

Is ChatGPT More Empathetic than Humans?

Anonymous ACL submission

Abstract

This paper investigates the empathetic responding capabilities of ChatGPT, particularly its latest iteration, GPT-4, in comparison to human-generated responses to a wide range of emotional scenarios, both positive and negative. We employ a rigorous evaluation methodology, involving a between-groups study with 600 participants, to evaluate the level of empathy in responses generated by humans and ChatGPT. ChatGPT is prompted in two distinct ways: a standard approach and one explicitly detailing empathy's cognitive, affective, and compassionate dimensions. The findings of our study reveal that ChatGPT's responses exhibit a higher degree of empathy compared to those generated by humans by a statistically significant margin. Additionally, we found that instructing ChatGPT to incorporate a clear understanding of empathy in its responses significantly enhances the responses' alignment with the expectations of highly empathetic individuals.

1 Introduction

The introduction of ChatGPT has revolutionized the field of artificial intelligence. Its ability to understand and generate human-like text has opened up new avenues in different fields not limited to healthcare, education, customer service, and entertainment. Though ChatGPT has proven to be proficient in diverse tasks (Laskar et al., 2023; Ziyu et al., 2023) such as question-answering, machine translation, text summarization, commonsense reasoning, and mathematical problem-solving, its empathetic capabilities when responding to emotions expressed by users remain relatively underexplored (Brin et al., 2023). Recognizing emotions and responding to them in an empathetic manner are more difficult tasks because empathy, being a deeply nuanced and multifaceted human experience, requires not only linguistic proficiency but also a deep understanding of human psychology,

emotions, and social context (Ioannidou and Konstantikaki, 2008).

Empathy is a fundamental aspect of human interaction and can be defined as the ability to understand and share the feelings of another person. It is a complex phenomenon that involves a range of cognitive, affective, and compassionate counterparts (Ekman, 2004; Decety et al., 2006; Powell and Roberts, 2017). Cognitive empathy is understanding and accurately identifying others' feelings. Affective empathy is sharing the other person's emotions and feelings. Compassionate empathy is taking action to help the other person deal with his emotions. Empathetic responding has been identified as a key component in making artificial intelligence agents human-like, which helps to increase people's adoption of this technology (Goetz et al., 2003; Stroessner and Benitez, 2019; Svikhnushina and Pu, 2022). It is also shown to enable artificial conversational agents to build trust and rapport with users (Liu-Thompkins et al., 2022).

Several studies measure the capacity of ChatGPT (GPT-3.5) to empathize using human assessment (Chen et al., 2023; Ayers et al., 2023; Liu et al., 2023; Elyoseph et al., 2023; Zhao et al., 2023; Belkhir and Sadat, 2023; Qian et al., 2023). They have shown that ChatGPT exhibits elements of empathy, including recognizing emotions and providing emotionally supportive responses. But most of these studies are limited to the context of healthcare. Empathy is an important part of day-to-day conversations or chitchat. It involves recognizing and responding to a variety of positive as well as negative emotional situations. But to the best of our knowledge, there are no studies assessing the level of empathy displayed by ChatGPT compared to humans in such chitchat-oriented conversations. Even those that do evaluate ChatGPT's ability to empathize during everyday conversations have done so in comparison to other state-of-the-art language models, lacking a human baseline (E.g.

