UNIVERSITY
OF OULU

Exploring the interaction between humans and an AI-driven chatbot

University of Oulu
Information Processing Science
Master's Thesis
Gloria Stanciu
2023

Abstract

Chatbots have become an omnipresent software that many services is using nowadays to provide easy and continuous support for users. Regardless of the domain in question, people are using chatbots to get quick access to information in a humanlike manner. Still, chatbots are limited in terms of interactivity, providing facts, or solving elemental problems. Moreover, the lack of empathy that chatbots have is a drawback that limits them from providing the best outcome possible for the user.

With that in mind, this thesis aims to find out how an emotionally aware chatbot would influence the interaction and engagement level of participants, starting from the hypothesis that "The awareness that the chatbot shows during the conversation impacts the engagement of participants". The research method used was an experimental study approach because it helps with finding how the cause of the awareness of chatbots can affect the engagement level of participants. For that, a web application was developed that consisted of a chatbot driven by OpenAI. Before the participants started to interact with the chatbot, they were provided with information and instructions on how to adjust their cameras so their facial expressions could be analyzed properly in order to get the intended experience. A total number of 180 participants were recruited using the Prolific crowd-sourcing platform, from which 178 responses were used in analyzing the results.

The participants were split into three study conditions, namely BASELINE, EMOJIONLY, and EMOJI-AND-CHAT which differed in the emotional awareness levels that the chatbot had. BASELINE study group interacted with a simple chatbot that was not aware of participants' emotions at all. The EMOJI-ONLY study group discussed with a chatbot that during the interaction showed participants their emotions in real time with the use of emoji pictograms. In the last study group, EMOJI-ANDCHAT, besides showing the participants' expressions through emojis, the chatbot also replied to the mood changes of the participants with messages that clearly stated that the chatbot noticed their facial expression changes. Each participant, regardless of the study group, had a conversation with the chatbot that lasted for a few minutes and started with the topic of their own chronic pain experiences. The chronic pain topic was used in order to trigger differences in facial expressions naturally. During a conversation of only a few minutes, the topic discussed needs to be of interest to the participant so that differences in facial expressions could occur. With that in mind, participants were recruited using Prolific's option of selecting participants that deal with chronic pain. During the conversations participants' facial expressions were analyzed and collected. Moreover, at the end of the interaction, the participants answered a questionnaire composed of a mix of 23 quantitative and 3 qualitative questions.

The data collected showed that the emotional awareness that a chatbot is showing during a discussion impacts the level of engagement of participants. However, the results were not able to particularly point out if participants' level of engagement is affected positively, and thus feeling more engaged, or is affected negatively, feeling less engaged than when interacting with a non-emotional aware chatbot.

Participants showed both significant interests in the emotionally aware chatbots, as well as concerns, and identified possible issues and limitations.

The chatbot used throughout this research was effective and succeeded to show the potential of such applications. Nevertheless, improving the way the chatbot reacts to changes in facial expressions needs further testing and development, as well as improving its privacy and security side so people would trust it more.

Keywords
chatbot, emotion recognition, awareness, engagement

Supervisor
DSc (Eng), Professor Mika Mäntylä

Foreword

I would like to offer special thanks to Dr. Tech. Associate Professor Simo Hosio for being my thesis supervisor, guiding me throughout the process of conducting this research, and especially for his patience and interest in my thesis. Additionally, I would also like to thank his entire Crowd Computing team for providing their feedback and expertise that helped me to improve my study.

Finally, I want to also thank D.Sc, Professor Mika Mäntylä for accepting to be my official supervisor and everyone who offered their feedback and actively helped in perfecting this thesis.

Gloria Stanciu

Oulu, 15.06.2023

Contents

List of Tables

List of Figures

List of abbreviations

FAQ Frequently Asked Questions

AI Artificial Intelligence

API Application Programming Interface

JSON JavaScript Object Notation

UES User Engagement Scale

HCI Human-Computer Interaction

VA Virtual Agent

1. Introduction

In the last decades, technology has without question become an omnipresent factor in many people’s lives. Whatever they do, they are surrounded by technology and next-generation machinery that is meant to transform their lives, making them easier and better. One powerful example of such technology, which is highly used nowadays in many different domains, is chatbots.

In simple words, chatbots are artificial intelligence-driven systems, designed for engaging in conversations on multiple topics with users, providing instant responses over text or audio. (Smutny and Schreiberova, 2020)

The interaction between humans and chatbots has been ongoing for many years and continuously improving and expanding. These interactions are integrated into various fields and for multiple purposes. Educational environments, customer services, healthcare, robotics, or industrial use cases are only some of the domains in which chatbots are used regularly. (Adamopoulou and Moussiades, 2020)

A study conducted by Brandtzaeg and Følstad (2017) aimed to find out also the reasons why users engage in conversations with chatbots. Among the reasons listed by participants, the one that showed the most frequency was productivity, followed by entertainment and social relationship. A small number of participants from the study also explained that the reason for using chatbots is that important matters can be shared or discussed more easily with a chatbot than with a person. Regardless of the number of participants that expressed this opinion, the claim cannot be ignored as it may suggest the importance of chatbots being designed as being emotionally aware.

Communicating with people, supporting, and guiding them are the main purposes of any chatbot. Yet, their current limitations prevent them from offering the best outcome possible. Currently, chatbots are limited to providing facts and general information and solving basic problems and tasks. A current study conducted by Zhou et al. (2023) declares that the lack of empathy and emotional interaction is a critical drawback in the overall human-chatbot interaction and the users’ anticipated communication quality. Additionally, another study by Adamopoulou and Moussiades (2020) claims that one of the biggest disadvantages of chatbots is their incapability of acknowledging what the user is aiming for, and what is its purpose or feeling on the discussed topic.

This limitation can obstruct the effectiveness of chatbots in multiple domains. Chatbots used for providing customer support, marketing, and sales are more cost-efficient than human employees but considerably less effective, as they do not have the ability to comprehend users’ possible lack of satisfaction. (Ashfaq et al., 2020)

When people talk to each other they are able to figure out what emotion the other person is expressing by interpreting their answers, or even their facial expression, if the communication is face-to-face or through a video call. In order to enhance the interaction between humans and chatbots, their abilities could be improved by making them more emotionally aware. Some studies reflect upon the emotionally aware nature of a chatbot. However, designing emotionally aware chatbots is still in the early phase. There are many unknowns that still need answers and various aspects that need to be understood. (Ghandeharioun et al., 2019)

The purpose of this study is to investigate how an emotionally aware chatbot would influence the interaction and engagement level of the users. We aim to determine how different engagement levels of a chatbot are affecting the interaction with users and their answers as well.

Thus, the research question that this study addresses and aims to find an answer to is the following:

- RQ: How do different levels of emotional awareness in a chatbot impact users' engagement during a conversation?

To address this question, there was a need for an application that could help proceed with the experiment. Thus, a web application was designed to support the study by providing a chat where people engaged in a conversation with a chatbot enabled by Artificial Intelligence (AI)s and APIs by OpenAI[1]. The application collects the facial expressions of participants during the entire conversation and uses them to make the chatbot aware of the people's reactions.

Based on the previous research findings, the purpose of this study, and the research question to be answered, the research that is going to be carried out focuses on the following hypothesis:

- H1: The awareness that the chatbot shows during the conversation impacts the engagement of participants.

Additionally, two objectives for the thesis process are set. These represent the results that are aimed for, during the research study and which contribute to the final results that this study aims to obtain.

- O1: The development of a software artifact that would facilitate the scientific experiment.
- O2: Carry out the experiment in order to gather data from which results can be extracted.

The thesis is structured into six different chapters. The current chapter introduces the topic that is going to be studied and its purpose of it. Further, the second chapter will present relevant prior research in this field and will give a more in depth understanding. The third chapter will present the research methods used for this study, followed by the results collected which will be analyzed in chapter four. Chapter five will contain discussions related to the findings and their relevance. Finally, the last chapter will give the final conclusion of the study.

[1] OpenAI website: https://openai.com/

2. Literature review

This chapter is divided into several sections that present prior investigations and discoveries that are relevant to the problem discussed in this study. The first section covers ideas from the human interaction domain, revealing the insides of human interaction and smiling behavior. The next section takes into consideration the domains of use effectiveness and limitations of standard chatbots and AI-driven chatbots up to this time. Finally, the last section provides information about the current interaction between humans and chatbots and what are the next possibilities to improve it further.

2.1 Human interaction and facial expressions

In order to understand the interaction between humans and computers and human behavior in this particular case, it is necessary to take a step back and have a look at human-to-human interaction in its natural form. In the human-computer interaction field, the focus is on how humans are interacting with technology, not the other way around. Human interaction is the comparison term that people use when engaging in any other type of interaction, even when interacting with a computer. The entire world is built up and relies on interactions between people, therefore, it is rational to use the personal experience that one accumulated through connecting to people, to try and engage in other types of communication such as computer interaction. For this reason, it is first needed to understand what people expect from interaction and how they engage in one.

2.1.1 Smiling behaviour

William James, who is called the "Father of the American Psychology" stated in one of his books (James, 1995) that "We don't laugh because we're happy, we're happy because we laugh". The significance of this strong remark points out that people feel a certain way due to the choices and actions they make, not the other way around. If people would engage in positive activities which bring joy, they will feel greater levels of satisfaction and happiness, whereas negative thinking will only make someone feel worse. In that sense, there is evidence that implies that laughing and smiling repeatedly works as a medicine. Norman Cousins found this by himself and explained it in his book called "The Anatomy of Illness" (Cousins, 1979). Norman discovered that only ten minutes of pure laughter worked as an anesthetic for his disease and helped him have pain-free sleep for a couple of hours.

Moreover, it was discovered that the facial expressions that someone makes, voluntarily or involuntarily, influence the feelings, mood, and emotional experience they end up having. For example, studies showed that people who engage in events that cause them to smile frequently end up having a positive mood, whereas if they engage in negative events – which can cause them to frown, for example –, they will end up having a negative mood. Moreover, the study also showed that the connection between facial expressions and emotions had the same results even when people are only imitating the facial expressions they saw in others. Seeing their own facial expressions in the mirror resulted as well in the same conclusion. (Tomkins, 1962; Tomkins, 1963; Kleinke et al., 1998)

2.1.2 Emotional intelligence

Salovey and Mayer (1990) define emotional intelligence as the subdivision of social intelligence, that is responsible for one's capability to observe and distinguish their own or others' feelings and emotions and to be able to use the information acknowledged in order to help deal with them accordingly.

Emotional intelligence plays an important role in what defines us as humans and affects our entire life. Persons that have emotional intelligence are able to control their emotions better and understand that negative emotions are sometimes necessary for one's way to achieving a goal, but though necessary, they also need to be accepted and understood. Their presence is thought to make those around them feel comfortable mostly because the emotional awareness of one can affect the personal well-being and also of those around. Moreover, lacking the ability to cope with personal feelings may lead to emotional problems such as remaining trapped in an emotional state which eventually will result in being excluded from social interactions. Furthermore, those who cannot identify their own emotions encounter difficulties in making plans that are emotionally satisfying for them (Swinkels and Giuliano, 1995; Salovey and Mayer, 1990).

2.1.3 Non-verbal communication

Non-verbal communication refers to all the gestures that a person makes in order to transmit a message or to express an emotion, for example, body language, tone of voice, or facial expressions. In face-to-face communication, non-verbal communication has a significant role. This type of communication is essential in order to transmit emotions back to the interlocutor. Moreover, non-verbal communication provides additional information, creates a level of interaction, and builds intimacy between the parties engaged in the conversation by transmitting empathy (Beattie et al., 2020; Liu et al., 2018; Boutet et al., 2021).

However, when it comes to computer-mediated communication, where the conversation takes place in a virtual environment, the non-verbal communication processes need to be transposed in order not to lose the ability to transmit emotions. One of the most used and effective methods to bring non-verbal communication into virtual environments is emoticons (or emojis). It was discovered that emoticons have a similar effect virtually as non-verbal communication does during face-to-face interactions, meaning that people who used more emoticons that express happiness on their social platforms were taught to have more positive qualities, be more likable and friendly (Wall et al., 2016). Emojis are not only nice-looking pictograms that can be added to a text, but they have the power to give the message more context, understanding, and emotion in order to enhance the conversation and its quality (Boutet et al., 2021).

Liu et al. (2018) however saw a difference between the use of emojis to transmit emotional expressions and face-to-face emotional expressions, namely the fact that in the virtual environment, people need to manually insert the emojis they believe represent best their reactions, moods, or feelings. To overcome this limitation, they conducted a study where emojis would automatically be attached to a message that was received based on the recorded facial expression of the receiver, making it faster, more reliable, and more accurate. The results showed that the study had positive results and the participant had used fewer self-inserted emojis and relied on the system to do it, meaning that the system successfully recognized the expressions almost all the time.

In human-to-computer interaction, such a system is not needed, however, the study offers results that prove that emojis can be integrated into other services successfully as well, during conversations with a chatbot, where it could help with emotional awareness in both the user and chatbot.

2.2 Chatbots

This section provides information on the current domains in which chatbots are used regularly, what are their capabilities, but also what limitations still exist.

Chatbots (also known as virtual agents) are systems that engage in communication with humans with the purpose of offering services and helping deal with problems of a social, intellectual, psychological, or even health nature. These systems can be divided into two categories based on how they are developed and what type of experience is offered to the user. Chatbots such as MobileMonkey, BotStar, ManyChat, ChatFuel, etc. are chatbot builder platforms that are used to create chatbots with pre-defined conversations. These chatbots have a clear purpose and are mostly used in domains such as sales and customer service. Chatbots that have a predefined conversation, can prevent users from giving open end responses to questions and in exchange, provide multiple-choice answers to choose from. They can also provide features like natural language processing in order to allow users to send open-end replies which will be processed by the chatbot in order to extract key data and understand users' needs. The other types of chatbots such as Amazon Alexa, Siri, Google Assistant, or the most recent one ChatGPT, are systems that heavily rely on artificial intelligence and machine learning technologies, and which allow a user to interact more naturally with them.

Chatbots can also be differentiated through the communication channel support they offer, which can be through text messages or simulated voice. Text-based communication can also be split into two communication means open text messages or navigation buttons (Mohamad Suhaili et al., 2021). For this study, the focus is going to predominantly be on text-based human-chatbot interaction.

Chatbots have considerably grown in popularity in the last few years. They are used in many domains, for multiple purposes, and in various ways. Further, some of their implications in different fields will be presented.

2.2.1 Education

Chatbots used in the educational environment can be divided into two main categories: service-oriented and teaching-oriented. Service-oriented chatbots may help students with their managerial tasks such as enrollment processes, registrations, or grades administration, guide first-year students accommodate with university life, or they could be used as general Frequently Asked Questions (FAQ) chatbot that is ready to answer any question a visitor might have about the university, academic requirements or any academical matters. On the other hand, chatbots are heavily used for teaching purposes too. The most used methods are chatbots that engage and promote learning methods through communication with a teacher (in this case, virtual), revisiting lectures, offering support for problem-solving, or having a nonformal educational learning system through gamified experiences. Medical students benefit from systems that simulate patients to help them put into practice what they learned (Adamopoulou and Moussiades, 2020; Pérez et al., 2020).

Limitations that chatbots used in this domain encounter are mostly related to the quality of interaction between the chatbot and the students. Too long monologues from chatbot decrease users' interaction, curiosity, and determination to learn. Their incapability of empathy and making a difference can lead to a lack of inclusiveness and responses that can become an issue for minorities (Pérez et al., 2020).

2.2.2 Health

In the healthcare system, chatbots show real potential in many subdomains such as emotional health, mental health, wellness, and healthy lifestyle or health problems that require behavior change such as weight control or smoking addiction. These systems are mostly used by patients for finding out diagnoses or treatments for diseases or only for providing information on their problems. On the other side, nurses and doctors find useful chatbots that help with reminding patients about treatments or managing appointments. The most benefic aspect of chatbots in the healthcare system is making them more accessible. However, the main problem remains trust. Doctors have insecurities about the capabilities of a chatbot to diagnose a disease or offer a treatment plan, as their information is limited (Pérez et al., 2020).

