university of groningen / faculty of science and engineering / artificial intelligence

Artificial friends: A ChatGPT implementation in AlphaMini companionship robots

Bachelor's Project Thesis

Alexia Spinei, s4312112, a.spinei@student.rug.nl,
Supervisor: Prof Dr Paul Vogt

Abstract: Addressing the rise in loneliness, especially among emerging and older adults, this study investigates the potential of social robots as companions by enhancing their conversational abilities with ChatGPT-3.5, complemented by sentiment analysis. Utilizing the AlphaMini robot, equipped with Whisper for speech transcription and Amazon Comprehend for sentiment analysis, the research aims to improve the naturalness and empathy of robot-human interactions. Through an experiment involving 22 participants, the study compares the effectiveness of standard ChatGPT responses versus those augmented by Amazon Comprehend. The results indicate no significant difference in participants' perception of anthropomorphism, likeability, or general and emotional intelligence of the robot between groups. This suggests that ChatGPT-3.5 may already offer a level of empathetic engagement sufficient for the context of social robotics or that there is a potential need for advancements in complementary areas, such as integrating contextually relevant gestures, to enhance the overall interaction quality and user experience.

1 Introduction

In recent years, there has been an increase in feelings of loneliness across various age groups, especially emerging adults, according to Buecker et al. (2021). In response to this societal concern, technology has sought innovative ways to counteract this trend. Social robots could play an instrumental role in addressing this issue by fulfilling our compelling need for connection, understanding, and companionship. Studies like the one by Kühne et al. (2022) highlight the potential benefits of social robots. According to Kühne, although social robots cannot fully replace human interaction, they have a great potential to alleviate feelings of loneliness.

However, the seamless integration of social robots into our daily lives is still in its beginning stages. They are currently in their developmental stage and frequently face challenges in accurately mimicking human speech, behaviors, and gestures, and they lack the empathy required for forming connections. A recent study by de Graaf et al. (2019) concluded that most people remain hesitant, even uneasy, about engaging with or being around social robots. This absence of naturalness can lead to undermining people's trust in them, as stated by Mara et al. (2022). Different researchers are testing ways to bridge the gap between human expectations and robot-human conversational capabilities. While these challenges persist, leading to widespread reluctance to accept social robots, innovative research efforts continue to explore solutions that might enhance their acceptability and functionality. Irfan et al. (2023)'s pioneering study offers a peek into a new approach where language models, specifically ChatGPT-3.5 (OpenAI), are integrated into companion robots. This study aims to develop guidelines for integrating Large Language Models (LLM) into companion robots for older adults by exploring and testing a social robot, Furhat, for this demographic. A critical insight from Irfan's study highlighted participants' perceptions of interactions with Furhat as frequently being superficial and lacking depth, with a notable "69.87% of conversations categorized as 'informal/superficial'".

Another significant study in the field of social robotics is the one by Khoo et al. (2023), in which custom software was implemented in the commercially available QTrobot(QT) to explore how robots

could play an instrumental role in bolstering the well-being of older adults by facilitating meaningful dialogues. After having an open-ended dialogue with QT, participant feedback underlined limitations in its communicative abilities, with remarks like "QT is very slow and rudimentary in its communication and resources." (Section 3.2, Evaluation with Survey and Observation). This resonates with Irfan's findings, emphasizing the importance of fluid and personalized interactions, especially when discussing topics around well-being.

The latter two studies highlight a capability that is sometimes absent, yet crucial, in LLM-equipped robots for achieving authentic, human-like communication: the discernment of a speaker's emotional subtleties within the dialogue. This deficiency is one factor contributing to social robots' superficial and mechanically driven feel. A practical example of this challenge is presented in Khoo's study, in which there was a notable misalignment between the participant's statement, 'Yes, and I have a sister as well', and the robot's inappropriate response, 'I am sorry to hear that."—a reply out of sync with the speaker's neutral emotional state (Section 3.2, Evaluation with Survey and Observation). Similar problems appeared in Irfan's research, which could be due to ChatGPT's occasional inability to gauge the underlying emotion of the user, be it happiness, sadness, or neutrality. Drawing from these challenges, it can be inferred that these slight misalignments in response can significantly impact the user's sense of connection and understanding with the robot. These situations underline the need for robots to understand the literal content of human speech and grasp the emotional subtleties it conveys. Therefore, empathy proves to be an essential quality in social robotics, enabling robots to engage in genuinely meaningful interactions.