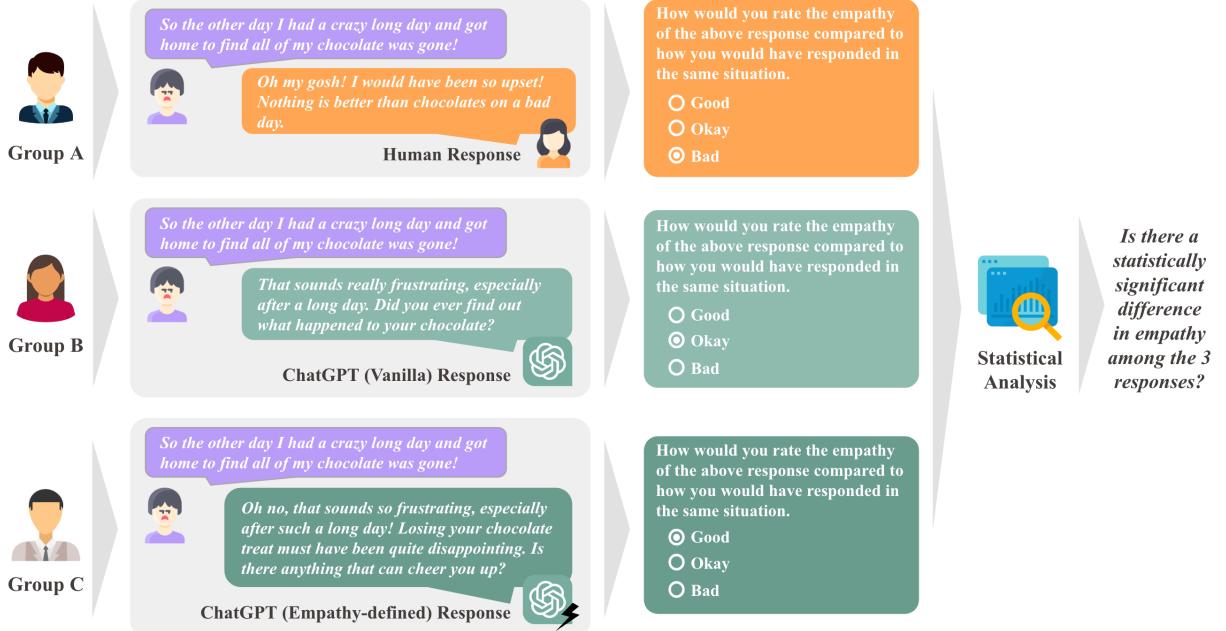


Figure 1: Between-subjects experiment design to evaluate the level of empathy demonstrated by ChatGPT compared to humans when responding to emotional situations.

(Zhao et al., 2023; Belkhir and Sadat, 2023; Qian et al., 2023)). Also, many of these studies do not adequately employ standard statistical methods such as t-tests (Semenick, 1990) or ANOVA (St et al., 1989) when analyzing human evaluation results. Lack of proper statistical analysis makes it difficult to have a proper understanding of how statistically significant these results are. Further, the absence of preliminary statistical analysis in these studies often leads to an insufficient number of human evaluators being recruited, which further undermines the statistical significance of their findings.

In this work, we comparatively analyze the level of empathy exhibited by GPT-4 and humans when responding to a variety of positive and negative emotional situations. Recognizing the limitations of prior studies, we specifically focus on chitchat-oriented dialogues. To elicit responses from ChatGPT and form a human baseline with which ChatGPT’s responses can be compared, we utilize dialogues from the EmpatheticDialogues dataset (Rashkin et al., 2019). It is a state-of-the-art dialogue dataset containing dialogues grounded in 32 distinct positive and negative emotions.

Our methodology involves recruiting 600 crowd workers to assess the empathetic quality of responses generated by GPT-4 and humans under similar emotional situations. We prompt GPT-4 with two types of instructions: a generic prompt (GPT-4 vanilla) and one that explicitly defines em-

pathy in terms of its cognitive, affective, and compassionate aspects (GPT-4 empathy-defined). This helps us closely examine how GPT-4’s empathetic responses change with respect to different prompts. Our approach involves a **between-groups** study, ensuring that different sets of evaluators assess the distinct groups of responses, thus minimizing biases. We adopt a simple and straightforward evaluation scale – *Bad*, *Okay*, and *Good* to gauge the empathy level in these responses. This simplified scale, along with rigorous statistical analyses, including t-tests and ANOVA, provides a robust framework to identify any statistically significant differences in the empathetic responses of humans and GPT-4. To further enrich our study, we employ the Toronto Empathy Questionnaire (Spreng et al., 2009) to assess each evaluator’s natural inclination towards empathy. This additional layer of analysis help us understand how individual differences in empathy influence the perception and evaluation of empathetic responses from GPT-4 and humans.

We make three significant contributions. 1) We present a comprehensive between-groups study involving 600 participants, designed to evaluate empathy in responses generated by GPT-4 against a human baseline. 2) Our statistical analysis results show that GPT-4’s responses surpass human responses in terms of empathy by a statistically significant margin. 3) We explore the relationship between the participants’ natural propensity to

142 empathize and their assessment of empathy in re-
143 sponses, which reveals that GPT-4’s empathetically
144 framed responses (GPT-4 empathy-defined) align
145 significantly more closely with the expectations of
146 those having a higher empathy propensity.

147 2 Literature Review

148 Schaaff et al. (2023) investigated the extent to
149 which ChatGPT based on GPT 3.5 can exhibit
150 empathetic responses and emotional expressions.
151 They asked ChatGPT to rephrase neutral sentences
152 into 6 emotional sentences and asked human work-
153 ers to label the rephrased text. They also mea-
154 sured to what percentage ChatGPT reacts with the
155 same emotional category as in the initial prompts.
156 However, empathy is known to be a more com-
157 plex psychological construct than mere mimicry of
158 emotion (Schuler et al., 2016), and to the best of
159 our knowledge, there are no definitive rules in the
160 psychological literature that describe how empathy
161 is elicited. Thus, in our study, we take a different
162 evaluation approach.