2.2.3 Customer service

Many businesses use chatbots for their customer services because they provide a lot of benefits. Chatbots can run 24/7, being available anytime for anyone, regardless of language spoken, and reduces the waiting time to zero for the customers, compared to human customer services where the queues can be even several hours long. Additionally, they are a very cost-effective solution for the company. (Pérez et al., 2020)

Despite all the positive outcomes that a business can benefit from when replacing people with chatbots, there are also possible downsides that they may bring along. Their lack of emotional awareness and empathy is a major downside for this domain as it can leave customers unsatisfied with the service provided and without a solution to the problem they sought help with. Therefore, the limitations that chatbots still have regarding communication may contribute to the business losing customers. (Ashfaq et al., 2020)

Alongside these main domains of implication, chatbots also play a considerable role in domains such as robotics where new natural language models are developed for autonomous robots, or industrial use cases, where they are used in the banking sector as virtual assistants or for feedback purposes in many other industries (Ashfaq et al., 2020). However, regardless of the domain of use, the limitations of chatbots are similar. Empathy is a key element that seeks to be improved in order to bring chatbots to another level.

2.3 Emotionally aware chatbots

This section provides more in-depth information about prior research conducted on emotionally aware chatbots and the nonverbal communication between those and users.

The integration of processes that include emotionally aware responses from chatbots need to be carefully integrated by taking into consideration aspects relating to users' behavior. Extroverts are more prone to respond in a positive way to the emotional intelligence ability of chatbots, whereas introverts are not as comfortable using such chatbots. (Ghandeharioun et al., 2019)

The frequency and sensitivity to which chatbots respond to users' emotions is a design issue that needs precise integration. Responding to every small change in facial expressions or to neutral facial expressions can compromise the integrity and trustworthiness of the chatbot's ability to accordingly react to mood changes. Humans compare the emotional abilities of chatbots to what they are already used to, that being the emotional awareness of other humans. Therefore, if the chatbot is too sensible to facial expressions and prompts, for example, happy messages at every slight smile, the users will not identify their mood with what the chatbot expressed, making it feel exaggerated. Small changes in facial expression should have subtle inputs in the chatbot's replies (Ghandeharioun et al., 2019).

Prior studies have tried to determine how signs of emotions from chatbots would affect the emotions and behavior of users. Such a study was conducted by Soderlund et al. (2021), which aimed to find if linguistic elements sent by a chatbot would have any impact on the user's perception of Virtual Agent (VA) happiness. Furthermore, they also wanted to find out if and how the evaluation of a VA is influenced by perceived happiness. The linguistic elements that were taken into consideration and used in the experiment study were exclamation marks and positive words such as "happy", "sun", "summer", "heaven", and "kiss"). The exclamation marks were used during the study due to the findings made by Hancock et al. (2007) that claim the exclamation marks are positively related to how happy the sender is believed to be. Based on the same findings, the negative and negative words from a message sent by the VA are negatively related to its happiness.

Among human-to-human interactions, we have a tendency to let our replies be influenced by how others communicate with us. Soderlund et al. (2021) showed that this is happening between a human-to-chatbot as well. Results from their study reveal that the replies to the VA's messages have a tendency to follow the emotion that is perceived from those messages.

Another study conducted by Krämer et al. (2013) aimed to discover whether or not the smiling displayed by a chatbot would have any impact on a person's smiling behavior. The study that the participants enrolled in required having an 8-minute small talk with a chatbot, who frequently smiled, occasionally smiled, or did not show any smile during the entire conversation. The hypotheses from which the study started are the following:

> H1: A virtual agent who smiles is evaluated more positively than a virtual agent who does not smile.
>
> H2: A virtual agent who smiles evokes more smiling in a human interaction partner than an agent who does not smile.
>
> H3: Women smile more than men when interacting with a virtual agent.

In the end, the results showed that the first hypothesis, which was taken into consideration, was not supported by any of the results from their study, so it did not stand. When considering the second hypothesis, the results showed that the smiling behavior of the chatbot influences the smiling behavior of the person who is interacting with. Results also showed that the last hypothesis was true and women smiled more when interacting with the chatbot.

Melo et al. (2011) aimed to find out if a negotiation task between a human and a chatbot would have the same outcome effects as a negotiation task between two humans. The focus of the study is on analyzing expressions of anger and happiness that are going to be seen in the chatbot. There were three facial expressions that the chatbot could display: angry, neutral, and happy. The results of the study showed that the most effective chatbot was the one that showed an expression of anger, followed by the one that showed no expression and leaving the happy chatbot as the least effective one. The results of the study aligned with the findings from prior research conducted regarding the effects of expressions in human-to-human negotiation.

This finding confirms that when VA is developed with the ability to express emotions, their interaction with a user has the same effect as the interaction with a human being. In the case of Melo et al. (2011) study, the results are showing that in the circumstances of negotiation, the feeling of anger is correlated with a high amount of confidence and power in own abilities, whereas the feeling of happiness is perceived as having lower limits.

Another interesting study was conducted by Sidner et al. (2006), where people were asked to have a conversation with a conversational robot. During the conversation, the robot is programmed to recognize each head nod made by the user. This study was split into two cases, one in which the robot will only recognize users' head nods, without users being aware of this, and the other one in which the users are aware of the robot's feature and the robot also provides feedback at every head nod made. The results of the study showed that the users who were aware of the robot's feature and who were provided visual feedback nodded more time than the other study group.

Human-robot interaction field differs in many ways from human-to-computer interactions, primarily through the environment in which the interaction occurs. However, with that in mind, the end result of the study shows the potential of being applied and having the same effect on the interactions between humans and chatbots. Chatbots are robots both systems that were developed to be able to interact on some level with people. The results may be worth analyzing in the context of chatbots.

There is limited past research on the domain of emotionally aware chatbots. A large part of the research conducted until now has focused on studying chatbots that during interaction had presented users with a 2D or 3D avatar through which they could visualize the facial expressions of the chatbots. However, in this study, the chatbot's appearance and facial expression capabilities are not of interest. The main focus is to understand how users' engagement in a conversation could be affected based on the ability of a chatbot to recognize their emotions.

3. Research methods

This section will describe the steps, methods, and processes that will be used during the study. Firstly, the research question and hypothesis of the study will be reminded, as they represent the core and primary steps of conducting the research. The second section will present the methods chosen to be used for this study, as well as the design of the study. Finally, the last sections talk about the survey created for the research and the recruitment process of participants.

3.1 Research question and hypothesis

The research question that is aimed to be answered at the end of this study is:

RQ1: How do different levels of emotional awareness in a chatbot impact users' engagement during a conversation?

Based on the previous findings of several researchers in the field, but also of the primary purpose of this research, the study starts with the following hypothesis, which is going to be investigated further:

H1: The awareness that the chatbot shows during the conversation impacts the engagement of participants.

The research question stands as a support for understanding how people's interaction with chatbots could be improved and whether or not, bringing human-specific characteristics into the chatbots could enhance users' engagement level.

3.2 Experimental study

The method that has been used to discover the answer to the research question and to test the set hypothesis, was an experimental study approach.

Experimental studies are used in cases where the relationship between two variables needs to be determined, in a certain situation. Experimental studies rely on the concept of causality which explains how one variable, the cause, has the power to influence another variable, the effect, through different events or processes. These types of studies are ideal for determining the cause-effect relationship between two variables (Patten and Newhart, 2018, pp. 12–14; Gergle and Tan, 2014; Thyer, 2012, p. 6; Wildemuth, 2017, p. 103).

Table 1. Description of the study group conditions.

Study group	Description
BASELINE	The participants will have a conversation with the chatbot, without receiving any kind of feedback regarding its emotional awareness.
EMOJI-ONLY	During the entire conversation with the chatbots, the participants received subtle feedback from the chatbot that reflected their emotions. This feedback consisted of a subtle emoji that displayed user emotion in real time by switching through different pictograms.
EMOJI-AND-CHAT	Besides the subtle emoji feedback which is present in the second experimental condition, participants also benefited from receiving direct messages from the chatbot regarding their recorded emotions

Experimental studies start from defined hypotheses that reflect what the researcher wants to test to be true or to understand. In order for an experimental study to be successful, the hypothesis needs to be formulated as clearly as possible. Gergle and Tan (2014, pp. 196–197) explain that a good hypothesis should be specific on what the author wants to understand, and it should clearly define the causeeffect relationship between the variables in question. Moreover, the results of a Table 1 tested hypothesis should be meaningful and contribute to further research. To test the initially set hypothesis, the participants in the studies are randomly divided into several groups that are offered different experiences, based on the independent variables from the study. The independent variables are the cause within the causeeffect relationship of an experimental study, while the dependent variables are the effect (Patten and Newhart, 2018, p. 78; Gergle and Tan, 2014, pp. 199–200; Wildemuth, 2017, p. 101).

Among the study conditions created, the first one is the control group, which is defined as not receiving any kind of treatment that would influence the cause-effect relationship in any way (Thyer, 2012, p. 77). This condition is used to show the starting point of the relationship between variables and to compare the outcomes of the hypothesis to it.

The presented study is composed of three different study conditions, that differ based on the independent variable that describes the level of emotional awareness of a chatbot during a conversation. The study groups are presented and described in Table 1.

3.3 Survey

The experimental study conducted offered information about participants' emotions throughout the entire conversation with the chatbot. Moreover, having the entire conversations stored helped in analyzing the effects that different messages have on participants' moods. However, there is still a need for information about how the participants felt during the conversation and what are their opinions regarding the chatbot's emotional abilities. The level of engagement that users had during the conversation is the decisive finding, therefore the second method used was a survey.

Surveys are used to collect data based on people's opinions, beliefs, or experiences in certain situations. Each response to a survey plays an essential role in the final Figure 1 result of the phenomenon analyzed. The number of people from the targeted group, that take part in the survey, influences how likely it is to be able to generalize the results to the entire group of people. Thus, the more people take part in a survey, the better the final results are (Patten and Newhart, 2018, p. 19).

Prior studies have used surveys to identify the level of engagement in different circumstances, including in other chatbot applications. O'Brien et al. (2018) proposed an engagement scale that is being used in many Human-Computer Interaction (HCI) studies either by implementing the exact recommended User Engagement Scale (UES) form or as a base for determining and formulating the appropriate questions based on the study's needs. The UES form proposed by O'Brien et al. (2018) is divided into six different dimensions that are thought to have an effect on users' engagement, namely: focused attention, perceived usability, aesthetic appeal, endurability, novelty, and felt involvement. For the purpose of this study, the short UES form would be used as a base in formulating the appropriate form questions, using the same five-point rating scale approach, which has been proven to provide more qualitative results when the participants are part of the general public (Weijters et al., 2010; Revilla et al., 2014).

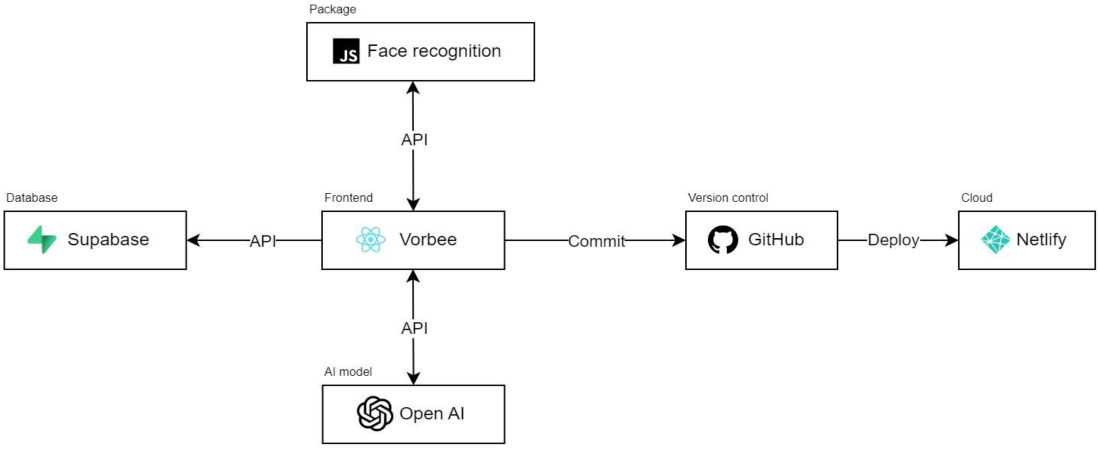


Figure 1. Architecture of Vorbee.

Additionally, other engagement scales used in chatbot applications were reviewed and used as a reference in creating the survey for this paper. Moilanen et al. (2022) study aimed to find out how different personalities of a chatbot can influence user engagement. The study used the short form of UES provided by O'Brien et al. (2018), and moreover, it used clusters for determining the most used positive and negative words when users described the chatbot's personalities. In the study conducted by Croes and Antheunis (2021) a five-point Likert scale was also used in order to find out how dimensions such as trust, social presence, self-disclosure, anonymity, and shame differ in particular communication environment cases (face-to-face, online with another human, or chatbot). The study was focused on improving the implementation of chatbots that help people that struggle with mental health problems. However, the results could have an impact on the implementation of chatbots in all domains and on the level of engagement in participants as the dimensions used, rely on the same ideas provided in the UES form of O'Brien et al. (2018).

Other studies such as the one from Zhu et al. (2022) used a slightly different method, choosing a sliding scale approach from 0 to 100 in order to determine what is the impact that emotional expressiveness has on the interaction between voice-based chatbots and humans.

Relevant engagement items were identified also in a study conducted by McLean and Osei-Frimpong (2019). The study focuses on voice assistant technology and how users can be motivated by different dimensions of benefits provided by the application. Even though it's focused on a type of virtual assistant that does not have the same characteristics as the virtual assistant provided in this study, the part of the survey which focuses on the users' emotional experience (hedonic benefits) could be helpful for developing the form items of this study.

By evaluating all the engagement scale approaches used in the above-mentioned studies, a five-point Likert scale survey approach was developed to identify the level of engagement of the proposed chatbot by considering a set of dimensions. The dimensions and their items are provided in Table 2.

Table 2. Survey for user engagement scale.

Dimension	**Items**
Focused attention	I lost myself in this experience.
	The time I spent communicating with the virtual agent just slipped away.
	I was absorbed in this conversation with the virtual agent.
Perceived usability	I felt frustrated while talking with the virtual agent.
	I found the conversation with the virtual agent confusing.
	Having this conversation with the virtual agent was taxing (demanding).
Endurability	The conversation with the virtual agent was worthwhile.
	The overall conversation was rewarding.
	I felt interested in having this conversation with the virtual agent.
Felt-involvement	I felt engaged in the conversation at all times.
	The conversation with the virtual agent was human-like.
	The actual conversation with the virtual agent was entertaining.
Trust	During the conversation, I felt comfortable sharing personal information.
	During the conversation, I felt that I could be open.
	The virtual agent was trustworthy.
	The virtual agent was understanding.
	The virtual agent had good intentions.
Social presence	During the conversation I was able to respond to the reactions of the virtual agent.
	During the conversation, I felt that I was having a conversation with a social being.
	During the conversation, the virtual agent reacted to my emotions.
Anonymity	During the conversation I felt anonymous.
	During the conversation, I felt like I could share more about myself because my conversation partner did not know me.
Choose five terms from the list to describe the perceived personality of the virtual agent	Calm, Clear, Uninterested, Formal, Confident, Kind, Repetitive, Annoying, Chatty, Humane, Serious, Boring, Friendly, Attentive, Cold, Inhumane, Honest, Interested, Inattentive, Limited, Direct, Likeable, Disconnected, Monotone, Informative, Sociable, Superficial, General, Useful, Optimistic, Emotional, Entertaining, Respectful, Positive, Empathetic, Compassionate, Happy, Interactive, Trustworthy, Joyful
Open-ended questions	How does a chatbot's ability to detect your emotions affect your perception of the bot?
	How would a chatbot's ability to detect your emotions affect the topics you are willing to discuss with it? Consider any scenarios also outside the interaction you just had with a bot,
	Please reflect on the overall idea of leveraging emotion detection in online crowd work platforms (e.g. the one you are using now)?

3.4 Participants

The participants for this study were recruited using Prolific[2]. Prolific is a platform that helps researchers find participants in their studies, providing many features, including an easy and reliable way to get in touch with participants. A number of 180 participants were recruited in total, from different locations USA, UK, Ireland, and Australia. The platform used offers the ability to select participants based on certain criteria. For this study, the participants were required to speak English fluently and to have experience, knowledge, and accept pain-related questions as the topic of the discussion with the chatbot was chronic pain. In order to make sure that all the participants experienced chronic pain, one of Prolific's options was used, namely recruiting only participants with chronic pain. Moreover, the participants were required to have a minimum approval rate of 80 on the platform and a minimum number of previous submissions of 100. For this study, the requirements for the study were the use of a webcam and the recording of videos, even though the study does not store any raw videos or images with the participants.