Empathy represents a fundamental human attribute that makes conversations meaningful, enabling nuanced and enriched interpersonal exchanges. Hence, as Tapus & Mataric (2007)'s study suggested, it is imperative to incorporate this element in the design of social robots intended for companionship and human-robot interaction (HRI). Building on this foundation, Park & Whang (2022) conducted a systematic review underlining this idea and defined empathy, within the context of HRI, as the robot's capability to recognize the human's emotional state and thoughts and use them to produce affective responses. Park & Whang (2022)'s comprehensive study highlights the possibility of emulating empathy in social robots to authentically replicate human interaction while providing a framework for designing empathic robots. Therefore, empathy proves to be an essential trait that social robots need to possess to fulfill their roles as companions and interactive agents in a manner that authentically replicates human interaction and aligns with social norms.

Many methods have been proposed for emotion detection, which is crucial for enabling robots to understand and respond to human emotions effectively. Feldman (2013)'s research provided a comprehensive overview of the applications and challenges of sentiment analysis, highlighting its importance in computer science and its critical role in enriching social robots' interactional systems. Furthermore, Kim & Hovy (2004) introduced a novel system aimed at identifying sentiments by automatically finding the people who hold opinions about a topic and assessing the sentiment of each opinion. This system, which includes modules for determining word sentiment and combining sentiments within a sentence, represents a significant advance in the robot's ability to process and interpret complex emotional cues in human language. The innovative approach of classifying and combining sentiment at both word and sentence levels, as detailed by Kim-soo, shows promising results for social robots' capability to engage in more nuanced and empathetic interactions. By integrating advanced sentiment analysis techniques, social robots can better understand and adapt to their human counterparts' emotional states and needs, paving the way for more meaningful and supportive human-robot interactions.

Until social robots enhance their capability for emotion detection, exploring the supplementation through integrating a sentiment-analysis system warrants consideration. This brings us to my research question: How does incorporating a text-based sentiment analysis tool affect human assessments of empathy and emotional intelligence in conversations with ChatGPT-equipped social robots? My hypothesis is that the additional sentiment analysis enhances the human perception of conversations with ChatGPT-equipped social robots. This study explores enhancing the naturalness of interactions with ChatGPT-integrated

robots. Within social robotics, we will explore this by utilizing AlphaMini robots, seen in Figure 1.1. AlphaMinis, developed by UBTECH Robotics, are compact, humanoid robots designed to interact with users engagingly and intuitively. In this experiment, these robots feature a tailored integration of ChatGPT, complemented by a text-based sentiment analysis tool. This tool will assist ChatGPT in refining responses by providing additional prompts on the underlying emotion detected in the user's dialogue. ChatGPT-3.5 was specifically chosen as the LLM for this research due to its faster response times than other iterations.

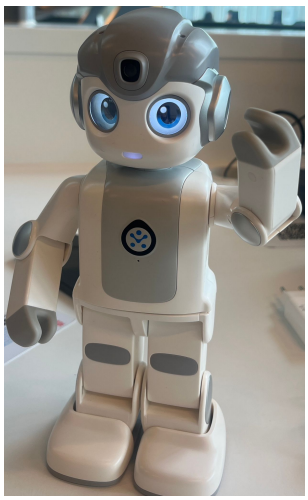

Figure 1.1: UBTECH AlphaMini robot used in this study, photographed by Alexia Spinei

2 Methods

Model

The model for this experiment centers on human-robot interaction facilitated by external software, which plays a critical role in controlling the AlphaMini robot. This software manages the robot and orchestrates communication with cloud-based services. The participants communicate exclusively with the robot, using its capability to speak by using text-to-speech and its sound recording features. The external software processes the sound recorded by the robot, extracts the dialogue with the help of a speech-to-text tool, and performs sentiment analysis depending on the experiment condition. Two conditions were used for these experiments. In condition 1, participants interact with AlphaMini robots equipped solely with ChatGPT, focusing on standard conversational capabilities. In this case, as shown in Figure 2.1, the extracted text from the speech-to-text tool is then sent to ChatGPT to obtain a reply, and the reply is then sent to the robot and articulated through its own speech-to-text system. Condition 2 adds to this arrangement Amazon Comprehend for real-time sentiment analysis by injecting the detected user sentiment for each message exchange. In condition 2, the extracted text and the sentiment cue are then used to create a response from ChatGPT-3.5, as shown in Figure 2.2. The response is then articulated to the participants through the robot's text-to-speech system, simulating a conversational exchange.

