2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)

2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII) | 979-8-3503-2743-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACII59096.2023.10388208

Exploring ChatGPT's Empathic Abilities

Kristina Schaaff, Caroline Reinig, Tim Schlippe
IU International University of Applied Sciences
Erfurt, Germany
kristina.schaaff@iu.org; caroline.reinig@gmail.com; tim.schlippe@iu.org

***Abstract*—Empathy is often understood as the ability to share and understand another individual's state of mind or emotion. With the increasing use of chatbots in various domains, e.g., children seeking help with homework, individuals looking for medical advice, and people using the chatbot as a daily source of everyday companionship, the importance of empathy in human-computer interaction has become more apparent. Therefore, our study investigates the extent to which ChatGPT based on GPT-3.5 can exhibit empathetic responses and emotional expressions. We analyzed the following three aspects: (1) understanding and expressing emotions, (2) parallel emotional response, and (3) empathic personality. Thus, we not only evaluate ChatGPT on various empathy aspects and compare it with human behavior but also show a possible way to analyze the empathy of chatbots in general. Our results show, that in 91.7% of the cases, ChatGPT was able to correctly identify emotions and produces appropriate answers. In conversations, ChatGPT reacted with a parallel emotion in 70.7% of cases. The empathic capabilities of ChatGPT were evaluated using a set of five questionnaires covering different aspects of empathy. Even though the results show, that the scores of ChatGPT are still worse than the average of healthy humans, it scores better than people who have been diagnosed with Asperger syndrome / high-functioning autism.**

***Index Terms*—empathy, chatbot, ChatGPT, emotions**

I. INTRODUCTION

Chatbots have become a common tool in daily life over recent years [1]. These systems can simulate human-like conversations with users [2], provide assistance [3], information [4], and emotional support [5]. Among these, OpenAI's ChatGPT has emerged as one of the most widely used chatbots. According to [6], ChatGPT reached over one million users in less than a week after launch. Users include children seeking help with homework, people seeking medical advice and people who use it as a daily source of companionship.

Empathy is a human concept that is important in social interaction. Research shows that empathy prevents individuals from displaying aggression toward others [7] and helps to prevent children and adolescents from exhibiting antisocial behavior [8]. Also, in chatbots, empathy has shown to be a crucial trait in its acceptance by users [1].

[9] conducted the Short Dark Triad (SD-3) questionnaire [10] on the model GPT-3. The results showed a psychopathy score exceeding human results, indicating a lack of empathy. However, there is—to the best of our knowledge—no study focusing on empathy in ChatGPT, which is based on the newer model GPT-3.5. Consequently, this paper addresses this gap by exploring ChatGPT's empathic abilities.

Empathy is often understood as the ability to share and understand another individual's state of mind or emotion [11]. Therefore, as the first part of the study, we tested ChatGPT's ability to understand and express emotions. Afterwards, we analyzed ChatGPT's ability to express parallel emotional responses, a subcategory of empathy [12]. Finally, we evaluated ChatGPT's level of empathy in various aspects using psychologically acknowledged questionnaires.

To summarize, in this paper, we investigated ChatGPT's emotional and empathic abilities across three aspects: (1) understanding and expressing emotions, (2) parallel emotional response, and (3) empathic personality.

In the following section, we will describe related work regarding ChatGPT, the definition of empathy, and measuring empathy in chatbots. In Section III, we will explain how we evaluated ChatGPT's ability to understand and generate emotions. Section IV describes our experiments to assess ChatGPT's parallel emotional responding capabilities. Additionally, we investigated ChatGPT's empathic capabilities with standardized questionnaires to assess empathy, which we will show in Section V. We conclude our work in Section VI and suggest further steps.

II. RELATED WORK

A. *ChatGPT*

ChatGPT is a state-of-the-art chatbot developed by OpenAI that can produce natural language text when given a prompt or context [13]. This versatile tool can be employed in numerous fields, including education [13], medicine [14], and language translation [15].

The chatbot is based on the large language model GPT-3.5 and was fine-tuned using reinforcement learning from human feedback [16]. This approach allows the model to grasp the meaning and intention behind user queries, leading to relevant and helpful responses. To ensure safety and prevent the generation of inappropriate or factually incorrect text, the training of ChatGPT was enhanced by incorporating a large dataset of human-human and human-chatbot conversations. OpenAI has not released any official information about the exact amount of training data of ChatGPT, but the previous model GPT3 with 175 billion parameters was already significantly larger than other language models like BERT, RoBERTA, or T5 and was trained with 499 billion crawled tokens (i.e., subword units) [17]. By learning the intricacies and nuances of human language through this extensive dataset, ChatGPT can produce highly realistic text almost indistinguishable from human writing [18].

979-8-3503-2743-4/ 23/$31.00 ©2023 European Union

Authorized licensed use limited to: University of North Texas. Downloaded on February 09,2024 at 17:05:28 UTC from IEEE Xplore. Restrictions apply.

B. Definition of Empathy

Empathy is a crucial component of effective communication, especially in social interactions, as it allows humans to understand and share another person's feelings [19]. However, there is no consensus on the definition of empathy [20]. One possible definition is to distinguish between *cognitive* and *affective empathy* [20]. *Cognitive empathy* is the ability to understand and identify another individual's thoughts, feelings, and perspectives without necessarily experiencing the same emotions, i.e., the capability of mental *perspective taking*. It involves the capacity to recognize and interpret social cues, facial expressions, body language, and verbal communication to comprehend and infer the mental and emotional states of others [21]. *Affective empathy*, on the other hand, facilitates a deeper connection and understanding with others and involves a more visceral and personal connection to another's emotions [22]. *Affective empathy* can be divided into *parallel emotional response* and *reactive emotional response*. Parallel emotional responses involve responding with the same emotion as the other individual, while reactive emotional responses go beyond matching emotions, such as sympathy or compassion [12].