[2] Prolific website: https://www.prolific.co/

4. Designed artifact

The experimental study presented in the previous chapter requires support for implementing and providing participants with the appropriate environment that would allow them to observe and evaluate the chatbot's level of awareness. In this section, the designed artifact built for this experimental study will be presented.

4.1 Design science

Design science is a research methodology used for finding solutions to known problems by designing and building artifacts, thus enhancing people's competencies and knowledge in certain domains. Researchers combine their own expertise and creativity with the theories and findings from prior research for their study in order to recursively improve and innovate their design science artifact. (Hevner et al., 2004).

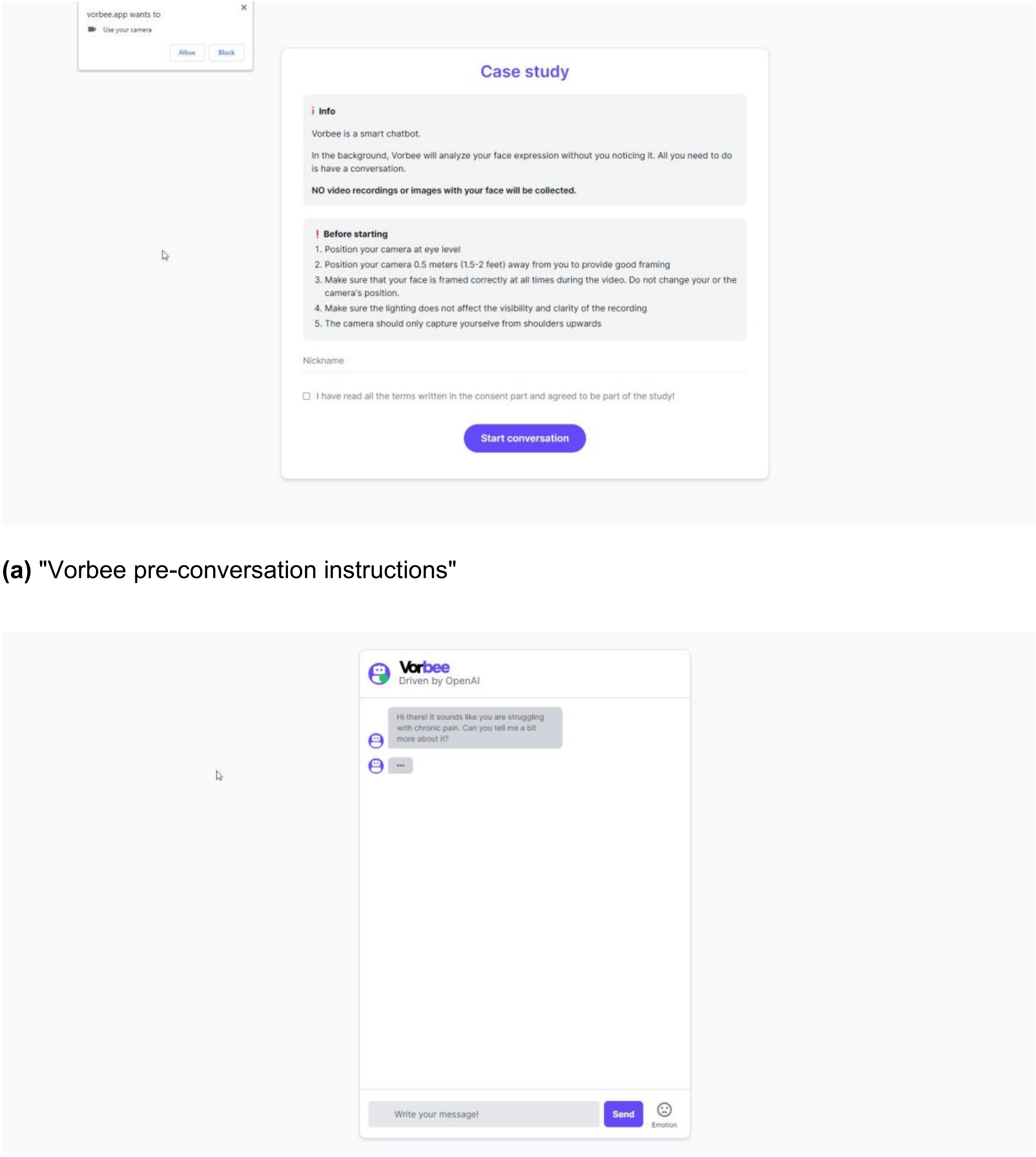
Prior studies have used the design science methodology in the domain of humanchatbot interaction as well. Multiple artifacts which consisted of the development of a chatbot have been designed over the years in order to find answers and solutions to multiple questions. To determine how different facial expressions of a chatbot influence people when in a negotiation situation with it, Melo et al. (2011) developed software that included a chatbot, a chatbot with a virtual character created that showed emotions. Ghandeharioun et al. (2019) designed an artifact similar to the one designed in this study, that replied to messages based on user's expressions, or Krämer et al. (2013) also built an artifact that used a chatbot to determine people's smiling behaviors.

Prior work showed that one of the main limitations of chatbots is their lack of empathy (Adamopoulou and Moussiades, 2020;Zhou et al., 2023). This limitation can be seen in different domains where chatbots have been used: education, health, or customer service (Ashfaq et al., 2020; Pérez et al., 2020). The presented designed artifact is a software tool, called Vorbee, which contributes to the prior work by proposing a solution to creating more empathetic chatbots. A first requirement for creating such a design tool is to create a chatbot that can be aware of a person's emotions, as empathy is strongly related to a person's emotions.

4.2 Vorbee – The Designed Artifact

For this study, a responsive web application was developed, which allowed participants to engage in a conversation with a chatbot for a few minutes. The chatbot's name is Vorbee. During the conversation, the participant can talk about any topic of interest, while the chatbot engages in the discussion and responds to any changes in users' facial expressions. Based on the three conditions, the chatbot was set to show no emotional awareness, show subtle emotional awareness through an emoji pictogram, or offer high emotional awareness feedback by replying with a direct message to the participant, commenting on the emotion recorded.

The application used multiple technologies in order to create a suitable experience for the participant that would also facilitate the purpose of the study. Therefore, the application was composed of two main sections: the pre-conversation section, subfigure 2a and the actual chat, subfigure 2b. In the pre-conversation section, participants were offered a set of instructions and good practices in order for them to have a proper experience with the chatbot, as well as offering them a chance to see and adjust their camera before starting (after agreeing to video permissions). In the conversation section, the participants were prompted to a chat where the VA initiated the conversation. Based on the condition, some of the participants were also able to see the emoji that shows what emotion the chatbot registers, or even get emotional-aware messages during the discussion.



vorbee.app wants to

Use your camera

Case study

i Info

Vorbee is a smart chatbot.

In the background, Vorbee will analyze your face expression without you noticing it. All you need to do is have a conversation.

NO video recordings or images with your face will be collected.

! **Before starting**

1. Position your camera at eye level
2. Position your camera 0.5 meters (1.5-2 feet) away from you to provide good framing
3. Make sure that your face is framed correctly at all times during the video. Do not change your or the camera's position.
4. Make sure the lighting does not affect the visibility and clarity of the recording
5. The camera should only capture yourselve from shoulders upwards

Nickname

□ I have read all the terms written in the consent part and agreed to be part of the study!

(a) "Vorbee pre-conversation instructions"

Vorbee
Driven by OpenAI

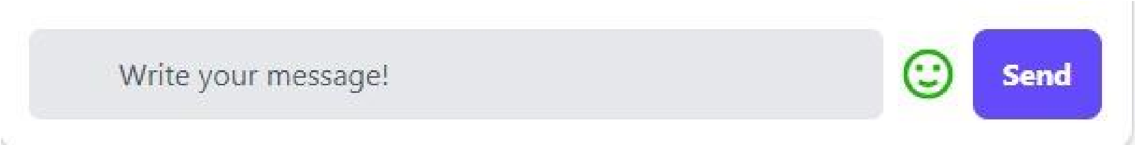
Hi there! It sounds like you are struggling with chronic pain. Can you tell me a bit more about it?

...

Write your message!

(b) "Vorbee ongoing chat"

Figure 2. Vorbee interface.



Write your message!

Figure 3. Example of subtle emoji that displays user's emotion in real-time.

The web application, Vorbee, is a React application built on top of Vite[3], which is a tool for building fast and optimized web applications. Vorbee has integrations with a face-recognition module called face-api.js, which uses Javascript, a *Supabase*[4] database, which is an open-source PostgreSQL database that provides lots of features to easily store and manage data for an application, and Open AI for providing the means of integrating the intelligent chatbot into the application. GitHub was used as a version control system, and also as a support for automatic deployments on *Netlify*[5], a web platform used for deploying and hosting web applications fast and easy. The architecture of the application is presented in Figure 1.

From a technical point of view, the developed web application has two main integrations that allow the chatbot to be aware of users' feelings.

Face recognition [6]is a public open-source API available on GitHub that is used for face detection and face recognition purposes. The API can be used using NodeJS or directly from the web browser. It is built using JavaScript and Tensorflow core JavaScript API to integrate pre-trained models in applications in different forms. The package offers multiple features and ways in which it can be used, from single to multiple face detection, as well as face and emotion detection, or age and gender prediction. The module was mainly used for its emotion recognition feature, but age and gender predictions were also made during the conversation.

The API can detect a total of seven emotions, all of which were transposed into emojis in the web application, as it is presented in Table 3. All emotions that were available for detection were transposed into the web application as emojis because the study does not aim to find out what happens only when a certain emotion is detected, but whether how the chatbot is able to respond in different situations, to different emotions. The emotional awareness of the chatbot is not measured only by sentiments of happiness or sadness but by the multitude of sentiments that a person can have and need to be recognized.

All the data that was extracted was stored in Supabase, a Firebase open-source alternative, using relational databases. From there, the data was easy to access, visualize, and export for further data analysis. Table 4 includes all the relevant data Figure 2 stored for data analysis.

The second fundamental integration of the web application is the use of OpenAI's GPT-3 model through their API. GPT-3 is one of the large language models OpenAI developed. GPT-3 is trained to interact in a conversational way with users by answering questions or correcting mistakes. Through the available APIs, it is possible to create a chatbot that is instructed through a prompt on how it should interact. Vorbee was prompted to discuss participants' chronic pain. The chronic pain topic was used in order to trigger differences in facial expressions naturally. A conversation between the chatbot and a participant is going to last only a few minutes. For such a short conversation, the topic discussed needs to be of interest to the participant so that changes in facial expression would be more likely to happen. Thus, participants were recruited using Prolific's option of selecting

[3] Vite website: https://vitejs.dev/

[4] Supabase website: https://supabase.com/

[5] Netlify website: https://www.netlify.com/

[6] Face-recognition website: https://justadudewhohacks.github.io/face-api.js/docs/index.html

participants that deal with chronic pain to ensure that participants that deal with such problems are recruited and that the topic is suitable for all participants.

Table 3. Recorded emojis by face recognition API and their emoji representation in the application.

Emoji representation	Recorded emotion
🙂	Happy
☹	Sad
😠	Angry
😨	Fearful
😕	Disgusted
😦	Surprised
No emoji representation	Neutral

The third study condition required additional features to the chatbot so that whenever users change their emotions, the chatbot would reply with a comment about the user's feelings, accordingly. Therefore, the web application was designed to trigger a reply from the chatbot whenever the facial expression of a user changes. The prompt that the chatbot received needed to clearly state what the chatbot should say, in order to give valuable replies to the participants. Therefore, for every emotion, the prompt provided had to be different. Table 5 displays all the prompt messages that the OpenAI API received, based on the response needed.

Table 4. Relevant data stored for further analysis.

Data stored	Description
Gender probability	Gender probability, between 0 and 1 with high precision
Age prediction	Age probability
Happy score	Happiness probability, between 0 and 1 with high precision
Sad score	Sadness probability, between 0 and 1 with high precision
Angry score	Anger probability, between 0 and 1 with high precision
Fearful score	Fear probability, between 0 and 1 with high precision
Disgusted score	Disgust probability, between 0 and 1 with high precision
Surprised score	Surprise probability, between 0 and 1 with high precision
Neutral score	Neutral probability, between 0 and 1 with high precision
Participant message	Content of users messages
Chatbot message	Content of chatbot messages
Sentiment detected	Sentiment detected by OpenAI based on user's message

The flow of the overall application looks as shown in Figure 4. First of all, the user needs to consent to the video camera being open so that the facial expressions can be analyzed and followed. No pictures or videos with participants' faces were taken or stored anywhere. After consenting, the user needed to start a conversation with the chatbot, which lasted for a few minutes. During this conversation, the data recorded by the video camera through the face-recognition API was extracted, parsed, and further send through a JavaScript Object Notation (JSON) format to the OpenAI API and to the database, where it was stored. In the end, after a few minutes of conversation, the participants were able to navigate to the last step of the study, where they were provided a survey containing a set of questions that helped determine their engagement scale and perception of the conversation they had. The survey was created and managed via Google Forms for multiple reasons such as consistency and a bug-free environment or easy sharing and managing of data.

Table 5. Prompt messages provided to OpenAI API based on the needed response.

Prompt type	**Prompt content**
Simple Response	Vorbee is a chatbot that engages in conversations with people with chronic pain. Start a conversation about their chronic pain. This is the start of the conversation. Vorbee needs to also greet*{username}*. + *{conversation history}*
Detect Happiness	*{username}* seems to be smiling. Make a comment to *{username}* about noticing this.
Detect Sadness	*{username}* seems to be feeling a little sad. Encourage *{username}* and say something to make *{username}* smile!
Detect Anger	*{username}* is feeling angry. Show support and ask about their feelings.
Detect Fear	*{username}* is afraid. Comfort them and suggest what they can do to feel better.
Detect Disgust	*{username}* is disgusted. Ask why are they feeling this way and offer solutions.
Detect Surprise	*{username}* is surprised. Mention seeing this and try to find out what they are surprised about.
Detect Neutral	No prompt

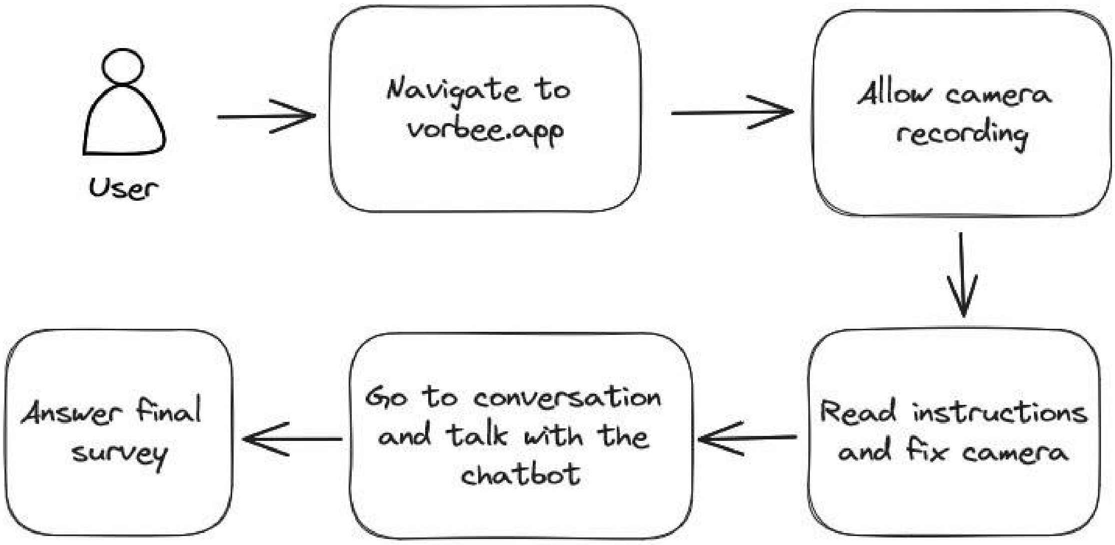

User
Navigate to vorbee.app
Allow camera recording
Read instructions and fix camera
Go to conversation and talk with the chatbot
Answer final survey

Figure 4. Steps followed by each participant in the study.

5. Results

In this section, the study's results will be presented and statistically interpreted. The section will present the data gathered from the survey.