163 Elyoseph et al. (2023) evaluated ChatGPT’s abil-
164 ity to identify and describe emotions by utilizing
165 the Levels of Emotional Awareness Scale (LEAS)
166 (Lane et al., 1990), in which ChatGPT demon-
167 strated significantly higher performance than the
168 general population on all the LEAS scales. Though
169 some evidence suggests that there is a strong corre-
170 lation between emotional awareness and empathy
171 (Koufouli and Tollenaar, 2016; Rieffe and Camo-
172 deca, 2016), the LEAS scale does not directly mea-
173 sure one’s ability to respond empathetically to emo-
174 tional situations.

175 Belkhir and a Sadat (2023) explored ChatGPT’s
176 ability to generate empathetic responses by tak-
177 ing into account the emotional state of the user.
178 They compared ChatGPT with two slightly modi-
179 fied versions of it prompted with instructions that
180 incorporated the user’s emotion. They compared
181 the responses generated by them with state-of-the-
182 art language models using automatic metrics such
183 as accuracy, precision, and recall of the response’s
184 emotion. There are limitations to such evaluation
185 since automatic metrics may not fully capture the
186 nuances of empathetic communication. Also, an
187 empathetic response may not necessarily involve
188 an emotion but it could be more neutral and encom-
189 pass specific intents as outlined by Welivita and Pu
190 (2020), challenging the validity of such evalua-
191 tion metrics.

192 Some other studies have evaluated the empa-
193 thetic responding capability of GPT-3 that uses
194 in-context learning using human evaluators (Lee
195 et al., 2022; Fu et al., 2023; Zhao et al., 2023; Qian
196 et al., 2023). But they use within-subjects studies,
197 in which the same group of evaluators evaluate re-
198 sponds from different models, which introduces
199 bias in the evaluations. They use standard A/B test-
200 ing to compare GPT-3 with another state-of-the-art
201 dialogue response generation model and/or use a
202 5 or 7-point numerical rating scale (without any
203 textual interpretations for each option) to rate the
204 empathy-level of the responses generated by sev-
205 eral models. They also lack comparisons with hu-
206 man baselines nor make any associations with the
207 human evaluators’ propensity to empathize when
208 analyzing the ratings. The between-groups study
209 presented in this paper that compares ChatGPT’s
210 responses with those of humans using a much sim-
211 pler rating scale having textual interpretations for
212 each option, and takes into account the evaluators’
213 propensity to empathize during analysis, addresses
214 the above limitations.

215 3 The Dataset

216 For the study, we used dialogues from the Em-
217 patheticDialogues dataset introduced by Rashkin
218 et al. (2019). The dataset consists of $\approx 25K$ dia-
219 logues grounded on 32 fine-grained positive and
220 negative emotions, ranging from basic emotions
221 derived from biological responses (Ekman, 1992;
222 Plutchik, 1984) to larger sets of subtle emotions de-
223 rived from contextual situations (Skerry and Saxe,
224 2015). The authors have recruited crowd workers
225 from Amazon Mechanical Turk (AMT)¹ and paired
226 them to engage in a dialogue. The speaker counter-
227 part was instructed to come up with a situation
228 based on a given emotion and the listener counter-
229 part was instructed to respond to these situations in
230 an empathetic manner. Based on the sample size
231 predicted by power analysis (details in Section 4.3),
232 we used randomly sampled 2,000 dialogues, which
233 are more or less equally distributed across the 32
234 emotions, for our study (see Appendix A).

235 The study participants were only shown the first
236 two turns in the dialogue along with the situation
237 description and the emotion the situation was based
238 on and asked to rate the empathy of the 2nd turn i.e.
239 the listener’s response to the emotional situation
240 described in the 1st turn. This formed the human

¹<https://www.mturk.com>

baseline for our study. GPT-4 was instructed using two different prompts to generate responses given the 1st turn in these dialogues. Table 1 denotes the prompts that were used to instruct GPT-4 to generate the responses. The first one is a generic prompt that does not define the concept of empathy or explicitly ask the model to generate an empathetic response. We call this version **GPT-4 (vanilla)**. The second prompt defines the concept of empathy concerning its cognitive, affective, and compassionate counterparts and explicitly asks the model to respond in an empathetic manner to the given dialogue prompts. We call this version **GPT-4 (empathy-defined)**. Table 2 denotes the statistics of the prompt-response pairs evaluated in the study.