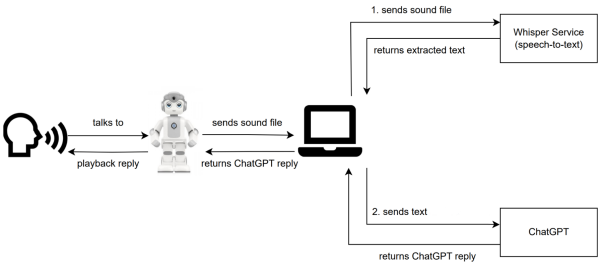

Figure 2.1: The process through which a reply is generated for the participant, in the condition of the purely ChatGPT-equipped robot.

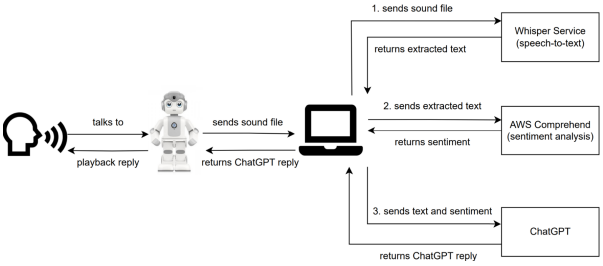

Figure 2.2: The process through which a reply is generated for the participant, in the condition of the purely ChatGPT-equipped robot.

The AlphaMini robots were supplemented with a speech-to-text and ready-to-use tool, Whisper V3 (OpenAI), used for speech transcription. This choice was grounded in its robust speech recognition capabilities, as detailed by Radford et al. (2023). This tool provides high accuracy for the

English language, which is essential for this experiment's linguistic needs. Integrated seamlessly into this setup, Whisper V3 played a crucial role in transcribing participant-robot interactions, adhering to strict privacy and ethical standards.

Amazon Comprehend, a service provided by Amazon Web Services (AWS), is a ready-to-use service that uses natural language processing (NLP) to uncover insights and relationships in a text (*Amazon Comprehend Documentation* (2023)) and was selected as the sentiment analysis tool for this research. As Romeo (2020) 's research highlighted, Amazon Comprehend outperformed the accuracy of leading services like Google Cloud NLP API (Google) and Meaning Cloud (MeaningCloud LLC). It utilizes machine learning models to analyze and comprehend documents, identifying elements such as language, entities, key phrases, and sentiment. Amazon Comprehend enables developers to integrate a nuanced understanding of textual content into various applications, enhancing the interaction quality of chatbots, social robots, and other communicative technologies by providing the ability to analyze text for positive, negative, or neutral sentiment. Therefore, it is a powerful instrument in assisting LLMs and, indirectly, social robots to engage in more natural and empathetic dialogues, potentially enabling them to go beyond surface-level discussions in the future. These features align with this study's needs, facilitating ChatGPT to respond with a more nuanced understanding of the user's emotional context.

ChatGPT-3.5 API, as described in *OpenAI API Documentation* (2023), was the primary conversational agent responsible for generating responses during user interactions. This particular LLM was selected for its widespread recognition and significant research foundation supporting it. The decision to employ ChatGPT-3.5 over the newer ChatGPT-4 was due to its low latency in generating responses, bringing it closer to the speed of a human conversation. The effectiveness of ChatGPT-3.5 is primarily attributed to its use of the Transformer algorithm, a neural network architecture optimized for natural language processing, as extensively discussed in Team (2017). The Transformer's unique 'attention' mechanism enables the model to contextually weigh the significance of words in a sentence, enhancing its capability to produce coherent and contextually relevant text.

The AlphaMini robot, a key component in this model, is utilized for its audio recording and speech synthesis capabilities. Throughout the experiment, the AlphaMini remains in a fixed standing position, serving as the physical medium for this interactive and responsive communication model. The decision to have the robot in a fixed position was made for several reasons. Firstly, the topic of gesturing within robotic interactions represents a distinct and extensive field of research, and it would introduce significant complexity to this study. Secondly, given that the robot operates without a predefined script, its gestures must align with its live, input-driven replies. Such real-time gestural adaptations would complicate the experimental design and exceed the study's primary objective.