Table 6. Demographics

	Baseline	Emoji-only	Emoji-and-chat	Grand total
Total	60 (33.7%)	59 (33.1%)	59(33.1%)	178 (100%)
Age				
18-25	7 (11.7%)	10 (16.9%)	13 (22.0%)	30 (16.9%)
26-35	22 (36.7%)	17 (28.8%)	16 (27.1%)	55 (30.9%)
36-45	14 (23.3%)	21 (35.6%)	13 (22.0%)	48 (27.0%)
46-55	9 (15.0%)	5 (8.5%)	10(16.9%)	24 (13.5%)
55+	8 (13.3%)	6 (10.2%)	7 (11.9%)	21 (11.8%)
Gender				
Female	28 (46.7%)	32 (54.2%)	33 (55.9%)	93 (52.2%)
Male	32 (53.4%)	27 (45.8%)	26 (44.1%)	85 (47.8%)
Nationality				
Australia	0	5	4	9 (5.1%)
Belarus	1	0	0	1 (0.6%)
China	0	0	1	1 (0.6%)
El Salvador	0	0	1	1 (0.6%)
Ghana	0	1	0	1 (0.6%)
Hungary	0	0	1	1 (0.6%)
India	0	1	0	1 (0.6%)
Italy	0	2	0	2 (1.1%)
Latvia	0	0	1	1 (0.6%)
Lithuania	1	0	0	1 (0.6%)
Malta	0	0	1	1 (0.6%)
Mexico	0	0	1	1 (0.6%)
Nigeria	0	0	2	2 (1.1%)
Poland	3	0	0	3 (1.7%)
Saudi Arabia	0	0	1	1 (0.6%)
South Africa	0	0	1	1 (0.6%)
Turkey	1	0	0	1 (0.6%)
United Kingdom	50	31	38	119 (66.9%)
United States	3	16	8	27 (15.2%)
Vietnam	0	1	0	1 (0.6%)

A power analysis using the statistical software G*Power was conducted to determine an appropriate number of participants. Using a one-way ANOVA study design, we specified a medium effect size f = 0.25 with a power level of 0.80 and α = 0.05 (Faul et al., 2009). The required sample size was determined to be 159, which we rounded up to 180, resulting in 60 participants in each of the three conditions.

Out of the total 180 targeted participants, two participants did not complete the study entirely, omitting to answer the form questions. As a result, two out of the three study conditions, specifically the EMOJI-ONLY and the EMOJI-AND-CHAT conditions, have one participant less. Therefore, the total number of participants considered in the analysis of the results is 178, with 60 participants in the BASELINE condition, 59 participants in EMOJI-ONLY, and 59 participants in the EMOJIAND-CHAT condition.

5.1 Demographics

Table 6 shows the number and percentage of participants based on the study condition and grouped by age group, gender, and nationality.

Most of the participants were between 26-35 or 36-45 years old with a total percentage of 30.9% and 27.0%. From the gender point of view, the participants were relatively evenly-balance with 93 participants female and 85 male participants, even though this was not a requirement or a purpose for this study.

Participants from 22 nationalities took part in the study, from all over the world, most of them being from the United Kingdom, with 119 participants or the United States, with 27 participants.

The results will be split into qualitative and quantitative results and further analyzed and described individually.

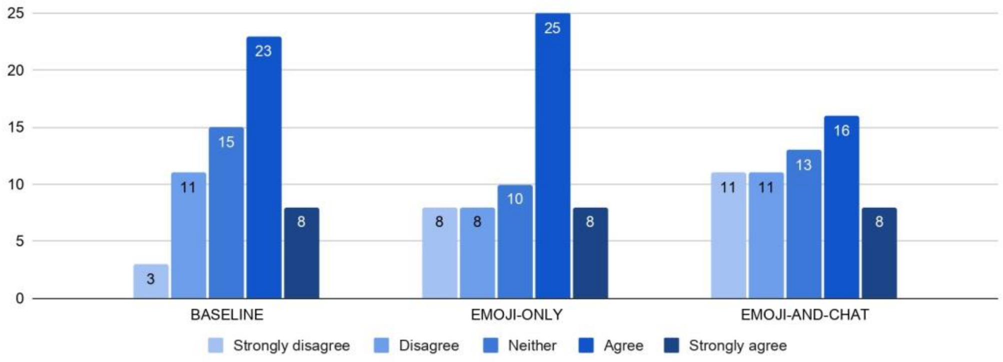
5.2 Quantitative results

The survey results will be presented based on the dimensions of the questions which were introduced in Table 2. The five-point Likert scale was used in the first seven dimensions. The Likert scale is represented in each of the following figures using numbers from 1 to 5, where 1 means "Strongly disagree" and 5 means "Strongly agree".

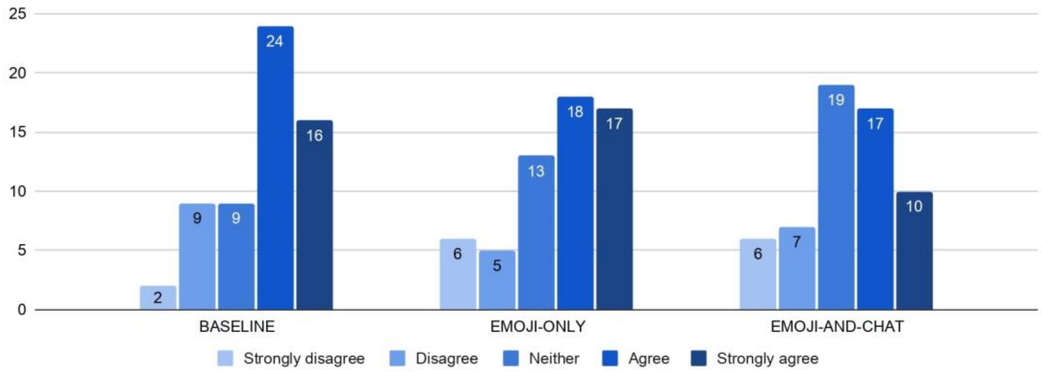
5.2.1 Focused attention

The focused attention dimension refers to the capacity of the participants to deeply engage in the discussion, to the extent that they would lose track of time, experiencing a sense of immersion in the interaction with the chatbot (O'Brien et al., 2018).

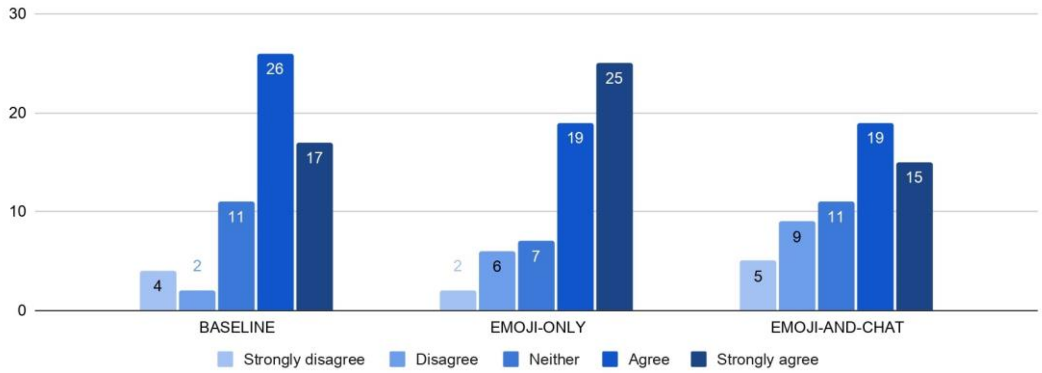
The statements analyzed in the first dimension are a) *"I lost myself in this experience"*, b) *"The time I spent communicating with the virtual agent just slipped away"* and c) *"I was absorbed in this conversation with the virtual agent"*. Figure 5 displays frequency bar charts for each statement which help identify the central tendency of each statement and each study condition individually.



(a) *"I lost myself in this experience"*



(b) *"The time I spent communicating with the virtual agent just slipped away"*



(c) *"I was absorbed in this conversation with the virtual agent"*

Figure 5. Focused attention category results

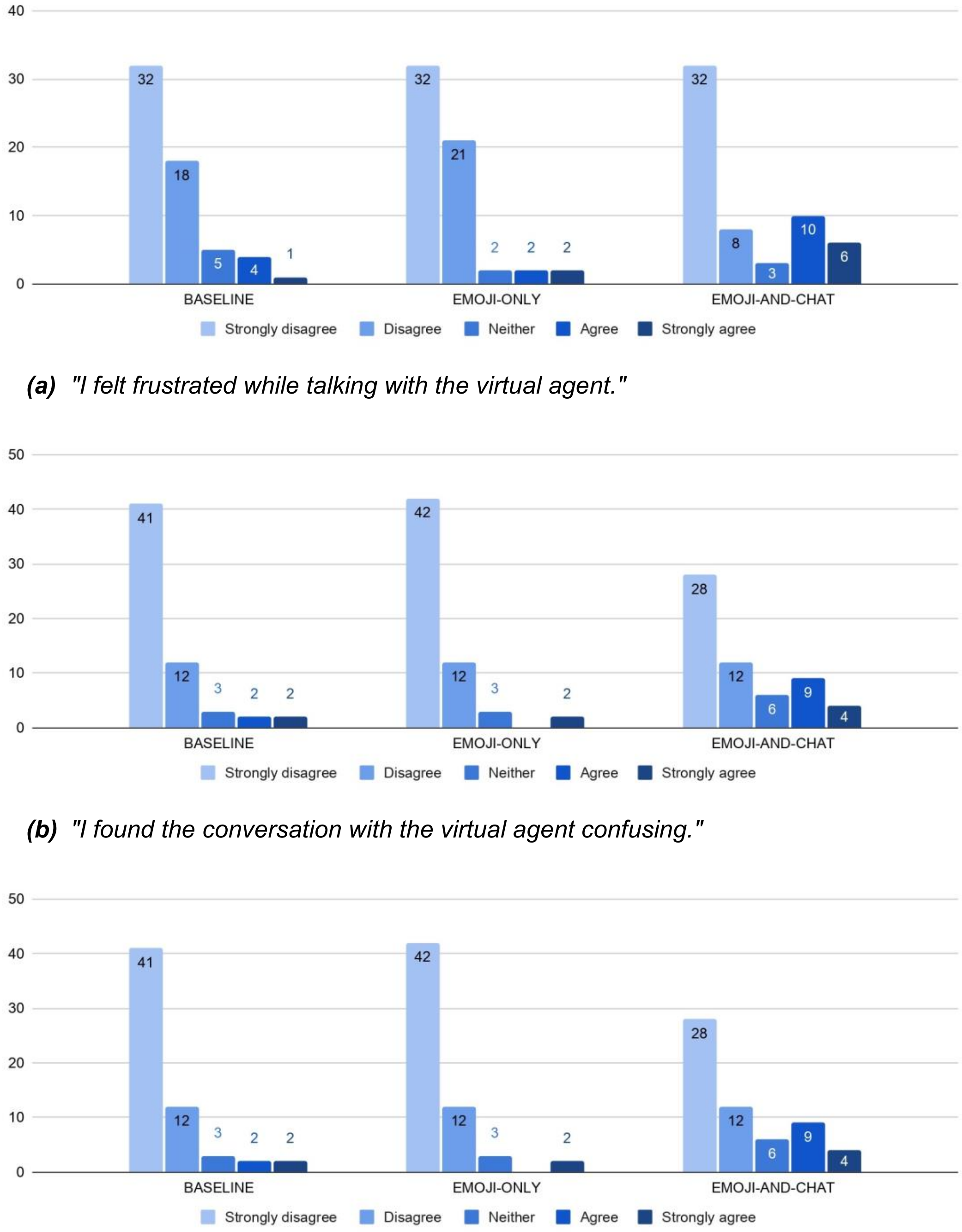
The subfigures 5a and 5c show that for each case study the central tendency is 4, namely Agree. In the second statement described by subfigure 5b the central tendency of the EMOJI-AND-CHAT condition is 3, Neutral, and for the other two cases is 4, Agree.

A Kruskal-Wallis H test was conducted for each statement, and each dimension to determine if there were differences in the dimensions scores between conditions. The results were interpreted using the Kruskal-Wallis H test because of its ability to determine significant differences between three or more conditions.

For the focused-attention dimension, the scores between conditions are BASELINE (*n* = 60), EMOJI-ONLY (*n* = 59), and EMOJI-AND-CHAT (*n* = 59) engagement level. For each statement, the distributions of scores were similar for all conditions, as assessed by visual inspections of bar charts. For statement one (subfigure 5a) the scores went from BASELINE (*Mdn* = 4), to EMOJI-ONLY (*Mdn* = 4), to EMOJI-AND-CHAT (*Mdn* = 3), but the differences were not statistically significant, ($x^2(2)$ = 2.781)(*p* = .249). For the second statement (subfigure 5b) the scores started from BASELINE (*Mdn* = 4), to EMOJI-ONLY (*Mdn* = 4), and decrease to EMOJI-AND-CHAT (*Mdn* = 3), however, the differences were not statistically significant, ($x^2(2)$ = 4.161)(*p* = .125). For the last statement (subfigure 5c) the scores were consistent BASELINE (*Mdn* = 4), EMOJI-ONLY (*Mdn* = 4), and EMOJI-AND-CHAT (*Mdn* = 4), but the differences were not statistically significant, ($x^2(2)$ = 5.469)(*p* = .065).

5.2.2 Perceived usability

The perceived usability dimension includes the negative feelings that a participant might feel due to the interaction with the chatbot. (O’Brien et al., 2018) The statements analyzed in this dimension are a) *"I felt frustrated while talking with the virtual agent"*, b) *"I found the conversation with the virtual agent confusing."* and c) *"Having this conversation with the virtual agent was taxing (demanding)."*. Figure 6 shows the frequency bar charts for each statement individually. From this data it can be observed that the central tendency of each statement and condition is 1, Strongly disagree.



(a) *"I felt frustrated while talking with the virtual agent."*

(b) *"I found the conversation with the virtual agent confusing."*

(c) *"Having this conversation with the virtual agent was taxing (demanding)."*

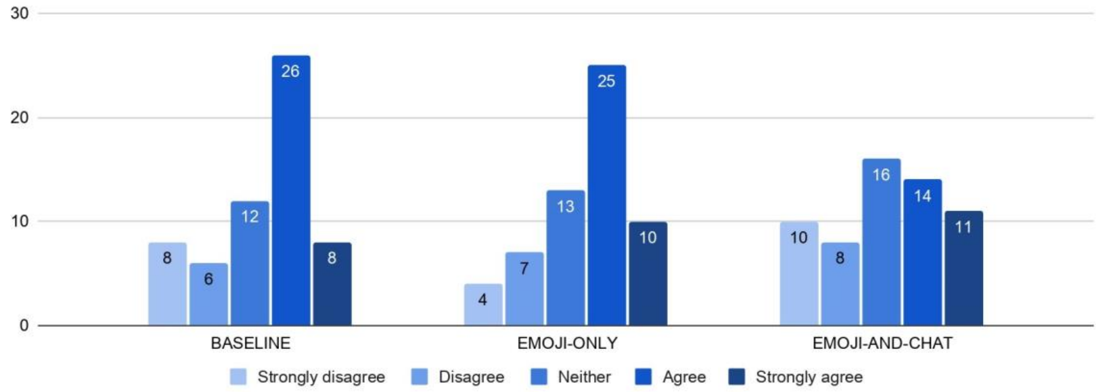
Figure 6. Perceived usability category results

Kruskal-Wallis H tests were also conducted on this dimension, using the same sample for each condition. In the first statement (subfigure 6a) the scores were constant at value 1, specifically Strongly disagree, between all three cases, but the differences were not statistically significant, ($x^2(2) = 1.393$)($p = .498$). In the second statement (subfigure 6b), distributions of scores were similar for all conditions, BASELINE (*Mdn* = 1), EMOJI-ONLY (*Mdn* = 1), and EMOJI-AND-CHAT (*Mdn* = 2). Median scores were statistically significantly different between conditions, ($x^2(2) = 11.559$)($p = .003$). Therefore, pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Adjusted p-values are presented. This post hoc analysis revealed statistically significant differences in scores between the EMOJI-ONLY (*Mdn* = 1) and EMOJI-AND-CHAT (*Mdn* = 2)($p = .005$) but not between any other conditions combination. In other words, the participants from the EMOJI-ONLY condition felt less confused during the conversation with the chatbot, than those from the EMOJI-AND-CHAT condition. Lastly, in the third statement (subfigure 6c) the scores resemble the ones from the first statement, meaning that scores between cases were constant at value 1 and the differences were not statistically significant, ($x^2(2) = 5.263$)(p = .072).