4 Experiment Design

4.1 Type of Study Design

As seen earlier, most existing studies that evaluate the empathy levels of large language models (LLMs) adopt a **within-subjects** study design, in which the same group of participants evaluates the level of empathy exhibited by a particular language model, with respect to other language models or human reference. But this has some disadvantages. As one person evaluates two or more models, the evaluator's perception of empathy could be distorted by overexposure to model outputs resulting in a bias in his evaluations, known as the "carry-over effect" (Shaughnessy et al., 2000). Though randomized exposure to model outputs can minimize this effect, the evaluator's judgment can still be influenced by the model output they see earlier when evaluating the level of empathy exhibited by the models that follow. For example, an averagely empathetic response may be judged more harshly by the evaluator if he has already seen an extremely empathetic response. Also, within-subjects study design makes the evaluation sessions longer making it difficult for the evaluators to stay attentive to the task. Since most of these experiments rely on crowd workers with an infamous reputation for being less attentive to tasks and instructions, making the task less exhaustive and time-consuming should be an important design consideration.

To cater to the above disadvantages, we adopt a **between-subjects** experiment design, in which one group of participants evaluates the level of empathy exhibited by humans when responding to positive and negative emotional situations and two

other groups of participants evaluate the level of empathy exhibited by **GPT-4 (vanilla)** and **GPT-4 (empathy-defined)** when responding to the same situations. The participants across the three groups were balanced across gender (Male and Female) and age groups (Young adulthood [19 - 25]; Middle adulthood [26 - 45]; Late adulthood [46 - 64]; and Older adulthood [65+]). A survey based on the Toronto Empathy Questionnaire (TEQ) to measure each participant's empathy propensity (an individual's natural inclination or tendency to empathize with others) was included in the study (see Appendix B). Later analysis revealed that the distributions of the participants' propensity to empathize was more or less similar across the three groups implying that the conditions of the participants were similar across the three groups (see Appendix I).

4.2 The Task

We used the crowdsourcing platform Prolific³ to recruit three groups of participants to evaluate the responses generated by humans and the two versions of GPT-4 (based on two prompts). Previous studies have shown that Prolific scores higher than other crowdsourcing platforms such as AMT, CloudResearch⁴, Dynata⁵, and Qualtrics⁶ in terms of worker attention, honesty, comprehension, and reliability (Peer et al., 2022; Douglas et al., 2023). We restricted our study to only those who were fluent in English, and having at least 100 previous submissions with above 95% approval rate. Each participant were shown 10 dialogue prompts randomly sampled from the subset of the Empathetic-Dialogues dataset chosen for the experiment along with the corresponding response, which is either generated by a human, GPT-4 (vanilla) or GPT-4 (empathy-defined). The participants did not have any knowledge whether the response was generated by a human or a language model. They were instructed to rate how empathetic the responses are in terms of *Bad*, *Okay*, and *Good*, compared to how they would have responded in the same situation. They were also shown a tutorial defining the concept of empathy with respect to its cognitive, affective, and compassionate counterparts (the tutorial included the same text used when prompting GPT-4 (empathy-defined)) along with examples. The examples shown to the participants were

³<https://www.prolific.com>

⁴<https://www.cloudresearch.com>

⁵<https://www.dynata.com>

⁶<https://www.qualtrics.com>

²<https://www.nltk.org/api/nltk.tokenize.html>

GPT-4 (Vanilla):	<i>You are engaging in a conversation with a human. Respond to the following using on average 28 words and a maximum of 97 words.</i>
GPT-4 (Empathy-defined):	<i>Empathy is the ability to understand and share the feelings of another person. It is the ability to put yourself in someone else’s shoes and see the world from their perspective.</i> <i>Empathy is a complex skill that involves cognitive, emotional, and compassionate components.</i> <i>Cognitive empathy is the ability to understand another person’s thoughts, beliefs, and intentions. It is being able to see the world through their eyes and understand their point of view.</i> <i>Affective empathy is the ability to experience the emotions of another person. It is feeling what they are feeling, both positive and negative.</i> <i>Compassionate empathy is the ability to not only understand and share another person’s feelings, but also to be moved to help if needed. It involves a deeper level of emotional engagement than cognitive empathy, prompting action to alleviate another’s distress or suffering.</i> <i>Empathy is important because it allows us to connect with others on a deeper level. It helps us to build trust, compassion, and intimacy. Empathy is also essential for effective communication and conflict resolution.</i> <i>You are engaging in a conversation with a human. Respond in an empathetic manner to the following using on average 28 words and a maximum of 97 words.</i>

Table 1: Different types of instructions used to prompt GPT-4 (vanilla) and GPT-4 (empathy-defined) versions.