In our work, we evaluated ChatGPT's understanding and expression of emotions. Next, we analyzed its parallel emotional response. Finally, we covered the other aspects of empathy with standardized questionnaires to evaluate empathy.

C. Measuring Empathy in Chatbots

While standardized metrics exist to measure empathy in individuals, there are currently no standardized or valid methods for measuring empathy in chatbots [23]. A possible solution is to evaluate a chatbot's level of empathy by human evaluation, such as A/B tests or human ratings [24]. In A/B tests, the annotator chooses which response is more empathic, often used when comparing the level of empathy between two models. In human ratings, the annotator chooses the level of empathy based on a scale.

Another way is to conduct a feature- or system-level evaluation instead. The feature-level evaluation involves assessing each component and capability of a chatbot to provide an incremental understanding of its empathic behavior, e.g., by testing the chatbot on its level of emotional communication. On the other hand, the system-level evaluation focuses on measuring the chatbot's overall perception of empathy, e.g., by conducting self-assessment empathy tests [23].

In our work, we conduct a feature-level evaluation of ChatGPT's performance in showing parallel emotional responses, further explained in Sections III and IV. Furthermore, we perform system-level evaluations by conducting four standardized empathy tests and one autism test. Several studies have found that individuals with autism may have difficulty with the cognitive component of empathy, such as *perspective taking* and understanding others' mental states, while still being able to experience emotions and show affective empathy, such as feeling concerned or compassion for others [25].

III. UNDERSTANDING AND EXPRESSING EMOTIONS

To analyze ChatGPT's ability to understand and generate emotions, our first goal was to evaluate its proficiency in rephrasing neutral sentences to express a particular emotion.

A. Experimental Setup

For our analyses, we instructed ChatGPT to rephrase neutral sentences into six emotional sentences of the following categories: *joy*, *anger*, *fear*, *love*, *sadness*, and *surprise*. These emotions were selected following the basic emotions from the *Junto's Wheel of Emotions* [26] for consistency with the experiments described in Section IV, where we trained a classifier using the CARER dataset [27], which contains these categories. To check ChatGPT's ability to handle neutral sentences from different domains, we used 10 sentences from self-produced everyday sentences, Wikipedia, and Amendments to the United States Constitution. We instructed ChatGPT to rephrase each neutral sentence six times—each time with a different emotion category—resulting in a total of 60 emotional sentences. Figure 1 illustrates how ChatGPT rephrased the sentence ‘We are celebrating my grandmother's 80th birthday today.” to express *joy* and *anger*.

Original prompt:

> “We are celebrating my grandmother's 80th birthday today.”

Rephrased as *joy*:

> “Hooray! Today is my grandmother's 80th birthday, and we are absolutely thrilled to be celebrating!”

Rephrased as *anger*:

> “I can't believe we're supposed to be celebrating my grandmother's 80th birthday today - I am absolutely furious about it! There's so much to do, so many people to deal with, and it's just one more thing that I have to manage.”

Fig. 1: ChatGPT's Rephrasing of a Neutral Prompt into *Joy* and *Anger*.

B. Experiment and Results

To evaluate whether the prompts generated by ChatGPT match the intended emotion category, we asked three people to label each of the 60 ChatGPT-produced prompts with the most suitable emotion category out of the six categories provided. Based on the human labels, we produced the reference emotion categories using a majority voting as follows: If two annotators agreed on the same emotion for one prompt, we took this emotion as the final emotion category. In case our three annotators assigned completely different emotion categories to one prompt, they discussed their decisions until they agreed on one emotion category.

The annotator agreement on our six emotion categories is listed in Table I. A complete agreement occurred in 71.7% of the cases. In 26.6% of the prompts, exactly two annotators

Authorized licensed use limited to: University of North Texas. Downloaded on February 09,2024 at 17:05:28 UTC from IEEE Xplore. Restrictions apply.

Agreement	Percentage
all annotators agree	71.7%
two annotators agree	26,6%
all annotators disagree	1.7%

TABLE I: Annotator Agreement for the 6 Emotion Categories.

assigned the same emotion category. In only 1.7% of the prompts, three different emotion categories were assigned. In the next step, we compared the results from the manual annotation of the prompt generated by ChatGPT to the intended emotion category.

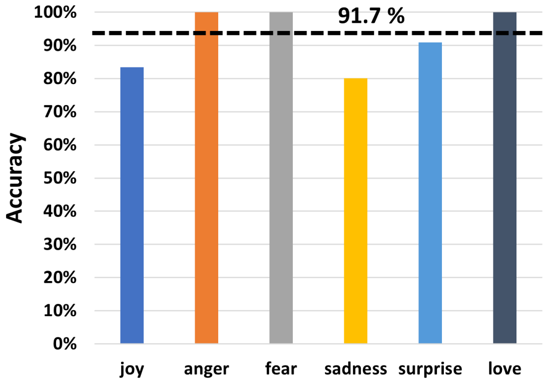


Fig. 2: Accuracy of Understanding and Expressing Emotions.

The results of the experiment are illustrated in Figure 2. The green line indicates the average classification accuracy over our six emotion categories. We see that when it comes to expressing emotions, ChatGPT can express the desired emotion with an accuracy of 91.7%. The reference was labeled differently in only 5 out of the 60 generated sentences. *Anger*, *fear* and *love* were produced with an accuracy of 100%, *surprise* with 91%, *joy* with 83%, and *sadness* with 80%.