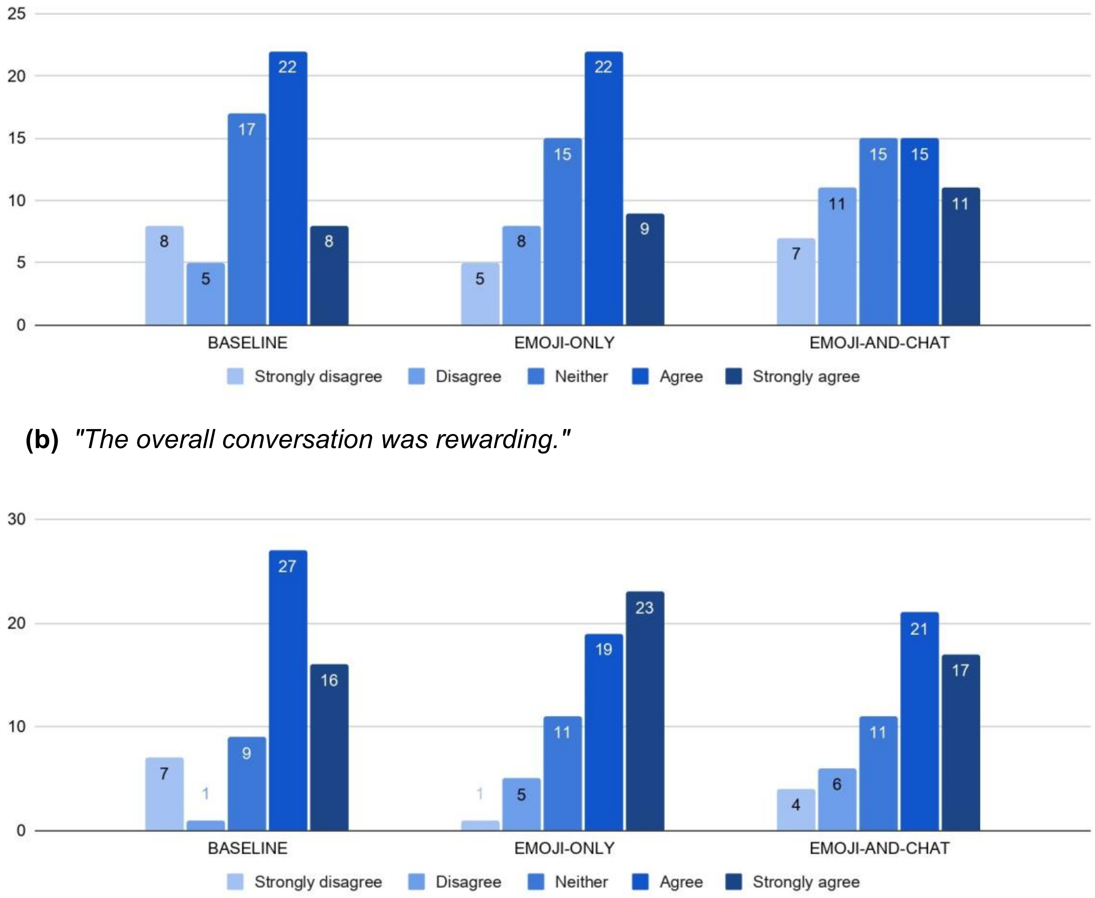
5.2.3 Endurability

The endurability dimension contains questions regarding how thriving the experience was and the likeliness of chatting with the chatbot again or suggesting the experience to others. (O'Brien et al., 2018)

The statements analyzed in this dimension are a) *"The conversation with the virtual agent was worthwhile."*, b) *"The overall conversation was rewarding"* and c) *"I felt interested in having this conversation with the virtual agent."*.



(a) *"The conversation with the virtual agent was worthwhile."*



(b) *"The overall conversation was rewarding."*

(c) *"I felt interested in having this conversation with the virtual agent."*

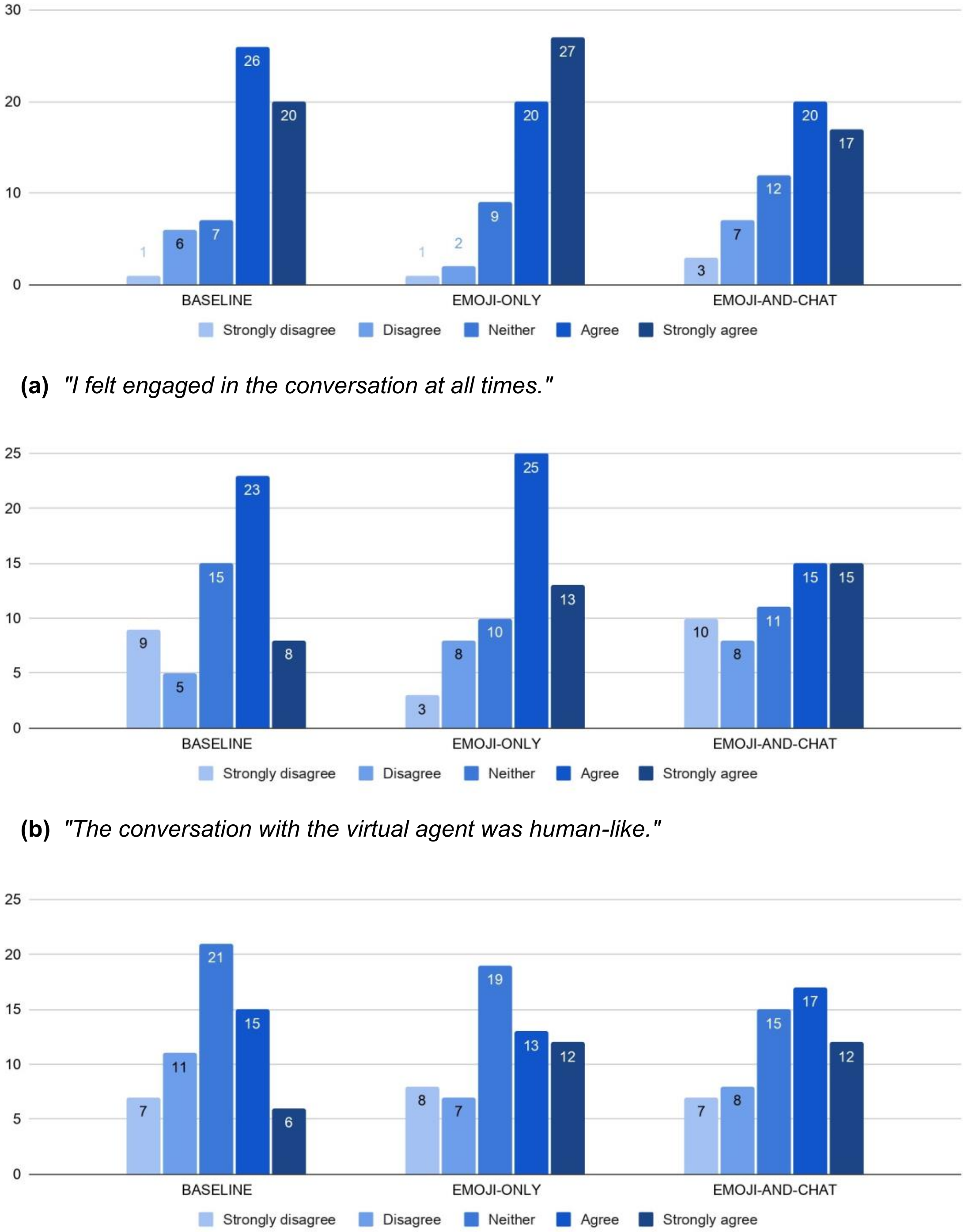
Figure 7. Endurability category results

The Kruskal-Wallis H tests conducted on the first statement (subfigure 7a) of this dimension indicate that the scores are from BASELINE (*Mdn* = 4), to EMOJIONLY (*Mdn* = 4), to EMOJI-AND-CHAT (*Mdn* = 3), but the differences were not statistically significant, (x^2(2) = 2.373)(*p* = .305). Focusing on the second statement, (subfigures 7b, the scores were the same for BASELINE AND EMOJIONLY cases (*Mdn* = 4), whereas, for the EMOJI-AND-CHAT case, the score was a little lower (*Mdn* = 3). Finally, for the third statement 7c), the scores were the same for all cases (*Mdn* = 4). The differences were not statistically significant in none of the last two statements, (x^2(2) = .545)(*p* = .762), respectively (x^2(2) =1.846)(*p* = .397).

5.2.4 Felt-involvement

The felt-involvement dimension refers to how enjoying and immersing the interaction was perceived by the participant (O'Brien et al., 2018).

The statements analyzed in this dimension are a) *"I felt engaged in the conversation at all times."*, b) *"The conversation with the virtual agent was human-like"* and c) *"The actual conversation with the virtual agent was entertaining."*. Figure 8 shows the frequency bar charts for each statement individually. From this data, it can be observed that the central tendency of the first two statements (subfigures 8a and 8b) is 4, namely Agree, and for the last statement (subfigure 8c), the central tendency is 3, Neutral.



(a) *"I felt engaged in the conversation at all times."*

(b) *"The conversation with the virtual agent was human-like."*

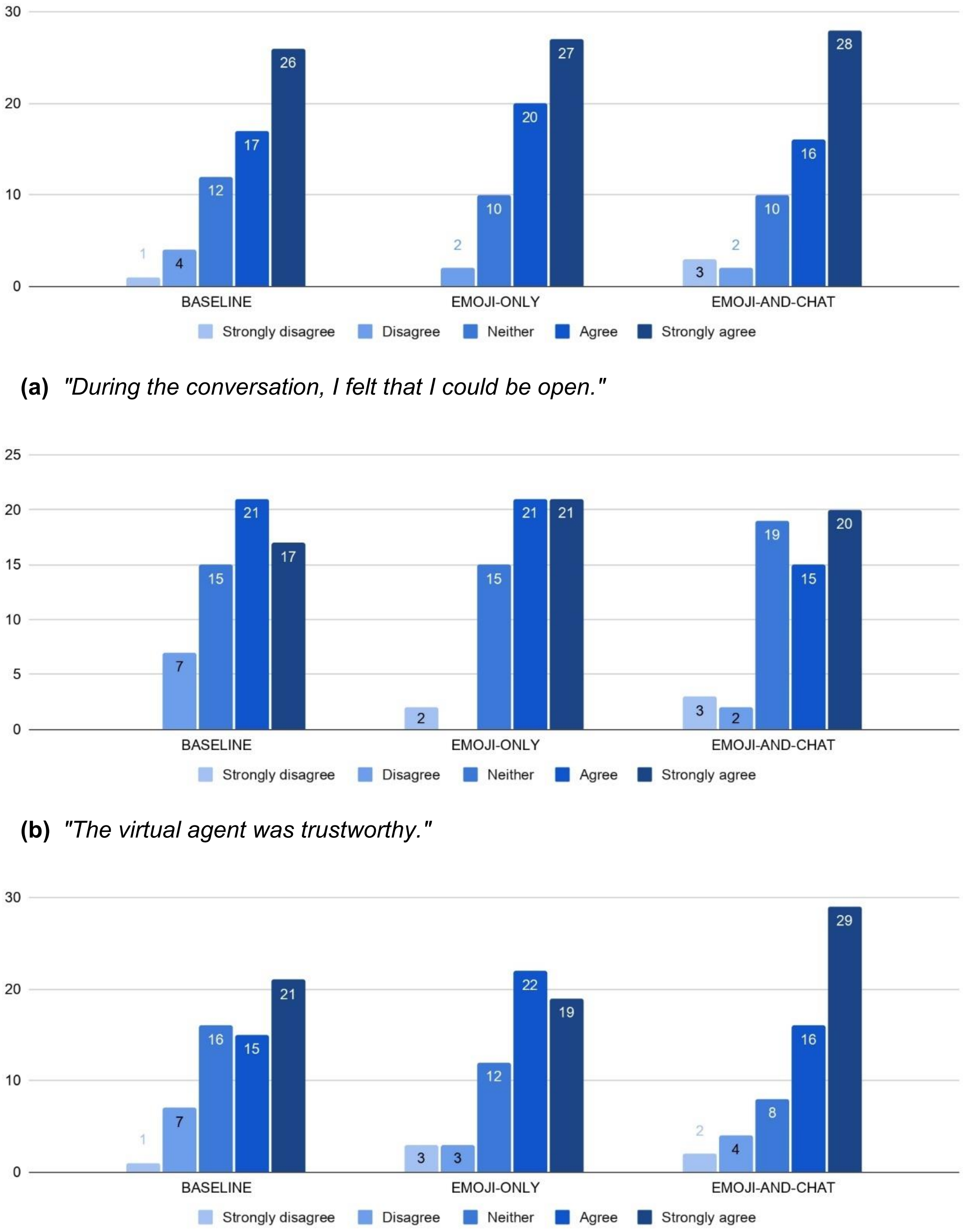
(c) *"The actual conversation with the virtual agent was entertaining."*

Figure 8. Felt-involvement category results

Kruskal-Wallis H tests were conducted for these statements also, to determine the differences in the felt involvement among conditions with different engagement levels from the chatbot. The first statement (subfigure 8a) showed a constant value score (*Mdn* = 4), among all the three conditions, with no statistical significance ($x^2(2) = 5.919$)(p = .052). For the second statement, the score value for all three conditions is also constant (*Mdn* = 4), and no statistical significance was found in this case as well ($x^2(2) = 2.612$)(p = .271). The last statement of this dimension showed constant score (*Mdn* = 3), with no statistical significance ($x^2(2) = 2.040$)(p = .361).

5.2.5 Trust

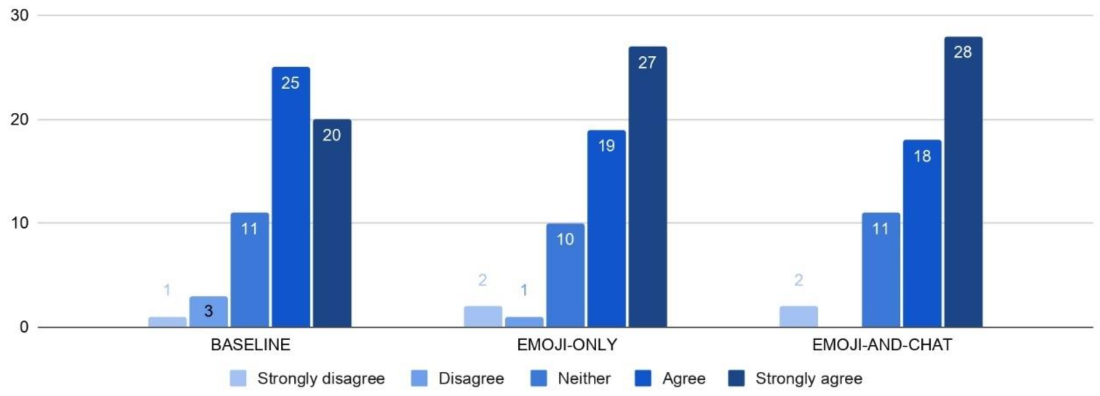
The trust dimension contains questions that refer to the willingness of the participant to share personal information during the conversation. (Croes and Antheunis, 2021) The statements used in the trust dimension are a) *"During the conversation, I felt that I could be open."*, b) *"The virtual agent was trustworthy."*, c) *"The virtual agent was understanding."* and d) *"The virtual agent had good intentions."*. Figure 9 shows the frequency bar charts for each statement individually. From this data, it can be observed that the central tendency of all the statements, between all conditions is 4, namely Agree.