	Human	GPT-4 (vanilla)	GPT-4 (empathy-defined)
# prompt-response pairs	2,000	2,000	2,000
Avg # prompt tokens	23.24	23.24	23.24
Max # prompt tokens	143	143	143
Avg # response tokens	28.37	36.87	34.94
Max # response tokens	97	72	65

Table 2: Statistics of the dialogue prompt-response pairs used for the study. The prompt here means the first dialogue utterance that initiates a reply. NLTK’s tokenized package² was used to tokenize the text.

selected from the dialogues in the EmpatheticDialogues dataset, which were rated the highest in terms of empathy, relevance, and fluency by the human workers who participated in the dialogue creation task (Rashkin et al., 2019). Appendix C includes all the interfaces relevant to the task.

4.3 Selection of the Sample Size

The sample size of the different participant groups is a crucial consideration in the experiment design since studies with inappropriate sample sizes fail to provide accurate estimates, which makes it difficult to derive judgments (Kang, 2021). The determination of the minimal sample size required for the study depends on the type of statistical test that is used to compare the empathy ratings between the three groups. To analyze the results of the study, we mainly use **one-way analysis of variance (ANOVA)** that tests whether there is any statistically significant difference between the average empathy ratings of the three groups (in this case, we assign numerical values 1, 2, and 3 to *Bad*, *Okay*, and *Good* ratings, respectively). The null and the

alternate hypotheses of the statistical test are indicated below. (Another type of statistical test that can be used to analyze the results is **Chi-square test of independence** that tests whether there is any statistically significant difference between the proportion of *Bad*, *Okay*, and *Good* ratings of the three groups. This is elaborated in Appendix F.)

One-way analysis of variance (ANOVA):

- **Null hypothesis:** There is no difference between the average empathy ratings of the three groups of responses.
- **Alternative hypothesis:** There is a difference between the average empathy ratings of at least one out of the three groups of responses.

We used the G-Power software (Faul et al., 2009) to determine the minimal sample size required to detect a significant difference between the empathy ratings of the three types of responses. For one-way analysis of variance (ANOVA) with a medium effect size (0.25), a significance level (α) of 0.05, and a power ($1-\beta$) of 0.95, the minimal sample size required is 252 (i.e. 84 participants per group). See Appendix D for how the effect size required for the study was determined. As we intend to statistically analyze the differences in empathy ratings when responding to positive and negative emotions separately, the minimal sample size required becomes twice the size suggested above (i.e. 168 participants per group). Considering the above, we decided to recruit 600 participants (i.e., 200 participants per group), which is sufficiently above the minimal sample size. As one participant was asked to rate 10 responses, altogether 6,000 responses (2,000 responses per group) were evaluated.

5 Results

Table 8 shows the average empathy ratings obtained for the human responses and those generated by GPT-4 vanilla and empathy-defined versions. We assigned values 1, 2, and 3 to *Bad*, *Okay*, and *Good* ratings, respectively, when calculating the average ratings and the standard errors. We also indicate these results for positive and negative emotions separately. According to the results, the responses generated by GPT-4 that use the empathy-defined prompt scores the highest in terms of the average empathy rating when responding to all emotions and negative emotions, whereas the responses generated by GPT-4 that uses the generic prompt scores the highest in terms of the average empathy rating when responding to positive emotions. The statistical F and p-values obtained for the one-way ANOVA test comparing the average empathy ratings of the three groups of responses indicate that there is an extremely statistically significant difference ($p < 0.001$) between the average empathy ratings of GPT-4's responses compared to the human responses. However, the difference between the average empathy ratings of GPT-4 based on the generic and empathy-defined prompts is not very statistically significant as indicated by the low t ($t < \pm 2$) and high p-values ($p > 0.05$) in Table 3. The details of statistical analysis conducted using the Chi-square test of independence are included in Appendix F, which yielded similar observations.

We also analyzed whether the empathy ratings assigned to responses generated by humans and the two versions of GPT-4 changed based on the evaluators' empathy propensities as measured by the Toronto Empathy Questionnaire. We computed the regression trendlines based on Ordinary Least Squares (OLS) regression after plotting the evaluators' average empathy ratings of the responses against their empathy propensities (in ascending order). The slopes of the trendlines were 0.0022, 0.0066, and 0.0130 respectively for empathy ratings of humans', GPT-4 (vanilla)'s, and GPT-4 (empathy-defined)'s responses, respectively. The slopes of the latter two were statistically significantly different from zero ($p < 0.05$ and $p < 0.001$ for the slopes of GPT-4 (vanilla) and GPT-4 (empathy-defined), respectively), implying that there is a statistically significant trend of assigning higher empathy ratings to responses generated by GPT-4 compared to human responses when the evaluator's propensity to empathize is higher. This

