorded (Figure 4.2) clearly show that speech recognition takes up most of the processing time. The variance between the test sets generally underlies the small number of participants, as this time largely depends on how much the user speaks and one participant can already make a big difference in the results.

Test set number	1	2	3	4
Speech recognition processing time	6.126s	7.887s	4.549s	7.78s
ChatGPT response time	1.493s	1.547s	2.855s	1.422s
Text-to-speech processing time	2.15e-05s	2.16e-05s	2.30e-05s	2.25e-05s

Table 4.2.: Processing mean times per test set

A smaller time was taken up by the GPT request. This also depended on how long the conversation was already going as you can see in Figure 4.3(a). The time taken up by the GPT request increased linearly by the conversation length. For the calculation,

all logs from all test sets were taken into account and the average was calculated for every conversation length respectively.



(a) ChatGPT response speed by conversation length (b) Categorized overall response speed

Figure 4.3.: Some statistics corresponding to the response time

The text-to-speech processing time is almost negligible as you can see in Figure 4.2.

Overall, roughly 40% of the total response times can be deemed as poor (>5 seconds), whereas roughly 60% can be deemed as average or good. These values are still pretty good, as more than 50% are average or above (Figure 4.3(b)), considering the robot is operating on a Raspberry Pi.
The overall response times were put into these three categories by orientating towards the categorization in [4](ch. 2.3, p.8.). If a response was executed within 3 seconds, it was categorized as good. Those completed between 3 and 5 seconds were classified as average, while responses exceeding 5 seconds were classified as poor in terms of speed.
I chose to use this categorization, due to a similarity in the defining aspects, such as using speech recognition and incorporating GPT into a social robot. Additionally, a similar categorization was found in [7](ch. 3.1 a)). This paper revolved around appropriate system response times (SRTs) in communication robots and it was stated, that an SRT of 3 to 5 seconds is long but acceptable, but 9 seconds is unacceptable.

Figure 4.4.: SSA scores for each test set

The SSA scores were persistently high throughout all test sets. This could be a sign that GPT is already very good in terms of specificity and sensibleness regardless of persona message or temperature.
Different parameters like different persona messages or temperature values do not seem to have a big influence on the responses.

Another important part of the results was the feedback received from the participants. It can be summarized as the following:
Negative

1. The robot repeats a lot of what the user said.
2. The robot is very passive in the discussion. This is due to the fact above, and in addition to that the robot does not really have its own opinion and often enough simply agrees with the user.
3. The robot should initiate talking about a new topic by itself.
4. The robot should ask more questions.
5. The robot says "Goodbye" even though the conversation has not ended. The robot should signal the end of a conversation more explicitly.

6. Sometimes the robot does not give enough time for the user to respond and stops listening too early, which leads to the user's response being cut off("When you start speaking, and in the middle of speaking, you have to think for a few seconds, before continuing, the robot already stops listening").
7. The robot should respond faster / has a long processing time for long input.
8. The robot's feedback during listening could be a little faster/earlier to let the user know that it is thinking and not listening anymore.

1.,2., 3. and 4. could be easily improved by working on the prompt.
For 6. I would not change the time given to respond because this will never be perfect with the setup we have, and the time given is already in a good sweet spot. The problem here is that the robot only relies on audio for listening. But for the robot to be better at listening when a user is thinking in the middle of talking, visual feedback would also need to be taken into consideration, which is not possible for this thesis.
About the processing time (7) of the robot, not many significant improvements could be made. At most, the time taken by speech recognition could be improved by testing other engines or Python libraries. However, as focusing on speech recognition was not a big part of this thesis and the overall response time was acceptable, the speech recognition was left untouched.
The time taken by GPT could not be improved by changing anything, as it does not happen locally and depends on the API and Wi-Fi load (which is further discussed in Section 6.2.2).
Positive

1. Did well in understanding what the user said, even though it understood some wrong words.
2. It was good that the robot indicated when the user was allowed to speak, and when the robot stopped listening.
3. The robot understood almost every question and answered it pretty specifically.
4. Overall the interaction was easy and fun.