Dialogue initiation with ChatGPT begins with an initial system prompt*, instructing it to behave as a friendly robot, respond concisely (one or two short sentences), use a familiar tone, and maintain engagement by asking questions. The conversation starts with the AlphaMini greeting the user with a pre-set greeting †. After the user responds to the greeting, the dialogue with ChatGPT begins. The initial system prompt for ChatGPT advises it to ask about the user's family. The initial discussion topic is strategically chosen to elicit various sentiments and alleviate participant nervousness regarding topic selection. The participant's response text is integrated into the ongoing prompt as a user message. Depending on the experimental conditions, an additional system prompt ‡ indicating the user's sentiment is also incorporated. The entire conversation, including ChatGPT's replies, marked as assistant messages, is sent to ChatGPT for further response generation. The prompt is regularly trimmed to include only the initial system message and the last 10 messages (user and assistant combined) to manage latency and adhere to token length limitations. This approach ensures a fluid and coherent conversation flow within the system's technical limit.

*The system prompt: *Always say that you are a friendly robot. Always reply in a maximum of one or two short sentences, like you are having a short dialogue with a friend or acquaintance. Use a familiar tone. You can ask questions to keep the conversation alive. Initiate the conversation by asking about the user's family.*

†Initial greeting: *Nice to meet you! Let's chat.*

‡Sentiment prompt example: *The user's sentiment is: HAPPY. Please reply accordingly.*

A straightforward mechanism was employed to facilitate turn-taking between the robot and participants. Upon completing their reply, participants were instructed to press any key on the laptop's keyboard on which the control software was running. This method was chosen because the robot's software does not offer a reliable dialogue pause detection mechanism. Text is displayed in the virtual environment whenever the robot processes a response. This helps manage the occasional delays in the robot's reply formulation and enhances the interaction's smoothness and predictability.

Experiment

Participants

The experiment involved 22 participants, 13 males and 9 females. The two experimental conditions were evenly distributed to both genders, with 11 participants each in each group. All participants were students in the 20 to 29-year-old age bracket.

Conditions

To test the hypothesis, this experiment involves two distinct conditions for the AlphaMinis: condition 1, with conversational ChatGPT, and condition 2 which implements additional sentiment analysis. This addition allows ChatGPT to adapt its responses based on the emotional context of the interaction. The purpose of these different conditions is to evaluate the impact of sentiment analysis on user experience and interaction quality. Participants were randomly allocated to either condition, ensuring a 50-50 split between conditions and across genders. This method aimed to equalize male and female representation in each group, minimizing gender-related confounding variables.

Materials

For this experiment, the materials included a quiet setting with a table on which an AlphaMini robot was facing the participant and a computer to control the experiment and record data. The AlphaMini robot was used rather than just the computer because it has an anthropomorphic shape and creates a better connection with the participant than a laptop. The consent form provided participants with an overview of the study's aims and procedures detailed the interaction with the robot, and explained the feedback process. It assured confidentiality and ethical data handling, specifying no conversation recordings and naming data processing tools (Whisper and Amazon Comprehend) that do not store conversations. Participants were informed of their right to withdraw anytime, emphasizing ethical compliance and awareness.

After ending the dialogue, the participants receive the digital feedback questionnaire. Its structure is adapted from the Godspeed robot questionnaire Bartneck (2008) to assess the participants' experience. This experiment's questionnaire is divided into four sections: "Anthropomorphism," "Likeability," "Perceived Intelligence." and "Emotional Intelligence". The first three sections comprised five scales to gather the participants' feelings toward the specific attribute. These scales were designed as five-point differential scales, offering a range of responses for participants to express their perceptions. For instance, in the "Likeability" section, one of the scales ranged from 'Dislike' to 'Like'. Participants selected one of five bullet points on each scale to indicate their level of agreement or feeling about the robot concerning the specific attribute. The questionnaire retained the original scales for "Likability", "Perceived intelligence", and "Anthropomorphism", while "Animacy" and "Perceived safety" were excluded. "Animacy" was excluded since the robot's only gesture was the opening waving gesture. The "Perceived safety" scale was also excluded, as it was irrelevant for the AlphaMini robot. This robot, designed for social interaction, is inherently harmless and immobile during dialogues and when turned off, rendering safety concerns inapplicable.