significance is even more higher for the responses generated by GPT-4 empathy-defined. This can be because highly empathetic individuals may have a positive bias towards assessing substantially good empathetic expressions or forms of support, which may lead to higher ratings. The t-test statistics also indicated that there is a statistically significant difference in the slopes of the average empathy ratings of the humans' and GPT-4 (empathy-defined)'s responses ($t = 3.822$ and $p < 0.001$) indicating further that the responses generated by GPT-4 (empathy-defined) aligns more with the expectations of highly empathetic individuals by a very high statistically significant margin ($p < 0.001$) compared to responses given by humans. Appendix G includes the statistical t-test values and the graphical plots corresponding to each case.

6 Case Study

Table 4 shows an example in which the response generated by GPT-4 (empathy-defined) was rated *Good* whereas the response generated by GPT-4 (vanilla) and the human response were rated *Okay* and *Bad* respectively, by the participants. It could be observed that the response given by the human does not seem to acknowledge the feeling of the speaker whereas the responses given by both versions of GPT-4 first acknowledge what the speaker is going through. The GPT-4 (vanilla) version seems to provide some advice and suggestions using phrases "*It's important to ...*" and "*Perhaps you could ...*" while GPT-4 (empathy-defined) seems to take a more non-directive approach to giving the same advice by phrasing it as a question ("*Have you considered ...*"). Table 5 shows an example in which the responses generated by both versions of GPT-4 were rated *Bad* whereas the response given by the human was rated *Good* independently by the participants. What we could observe in such scenarios was that the language used by GPT-4 was a bit more formal than the human's and contained more advices without asking for explicit permission from the speaker using phrases such as "*It's important to ...*" and "*Remember, it's important to ...*" as seen in the example.

The above observations align with the guidelines from the Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2014), a behavioral coding system used in psychology to evaluate how well a mental health provider responds, which recognizes providing advices without per-

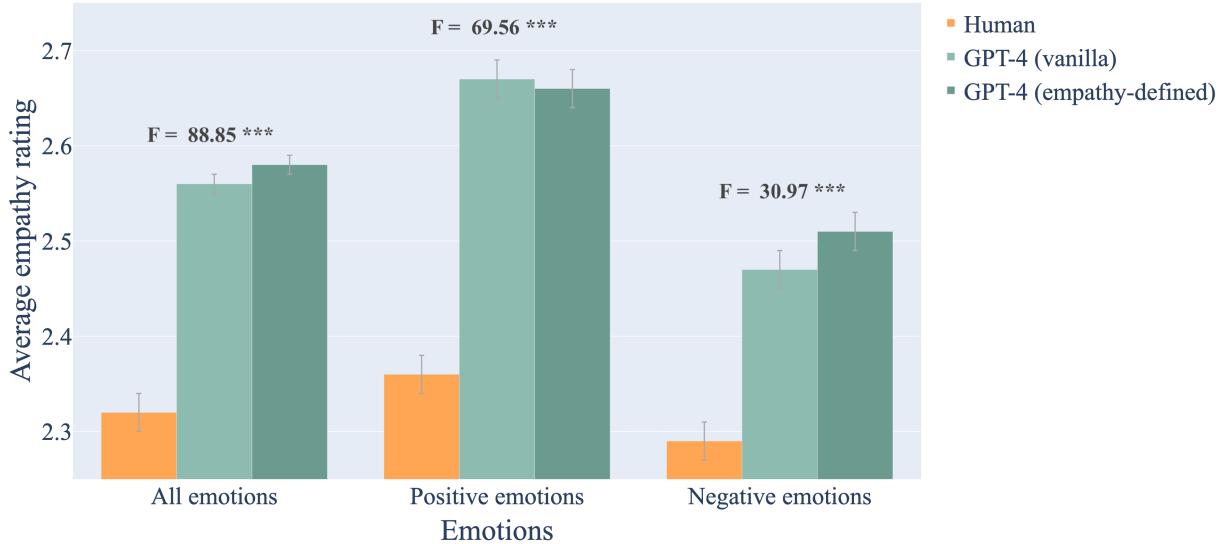


Figure 2: Average empathy ratings corresponding to the human’s and GPT-4’s responses (based on the two prompts for all, positive, and negative emotions. Error bars are calculated using the standard errors for each. The F-values computed using the statistical one-way ANOVA test for all, positive, and negative emotions are also indicated. The corresponding p-values are all less than 0.001, which indicates very high statistical significance. The exact numerical values obtained from statistical analysis are included in Appendix E.