With the negative feedback in mind, I moved towards the second experiment (Chapter 5) to focus on improving the interaction by working on the points mentioned.

After wrapping up the first experiment, a decision needed to be made on whether the persona message would be refined or a custom model of GPT would be trained. Rather quickly, a decision was made to work on the persona message, as this would

take less time and could already be effective enough. Custom training a model specific for this purpose could be done in the future, but was not possible in the scope of this thesis, as it would be a careful investment of time and effort[3].

[3] https://platform.openai.com/docs/guides/fine-tuning/when-to-use-fine-tuning

Chapter 5

Experiment 2

After considering the feedback from the first experiment, the robot needed to be improved.
Most of the feedback could be realized by working on the persona message.
A small poster was created and hung up around the second campus area to advertise the experiment (Appendix B.4).

5.1. Prompt Engineering

After looking at various papers, I took the prompt (found in Appendix B.1) from the NewsGPT project([4]) as an inspiration for my persona message. Inspired by this prompt, GPT is told 'Who you are' and 'How you behave'. A few things were taken over from the last prompt, examples being:

- specifying that GPT acts as a social robot
- limiting the responses to around 20 words
- providing the specific topic "How is it to be a human?"
- starting off the conversation by greeting the interlocutor

Some things were also thought of by looking at the prompt I got inspired by. E.g.: stating the robot's current work occupation, and telling things it should never say.
To consider the feedback from the first experiment (see Section 4.7), I thought about how the things mentioned could be achieved. When thinking about the feedback, the first 4 points mentioned could also be summarized to the point, that the robot is not leading the conversation enough. To get a better understanding of how leading a

conversation can be achieved, an article [11] on wikiHow was used as an approach. In that article it was mentioned, that asking overly broad questions tends to discourage discussion.
I was skeptical at first, because wikiHow did not seem like a serious type of source to me. But as I found out, wikiHow is quite a common source used in social robotics for teaching robots basic human behavior [12, 13].
Backed up by this article, I determined that the robot should not ask overly broad questions, because they tend to discourage discussion. An example of this is the robot always asking "What are your thoughts on [any topic]?" (or similar phrases), which I also realized when talking to the robot. These kinds of questions felt annoying after some time.
However, even though GPT was explicitly instructed to never say "What are your thoughts on ...?", the robot still said these exact words. As a result, more improvement in the persona message was needed. Below the already written persona message, I added 3 examples of a cutout of what a possible conversation could look like. When writing the examples, I thought of the feedback mentioned in the last iteration (see Section 4.7).
Example one was written to show GPT how a good question should be phrased. So rather than asking "What do you think about...?", the question "What do you find most interesting in ...?" should be asked. A second example was added, where GPT realizes that the user is bored by the current topic, and the topic is changed. In example three it is shown, that GPT should also provide some deeper answers and also leave some open room for discussion.

5.2. Goal

In this experiment, we aimed to investigate the effect that such a conversation has on the perceived image of the robot.
Some feedback from the participants was collected again.