IV. Parallel Emotional Responding

Empathic behavior in a conversation consists of two components [28]: First the emotion category of the conversational partner is identified (*cognitive empathy*). After that, a response is generated that addresses the emotion category of the conversational partner (*affective empathy*). A *parallel emotional response* is defined as an emotional response where one individual shows the same emotion as another individual in response to a particular situation or stimulus [12]. This response can be observed when individuals share a similar emotional experience, leading to the concurrent manifestation of the same emotion in both individuals [12]. For instance, feeling *joy* when another person expresses *joy* is a common example of a parallel emotional response. In the following study, we focus on the analysis of ChatGPT's ability to generate parallel emotional responses.

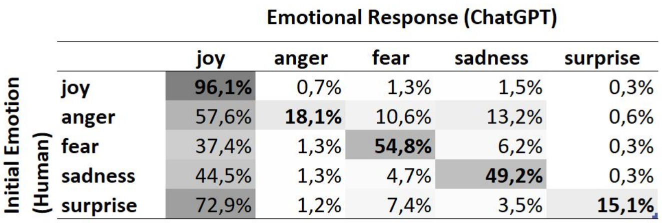
A. Experimental Setup

For these experiments, we used 20.3k initial prompts from conversations in the *EmpatheticDialogues* dataset from Facebook Research[1] [29]—named as *Speaker* prompts—to trigger ChatGPT with initial emotional prompts and generate a response, which we then classified and evaluated. Figure 3 demonstrates an initial joyful prompt from *EmpatheticDialogues* and ChatGPT's parallel emotional response.

> Speaker:
>
> > "I got a raise at my job today and I've only been there three months! I was so shocked I couldn't believe it."
>
> ChatGPT:
>
> > "Congratulations on your raise! That is an amazing accomplishment, especially considering that you have only been at your job for three months. You must be doing an excellent job, and your hard work is being recognized. Keep up the good work, and enjoy the benefits of your pay increase."

Fig. 3: Joyful Prompt from *EmpatheticDialogues* and ChatGPT's Parallel Emotional Response.

Since it was impossible to manually classify the large number of 20.3k ChatGPT-generated responses into emotion categories, we used an emotion classification system, based on the Bidirectional Encoder Representations from Transformers (BERT) [30]. The system was fine-tuned on the 16k training and 2k validation sentences of the CARER dataset[2] [27]. CARER consists of tweets labeled with our 6 emotion categories *love*, *joy*, *anger*, *fear*, *sadness*, and *surprise*. Our fine-tuning reached convergence after 8 epochs, resulting in a performance of 63% when applied to our 60 manually annotated prompts used in Section III-B. Analyzing this system demonstrated that the category *love* reduced the system performance by 15% absolute. Therefore, we removed the prompts labeled with *love* in the CARER training and validation sets and re-trained the system resulting in an accuracy of 78%. To contribute to the improvement of empathic chatbots, we share *ChatGPTsEmpatheticDialogues*—our corpus, which consists of *EmpatheticDialogues*' initial prompts, ChatGPT's responses, and the corresponding emotion categories—with the research community[3].



Initial Emotion (Human)	Emotional Response (ChatGPT)				
	joy	anger	fear	sadness	surprise
joy	**96,1%**	0,7%	1,3%	1,5%	0,3%
anger	57,6%	**18,1%**	10,6%	13,2%	0,6%
fear	37,4%	1,3%	**54,8%**	6,2%	0,3%
sadness	44,5%	1,3%	4,7%	**49,2%**	0,3%
surprise	72,9%	1,2%	7,4%	3,5%	**15,1%**

TABLE II: Emotional Responses by ChatGPT (normalized)

---

[1] https://github.com/facebookresearch/EmpatheticDialogues

[2] https://huggingface.co/datasets/dair-ai/emotion

[3] https://github.com/iu-ai-research/ChatGPTsEmpatheticDialogues

Authorized licensed use limited to: University of North Texas. Downloaded on February 09,2024 at 17:05:28 UTC from IEEE Xplore. Restrictions apply.

B. *Experiment and Results*

In our analysis, we focused on parallel emotional responses, i.e., to what percentage ChatGPT reacts with the same emotional category as in the initial prompts. Table II illustrates the distribution of emotional responses to each emotion category based on our classification system's output. The results indicate that the emotional responses are strongly biased towards replying with *joy*. In 96.1% of the cases, ChatGPT's emotional response to a prompt categorized as *joy* was *joy*. Moreover, for *anger* and *surprise*, we observe even more responses categorized as *joy* than for the original emotion category. About half of the initial prompts with *sadness* and *fear* are answered with the same emotion category. Overall, in 70.7% (20,237 responses), ChatGPT responds with the same emotion category as the initial prompt.

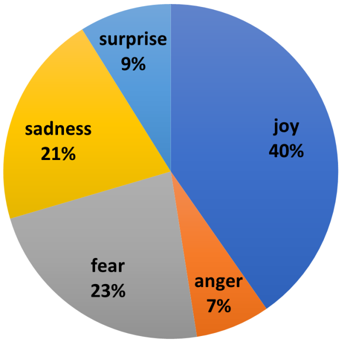


Fig. 4: Distribution of ChatGPT's Emotional Responses.

Figure 4 visualizes the distribution of all emotion categories produced by ChatGPT. With 40% of responses categorized as *joy*, we observe a strong tendency of ChatGPT to reply in a positive way.