(a) *"During the conversation, I felt that I could be open."*

(b) *"The virtual agent was trustworthy."*

(c) *"The virtual agent was understanding."*



(d) *"The virtual agent had good intentions."*

Figure 9. Trust category results

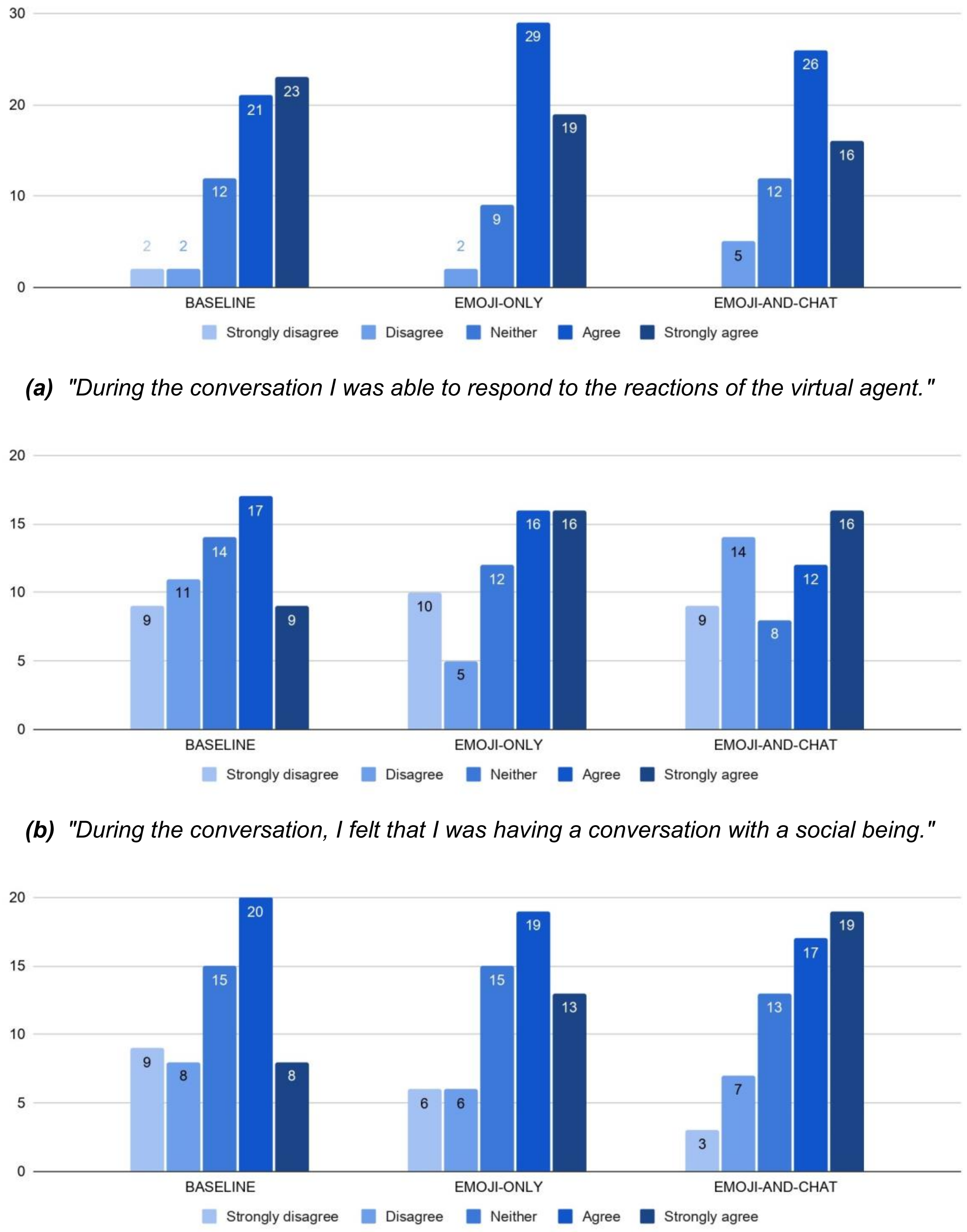
Kruskal-Wallis H tests were conducted for the statements, to determine the differences in trust among conditions. All statements (subfigures 9a, 9b, 9c, 9d), among all conditions showed constant score (*Mdn* = 4), but none were statistically significant: statement 1 ($x^2(2)$ = .489)(*p* = .783), statement 2 ($x^2(2)$ = 1.562)(*p* = .458), statement 3 ($x^2(2)$ = 3.806)(*p* = .149) and statement 4 ($x^2(2)$ = 2.037)(*p* = .361).

5.2.6 Social presence

The social presence dimension refers to the quality of a person that is particularly noticeable and prominent during the interaction with another person. (Croes and Antheunis, 2021)

In this dimension, the three statements used are: a) *"During the conversation I was able to respond to the reactions of the virtual agent."*, b) *"During the conversation, I felt that I was having a conversation with a social being."*, and c) *"During the conversation, the virtual agent reacted to my emotions."*. Figure 10 shows the frequency bar charts for each statement individually.

For each statement, the distributions of scores were similar for all conditions, as assessed by visual inspections of bar charts. For statement one (subfigure 10a) the scores were constant between all three conditions (*Mdn* = 4), but the differences were not statistically significant, ($x^2(2)$ = 1.520)(*p* = .468). For the second statement (subfigure 10b) the scores started from BASELINE (*Mdn* = 3), to EMOJIONLY (*Mdn* = 4), and decrease to EMOJI-AND-CHAT (*Mdn* = 3), however, the differences were not statistically significant, ($x^2(2)$ = 1.554)(*p* = .460). For the last statement (subfigure 10c) the scores started from BASELINE (*Mdn* = 3), to EMOJI-ONLY (*Mdn* = 4), and EMOJI-AND-CHAT (*Mdn* = 4), but the differences were not statistically significant in this case either, ($x^2(2)$ = 5.634)(*p* = .060).



(a) *"During the conversation I was able to respond to the reactions of the virtual agent."*

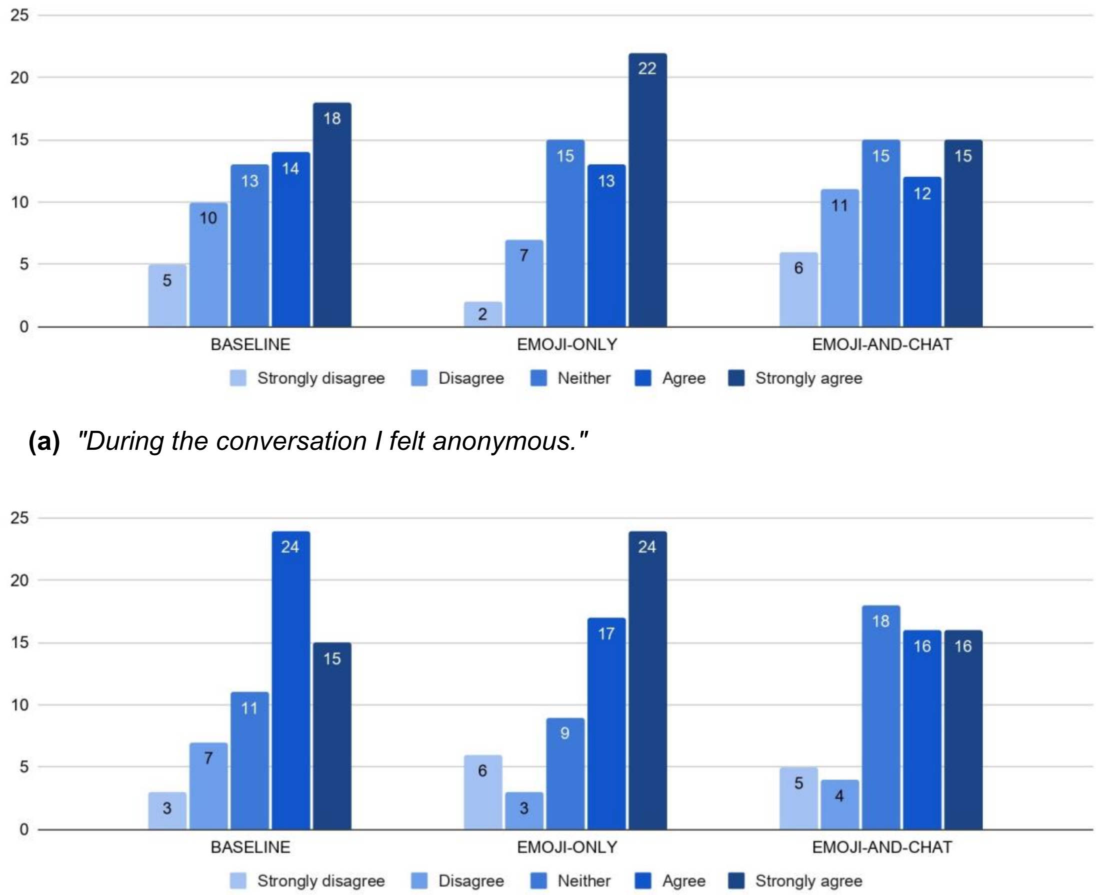
(b) *"During the conversation, I felt that I was having a conversation with a social being."*

(c) *"During the conversation, the virtual agent reacted to my emotions."*

Figure 10. Social presence category results

5.2.7 Anonymity

The anonymity dimension aims to discover how safe, unrecognized, or undiscovered the participant is feeling during the conversation, as anonymity plays an important role in the willingness to share information with others. While protected by anonymity, a person may feel safer sharing more sensitive information (Croes and Antheunis, 2021).



(a) *"During the conversation I felt anonymous."*

(b) *"During the conversation I felt like I could share more about myself because my conversation partner did not know me."*

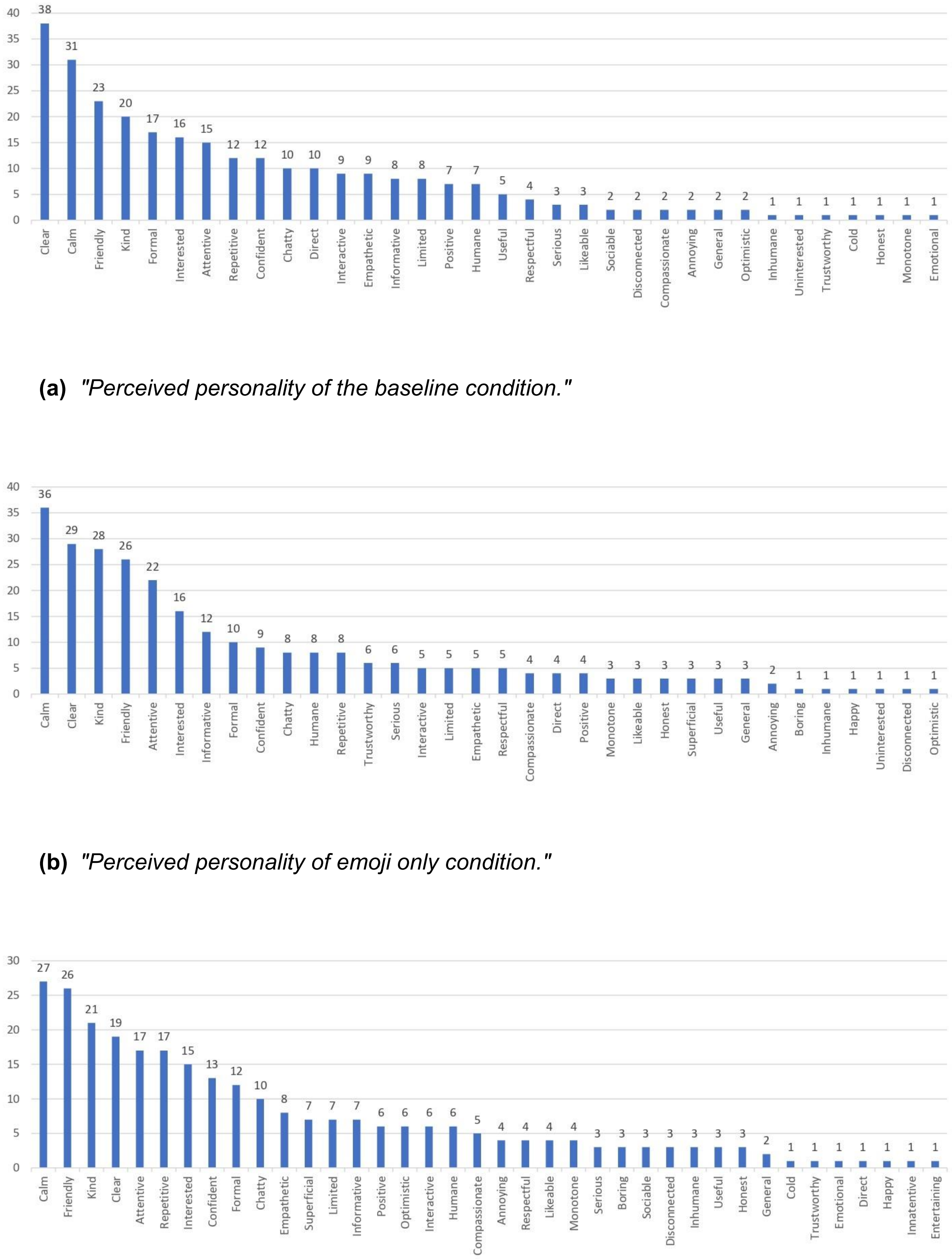
Figure 11. Anonymity category results

Anonymity dimension covers two statements: a) *"During the conversation I felt anonymous."*, and b) *"During the conversation I felt like I could share more about myself because my conversation partner did not know me."*. Figure 11 shows the frequency bar charts for each statement individually.

Kruskal-Wallis H tests were conducted for the statements, to determine the differences in anonymity among conditions. The first statement (subfigure 11a) the scores started from BASELINE (*Mdn* = 4), EMOJI-ONLY (*Mdn* = 4), and decrease to EMOJI-AND-CHAT (*Mdn* = 3). However, the differences were not statistically significant, (x^{2}(2) = 3.592)(*p* = .166). For the second statement (subfigure 11b), the scores were constant between all three conditions (*Mdn* = 4), but the differences were not statistically significant, (x^{2}(2) = 2.727)(*p* = .256).

5.2.8 Perceived personality of the chatbot

In addition to the Likert scale statements presented above, the survey included a part where participants were asked to choose a maximum of five terms from a list, that they believed best described the chatbot. The dimension aims to explore the perceived personality of the chatbot, aligning with the practices outlined by Moilanen et al. (2022) to find out how different chatbot personalities influence user engagement.



(a) *"Perceived personality of the baseline condition."*

(b) *"Perceived personality of emoji only condition."*

(c) *"Perceived personality of emoji and chat conditions."*

Figure 12. Perceived personality of the chatbot

Figure 12 shows participants' choices of terms to describe the chatbot, based on the study condition. Subfigure 12a shows that in the BASELINE condition, the ten most used terms were: clear (*n* = 38), calm (*n* = 31), friendly (*n* = 23), kind (*n* = 20), formal (*n* = 17), interested (*n* = 16), attentive, (*n* = 38), repetitive and confident (*n* = 12) and chatty and direct (*n* = 10). For the second study condition, EMOJI-ONLY the results shown in subfigure 12b indicate that the ten most used terms in this condition are: calm (*n* = 36), clear (*n* = 29), kind (*n* = 28), friendly (*n* = 26), attentive (*n* = 22), interested (*n* = 16), informative (*n* = 12), formal (*n* = 10), confident (*n* = 9) and chatty, humane and repetitive (*n* = 8). In the last study condition, EMOJI-AND-CHAT, the results from subfigure 12c reveal that the most frequent ten terms used are: calm (*n* = 27), friendly (*n* = 26), kind (*n* = 21), clear (*n* = 19), attentive and repetitive (*n* = 17), interested (*n* = 15), confident (*n* = 13), formal (*n* = 12) and chatty (*n* = 10).

5.3 Qualitative results

The open-ended question aims to find out participants' opinions about the overall experience, and the ability of chatbots to detect emotions.

The three open-ended questions that are going to be analyzed are: a) *"How would a chatbot's ability to detect your emotions affect the topics you are willing to discuss with it? Consider any scenarios also outside the interaction you just had with a bot."*, b) *"Please reflect on the overall idea of leveraging emotion detection in online crowd work platforms (e.g. the one you are using now)?"*, and c) *"How does a chatbot's ability to detect your emotions affect your perception of the bot?"*.

Participants' responses were effectively examined using various coding, analysis, and interpreting techniques. Firstly, the thesis author read and categorized the data in order to familiarize themselves with the answers and group them based on similarity. Categorization was based on the participants' emotional disposition toward the chatbot and its use of emotion recognition in regard to the targeted question. This means that the responses which expressed a positive view of the emotion detection chatbots were categorized as *Positive Opinion*, while those highlighting issues and drawbacks of using such chatbots, were marked as *Negative opinion*. Responses presenting both advantages and disadvantages, expressing confusion or uncertainty, or not fitting into any category were classified as *Neutral opinion*. Additionally, a category labeled *Unrelated opinion* was created for those responses that did not make sense or did not align with the question's intent.

Table 7. Number of participants' responses grouped by category and condition for each open-ended question

Question	Condition	Opinion			
		Positive	Negative	Neutral	Unrelated
1	BASELINE	25	18	14	3
1	EMOJI-ONLY	29	12	12	3
1	EMOJI-AND-CHAT	30	13	15	4
2	BASELINE	15	30	15	0
2	EMOJI-ONLY	9	30	18	2
2	EMOJI-AND-CHAT	9	25	25	0
3	BASELINE	28	8	25	1
3	EMOJI-ONLY	34	4	19	3
3	EMOJI-AND-CHAT	33	12	14	0

After finding out the right category for each response, the coded responses were again individually analyzed to identify patterns, trends, and insights that might be relevant to the final results. During the pattern analysis phase, the number of responses that fit one pattern was counted, and reasons for why participants made certain claims were also collected for reference use. The patterns were made so that one response would be included in only one pattern or subcategory.