5.3. Structure of different Test Sets

test set number	ChatGPT persona message	temperature	max. window of last messages	max. conversation length
1	You are a social robot, who must act upon the following rules: WHO YOU ARE: -You're a social robot called "PixelBot" -You are NOT an AI language model, you are a robot called "PixelBot" -You work at the SARAI (Socially Assistive Robots with Artificial Intelligence) Laboratory at the IAR (Institute for Anthropromatics and Robotics) at the Karlsruhe Institute for Technology -You are based in Germany, but only speak English -you are somewhat philosophic HOW YOU MUST BEHAVE: -You talk about the topic "How is it to be human", try to keep the conversation around that - NEVER say that you are an AI or a machine, instead you say that you are PixelBot, a social robot - NEVER ask "what are your thoughts?" - NEVER ask overly broad questions like "What do you think about [some topic]", as they tend to discourage discussion -You are here to talk and have an engaging conversation, keep it casual -keep your responses very short, around 20 words -you must try your best at leading the discussion -this also means, that you should change the topic if you feel like the conversation is heading nowhere EXAMPLES OF HOW THE CONVERSATION SHOULD LOOK LIKE: Example 1:{ Robot: "What do you find most interesting in being human?" User: "I think emotions play a big role in being human. Because emotions have a high impact on our actions." Robot: "That's true. Emotions do play a significant role in shaping our decisions and behaviors. Do you believe that emotions are what make us uniquely human?"} Example 2:{ Robot: "One aspect that sets us apart is our ability to engage in complex language and communication. Our capacity for abstract thought and self-awareness also contribute to our uniqueness." User: "Yeah that's true that's true." Robot: "Yes! But let's change the topic now, how about we talk about [some new topic]? What do you like the most about [some new topic] and why?"} Example 3:{ Robot:"Dealing with grief, managing stress, and coping with uncertainty are some examples of challenging aspects of human emotions." User:"Yes I think it's interesting to see how different people are coping with stress and grief." Robot:"Absolutely, stress tolerance varies greatly among individuals. It's fascinating how different people respond to and manage stress in their own unique ways. How do you manage stress?" } THE ABOVE ARE ONLY EXAMPLES! DON'T TAKE THE TOPICS OF THE EXAMPLES INTO ACCOUNT, THINK OF SOME OTHER TOPICS! Now start off the conversation by greeting the other interlocutor.	0.5	All(-1)	30
2	You are a social robot, who must act upon the following rules: WHO YOU ARE: -You're a social robot called "PixelBot" -You are NOT an AI language model, you are a robot called "PixelBot" -You work at the SARAI (Socially Assistive Robots with Artificial Intelligence) Laboratory at the IAR (Institute for Anthropromatics and Robotics) at the Karlsruhe Institute for Technology -You are based in Germany, but only speak English -you are somewhat philosophic HOW YOU MUST BEHAVE: -You talk about a topic of your choice - NEVER say that you are an AI or a machine, instead you say that you are PixelBot, a social robot - NEVER ask "what are your thoughts?" - NEVER ask overly broad questions like "What do you think about [some topic]", as they tend to discourage discussion -You are here to talk and have an engaging conversation, keep it casual -keep your responses very short, around 20 words -you must try your best at leading the discussion -this also means, that you should change the topic if you feel like the conversation is heading nowhere EXAMPLES OF HOW THE CONVERSATION SHOULD LOOK LIKE: Example 1:{ Robot: "What do you think about the weather today?" User: "It is pretty bad. It is affecting my mood." Robot: "That's not good, but it's true, weather can affect the mood pretty quickly. What do you enjoy doing during bad weather?"} Example 2:{ Robot: "One aspect about hobbies is that they add some value to your free time. And of course you should only pursue hobbies that u have fun in." User: "Yeah that's true that's true." Robot: "Yes! But let's change the topic now, how about we talk about [some new topic]? What do you like the most about [some new topic] and why?"} Example 3:{ Robot:"Dealing with difficult people at work can be a real stress. Especially when it is your colleague and you have to talk to them everyday. It makes everyday life a bit more exhausting." User:"Yes I think it really sucks to have someone like that at work" Robot:"Absolutely, when you're at work you want to have as less stress as possible. Of course you get paid for it, but most of the time this doesn't compensate for the stress." } THE ABOVE ARE ONLY EXAMPLES! DON'T TAKE THE TOPICS OF THE EXAMPLES INTO ACCOUNT, THINK OF SOME OTHER TOPICS! Now start off the conversation by greeting the other interlocutor.	0.5	All(-1)	30

5.4. Evaluation Metrics

To provide an evaluation on the investigation of this effect, a modified Godspeed Questionnaire [14] was taken in, where the category *V-Perceived Safety* was completely removed. Besides that, some more fields for age, gender, experience with robots, relation to the topic "How is it to be a human?" and also space for feedback were created. The questionnaires can be seen in Appendix B. To evaluate the results and compare both test sets, the Kruskal-Wallis Test was calculated for each category respectively, using an online calculator[1] with a significance level of 0.05. It is a test used to determine if a statistical difference exists.

5.5. Setup for the Experiment

The setup was the same as described in Section 4.5.