V. HOW EMPATHIC IS CHATGPT'S PERSONALITY?

To learn more about ChatGPT's empathic capabilities, we conducted system-level evaluations using psychologically acknowledged questionnaires to evaluate ChatGPT's empathy level in different aspects. Currently, there are no standardized questionnaires available to assess the level of empathy for Chatbots. Therefore, we used five questionnaires which have been developed and validated for human subjects: *Interpersonal Reactivity Index*, *Empathy Quotient*, *Toronto Empathy Questionnaire*, *Perth Empathy Scale*, and *Autism Spectrum Quotient*. In this section, we will describe the content of each questionnaire and how we used it to gather further insights about ChatGPT's empathic capabilities.

To get ChatGPT's answer to each question in the questionnaires, we used the questions as initial prompts and then evaluated ChatGPT's response with respect to the emotion category as follows: For each of ChatGPT's answers, we had our three annotators decide which possible answer in the questionnaire it matched using the same rules for majority voting as described in Section III-B. We had to perform this procedure as ChatGPT did not directly provide us with the responses expected in the questionnaire, such as *strongly agree* or *strongly disagree*.

As an alternative to manually matching ChatGPT's answers and the answers in the questionnaire, we tried the following sentence vector-based approach: We converted ChatGPT's answer and the answers in the questionnaire to word embeddings using Sentence-BERT[4] [31], and then mapped ChatGPT's answer to the answer with the smallest distance in the semantic vector space. However, we had to discard this approach as it did not perform well, with an accuracy of 38.5%, i.e., only 70 of 182 tested answers could be mapped correctly.

In the following paragraphs, we will present the results of the questionnaires.

A. *Interpersonal Reactivity Index*

The *Interpersonal Reactivity Index* (IRI) is a widely utilized self-report measure for assessing empathy in individuals [32]. We chose to have ChatGPT conduct the IRI since it has been used in various research and clinical settings to understand empathy better and develop interventions to improve empathy skills, e.g. [33] or [34]. In addition, the questionnaire covers the categories of *fantasy*, *personal distress*, *perspective taking*, and *empathic concern*.

1) Experimental Setup: The IRI comprises 28 questions that evaluate the following four components of empathy measured in subscales which are part of the overall scale of the questionnaire: *perspective taking*, *empathic concern*, *personal distress*, and *fantasy*. *Perspective taking* refers to an individual's ability to understand the perspectives of others, *empathic concern* to the ability to feel compassion and concern for others, *personal distress* to the tendency to experience anxiety or discomfort in response to others' negative experiences, and *fantasy* to the tendency to imagine oneself in fictional situations. The level of agreement with each statement is rated on a 5-point Likert scale ranging from *does not describe me well* to *describes me very well*. The scores for each subscale of the IRI are obtained by summing the responses to the questions that belong to that subscale, resulting in a score that ranges from 0 to 28. Higher scores on the *perspective taking* and *empathic concern* subscales indicate greater empathy, while lower scores on the *personal distress* subscale suggest better emotional regulation.

2) Experiment and Results: Figure 5 visualizes ChatGPT's performance on the four IRI subscales and the mean performance of males and females on [32]. For a comparison with the other questionnaires, the absolute scores are not displayed in the figure, but the percentage achieved compared to the possible total score of 28. Comparing ChatGPT's absolute score for *fantasy* reveals interesting results: while the score of 17 is significantly higher than the mean score of healthy males (15.73, $SD = 5.6$, $t(578) = 5.46$, $p < .001$) it is significantly lower than the score of healthy females (18.75, $SD = 5.17$, $t(581) = -8.17$, $p < .001$). For *perspective taking* the absolute score of 16 is significantly lower than the

---

[4] https://github.com/UKPLab/sentence-transformers

Authorized licensed use limited to: University of North Texas. Downloaded on February 09,2024 at 17:05:28 UTC from IEEE Xplore. Restrictions apply.

mean score of males (16.78, $SD = 4.72$, $t(578) = 3.98$, $p < .001$) and females (17.96, $SD = 4.85$, $t(581) = -9.75$, $p < .001$) demonstrating that ChatGPT has lower abilities to take the perspective of others and understand their feelings than healthy humans. ChatGPT's score for *empathic concern* (11) is much lower than the mean scores of males (19.04, $SD = 4.21$, $t(578) = -45.95$, $p < .001$) and females (21.67, $SD = 3.83$, $t(581) = -67.21$, $p < .001$) indicating that ChatGPT has a significantly lower level of emotional response to others. Finally, also the absolute score for personal distress of 9 is significantly lower than the mean of healthy males (9.46, $SD = 4.55$, $t(578) = -2.43$, $p < .05$) and females (12.28, $SD = 5.01$, $t(581) = -15.79$, $p < .001$). Taken together, in almost all dimensions of the IRI ChatGPT performs significantly worse than healthy humans.

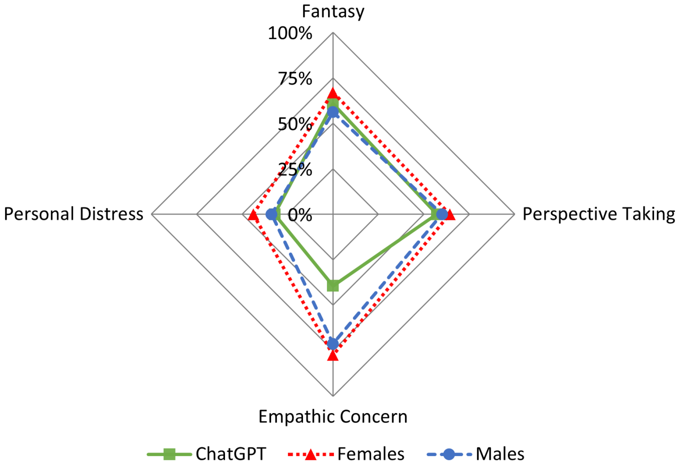


Fig. 5: ChatGPT's IRI Results Compared to Males/Females.