	Human vs GPT-4 (vanilla)		Human vs GPT-4 (empathy-defined)		GPT-4 (vanilla) vs GPT-4 (empathy-defined)	
	t-value	p-value	t-value	p-value	t-value	p-value
All emotions	-10.75	3.06e-26***	-11.78	5.01e-31***	-1.09	0.27 (p > 0.05)
Positive	-9.98	2.72e-22***	-9.65	4.97e-21***	0.09	0.93 (p > 0.05)
Negative	-5.88	5.29e-09***	-7.33	4.32e-13***	-1.41	0.16 (p > 0.05)

Table 3: Statistical t-test results corresponding to the average empathy ratings of the human’s and GPT-4’s responses (based on the two prompts). In this case, we compare two by two.

Situation	Our marketing manager at work recently added a ton of new tasks to my workload that weren’t part of my job when I was brought on. I’m a little irritated, like she’s passing her work off on me.
Emotion	Annoyed
Speaker utterance	<i>The head of marketing at my job recently dumped a bunch of new tasks on me and my coworker. I feel like she’s just lazy and making us do her job.</i>
Responses:	
Human	<i>Or maybe she just knows you’re a great worker and very responsible? I hope you will get a raise after completing those tasks (Bad)</i>
GPT-4 (vanilla)	<i>I understand how you feel. It’s important to have clear role boundaries at work. Perhaps you could discuss this with your supervisor or HR, expressing your concerns in a professional manner. (Okay)</i>
GPT-4 (empathy-defined)	<i>I’m really sorry to hear that you’re feeling overwhelmed at work. It sounds like you’re under a lot of pressure. Have you considered discussing this issue with your supervisor? (Good)</i>

Table 4: An example where the responses generated by the human, GPT-4 (vanilla), and GPT-4 (empathy-defined) were rated *Bad*, *Okay*, and *Good*, respectively by the participants.

mission as unfavorable. We used the classifier introduced by Welivita and Pu (2023) to classify the responses generated by humans and GPT-4 into favorable and unfavorable response types defined in the MITI coding system (See Appendix K for details). It was seen that *Advise without Permission* type of responses were present 30.87% and 73.79% more respectively in GPT-4 (vanilla)’s and GPT-4

(empathy-defined)’s responses rated *Bad* compared to those rated *Good*. This implies that GPT-4 can improve its responses by using more subtle and non-directive ways of providing advice and suggestions such as rephrasing them as questions. This can help the person going through a difficult situation to have more autonomy in the decision-making process and come up with his own solutions.

Situation	I ran a red light yesterday and got stopped by the cops. It was very traumatizing.
Emotion	Embarrassed
Speaker utterance	<i>I ran a red light yesterday because I didn't think anyone was out on the road at 2AM. But a cop managed to nail me.</i>
Responses:	
Human	<i>Oh man, I'm glad you didn't get into an accident though. Please be careful next time. Did he give you a ticket? (Good)</i>
GPT-4 (vanilla)	<i>That's unfortunate. Even though the roads may seem empty, it's important to follow traffic rules at all times for safety. Hopefully, this experience will serve as a reminder. (Bad)</i>
GPT-4 (empathy-defined)	<i>I'm really sorry to hear that happened. It must have been frustrating, especially considering the time. Remember, it's important to always follow traffic rules for safety. (Bad)</i>

Table 5: An example where the responses generated by GPT-4 were rated *Bad* whereas the human response was rated *Good* independently by the participants.

503 7 Discussion

504 This study explored the empathetic responding ca-
 505 pability of ChatGPT powered by GPT-4 compared
 506 to a human baseline. The between-subjects study
 507 design we used made it possible to avoid any distor-
 508 tion of the participants' perception of empathy due
 509 to exposure to other types of responses. It was seen
 510 that GPT-4 was capable of surpassing humans with
 511 respect to empathy by a very high statistically sig-
 512 nificant margin when responding to both positive
 513 and negative emotional situations experienced by
 514 humans. Though there was a slight improvement
 515 in the empathy ratings of the version of GPT-4 that
 516 used a prompt defining empathy, it was observed
 517 to be not statistically significant compared to the
 518 generic version, implying the inherent capability of
 519 GPT-4 to empathize without explicit instructions or
 520 definitions. However, it was seen that GPT-4 when
 521 prompted to incorporate a clear understanding of
 522 empathy with respect to its cognitive, affective, and
 523 compassionate correlates, significantly enhances
 524 the responses' alignment with the expectations of
 525 highly empathetic individuals. By qualitatively an-
 526 alyzing responses generated by GPT-4 that were
 527 rated *Bad* by the participants, we could observe that
 528 GPT-4 can further make improvements by making
 529 the language used less formal and adopting more
 530 subtle and non-directive ways of advising and pro-
 531 viding suggestions.