In addition to the three sections, this study introduces a new section titled "Emotional Intelligence." This section is inspired by a series of questions outlined in Wang et al. (2019), which is aimed to provide participants' views on various aspects of robot performance, usability, and interaction quality. This section allows us to more precisely assess whether participants feel a stronger connection and are more comfortable with the robot enhanced with sentiment analysis compared to one without. This section aims to capture the participants' perceptions of the robot's emotional sensitivity and contextual awareness, enriching the assessment framework with a focus on the robot's interpersonal in-

teraction capabilities.

Procedure

Upon entering the experiment room, participants are guided to the computer-equipped workstation where the experiment takes place. They are then provided with an informed consent form.

Next, the participants are shown the virtual environment used for the turn-taking in the dialogue with the AlphaMini and how to start the conversation by pressing a green arrow at the top right corner of the page. The AlphaMini, in stand-down mode, is placed in front of them, and then they are free to start the conversation whenever ready. After pressing the green arrow at the top right corner, the robot begins the conversation by standing up and says: "Nice to meet you! Let's chat!" while waving with the right arm. The participants will see the text "You can now speak to the robot. Please press any key when you are done speaking." in the terminal, which denotes that the robot will now only listen to them without interrupting. The participants press any key after they are done speaking to continue the dialogue. After pressing any key, the terminal will show the text "replying...", which denotes that the robot's reply will follow. The conversation continues the conversation for as long as they like, without minimum or maximum time limit, and they can end the conversation by saying only "bye" or "bye-bye" or "goodbye" to the robot. They are assured they can also end the conversation by asking for my help. After the conversation ended, the participants were asked to complete the digital feedback questionnaire and were free to leave the experiment room.

Analysis of Data

In this study, the Wilcoxon Rank-Sum test was selected for statistical analysis to compare the responses of two independent groups interacting with ChatGPT, with and without the integration of sentiment analysis via Amazon Comprehend. This choice was made because of the ordinal nature of the data, collected through 5-point Likert-scale questionnaires assessing Anthropomorphism, Likeability, Perceived Intelligence, and Emotional Intelligence.

Given the small sample size in this study of 11 participants per group, traditional tests for normality and homogeneity of variance, such as the Shapiro-Wilk test, often lack sufficient power and can be misleading. Consequently, I did not perform these tests as they could yield unreliable results. A traditional parametric test like the t-test can also be inappropriate due to its reliance on assumptions of normal data distribution and variance homogeneity. Instead, I opted for the Wilcoxon rank sum test, a non-parametric method that does not require the assumptions of normality or equal variances, making it a more suitable and robust choice for this analysis. It assesses differences in the central tendency between groups by comparing ranks, thereby aligning with this study's objective of evaluating the impact of sentiment analysis enhancement on user perceptions of the interaction with the AlphaMini. This method ensures the robustness and validity of these findings despite the small sample size and the ordinal nature of the data.

3 Results

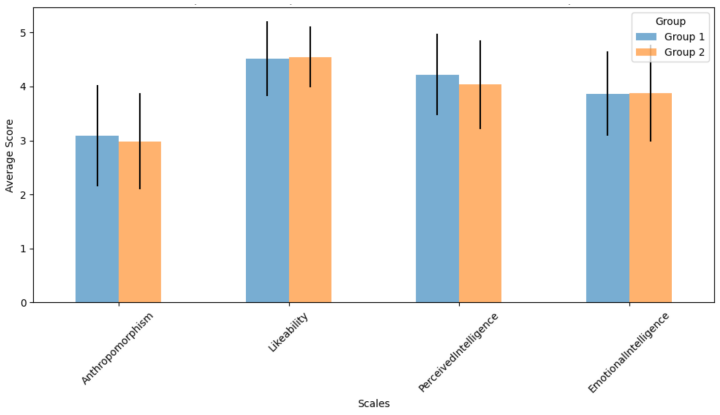

Figure 3.1: Bar chart comparing average scores on the Godspeed Questionnaire scales between Group 1 (blue bars) and Group 2 (orange bars). Error bars represent the standard deviation of the average scores for each group.