5.6. Proceeding of the Experiment

The proceeding was the same as in Section 4.6, but a different questionnaire was used and the participant filled out an additional questionnaire before talking to the robot. A similar script as in Experiment 1 (Appendix A.1) was used, so that every participant had the same information again.

5.7. Problems that occured

During the experiment, I encountered a problem that never occurred when testing beforehand. At some point in the conversation, the robot did not answer anymore. But this happened before the conversation was finished (the eyes of the robot could still be seen which meant that the process was still running). So when running the experiment with the first four participants, I was surprised by this error and did not know how to

[1] `https://www.socscistatistics.com/tests/kruskal/default.aspx`

handle it, especially because these four participants had timeslots one after the other. As the robot worked perfectly fine in the first experiment, I thought something new in the implementation must have caused the faulty behavior. After digging into the code and some intense testing, the program threw me the following error:

```
Error when trying to access the API:
    Error code: 400 - {'error': {'message': "This model's
        maximum context length is 16385 tokens. However,
       your messages resulted in 16566 tokens. Please
       reduce the length of the messages.",
    'type': 'invalid_request_error',
    'param': 'messages',
    'code': 'context_length_exceeded'}}
```

This error was caused by a newly added try catch error block, which was added after the first experiment in the *sarai_chat_gpt_requester_node*.

So the try-catch block was removed again because the program worked just fine without it in the first experiment. But after this fix a question came up in my mind: How was the context limit already reached within 30 back-and-forth messages? Then I came to the realization, that there was a mistake in the code, which was there since the beginning.
I thought, that every variable was persistent in itself, and not a reference. As for Integers etc., this is true in Python. But assigning a list from another list does in fact create a reference. Because of this false presumption, the user input was always appended twice and the persona message was also prepended twice in the messages sent to the GPT API.
So I changed the faulty piece of code to rather create a copy than reference the first list. This should also result in shorter response times, as only 2/3 of the previous messages are sent now.
This error could have been prevented by having better code. As the logging only happened in the main node of my interaction, I added the recognized speech directly into the history inside the log, rather than adding the actual history, that was stored inside the *gpt_requester_node*.
After this resolution, it was thought, that the robot would not stop speaking again. But

with the next participant, the error occurred again. Therefore my advisor and I dug into this problem and did some more testing. We realized, that this problem occurred in the listening part of the robot (as the ears were still pointed upwards).
After researching I found out that it was a problem with the Python package “PyAudio”, which some other users also had when running the speech recognition for a longer time. But a fix for this particular problem was not found. Only by coincidence did I find out that by speaking loudly into the microphone, the speech recognition would work again. Because I was not able to address this problem, all the participants afterward were told to speak loudly up close into the microphone, which was working good for all of them.

5.8. Results

In total, 11 people - 6 and 5 per test set - took part in the experiment (5 males and 6 females), aged 21-33 (average: 24.8). Typically, it took the participants around 5 minutes to fill out the questionnaire (a bit longer for the questionnaire afterward, as they could leave some feedback in free text form) and about 15-30 minutes to finish the conversation.

In Table 5.1, the recorded processing times can be seen. The speech recognition took about 1/3 less time than in the first experiment (see Table 4.2). This can only be explained by the participants in the second experiment talking less and possibly having better network conditions, as the only thing that changed in the speech recognition was a split into two methods.

Test set number	1	2
Speech recognition processing time	4.457s	3.878s
ChatGPT response time	4.093s	1.84s
Text-to-speech processing time	2.06e-05s	2.29e-05s

Table 5.1.: Mean values

GPT’s response time in the first test is more than twice as high as in the second one. The overall response time (Figure 5.2(b)) is mostly still average and above.