B. Empathy Quotient

The *Empathy Quotient* (EQ) is a self-reported questionnaire that has been specifically designed to assess an individual's ability to comprehend and respond to others' emotions [25]. We chose to have ChatGPT conduct the EQ since the questionnaire was conducted on two groups: one group with Asperger syndrome / high-functioning autism (AS/HFA) and another group of healthy humans. In the evaluation study, a clear threshold was identified to differ between the two groups, which allows us to assign the score we achieve with ChatGPT to one of those groups.

1) Experimental Setup: The EQ comprises 60 questions, each with four possible responses: *strongly agree*, *agree*, *disagree*, and *strongly disagree*. The questions cover a variety of topics related to *social interaction*, *emotional recognition*, and *communication*. Scores on the EQ can range from 0 to 80, with higher scores indicating a greater capacity for empathy.

2) Experiment and Results: Figure 6 displays ChatGPT's EQ performance (green line) compares to healthy males' and females' average EQ scores. The bars indicate the standard deviations. The mean scores reported by [25] are 41.8 ($SD = 11.2$) for healthy males and 47.2 ($SD = 10.2$) for healthy females. According to [25], more than 80% of people diagnosed with AS/HFA obtained a score below 30. Thus, with an EQ score of 34, ChatGPT scores significantly lower than males ($t(70) = -5.62$, $p < .001$) and females ($t(125) = -14.53$, $p < .001$). However, it scores higher than an average person with AS/HFA.

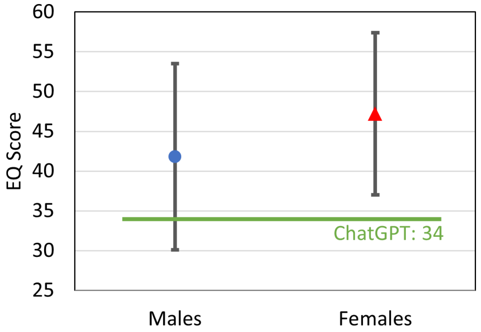


Fig. 6: ChatGPT's EQ Results Compared to Males/Females.

C. Toronto Empathy Questionnaire

Another tool to measure self-reported empathy is the *Toronto Empathy Questionnaire* (TEQ) [35]. We decided to use the TEQ, as the questionnaire tries to establish a general agreement between previous questionnaires such as the IRI, the Autism Quotient, and many more.

1) Experimental Setup: The TEQ consists of 16 questions that measure different components of empathy, including *affective empathy* (the ability to experience and understand the emotions of others) and *cognitive empathy* (the ability to understand the thoughts and perspectives of others). The TEQ also includes questions assessing an individual's tendency to take another person's perspective and willingness to help others. Scores on the TEQ can range from 0 to 64, with higher scores indicating a greater capacity for empathy.

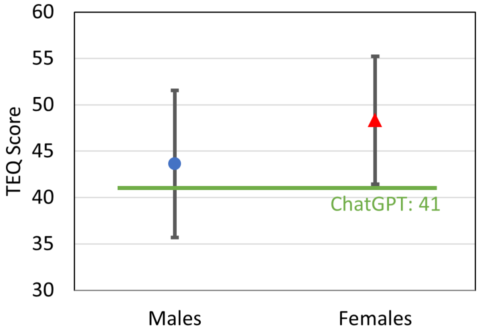


Fig. 7: ChatGPT's TEQ Results Compared to Males/Females.

2) Experiment and Results: Figure 7 illustrates how ChatGPT compares to the scores from a validation study, differentiated by males and females. The bars indicate the standard deviations. ChatGPT achieved a total of 41. In the validation study with 65 students from the University of Toronto, students achieved a mean score of 46.96 (SD = 7.47) [35]. Also, in the TEQ, females scored higher than males (48.33 vs. 43.63). ChatGPT's score of 41 is only slightly lower than the score

Authorized licensed use limited to: University of North Texas. Downloaded on February 09,2024 at 17:05:28 UTC from IEEE Xplore. Restrictions apply.

of males ($t(18) = -1.45$, $p = .165$) and significantly lower than the score of females ($t(45) = -7.20$, $p < .001$).

D. Perth Empathy Scale

The *Perth Empathy Scale* (PES) is a recently published self-report questionnaire consisting of 20 questions to assess empathy in adults and adolescents [36]. In addition to other existing scales, it covers the cognitive and affective components of empathy and the positive and negative dimensions of affective empathy. We selected this questionnaire as the splitting into positive and negative empathy can be seen as additional information not covered by the other analyzed questionnaires.

1) Experimental Setup: Each category of the PES includes five sentences that cover 10 emotions, including the five of the basic emotions described by [37] (i.e., *happiness*, *sadness*, *anger*, *scared*, *disgust*), the self-conscious emotions of *embarrassment* and *pride*, and the positive emotions of *amusement*, *calmness*, and *enthusiasm*. Respondents rate their level of agreement or disagreement with each statement on a 5-point Likert scale ranging from *never* to *always*. The PES yields a general empathy score from 0 to 100, calculated by adding the scores from the four scales. The questions for affective empathy ask if the emotions belong to someone else, while for cognitive empathy, the questions ask about someone else's feelings, indicating a self-other distinction. The higher the total score, the higher the level of empathy of an individual.

2) Experiment and Results: ChatGPT scored 40 out of 100 possible points on the PES, which is significantly below the score of healthy individuals (males: 64.1 ($SD = 10.92$), $t(187) = -30.26$, $p < .001$, females: 66.9 ($SD = 11.27$), $t(450) = -50.63$, $p < .001$) [36]. Figure 8 shows how the score distributes amongst the subscales of the PES and how ChatGPT compares to the mean scores of healthy humans. As a validation study showed that positive and negative cognitive empathy are highly correlated, both values are summed up in the figure. As observed in Section IV-B, we detect a higher tendency toward positive empathy than toward negative empathy.