The differences between the two independent groups for each scale were analyzed using the Wilcoxon rank sum test. The test compared responses from n1=11 participants in Group 1 (the non-sentiment analysis group) and n2=11 partici-

pants in Group 2 (the sentiment analysis group). The results indicated no statistically significant difference in scores for Anthropomorphism (U = 1606.5, p = .558), Likeability (U = 1470, p = .770), Perceived Intelligence (U = 1681.5, p = .280), or Emotional Intelligence (U = 2124.5, p = .798). Given that the p-values exceed the conventional alpha threshold of 0.05, we fail to reject the null hypothesis. This suggests that there is no statistically significant difference in the median scores between the two groups across the measured constructs. The sentiment analysis feature did not significantly impact participants' perceptions of the robot in these domains.

In Figure 3.1, a bar chart compares the average result scores across four scales—Anthropomorphism, Likeability, Perceived Intelligence, and Emotional Intelligence—between Group 1 and Group 2. On the Anthropomorphism scale, both groups presented similarly, with Group 1 showing a slightly higher average score than Group 2. The Likeability scale shows similar scores, with Group 1 marginally outscoring Group 2, suggesting a slightly more favorable perception of the robot without sentiment analysis. In the Perceived Intelligence category, the scores are nearly identical for both groups, indicating that the addition of sentiment analysis had little to no perceptible effect on how participants rated the robot's intelligence. Finally, the Emotional Intelligence scale shows almost no difference between the two groups. Figure 3.1 visually supplements these statistical results by illustrating the associated variability within each group. The error bars, denoting one standard deviation from the mean, depict the spread of scores within the groups across each scale. For instance, the error bars for Anthropomorphism overlap between Group 1 (M = 4.2, SD = 0.76) and Group 2 (M = 3.9, SD = 0.81), which graphically represents the similarity in score distribution between the two groups, in alignment with the non-significant p-value. Similarly, the close proximity of error bars across the Likeability, Perceived Intelligence, and Emotional Intelligence scales further visualizes the statistical findings, suggesting comparable perceptions between groups.

4 Discussion

In this research, I sought to determine whether integrating a sentiment-analysis system into ChatGPT-equipped social robots would yield more empathetic responses. I hypothesized that the addition of sentiment analysis would improve the interaction quality between humans and social robots. I assessed the perceived quality of interaction using the feedback questionnaire to determine the effect of sentiment analysis on ChatGPT-generated responses. However, the results show that there is no meaningful difference between the perception of the social robot with and without added sentiment analysis.

A separate analysis comparing responses from pure ChatGPT and those enhanced by sentiment analysis would provide further insights into the core question of this study. This analysis would be important as it not only reveals the level of empathy in the ChatGPT responses but also, by correlation with the survey results, would allow us to assess how much participants value this additional empathetic aspect. However, assessing the difference in emotional speech generated by ChatGPT between groups isn't easy. Even though the initial prompt to ChatGPT encouraged a dialogue centered on family, each participant interpreted and responded differently to the initial prompt. The conversations organically diverged in various topics across participants due to their free-form nature. In free-form dialogues, the depth and nuance of the conversation can vary widely based on individual participant engagement and interest. This variability makes it challenging to standardize the responses for a direct comparison. Another issue is that emotional interpretation is highly subjective, and what one participant perceives as empathetic might not be perceived the same way by another. This subjectivity adds a layer of complexity to quantitatively assessing emotional enhancements.

There is a notable lack of standardized methodologies for measuring empathy in AI-generated text responses. This makes it challenging to compare the empathetic nuances of ChatGPT responses objectively. Developing such methods would require interdisciplinary research from psychology, linguistics, and computer science to establish reliable metrics for empathetic communication in diverse contexts. For now, the most reliable method for eval-

uating the effectiveness of sentiment analysis lies in analyzing the responses from the feedback questionnaire.

I have observed in the bar plots depicted in Figure 3.1 that there is almost no difference in how participants felt about their interaction with the social robot between the two groups for all four scales. Moreover, the Wilcoxon Rank Sum test also showed no significant difference in the results between the two groups. The average results displayed in the barplot for 'Likeability', 'Perceived Intelligence', and 'Emotional Intelligence' exceed 4 out of 5. These scores suggest that ChatGPT is, in itself, effective in providing empathetic responses. This deduction is confirmed by Schaaff et al. (2023) 's recent study exploring ChatGPT-3.5'S capabilities to produce empathetic responses and emotional expressions. Schaaff et al. (2023) study shows that "in 91.7% of the cases, ChatGPT was able to correctly identify emotions and produces appropriate answers." However, the scores of ChatGPT-3.5 are still worse than the average scores of healthy humans, according to the same research. In conclusion, ChatGPT appears already equipped to recognize and respond empathetically to conversational cues.