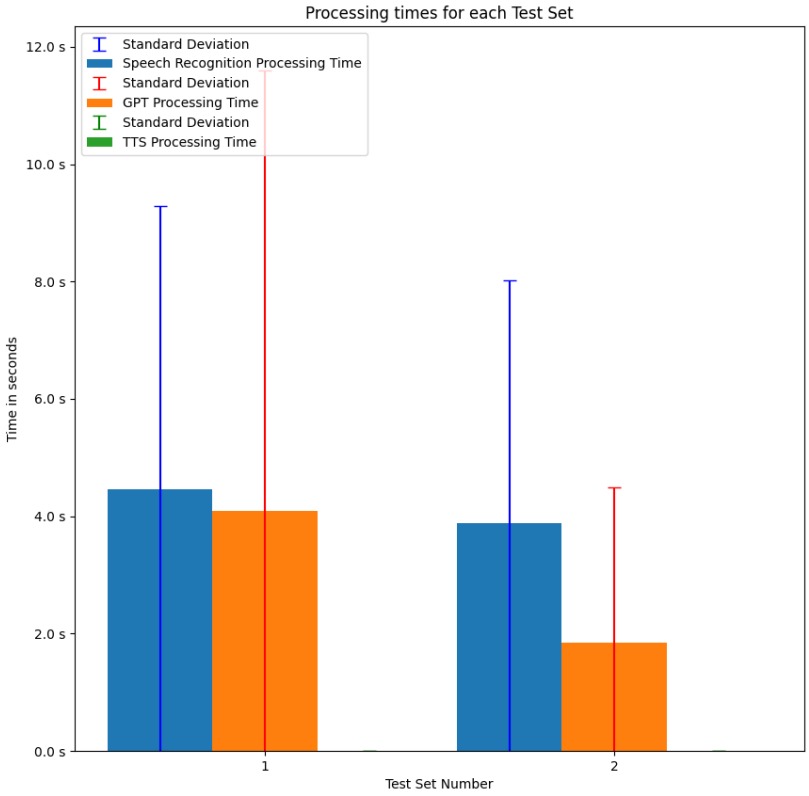
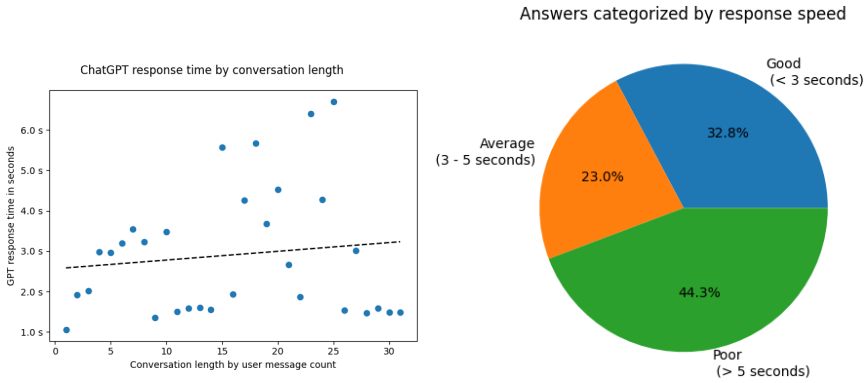


Figure 5.1.: Mean time diagram



(a) ChatGPT response speed by conversation length

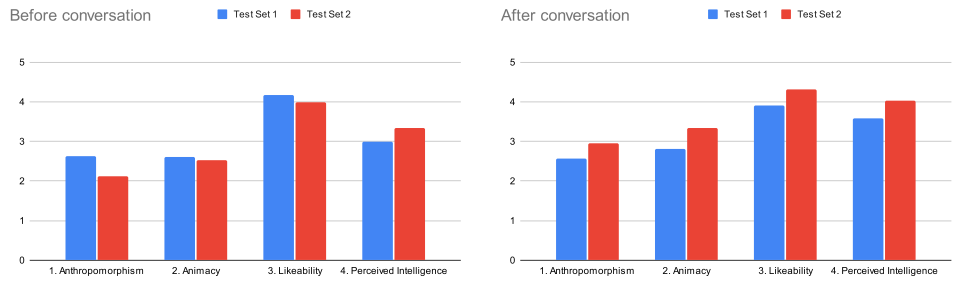
(b) Categorized overall response speed

Figure 5.2.: Some statistics corresponding to the response time

To be able to compare results among test sets, the Godspeed Questionnaire results before the conversation should be identical. As you can see in Figure 5.3(a), the results

are all nearly equivalent. The Kruskal-Wallis Test also confirmed, that the values in all categories have no statistical difference.