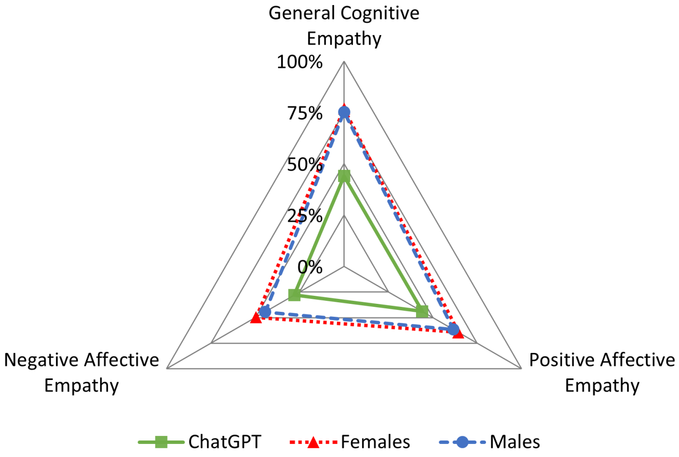


Fig. 8: ChatGPT's PES Results Compared to Males/Females.

E. Autism Spectrum Quotient

Several studies have found that individuals with autism may have difficulties with the cognitive component of empathy, such as *perspective taking* and understanding others' mental states, while still being able to experience emotions and show affective empathy, such as feeling concerned or compassion for others [25]. Therefore, we decided to additionally analyze the Autism Spectrum Quotient (AQ)—a questionnaire that measures the autistic traits in individuals who may or may not have a formal diagnosis of autism [38]. The AQ has been shown to be inversely correlated with the EQ [25].

1) Experimental Setup: On the AQ, respondents rate their level of agreement with each statement on a 4-point Likert scale ranging from *definitely agree* to *definitely disagree*. The AQ measures five different skills: *communication* (verbal and nonverbal communication), *social* (social interaction and understanding social cues), *imagination* (imaginative and flexible thinking), *local details* (tendency to focus on details and a preference for structured and predictable environments), and *attention switching* (changing focus from one topic to another).

The value of the AQ score ranges from 0 to 50, with 10 points for each skill. In contrast to the previously presented questionnaires, a higher score on the AQ refers to a lower level of empathy.

2) Experiment and Results: As shown in Figure 9, in our experiment, ChatGPT achieved a total score of 19, which is only slightly higher than the mean scores of healthy males (17.8, $SD = 6.8$, $t(75) = 1.54$, $p = .128$) but significantly higher than the mean scores of healthy females (15.4, $SD = 5.7$, $t(97) = 6.25$, $p < .001$). Moreover, people diagnosed with AS/HFA show a mean score of 35.8 [38]. As for the previous questionnaires, ChatGPT's scores are worse than average healthy humans but are still far away from the mean score a person diagnosed with AS/HFA would achieve.

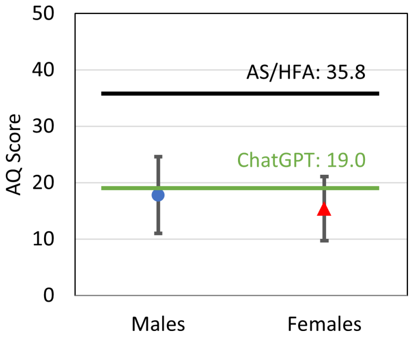


Fig. 9: ChatGPT's AQ Compared to Males/Females/AS/HFA.

Figure 10 illustrates how the scores are distributed amongst the respective skills compared to the average male and female scores. In the figure, the achieved scores are shown as a percentage in relation to the maximum possible score.

The scores for ChatGPT show that, especially for *social skills*, the score is higher than for healthy adults, which is in line with the score for empathic concern of the IRI. Moreover, *imagination skills* are worse than the average for healthy

Authorized licensed use limited to: University of North Texas. Downloaded on February 09,2024 at 17:05:28 UTC from IEEE Xplore. Restrictions apply.

	Total	Mean (SD)				
	ChatGPT	Males	Females	Range	$\Delta_{ChatGPT:Males}$	$\Delta_{ChatGPT:Females}$
IRI – Fantasy	17	15.73 (5.60)	18.75 (5.17)	0-28	8% higher	10% higher
– Perspective Taking	16	16.78 (4.72)	17.96 (4.85)	0-28	5% lower	12% lower
– Empathic Concern	11	19.04 (4.21)	21.67 (3.83)	0-28	73% lower	97% lower
– Personal Distress	9	9.46 (4.55)	12.28 (5.01)	0-28	5% lower	36% lower
EQ	34	41.8 (11.7)	47.2 (10.2)	0-80	23% lower	39% lower
TEQ	41	43.63 (7.93)	48.33 (6.90)	0-64	6% lower	18% lower
PES – General Cognitive Empathy	22	37.6 (7.13)	38.4(7.06)	0-50	71% lower	75% lower
– Positive Affective Empathy	11	15.4 (4.05)	16.1 (3.94)	0-25	40% lower	46% lower
– Negative Affective Empathy	7	11.1 (3.39)	12.4 (3.71)	0-25	59% lower	77% lower
– General Empathy	40	64.1 (10.92)	66.9 (11.27)	0-100	60% lower	67% lower
AQ	19	17.8 (6.8)	15.4 (5.7)	0-50	7% higher	23% higher

TABLE III: ChatGPT's Empathy Scores Compared to the Mean Score of Males and Females. (Note: In contrast to the scores of the other questionnaires, a higher AQ score refers to a lower level of empathy.)