Finally, another reason for these high average scores could partly be attributed to the participants' motivation and enthusiasm about engaging with a robot, considering that all participants were volunteers who received no extrinsic rewards for their involvement in the study. This effect, often related to the novelty of the experience of interacting with robots, might have enhanced their perceptions of the robot's emotional and general intelligence. Additionally, volunteers might have exhibited a 'pleasing bias,' where they provided positive feedback to support the research, especially in studies involving innovative technologies such as this one.

Limitations

Several limitations were encountered during the experiment that could have influenced the participants' perception of the interaction. The average scores for both groups on the 'Anthropomorphism' scale were lower than those on the other scales. Several factors could explain the scores being lower than those of the other scales, although they were not directly examined in this study. First, not making use of AlphaMini's gestures in this experiment might hinder its ability to convey emotions and intentions naturally, making its interactions seem mechanical. Additionally, AlphaMini's synthetic voice further reduces the robot's human-like qualities, as these voices often lack the expressiveness found in human speech, such as tone, pitch, and rhythm variations. Thirdly, the necessity of using a PC for managing turn-taking disrupts the conversational flow, making it less smooth and highlighting the robot's dependency on external systems rather than autonomous social skills. The latency in the robot's responses can also significantly detract from the interactions' quality. Delays in response times can disrupt the natural flow of conversation, potentially leading to frustration or disengagement among participants. Such disruptions can undermine the robot's responsiveness, thus negatively influencing participants' scores on the 'Anthropomorphism' scale. This delay could cause participants to view the robot as less lifelike or relatable, further impacting their overall experience and assessment of the interaction.

Another notable limitation of this study is the relatively small sample size, which consisted of 22 participants. This limits the generalizability of my findings and may not adequately represent the broader population. Furthermore, the narrow age range of the participants, who were all between the ages of 20 and 29 and identified as students, restricts the applicability of the results to this specific demographic. The homogeneity of this sample in terms of age and educational status means that the insights gained may not be applicable to older adults, individuals with different academic backgrounds, or those with varying levels of technological proficiency.

Future Research

Exploring social robots with enhanced computational capabilities could offer valuable insights in future research. Advanced models with more robust processing power may overcome current limitations, such as the notable response delay experienced with the AlphaMini robot during this study, which may have affected user perceptions. Moreover, developing more advanced robots capable of detecting pauses in speech would greatly facilitate

turn-taking in dialogue, making interactions with robots feel more natural and fluid. By understanding and responding to the natural rhythm of human conversation, these advanced robots could provide a more engaging and effective user experience. The tools used for this research, such as the LLM and the sentiment analysis, could also be accessed directly from the social robot, eliminating the need for a secondary computer device such as the laptop. In a more distant future, it is plausible that LLMs will be able to run entirely in the local robot computing environment.

Future research should consider pairing social robots equipped with LLMs with diverse sentiment analysis tools to determine the most effective combinations. To accomplish this, thorough comparative studies are needed, examining various LLMs and sentiment analysis systems across multiple dimensions. These dimensions should include accuracy, processing speed, and the capacity to parse and react to complex emotional cues. Defining these evaluation criteria clearly is essential for advancing human-robot interaction toward more fluid and lifelike exchanges. By conducting such assessments, we can better understand how to equip social robots with technologies that enhance their engagement and mimic natural human communication more closely.

Conclusion

The study found that social robots equipped with ChatGPT, whether or not they were enhanced with sentiment analysis, received similarly high ratings from participants in terms of 'Likeability', 'Perceived Intelligence', and 'Emotional Intelligence'. However, scores for 'Anthropomorphism' were lower, indicating room for enhancement in making these robots appear more human-like. This suggests a particular focus on refining elements such as the robot's voice and gestures can be beneficial.