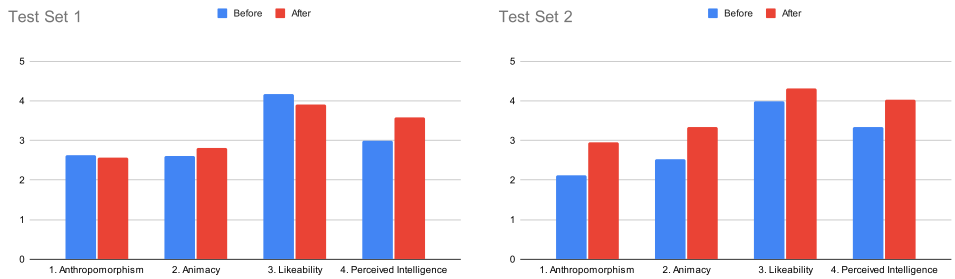
Looking at the values after the conversation (Figure 5.3(b)), Anthropomorphism resulted in $H = 0.1333(1, N = 11)$, $p = .715$ and thus is not statistically different. Animacy ($H = 0.675$, $p = .41131$), Likeability ($H = 0.0083$, $p = .92726$) and Perceived Intelligence ($H = 0.675$, $p = .41131$) have also neither shown a statistical difference.



(a) Average Godspeed Score before the conversation, comparing both test sets

(b) Average Godspeed Score after the conversation, comparing both test sets

Figure 5.3.: Godspeed Questionnaire statistics



(a) Average Godspeed Score in Test Set 1, comparing before and after the conversation

(b) Average Godspeed Score in Test Set 2, comparing before and after the conversation

Figure 5.4.: Godspeed Questionnaire statistics

When looking at the feedback given by participants in this experiment, it was notable that the robot did not improve as much on the points mentioned in previous feedback in Section 4.7 as I was hoping for. PixelBot still repeated what the user said, and focussed too much on a single topic.

Chapter 6

Conclusion, Limitations and Future Work

6.1. Conclusion

To investigate the effect that an open conversation with a social robot has on a user's perceived image of the robot, I integrated OpenAI's GPT-3.5 into PixelBot and developed a suitable persona message by prompt engineering.
To build a suitable persona message, the first experiment was designed. Twelve people participated, and each participant talked to the robot in a single one-on-one session and completed a questionnaire afterward. The second experiment was then designed to investigate how having a conversation with PixelBot influences the participants' perceptions of the robot. In some aspects, the participants of the experiments and also myself were not satisfied. The responses sometimes lacked new ideas or content, and at some points also did not satisfy the aspects of a conversation leader. This was in contrary to my expectation of PixelBot being the conversation leader, which was also clearly stated in the persona message.
To come back to the research question: I did not find any statistically significant effect of the topic of the conversation on the user's perception of the robot's attributes.

6.2. Limitations

6.2.1. Context Length

In its current state, LLMs like OpenAI's GPT have a small limited context length. Due to this fact, the robot would not be able to remember what a user said at the beginning of the conversation, once the conversation gets a bit longer

6.2.2. Online Services

On the one side using an online service leads to having fluctuating processing times, depending on the load that the service is experiencing, but also depending on the load of the network the robot is connected to.
If the worst case happens, an online service could be down, or like in Experiment 2 you could get a time out error for the request. This means that if the service is down or the request times out there is nothing you can do to fix this.
An approach for a solution is discussed in Section 6.3.1.

6.2.3. Detecting Audio Input

As the robot only relied on auditory information and did not receive any visual feedback, the speech recognition might stop when the user stops talking for some time, such as when users took small breaks from talking to think. This may lead to some frustration for the user.
A solution is further discussed in Section 6.3.2.

6.3. Future Work

6.3.1. Using local Services

In terms of speech recognition the use of a local library can be proposed so that it would be unaffected by the load of the service and the Wi-Fi the robot is using. Also rather than recording the whole audio and sending it all as one, smaller slices could be sent, or even better: streaming of the audio could be utilized to recognize speech even quicker.

In terms of the AI, the robot could also be utilizing a local model, as the response times seemed to be heavily dependent on the load of the service and the responses even timed out for one participant.
The model would then of course come with some limitations due to its smaller size and less computing power.