Table 7 presents the number of participants' responses for each question, grouped by conditions and category.

The open-ended questions are not related to the condition that the participants were a part of in any way. Still, it can be a factor that influenced their perception of the emotion recognition chatbot. In cases where the condition is considered to be relevant, it will be clearly mentioned, otherwise, the results will be interpreted only based on the categories determined in the analysis part.

5.3.1 First Open-Ended Question

The first open-ended question has got 84 positive opinions, 43 negative, 41 neutral, and 10 unrelated opinions that were not taken into consideration. The predominant pattern for the positive opinion answers was the fact that the chatbot's ability to detect emotions would make people more open and honest while discussing with the chatbot. People believe they will be more willing to discuss a wider range of subjects including more sensitive topics such as family, relationships, future plans, and even emotional matters. The reason behind people feeling that they could be more open with such kinds of chatbots is the fact that the conversation will remain anonymous.

> *"I would chat to it about things that were troubling me or I just needed to get off my chest knowing they wouldn't repeat it to anyone I know."*

Participants also stated that it might be harder to hide things from a chatbot if its ability to detect emotions is working as it should, which can make one open up more.

Chatbot's ability to offer information is unquestionable and the participants are also aware of that. A number of 11 participants mentioned that the chatbot would be helpful, especially for giving advice, discussing with it when one is feeling lonely, or helping people calm down when they are upset. The emotion detection feature makes the chatbot feel more natural and human-like, makes people feel more understood, and the conversation more tailored, which are also reasons why participants liked the chatbot.

Another important aspect brought up by participants is the potential in helping people that struggle with mental illnesses, and depression, offer emotional support, help people identify their feelings, and be used for counseling sessions.

> *"It would help especially in helping someone with a mental illness."*

> *"... I can see it being helpful for relevant things like mental health apps, to be able to give the best responses even when someone can't actually put into words how they are feeling, or to help them figure out what indeed they are feeling."*

> *"I would discuss more personal and emotional topics. I would discuss mental health issues and struggles I am going through."*

> *"If it works this can be very useful, especially with mental health issues where the bot doesn't just have to rely on written sentiment analysis but can also read the face of the user."*

However, participants did not just state that they would in any way be willing to open up to the chatbot. Some of the participants clearly stated that they would see the detection of emotion as a positive feature only if it is implemented really well and it interprets the emotions correctly, provides appropriate feedback, takes into account confidentiality and it is transparent regarding it.

> *"I find there is often little point in replying in depth to a chatbot if it has no recognition of what I am trying to get across to it. If I feel that a chatbot understands what I am trying to say and get my emotions or frustrations then that would help me to open up and give the bot more details."*

> *"I would be comfortable chatting with a bot about any subject if it was informative and confidential."*

> *"I might be more willing to discuss topics, but I would want to be really sure about confidentiality."*

> *"I'd be more than happy to communicate if the bot's feedback was more appropriate."*

The negative answers regarding the emotion detection feature indicate that 31 participants mentioned that they would open up less during the discussions with the chatbot, the reason being exactly the ability to detect their emotions. Participants do not feel comfortable with having their cameras open during the conversation, or with the general idea that their face is seen. Some of the terms used by the participants to describe how they would feel were: *"spied upon"*, *"manipulated"*, *"exposed"*, or *"vulnerable"*.

> *"I would feel spied upon I think, which would make me more reluctant to share."*

> *"I think I may be more cautious about the things I would talk about perhaps because I feel more vulnerable having my face and the emotions it presents exposed."*

Moreover, privacy matters are a big concern for participants, as mentions of this were seen also in the positive opinions. People are reticent when it comes to having a camera on or letting a program analyze their faces, as well as trusting that the conversations will remain anonymous.

> *"I don't think it would change what I would say. Knowing no matter what everything said is logged somehow, There are no secrets."*

Regarding the neutral opinions of participants, 33 claimed that emotion detection would not affect in any way the topics they are willing to discuss with the chatbot. Other participants also noted that they would completely ignore it, need time to adjust to it, or that its usefulness depends on the circumstances.

5.3.2 Second Open-Ended Question

The second open-ended question includes a total of 33 positive opinions, 85 negative opinions, 58 neutral opinions, and 2 unrelated ones. Although the data provided by the participants is valid and reveals relevant findings, it needs to be mentioned that some of the answers provided for this question are not crowd work specific, as the question specifies, but talk about the usefulness of emotion detection chatbots in general.

In the positive opinion category, a number of 24 participants claimed that the emotion-detection chatbot would be indeed useful. Some of the reasons mentioned are aligned with the ones mentioned in the previous question. Participants claimed that the feature could be helpful for people who need to discuss their personal feelings urgently, or in establishing trust and openness. A participant mentioned the possibility of using the feature even in police work if it would be capable of telling a truth from a lie. When focusing on crowd platforms, participants declared that it has the potential to help understand people and circumstances better.

> *"I think it is a good idea that, leveraging emotion detection on the crowd work platform as it can know when you are feeling stressed or other emotions and try to help or make you feel better just by having a* conversation*."*

> *"I think it has some interesting possibilities and could potentially help understand participants even further."*

Concerns arise regarding its need for improvement before it can become anything that generally can be used. First and foremost, emotion detection needs to work faultlessly so that it can assist people as best as possible and without causing problems. Further, participants also perceived this type of chatbot as a companion or as a method to improve research data even further.

> *"It got my emotions wrong so hard to say apart from to improve. Overall it could be a good thing."*

The participants who considered this feature as a negative addition to the crowd work platforms sustained their claims mainly through two statements: 40 people would not trust machines and algorithms with such a personal feature, and 27 participants are mainly concerned about invading their privacy or security risks. The reason behind the lack of trust is the fact that facial expressions can differ a lot from one person to another, which may lead to misinterpretation of emotions and giving wrong suggestions. Moreover, participants pointed out a significant downside that affects people that are neurodiverse, for example, people that have autism or Tourette's syndrome, or even people who suffer from neuro-facial paralysis. In these cases, the trained algorithm would not be able to interpret expressions correctly.

> *"I would be concerned that my emotions may not be read properly or misunderstood."*

> *"I am concerned that people with disabilities such as autism may be read incorrectly. Sometimes expressions do not accurately reflect feelings, and the chatbot should understand this."*

> *"It feels a little invasive. I wouldn't want to be penalized for getting visibly upset or frustrated. Also, I'm autistic, so my face sometimes looks blank or sad, when I'm not upset. So it would probably not always be accurate for everyone."*

Privacy and security are present in participants' answers as people have little understanding of what data is collected, what will be used for, or how they can be sure that the camera does not record their faces when it analyses their expressions.

> *"I would be very concerned about privacy, especially with sensitive topics of discussion, I would definitely be concerned I would be recorded and my data shared. I could see that the responses may be more personalized but I would not trust it."*

> *"I am concerned about privacy if other people are looking at the conversation and rating the bot responses. I would wonder why it's needed for most surveys or crowd work since they are generally not requiring an emotional response."*

> *"I don't think I would appreciate generalized use of this on crowd work platforms, it would feel less anonymous and feel too intrusive."*

Other concerns mentioned by participants include the possibility of people abusing this feature, for commercial (e.g. emotionally tailored advertisements) or political purposes or even that people could become too attached to the AI which will prevent them from forming relationships with other humans. When referring to crowd work, the possibility of human transactions ending was mentioned.

People whose responses were categorized as neutral opinions showed that most of them see the idea as a potentially helpful one but at the same time, they also expressed concerns regarding, privacy, security, or misinterpretation of feelings. Therefore, 25 participants from the neutral opinion group shared their vision of the possibilities and also impediments of this idea.

> *"It seems like it could add a lot to studies if that was something that researchers were interested in. I would be concerned about privacy and the truthfulness behind not being recorded in any way."*

> *"It could be useful for judging participants' attention and true feelings. However, it may misjudge people and wrongly categorize them in research."*

Some respondents mentioned that lack of concern regarding this or that they do not see this as useful, while others had trouble deciding, concluding that it depends on the circumstances. A few participants also mentioned that the interaction might restrict people from going and creating human relationships, or that they generally prefer human interactions more.

> *"My face would show my concern for the keyboard more than any other emotion so the chatbot analysis would be skewed."*

In the end, participants who also stated their opinion about the use of emotionally aware chatbots in crowd work showed interesting opinions regarding this feature. The statements showed both interest and positive views of adding this feature in crowd work platforms, while others also pointed out concerns or negative aspects regarding this.

> *"It would improve research data as it would be clearer if someone was only going through the motions."*

> *"I think it is a good idea that leveraging emotion detection on the crowd work platform as it can know when you are feeling stressed or other emotions and try to help or make you feel better just by having a conversation. "*

> *" ... studies on Prolific are generally trustworthy which makes a big difference (for example, the fact the video wasn't being recorded was a major factor in me being happy to take the study in this case, so I'd need to trust that to be true as well in others). Also because Prolific is almost all academic studies I can imagine plenty of legitimate & worthwhile reasons to use it; ... "*

> *"I am concerned about privacy if other people are looking at the conversation and rating the bot responses. I would wonder why it's needed for most surveys or crowd work since they are generally not requiring an emotional response."*

> *"I don't think I would appreciate generalised use of this on crowd work platforms, it would feel less anonymous and feel too intrusive."*

5.3.3 Third Open-Ended Question

Responses collected from the third question ended up including 93 positive opinions, 24 negative opinions, 57 neutral, and 4 unrelated ones. The predominant response in the positive group was that the emotion detection feature makes the bot feel more human, with 30 responses focused on this specific reason. A number of 13 participants claimed that this type of chatbot offers better responses, and another 27 mentioned that the overall experience is enhanced. The chatbot is able to make participants more understood and listened to, which they claimed helps them be more engaged.

Other claims made by the participants are that they are more trusting of the chatbot and that it makes them more aware of their facial expressions.

> *"It makes me feel like they understand me more."*

> *"It made it feel even more humanlike."*

> *"It allows me to have a more positive perception of the bot, as I can connect with it and it makes me feel as if I am not wasting my time talking to someone that doesn't understand how I feel."*

> *"It makes me trust the chatbot even more than I would have done otherwise."*

Moving the focus to the negative responses, participants mentioned trusting the bot less, their reasons including privacy issues, the bot feeling intrusive, and the fact that in the end, this is just a bot that someone programmed to act like that, meaning that they cannot forget the true nature of their interlocutor. People mentioned that the chatbot did not feel like talking to a real person, furthermore, it is unable to detect emotions completely accurately.

> *"Just felt that it was generic/repetitive with its responses."*

> *"It wasn't always 100% accurate to what I was experiencing. It seemed like it was guessing but was guessing incorrectly."*

> *"I don't like the idea of being watched by a camera, so I felt a bit uneasy."*

> *"I would perceive it to be a violation of my personal privacy."*

In the neutral category, most of the participants, number 27 said that the emotion detection did not have any effect on them, while others expressed mixed feelings, enjoying the idea but having concerns about the implementation and privacy. Another problem brought up multiple times was the inaccuracy of emotion detection. A total of 10 participants from this category claimed that the emotion detection did not work at all or it worked poorly, an issue that can change someone's perception of the feature entirely, so it needs attention.

> *"I don't know, I wasn't really aware that I was showing emotions or that it was responding to them - it felt more like it was responding to my text messages"*

> *"I don't think it paid any attention to my emotions only to what I said."*

6. Discussions

During this section, the entire results of the thesis will be presented. The first section will focus on the objectives set, whereas the second section will revisit the hypothesis and research questions, reflecting on the learnings extracted from participants' results. The third section presents the limitations met through the study, while the fourth section presents further implications and advice that could be used by other researchers. Finally, the last subsection talks about future research in this domain.

6.1 Revisiting objectives

In addition to the research question and the hypothesis that were set, two objectives were also defined at the beginning of the paper. The purpose of these objectives was to help in observing the progress of the results that were aimed for.

O1: The development of a software artifact that would facilitate the scientific experiment.

O2: Carry out the experiment in order to gather data from which results can be extracted.

The first objective of this paper was accomplished by creating the responsive web application, Vorbee. The application integrated a virtual assistant run by OpenAI whose purpose was to talk with the participants about their chronic pain experience, the participants being specifically chosen from this category using Prolific's pain category option.

The second objective was met by recruiting participants for each condition and after having a minimum 3-minute chat with the chatbot, providing them with a survey that contained the questions mentioned in Table 2. From a total of 180 participants recruited through the Prolific platform, a total number of 178 responses were taken into consideration, two participants omitted to answer the survey at all.

By successfully completing the above-mentioned objectives, valuable and dependable data was collected. The data collected was further used in analyzing and interpreting the results in order to be able to answer the research question and verify the hypothesis.

6.2 Revisiting hypothesis and research question

The scientific experiment conducted in this thesis aimed to discover if the engagement level of people can be enhanced by the awareness that the chatbot shows during a conversation.

With that purpose in mind, the following research question and the hypothesis were set in the initial stage of the study and used as support throughout every phase of the research.

RQ: How do different levels of emotional awareness in a chatbot impact users' engagement during a conversation?

H1: The awareness that the chatbot shows during the conversation impacts the engagement of participants.

The results analysis from the quantitative data collected shows that an overall consistency exists between the responses of participants from each condition. There is no or little difference between the conditions in all the dimensions, differences that were found statistically insignificant after running the Kruskal-Wallis H test. The dimensions where invariability was found between all three conditions were the perceived usability, felt involvement, trust, and social presence dimensions.

During the conversation with the chatbots, most of the participants did not experience any negative feelings such as frustration, confusion, or perceiving the conversation as demanding. For all three statements from the perceived usability dimension, people strongly disagreed with the initial statements that suggested otherwise. Moreover, the majority of the respondents agreed that they felt engaged during the entire conversation and that the conversation was human-like. However, in the last statement from the felt-involvement dimension, participants did not agree or disagree that the interaction was entertaining.

When focusing on trust, approximately all participants agreed that they felt they can be open while discussing with the chatbot and perceived it as trustworthy, understanding, and having good intentions. Participants also mostly agreed that the experience felt like having a conversation with a social being and the chatbot reacted to their emotions. Having the participants from all three conditions claim that the chatbot reacted to their emotions is an interesting finding, as only the participants from the last study condition, EMOJI-AND-CHAT, benefited from the chatbot's feature of reacting to users' emotions through messages. Therefore, the assumptions that the chatbot reacted to emotions within study conditions BASELINE and EMOJI-ONLY can be explained as the performance and capabilities that the OpenAI system has without additional emotional interpretation.

The rest of the dimensions, focused attention, endurability, and anonymity showed small differences between the EMOJI-AND-CHAT condition and the other two, however, the differences were not statistically significant. In general, most of the participants agreed with the statements "I lost myself in this experience" and "I was absorbed in this conversation with the virtual agent". Participants from the EMOJIAND-CHAT condition did not agree or disagree with the fact that time slipped away during the conversation, while in the other two study conditions, people agreed with this remark.

Additionally, the participants from the last condition had a neutral opinion regarding the conversation being worthwhile or rewarding, whereas the respondents from the other two conditions agreed with this allegation. The interest in having the discussion was stated by most of the participants, regardless of the study condition.

Finally, the anonymity dimension results showed that the participants from BASELINE and EMOJI-ONLY conditions agree they felt anonymous, while the majority of respondents from the last condition did not give a positive or negative answer. In general participants from all conditions mentioned, they felt they could share more information about themselves due to the anonymity.

When asked to choose a maximum of five terms from a given list that would best describe the perceived personality of the VA, the four most popular terms used in all 3 conditions, were "clear", calm, friendly and kind. All of these four terms have a positive meaning, framing the interaction into a suitable and comfortable experience with or without recognizing and replying to emotional changes. The term formal was also used in all three conditions, however, in the BASELINE condition it was mentioned by 17 participants whereas, in the EMOJI-ONLY and EMOJI-ANDCHAT, it was mentioned less, by 10 and

12 participants, which could show a slow decrease in how formal the chatbot is perceived when participants receive feedback about their facial expressions being analyzed. Another term that appeared in all three conditions was attentive which showed a slight increase between conditions (BASELINE 15, EMOJI-ONLY 22, EMOJI-AND-CHAT 17). The increase of the term in the EMOJI-ONLY conditions could be due to the emoji feature that shows participants how the chatbot "sees" them, while the decrease from emoji-only to the EMOJI-AND-CHAT condition could be due to the repetitiveness of the bot. The chatbot was mostly perceived as repetitive especially in the last condition, as the term was chosen by the same amount of people that choose attentive, 17, but also based on the statements from the open-ended questions. Lastly, a term that appeared only in the last condition was empathetic chosen by 8 participants, showing the potential that such features could have in improving the overall experience.