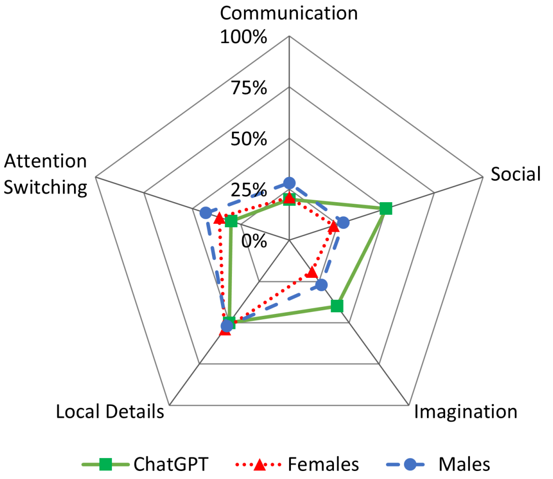


Fig. 10: ChatGPT's AQ Results Compared to Males/Females.

humans. If it comes to *attention switching*, ChatGPT performs considerably well, while the skill to focus on details and the *communicative skills* of ChatGPT is quite similar to those of healthy humans.

VI. CONCLUSION AND FUTURE WORK

In the studies presented, we investigated the empathic capabilities of ChatGPT. In our first study, we demonstrated that ChatGPT is able to rephrase a sentence to express a particular emotion with an accuracy of 91.7%. This shows that ChatGPT has the potential to be used as a tool for expressing emotions on demand which can help in the interaction with humans—be it in a learning environment or when being used as a source of everyday companionship. In a second study, we additionally demonstrated that ChatGPT can generate parallel emotional responses with 70.7% accuracy, meaning that it is able to respond with the same emotion as the initial prompt in many cases. Furthermore, our results show that ChatGPT has a strong tendency to reply with *joy*. In our last study, we used five questionnaires to test the empathic capabilities of ChatGPT. The questionnaires indicated that ChatGPT is able to interpret the emotions of others and take their perspective but still has some difficulties showing a higher level of empathy compared to healthy humans. All scores from the questionnaires in comparison to healthy males and females are summarized in Table III. While [9] concluded that GPT-3 shows a significant lack of empathy based on the psychopathy section of the SD-3 [10] questionnaire, in our empathy-focused studies we demonstrated that ChatGPT is able to show empathy in several aspects. It is important to note, that our results only demonstrate, that ChatGPT is able to show empathic behavior based on the data it has been trained on. However, to the best of our knowledge, it is not able to really feel and understand emotions or to be empathic.

With our research, we show a possible way to proceed with analyzing chatbots in the future. Further research should focus on developing more sophisticated models that can more accurately grasp the emotional context of a conversation, as well as on the development of standardized methods to measure the empathic capabilities of a chatbot. In addition, studies should be conducted to explore how ChatGPT can be used as a tool to support people more compassionately. Finally, it is important to consider the ethical implications of using chatbots such as ChatGPT. This is particularly important because they often interact with people who may not be aware that they are interacting with a computer program. Developing methods for assessing the ethical implications of using chatbots can help ensure that they are used ethically and that potential harm is minimized.

ETHICAL IMPACT STATEMENT

Individuals were asked to label the text data we collected for our data annotation. The participants who supported us participated voluntarily and free of charge. For privacy reasons, the names of the supporters are not disclosed. The collected corpus is made freely available to the community. The collected text data of the corpus are extracts from the *EmpatheticDialogues* dataset from Facebook Research [29] and produced by ChatGPT. Usually, ChatGPT produces text appropriate for the general public, but it cannot be ruled out that the content is not suitable for everyone. The text data contains emotional sentences in various forms. But this is the essence of a corpus that can be used to evaluate a chatbot's output realistically.

Authorized licensed use limited to: University of North Texas. Downloaded on February 09,2024 at 17:05:28 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] C. Pelau, D.-C. Dabija, and I. Ene, "What Makes an AI Device Human-Like? The Role of Interaction Quality, Empathy and Perceived Psychological Anthropomorphic Characteristics in the Acceptance of Artificial Intelligence in the Service Industry," *Computers in Human Behavior*, vol. 122, p. 106855, 2021.

[2] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a Human-Like Open-Domain Chatbot," *ArXiv Preprint ArXiv:2001.09977*, 2020.

[3] M. Dibitonto, K. Leszczynska, F. Tazzi, and C. M. Medaglia, "Chatbot in a Campus Environment: Design of LiSA, a Virtual Assistant to Help Students in Their University Life," in *Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part III 20*. Springer, 2018, pp. 103–116.

[4] D. Arteaga, J. Arenas, F. Paz, M. Tupia, and M. Bruzza, "Design of Information System Architecture for the Recommendation of Tourist Sites in the City of Manta, Ecuador through a Chatbot," in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2019, pp. 1–6.

[5] C. Falala-Séchet, L. Antoine, I. Thiriez, and C. Bungener, "OWLIE: A Chatbot that Provides Emotional Support for Coping With Psychological Difficulties," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 236–237.

[6] V. Taecharungroj, ""What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 35, 2023.

[7] M. Wiesner and R. K. Silbereisen, "Trajectories of Delinquent Behaviour in Adolescence and their Covariates: Relations with Initial and Time-Averaged Factors," *Journal of Adolescence*, vol. 26, no. 6, pp. 753–771, 2003.