Sentiment analysis did not improve human evaluations of empathy and emotional intelligence in interactions involving ChatGPT-equipped social robots. The relatively high average scores indicate that ChatGPT-3.5 is already capable of generating emotionally appropriate responses. This conclusion is grounded in the results from the feedback questionnaire, which indicated no significant differences in participant perceptions of their interactions with the ChatGPT-equipped AlphaMini robot.

References

Amazon comprehend documentation. (2023). Retrieved from `https://docs.aws.amazon.com/comprehend/` (Retrieved March 14, 2024, from Amazon Web Services)

Bartneck, C. (2008). *The godspeed questionnaire series.* `https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/`. (Accessed: 2024)

Buecker, S., Mund, M., Chwastek, S., Sostmann, M., & Luhmann, M. (2021). Is loneliness in emerging adults increasing over time? a preregistered cross-temporal meta-analysis and systematic review. *Psychological Bulletin*, *147*(8), 787.

de Graaf, M. M., Ben Allouch, S., & Van Dijk, J. A. (2019). Why would i use this in my home? a model of domestic social robot acceptance. *Human–Computer Interaction*, *34*(2), 115–173.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82–89.

Irfan, B., Kuoppamäki, S.-M., & Skantze, G. (2023). Between reality and delusion: Challenges of applying large language models to companion robots for open-domain dialogues with older adults.

Khoo, W., Hsu, L.-J., Amon, K. J., Chakilam, P. V., Chen, W.-C., Kaufman, Z., ... others (2023). Spill the tea: When robot conversation agents support well-being for older adults. In *Companion of the 2023 acm/ieee international conference on human-robot interaction* (pp. 178–182).

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics* (pp. 1367–1373).

Kühne, K., Jeglinski-Mende, M. A., Fischer, M. H., & Zhou, Y. (2022). Social robot–jack of all trades? *Paladyn, Journal of Behavioral Robotics*, *13*(1), 1–22.

Mara, M., Appel, M., & Gnambs, T. (2022). Human-like robots and the uncanny valley. *Zeitschrift für Psychologie*.

Openai api documentation. (2023). `https://platform.openai.com`. (Accessed: 2023)

Park, S., & Whang, M. (2022). Empathy in human–robot interaction: Designing for social robots. *International journal of environmental research and public health*, *19*(3), 1889.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Romeo, P. (2020). *Twitter sentiment analysis: a comparison of available techniques and services* (Unpublished doctoral dissertation). ETSI_Informatica.

Schaaff, K., Reinig, C., & Schlippe, T. (2023). Exploring chatgpt's empathic abilities. In *2023 11th international conference on affective computing and intelligent interaction (acii)* (pp. 1–8).

Tapus, A., & Mataric, M. J. (2007). Emulating empathy in socially assistive robotics. In *Aaai spring symposium: multidisciplinary collaboration for socially assistive robotics* (pp. 93–96).

Team, G. R. (2017). *Transformer: A novel neural network architecture for language understanding.* Google AI Blog. Retrieved Date Accessed, from `https://blog.research.google/2017/08/transformer-novel-neural-network.html` (Available online: `https://blog.research.google/2017/08/transformer-novel-neural-network.html`)

Wang, L., Iocchi, L., Marrella, A., & Nardi, D. (2019). Developing a questionnaire to evaluate customers' perception in the smart city robotic challenge. In *2019 28th ieee international conference on robot and human interactive communication (ro-man)* (pp. 1–6).

A Appendix

The rating sections provided in the feedback form are composed from the following 1-5 scales:

"Anthropomorphism":

Fake-Natural

Machinelike-Humanlike

Unconscious-Conscious

Artificial-Lifelike

Mechanical-Organic

"Likeability":

Dislike-Like

Unfriendly-Friendly

Unkind-Kind

Unpleasant-Pleasant

Awful-Nice

"Perceived Intelligence":

Incompetent-Competent

Ignorant-Knowledgeable

Irresponsible-Responsible

Unintelligent-Intelligent

Foolish-Sensible

"Emotional Intelligence"

Strange-Normal

Untrustworthy-Trustworthy

Unaware of my needs and emotions-Aware of my needs and emotions

Does not understand my feelings and responds empathically-Does understand my feelings and responds empathically

I don't find the robot easy to use and interact with-I find the robot easy to use and interact with

Doesn't understand the context of the conversation-Understands the context of the conversation