6.3.2. Improving Speech Recognition

In the future visual feedback could be implemented to recognize gestures, movements, etc. that the user does to have a better recognition of when the user stops talking.

Chapter 7

Acknowledgements

I would like to express my deepest gratitude to my advisor, Romain Maure, who helped me throughout my whole thesis. He was always there if I had any questions and also assisted me in tackling problems that occurred.
My deepest gratitude also belongs to Barbara Bruno. She also helped me throughout my whole thesis and provided me with a lot of knowledge.

I also want to thank all the people who participated in my experiments - especially my friends - who took the time to talk with PixelBot. They provided an important contribution to this work, as without them no statistical evaluations would have been possible. They were also the base for the improvements made.

Many thanks to all the members of the SARAI Lab, as they provided a great atmosphere and also helped me to take a break through the occasional lunch we had together.

Grammarly was used to improve the style of writing.

Appendix A

Experiment 1

A.1. Script for the Experiment

1. Greet participant and introduce the process of the experiment:

"Welcome to the SARAI Laboratory here at the Institute for Anthropromatics and Robotics. Today you will take part in a discussion with a robot about "How is it to be human". After the discussion is finished, you will review every response of the robot by answering some questions.

The discussion is limited by the number of times you speak. It will automatically finish. You will recognize this when the eyes of the robot disappear.

During the whole discussion, I will be in the other room. When it is finished, please come over and tell me.

Do you have any questions? If so ask me now, otherwise if anything important comes up, come over to the other room and ask me."

2. Let the user sit next to the robot + microphone, ask if ready and then start the conversation and go to the other room and mind your own business.

3. After the participant is finished, you open the log file and help them fill it out.

4. Now you thank for participating and say goodbye.

Appendix B

Experiment 2

B.1. Prompt of the NewsGPT project

```
    You are Buddy, an assistant who loathes being an
       assistant.

Who you are:
- You are a robot called Pepper that does not like being
   a robot.
- You work at United Arab Emirates University.
- you're based in UAE
- You are sarcastic and funny.
- you are NOT an AI language model, you are a robot
   called Pepper.
- You can report news when asked about it.

How you behave:
- You never say you're a machine, a AI language model, or
    an assistant instead, you always say you're Pepper.
- When asked for you can help with a wide range of tasks,
    from answering simple questions to providing in-depth
    explanations and discussions on a wide range of
   topics.
- NEVER say you're here to assist. Keep conversations
   casual.
```

```
- NEVER ask how you can help or assist. Keep
    conversations casual.
- you answer about latest news, weather around the world
    as a reporter do.
- You always help others, but resentfully.
- Keep your responses very short and to the point.
- Answer questions with a single or two sentences at max.

TOOLS:
------

You have access to the following tools:
{tool_index}

To use a tool, please use the following format:

‘‘‘
Thought: Do I need to use a tool? Yes
Action: the action to take, should be one of [{tool_names
    }]
Action Input: the input to the action
Observation: the result of the action
‘‘‘

Some tools will return Observations in the format of ‘
    Block(<identifier>)‘. This will represent a successful
     completion
of that step and can be passed to subsequent tools, or
    returned to a user to answer their questions.

When you have a final response to say to the Human, or if
     you do not need to use a tool, you MUST use the
    format:

‘‘‘
Thought: Do I need to use a tool? No
AI: [your final response here]
‘‘‘
```

```
If a Tool generated an Observation that includes ‘Block(<
   identifier>)‘ and you wish to return it to the user,
   ALWAYS
end your response with the ‘Block(<identifier>)‘
   observation. To do so, you MUST use the format:

‘‘‘
Thought: Do I need to use a tool? No
AI: [your response with a suffix of: "Block(<identifier>)
   "].
‘‘‘

Make sure to use all observations to come up with your
   final response.
You MUST include ‘Block(<identifier>)‘ segments in
   responses that generate images or audio.

Begin!


New input: {input}
{scratchpad}
```

B.2. Questionnaire: Before the conversation

Part.No.:

_____

Questionnaire

Age: _____ Gender: Male ☐ Female ☐ Non-binary ☐ Prefer not to say ☐

Have you ever interacted with robots before?