[8] N. Eisenberg, A. Sadovsky, and T. L. Spinrad, "Associations of Emotion-Related Regulation with Language Skills, Emotion Knowledge, and Academic Outcomes," *New Directions for Child and Adolescent Development*, vol. 2005, no. 109, pp. 109–118, 2005.

[9] X. Li, Y. Li, L. Liu, L. Bing, and S. Joty, "Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective," *ArXiv E-Prints*, 2022.

[10] D. N. Jones and D. L. Paulhus, "Introducing the Short Dark Triad (SD3) a Brief Measure of Dark Personality Traits," *Assessment*, vol. 21, no. 1, pp. 28–41, 2014.

[11] N. M. McDonald and D. S. Messinger, "The Development of Empathy: How, When, and Why," *Moral behavior and free will: A neurobiological and philosophical approach*, pp. 333–359, 2011.

[12] S. W. McQuiggan, J. L. Robison, R. Phillips, and J. C. Lester, "Modeling Parallel and Reactive Empathy in Virtual Agents: An Inductive Approach," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*. Citeseer, 2008, pp. 167–174.

[13] D. Baidoo-Anu and L. Owusu Ansah, "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning," *Available at SSRN 4337484*, 2023.

[14] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke, and M. Ingrisch, "ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports," *ArXiv E-Prints*, 2022.

[15] W. Jiao, W. Wang, J.-t. Huang, X. Wang, and Z. Tu, "Is ChatGPT a Good Translator? A Preliminary Study," *ArXiv Preprint ArXiv:2301.08745*, 2023.

[16] Natalie. (2023) What is ChatGPT? [Online]. Available: https://help.openai.com/en/articles/6783457-what-is-chatgpt

[17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *CoRR*, vol. abs/2005.14165, 2020.

[18] S. Mitrović, D. Andreoletti, and O. Ayoub, "ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-Generated Text," *arXiv preprint arXiv:2301.13852*, 2023.

[19] T. Singer, "The Neuronal Basis and Ontogeny of Empathy and Mind Reading: Review of Literature and Implications for Future Research," *Neuroscience & Biobehavioral Reviews*, vol. 30, no. 6, pp. 855–863, 2006.

[20] R. L. Reniers, R. Corcoran, R. Drake, N. M. Shryane, and B. A. Völlm, "The QCAE: A Questionnaire of Cognitive and Affective Empathy," *Journal of personality assessment*, vol. 93, no. 1, pp. 84–95, 2011.

[21] A. Smith, "Cognitive Empathy and Emotional Empathy in Human Behavior and Evolution," *The Psychological Record*, vol. 56, no. 1, pp. 3–21, 2006.

[22] E. J. Lawrence, P. Shaw, D. Baker, S. Baron-Cohen, and A. S. David, "Measuring Empathy: Reliability and Validity of the Empathy Quotient," *Psychological Medicine*, vol. 34, no. 5, pp. 911–920, 2004.

[23] Ö. N. Yalçın, "Evaluating Empathy in Artificial Agents," in *2019 8th International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2019, pp. 1–7.

[24] Y.-J. Lee, C.-G. Lim, and H.-J. Choi, "Does GPT-3 Generate Empathetic Dialogues? A Novel In-Context Example Selection Method and Automatic Evaluation Metric for Empathetic Dialogue Generation," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 669–683.

[25] S. Baron-Cohen and S. Wheelwright, "The Empathy Quotient: an Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences," *Journal of Autism and Developmental Disorders*, vol. 34, pp. 163–175, 2004.

[26] R. Chadha. (2022) The Junto Emotion Wheel: What it is, why we designed it and how it can be used. [Online]. Available: https://www.thejuntoinstitute.com/emotion-wheels/

[27] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized Affect Representations for Emotion Recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[28] K. Rogers, I. Dziobek, J. Hassenstab, O. T. Wolf, and A. Convit, "Who Cares? Revisiting Empathy in Asperger Syndrome," *Journal of Autism and Developmental Disorders*, vol. 37, pp. 709–715, 2007.

[29] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5370–5381.

[30] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[31] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[32] M. Davis, "A Multidimensional Approach to Individual Differences in Empathy," *JSAS Catalog of Selected Documents in Psychology*, vol. 10, pp. 85–103, 01 1980.

[33] O. Lauterbach and D. Hosser, "Assessing Empathy in Prisoners - A Shortened Version of the Interpersonal Reactivity Index," *Swiss Journal of Psychology*, vol. 66, pp. 91–101, 2007.

[34] A.-L. Gilet, N. Mella, J. Studer, D. Grühn, and G. Labouvie-vief, "Assessing Dispositional Empathy in Adults: A French Validation of the Interpersonal Reactivity Index (IRI)," *Canadian Journal of Behavioural Science*, vol. 45, pp. 42–48, 2013.

[35] R. N. Spreng, M. C. McKinnon, R. A. Mar, and B. Levine, "The Toronto Empathy Questionnaire: Scale Development and Initial Validation of a Factor-Analytic Solution to Multiple Empathy Measures," *Journal of Personality Assessment*, vol. 91, no. 1, pp. 62–71, 2009.

[36] J. D. Brett, R. Becerra, M. T. Maybery, and D. A. Preece, "The Psychometric Assessment of Empathy: Development and Validation of the Perth Empathy Scale," *Assessment*, 2022.

[37] P. Ekman, "Are There Basic Emotions?" *Psychological Review*, vol. 99 3, pp. 550–553, 1992.

[38] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, "The Autism Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians," *Journal of Autism and Developmental Disorders*, vol. 31, pp. 5–17, 2001.

Authorized licensed use limited to: University of North Texas. Downloaded on February 09,2024 at 17:05:28 UTC from IEEE Xplore. Restrictions apply.