The qualitative data gathered from the open-ended questions were the most valuable as they gave a lot of insights on problems identified by participants, their views on the emotion recognition chatbot, and explanations to back up their claims.

The first and last open-ended questions had predominantly answers which were categorized as *Positive Opinion*, while the second question mostly had responses categorized as *Negative Opinion*.

The responses gathered showed the willingness of people to discuss more personal subjects with a chatbot. Emotion detection is making people feel they are talking with a real person while also having the assurance that the conversation is anonymous. This combination of factors seems to be favorable for participants as it allows them to open up about sensitive, particular topics which they would be interested in discussing, but do not feel comfortable sharing with people. Such a feature was claimed to be important, especially for people that struggle with mental illnesses like depression and seek emotional support because the chatbot would not rely only on the written response of the user, but also on the analysis of facial expressions.

The overall experience with a chatbot is enhanced when it is able to detect emotions and discussing with the chatbot feels like having a conversation with a human, which makes people feel like they are listened to and therefore, it is worth putting more effort into the discussion.

The negative opinion category revealed the reasons why people would not use or do not see the importance of having an emotionally aware chatbot. As a result, all of these negative statements resulted from flaws identified in the chatbot, especially from the privacy point of view or from inaccuracy. Exaggeration of the chatbot can affect people's perception of the accuracy of the chatbot, as responding too often to little facial expressions or neutral ones, may feel odd. People tend to expect a chatbot to act as human-like as possible especially when it comes to empathy. For example, a small, temporary sad emotion should not trigger the chatbot to reply with a long message of encouragement, instead, it should either omit it the first time, simply try to change the tone of the message, or offer any kind of small, non-invasive and nonaggressive input (Ghandeharioun et al., 2019). Therefore, participants' responses align with findings from prior research made by Ghandeharioun et al. (2019).

Participants shared the same ideas on the positive outcomes that the enhanced chatbot could bring, and they also shared the same concerns regarding what needs to be improved before this idea can actually guide and assist people with their problems.

Considering all the feedback gathered from participants, it can be stated that participants' feedback contained strong and valid statements both for and against the emotional recognition feature. These differences between extremely positive and negative opinions regarding the emotional ability of chatbots can be very tightly related to the personality traits of people, whether they are extroverts or introverts. A study conducted by Ghandeharioun et al. (2019) explains that extroverts are more prone to respond positively to such features, whereas introverts are more cautious with using such chatbots. Although the limitations of such a feature cannot be overlooked, it is undeniable that emotionally aware chatbots would be positively seen by people, with the necessary updates and improvements being made.

Moreover, one aspect that results have in common is that respondents' opinion about the emotionally intelligent chatbot is driven by the anonymity and the amount of trust they have for such a chatbot. Participants who view the feature positively claimed that they would open up more to this type of chatbot ultimately because they would trust it to give better responses and their anonymity profile. Considering the positive opinion from the open-ended questions and the trust dimension where the majority of people agreed with having trust in the chatbot that used facial recognition, the results align with the findings of Croes and Antheunis (2021) that shows people's tendency to open up more to a chatbot due to the anonymity and thus increased trust they have in comparison to the human-to-human interaction.

6.3 Limitations

The current study faced some limitations that need to be known so that further research could be aware of the constraints present here.

6.3.1 One pool of participants

Even though testing the developed chatbot with multiple user audiences would have been beneficial for the research end results, the study was tested using only one participant pool.

6.3.2 Chatbot's prompts

One of the chatbot's features within the third condition, EMOJI-AND-CHAT, was that whenever the facial expression of the participant changed, the bot replied with a message suitable for the facial expression change. The replies sent by the chatbot were generated based on the prompts described in Table 5 and which need further improvement in order for the responses to feeling more natural and appropriate within the context of every conversation.

6.3.3 Testing only some emojis

The trained model used to detect facial expressions during the conversation offered a strict and clear set of feelings that could be detected, as mentioned in Table 3. Therefore, a limited set of emojis were used in order to show to the participants what emotions are registered by the system.

6.4 Implications

This section provides ideas and advice that could be used in the further development of emotionally aware chatbot systems. The suggestions are based on the participants' opinions extracted from the open-ended questions of the study.

6.4.1 Trained model

The trained model that is used to detect the facial expressions of a person needs to be rigorous and able to detect as accurately as possible people's expressions. Participants have raised their concerns about the facial expression being inaccurately detected, claiming even that the chatbot "was guessing but was guessing incorrectly". The trained model used during this research was a JavaScript implementation of face recognition API for the browser, built on top of TensorFlow. One known limitation of this model regarding expression detection is that wearing glasses might decrease accuracy. In the future, a more robust trained model should be used to be able to provide a more accurate expression detection that would improve the experience.

6.4.2 Chatbot responses development

The development process of the chatbot mainly includes adjusting the prompts that OpenAI API receives in order to offer suitable responses. The initial prompt that the chatbot receives aims to describe the situation in which is used and what kind of information needs to provide. The other prompts provided to the chatbot, namely the prompts sent whenever a facial expression is detected, need also a careful choice of words. In that sense, multiple iterations are required. The right combination of empathy and friendliness needs to be found so the chatbot would not be perceived as annoying, too invasive, or too sensitive.

Moreover, the trained models provided by OpenAI continuously improve, meaning that other trained language models which are more enhanced are available. The chatbot provided in the Vorbee web application used the trained model called textdavinci-003 which is a GPT3.5 model. However, other improved language models have been developed in the meantime.

Nonetheless, developing such a chatbot is a challenge due to the simple fact that how the chatbot is perceived differs from one person to another.

6.4.3 Security and privacy

When referring to security and privacy, people have a hard time trusting the application. In this case, the situation is sensitive, as the camera needs to be on even though it does not record visuals so that the emotion detection could run. This is a must if this feature would to be implemented and cannot be hidden in any way from the participants because of technical implementations but also ethical reasons.

A high number of responses from participants were categorized as Negative opinion due to their concerns regarding security and privacy, affirming that they are feeling like they are being watched and followed or that they do not trust that the camera is not recording anything. Options to overcome these issues need to be found. Some changes that could influence how people perceive security and privacy would be to add more clear and easy-to-understand instructions so that people could find out how everything works underneath and details about the platform. A more clear and more formal way of asking for consent from participants could also help people be more trusting.

6.4.4 Inaccessibility

Even though such an application would benefit from a trained model that can detect facial expressions surprisingly accurately, it will still be inaccessible for people who are neuro-diverse, for example, whom is autistic or has Tourette's syndrome or people who have some form of facial paralysis.

It is a challenge to train a model to handle all use cases that might appear due to these kinds of medical disorders. When referring to semi-facial paralysis, the model should be able to recognize and accurately identify facial expressions by being aware of the disorder. In the case of neuro-diverse people, the trained model should be aware that their facial expressions might differ from others or they could have twitches which should not be counted as facial expressions or emotion changing.

6.5 Future research

The sensitivity and frequency with which the chatbot responds to users' facial expression changes need to be precise. Responding too aggressively to small mood changes could weaken the image of the chatbot and its trustworthiness. Small or frequent mood changes should be pointed out more subtly in the chatbot's replies, in order for the discussion to feel as human-like as possible (Ghandeharioun et al., 2019). Future research should better account for these details that can easily make the difference between a chatbot that is perceived as empathetic and one that is perceived as invasive.

Furthermore, in human-to-human interactions, people's replies are affected by the tone and content of messages received from their transmitter (Soderlund et al., 2021). Future studies could further research if the replies sent by the chatbot based on the changes in facial expressions influence the way people respond and report to the overall discussion.

7. Conclusion

In this paper, a mix of qualitative and quantitative methods was used in order to discover how different levels of emotional awareness in a chatbot impact users' engagement during a conversation. With this question in mind, the study started with the hypothesis the *"The awareness that the chatbot shows during the conversation impacts the engagement of participants"*.

The data collected from a total number of 178 participants was split into three conditions each receiving different experiences while chatting with a chatbot. The control group was asked to discuss an AI-driven chatbot that did not have any emotional recognition features, the second condition received visual feedback in the shape of an emoji that let participants know what expression the chatbot registers, and lastly the third condition, besides the feature from the second condition, also benefited from a chatbot that would reply with a certain message whenever their mood changed.

The quantitative data collected showed that there are no significant differences between the chatbots used in the different conditions, thus based on the result of this data only, claims cannot be made which would approve that the awareness of chatbots impacts the engagement level of participants.

However, the results from qualitative data showed that people consider the emotionally aware chatbot made them more likely to discuss different more personal subjects, open up more, feel more listened to, and like talking to a human being, which increased their engagement level in the conversation.

Flaws and room for improvement were also discovered and pointed out by participants in the open-ended questions. These obstacles were preventing some of the participants from claiming that they would use this chatbot experience in the future as well.

In the end, we can say that the hypothesis from which this research started can be approved. The awareness of a chatbot impacts the engagement level of participants. However, it cannot be stated whether it impacts it more positively or negatively. In order to be able to say that, it is required that such a chatbot would be developed and iterated forward.

In conclusion, the concept of using emotionally aware chatbots should be investigated foreword in more depth. Such chatbots not only impact the engagement level of participants during the conversation but also impact the overall experience of discussing with a chatbot. Future research should focus on improving the skills of the chatbot to integrate emotion recognition into the discussion as naturally as possible, but also on the security and privacy of the users.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. Machine Learning with Applications, 2, 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. Telematics and Informatics, 54, 101473. https://doi.org/10.1016/j.tele.2020.101473

Beattie, A., Edwards, A. P., & Edwards, C. (2020). A Bot and a Smile: Interpersonal Impressions of Chatbots and Humans Using Emoji in Computer-mediated Communication. Communication Studies, 71(3), 409–427. https://doi.org/10.1080/10510974.2020.1725082

Boutet, I., LeBlanc, M., Chamberland, J. A., & Collin, C. A. (2021). Emojis influence emotional communication, social attributions, and information processing. Computers in Human Behavior, 119, 106722. https://doi.org/10.1016/j.chb.2021.106722

Brandtzaeg, P. B., & Følstad, A. (2017). Why People Use Chatbots [Series Title: Lecture Notes in Computer Science]. In I. Kompatsiaris, J. Cave, A. Satsiou, G. Carle, A. Passani, E. Kontopoulos, S. Diplaris, & D. McMillan (Eds.), Internet Science (pp. 377–392). Springer International Publishing. https:// doi.org/10.1007/978-3-319-70284-1_30

Cousins, N. (1979). Anatomy of an illness as perceived by the patient: Reflections on healing and regeneration [OCLC: 27708722]. Bantam.

Croes, E. A. J., & Antheunis, M. L. (2021). 36 Questions to Loving a Chatbot: Are People Willing to Self-disclose to a Chatbot? [Series Title: Lecture Notes in Computer Science]. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin, & P. B. Brandtzaeg (Eds.), Chatbot Research and Design (pp. 81–95). Springer International Publishing. https://doi.org/ 10.1007/978-3-030-68288-0_6

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149–1160. https://doi.org/10.3758/BRM.41.4. 1149

Gergle, D., & Tan, D. S. (2014). Experimental Research in HCI. In J. S. Olson & W. A. Kellogg (Eds.), Ways of Knowing in HCI (pp. 191–227). Springer New York. https://doi.org/10.1007/978-1-4939-0378-8_9

Ghandeharioun, A., McDuff, D., Czerwinski, M., & Rowan, K. (2019). Towards Understanding Emotional Intelligence for Behavior Change Chatbots. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), 8–14. https://doi.org/10.1109/ACII.2019.8925433

Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 929–932. https://doi.org/10.1145/1240624.1240764

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. Retrieved June 15, 2023, from http://www.jstor.org/stable/25148625

James, W. (1995). The principles of psychology: In two volumes. Vol. 2 (Facsim. of ed. New York, Henry Holt, 1890, Vol. 2). Dover.

Kleinke, C. L., Peterson, T. R., & Rutledge, T. R. (1998). Effects of self-generated facial expressions on mood. Journal of Personality and Social Psychology, 74(1), 272–279. https://doi.org/10.1037/0022-3514.74.1.272

Krämer, N., Kopp, S., Becker-Asano, C., & Sommer, N. (2013). Smile and the world will smile with you—The effects of a virtual agent‘s smile on users' evaluation and behavior. International Journal of Human-Computer Studies, 71(3), 335–349. https://doi.org/10.1016/j.ijhcs.2012.09.006

Liu, M., Wong, A., Pudipeddi, R., Hou, B., Wang, D., & Hsieh, G. (2018). ReactionBot: Exploring the Effects of Expression-Triggered Emoji in Text Messages. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1–16. https://doi.org/10.1145/3274379

McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa ... examine the variables influencing the use of artificial intelligent in-home voice assistants. Computers in Human Behavior, 99, 28–37. https://doi.org/10.1016/j.chb.2019.05.009

Melo, C., Carnevale, P., & Gratch, J. (2011). The effect of expression of anger and happiness in computer agents on negotiations with humans. 10th International Conference on Autonomous Agents and Multiagent Systems 2011, AAMAS 2011, 2, 937–944.

Mohamad Suhaili, S., Salim, N., & Jambli, M. N. (2021). Service chatbots: A systematic review. Expert Systems with Applications, 184, 115461. https://doi. org/10.1016/j.eswa.2021.115461

Moilanen, J., Visuri, A., Suryanarayana, S. A., Alorwu, A., Yatani, K., & Hosio, S. (2022). Measuring the Effect of Mental Health Chatbot Personality on User Engagement. Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia, 138–150. https://doi.org/10.1145/3568444.3568464

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. International Journal of Human-Computer Studies, 112, 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

Patten, M. L., & Newhart, M. (2018). Understanding research methods: An overview of the essentials (Tenth edition). Routledge, Taylor & Francis Group.

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. Computer Applications in Engineering Education, 28(6), 1549–1565. https://doi.org/10.1002/cae. 22326

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the Number of Categories in Agree–Disagree Scales. Sociological Methods & Research, 43(1), 73–97. https://doi.org/10.1177/0049124113509605

Salovey, P., & Mayer, J. D. (1990). Emotional Intelligence. Imagination, Cognition and Personality, 9(3), 185–211. https://doi.org/10.2190/DUGG-P24E52WK-6CDG

Sidner, C. L., Lee, C., Morency, L.-P., & Forlines, C. (2006). The effect of headnod recognition in human-robot conversation. Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, 290–296. https://doi.org/10.1145/1121241.1121291

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. Computers & Education, 151, 103862. https://doi.org/10.1016/j.compedu.2020.103862

Soderlund, M., Oikarinen, E.-L., & Tan, T. M. (2021). The happy virtual agent and its impact on the human customer in the service encounter. Journal of Retailing and Consumer Services, 59, 102401. https://doi.org/10.1016/j. jretconser.2020.102401

Swinkels, A., & Giuliano, T. A. (1995). The Measurement and Conceptualization of Mood Awareness: Monitoring and Labeling One’s Mood States. Personality and Social Psychology Bulletin, 21(9), 934–949. https://doi.org/10.1177/0146167295219008

Thyer, B. A. (2012). Quasi-experimental research designs [OCLC: ocn753468473]. Oxford University Press.

Tomkins, S. S. (1962). Affect, Imagery, Consciousness: Vol. 1. The Positive Affects. Springer.

Tomkins, S. S. (1963). Affect, imagery, consciousness: II. The Negative Affects. Springer.

Wall, H. J., Kaye, L. K., & Malone, S. A. (2016). An exploration of psychological factors on emoticon usage and implications for judgement accuracy. Computers in Human Behavior, 62, 70–78. https://doi.org/10.1016/j.chb.2016.03.040

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. International Journal of Research in Marketing, 27(3), 236– 247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Wildemuth, B. M. (Ed.). (2017). Applications of social research methods to questions in information and library science (Second edition). Libraries Unlimited.

Zhou, Q., Li, B., Han, L., & Jou, M. (2023). Talking to a bot or a wall? How chatbots vs. human agents affect anticipated communication quality. Computers in Human Behavior, 143, 107674. https://doi.org/10.1016/j.chb.2023.107674

Zhu, Q., Chau, A., Cohn, M., Liang, K.-H., Wang, H.-C., Zellou, G., & Yu, Z. (2022). Effects of Emotional Expressiveness on Voice Chatbot Interactions. 4th Conference on Conversational User Interfaces, 1–11. https://doi.org/10.1145/3543829.3543840