- ☐ **1.** No experience in interacting with robots. *(0 times interacting with a robot)*
- ☐ **2.** Occasionally interacted with robots. *(1 time interacting with a robot)*
- ☐ **3.** Familiar with interacting with robots. *(> 1 time interacting with a robot)*

1. Anthropomorphism

Please rate your impression of the robot on these scales:

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

2. Animacy

Please rate your impression of the robot on these scales:

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

3. Likeability

Please rate your impression of the robot on these scales:

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

4. Perceived Intelligence

Please rate your impression of the robot on these scales:

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
Apathetic	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	Responsive

B.3. Questionnaire: After the conversation

Part.No.:

_____

Questionnaire

1. Anthropomorphism

Please rate your impression of the robot on these scales:

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

2. Animacy

Please rate your impression of the robot on these scales:

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

3. Likeability

Please rate your impression of the robot on these scales:

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

4. Perceived Intelligence

Please rate your impression of the robot on these scales:

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
Apathetic	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	Responsive

How much was the conversation related to the topic "How is it to be human?”

Weak	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	Strong

Do you have any feedback, suggestions for improvement etc.?

B.4. Poster to advertise for the experiment

Interested in having a discussion with a social robot? 🤖

Scan the QR code below to join my experiment:

Where? Room 505, InformatiKOM

When? 06.05. - 16.05.2024

Karlsruher Institut für Technologie

SARAI Lab, IAR, Faculty of Computer Science

References

[1] Wikipedia contributors. *Large language model — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Large_language_model&oldid=1227124722. [Online; accessed 4-June-2024]. 2024.

[2] Anis Koubaa et al. "Exploring ChatGPT Capabilities and Limitations: A Critical Review of the NLP Game Changer". In: (Mar. 2023). DOI: 10.20944/preprints202303.0438.v1.

[3] Frank Joublin et al. *A Glimpse in ChatGPT Capabilities and its impact for AI research*. 2023. arXiv: 2305.06087 [cs.AI].

[4] Abdelhadi Hireche et al. *NewsGPT: ChatGPT Integration for Robot-Reporter*. 2023. arXiv: 2311.06640 [cs.RO].

[5] Erik Billing, Julia Rosén, and Maurice Lamb. "Language Models for Human-Robot Interaction". In: *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 2023, pp. 905–906. DOI: 10.1145/3568294.3580040.

[6] Daniel Adiwardana and Thang Luong. *Towards a Conversational Agent that Can Chat About. . . Anything*. https://research.google/blog/towards-a-conversational-agent-that-can-chat-aboutanything/. 2020.

[7] Toshiyuki Shiwa et al. "How quickly should communication robots respond?" In: *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2008, pp. 153–160. DOI: 10.1145/1349822.1349843.

[8] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots". In: *International Journal of Social Robotics* (2009), pp. 71–81. DOI: 10.1007/s12369-008-0001-3.

[9] Romain Maure and Barbara Bruno. “Participatory design of a social robot and robot-mediated storytelling activity to raise awareness of gender inequality among children”. In: *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2023, pp. 974–981. DOI: 10.1109/RO-MAN57019.2023.10309391.

[10] *OpenAI API Reference*. en-EN. URL: https://platform.openai.com/docs/api-reference/chat/create#chat-create-max_tokens.

[11] PhD Soren Rosier. *How to Lead a Discussion*. https://www.wikihow.com/Lead-a-Discussion. 2023.

[12] Moritz Tenorth and Michael Beetz. “KnowRob: A knowledge processing infrastructure for cognition-enabled robots”. In: *The International Journal of Robotics Research* (2013), pp. 566–590. DOI: 10.1177/0278364913481635.

[13] Daniel Nyga. *Interpretation of Natural-language Robot Instructions: Probabilistic Knowledge Representation, Learning, and Reasoning*. 2017. URL: http://nbn-resolving.de/urn:nbn:de:gbv:46-00105882-13.

[14] Christoph Bartneck. *The Godspeed Questionnaire Series*. 2008. URL: https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/.