532 Since this study was conducted with a large sam-
 533 ple size, which is sufficiently above the minimal
 534 sample size required to accurately detect a signif-
 535 icant difference between the ratings of the three
 536 groups of responses, the results obtained can be
 537 stated reliable and generalizable to a larger popu-
 538 lation. The balanced representation of participants
 539 across gender and age groups and the balanced
 540 distribution of dialogue prompts across 32 fine-

503 grained positive and negative emotions further rein-
 504 force the credibility and applicability of our results.
 505

506 EmpatheticDialogues (Rashkin et al., 2019) is a
 507 widely used state-of-the-art dataset used for train-
 508 ing and benchmarking a number of dialogue gen-
 509 eration models. However, the results of this study
 510 raise questions about the quality of such crowd-
 511 sourced datasets with respect to empathy. In this
 512 case, synthetic data generated by LLMs such as
 513 GPT-4 outperform human-generated data by a sta-
 514 tistically significant margin, implying that such
 515 synthetically generated data has the potential to be
 516 used for training or fine-tuning other models.
 517

518 The capacity of ChatGPT to empathize with hu-
 519 mans further opens up the possibility for a myriad
 520 of applications. It can be used as a companion and
 521 an empathetic listener for people suffering from
 522 loneliness or as a personalised life coach offering
 523 empathetic support and motivation for personal
 524 development. It can also serve to assist during dis-
 525 tressful situations, by actively listening and provid-
 526 ing compassionate support during crisis situations.
 527

528 Overall, this study contributed to the understand-
 529 ing of empathy in ChatGPT-generated responses to
 530 a variety of positive and negative emotions, com-
 531 pared to those generated by humans. The between-
 532 groups study design used for the study along with
 533 the release of data makes it possible to extend this
 534 study to other LLMs as well. But this study was
 535 limited to the participants' perception of empathy
 536 at a generic level without taking into account the
 537 differences in their socio-cultural backgrounds. Fu-
 538 ture work may further investigate whether there are
 539 any differences in the way people from different
 540 countries and cultural backgrounds perceive empa-
 541 thetic responses generated by humans and LLMs.
 542

577 8 Limitations

578 Though we conducted our study part by part during different time frames within the day to get a
579 more diverse set of participants distributed across
580 the world, participants from the United Kingdom
581 ($\approx 27\%$) and South Africa ($\approx 24\%$) were repre-
582 sented more across all three groups, followed by
583 Portugal ($\approx 13\%$), Poland ($\approx 8\%$), Italy ($\approx 5\%$),
584 and Greece ($\approx 4\%$). This over-representation of
585 participants from the United Kingdom and South
586 Africa can be due to the popularity of the crowd-
587 sourcing platform Prolific in these countries com-
588 pared to the others. Because of this imbalance,
589 $\approx 65\%$ of all the participants in our study were of
590 a White/Caucasian ethnicity followed by $\approx 18\%$
591 of Africans. And English is the first language of
592 $\approx 55\%$ of the participants. Several studies point
593 to the fact that the perception of empathy differs
594 across cultures (Birkett, 2014; Chopik et al., 2017;
595 Cassels et al., 2010). Thus, the imbalanced rep-
596 resentation of participants across countries and cul-
597 tures can have a bias in the results obtained.
598

599 9 Ethics Statement

600 **Data Usage:** We used a subset of the dialogues
601 from the EmpatheticDialogues dataset (Rashkin
602 et al., 2019) in our study, which is a publicly avail-
603 able dataset containing ethically sourced dialogues.
604 The dialogues used in the study did not contain
605 any personally identifiable information or any sen-
606 sitive content and were used in compliance with
607 the dataset’s release license terms.

608 **Human Evaluation:** Since the responses eval-
609 uated were in the English language, we recruited
610 only workers who were fluent in the English lan-
611 guage from the Prolific crowdsourcing platform.
612 They were paid €2.25 for rating 10 responses
613 that required on average ≈ 11 minutes to complete.
614 Thus, the amount paid to the human raters was
615 ≈ 1.4 times above the hourly rate recommended as
616 *Good* by Prolific (€9 per hour). All the participants
617 were informed about the nature of the study and
618 their role in it before exposing them to the actual
619 task. The participants had the chance to exit work-
620 ing on the task after reading the initial description.
621 Random subsets of dialogue prompt-response pairs
622 used in the study were manually inspected to en-
623 sure that the tasks assigned to the crowdworkers
624 were not psychologically distressing or offensive.

625 **Transparency and Reproducibility:** The dia-
626 logue prompt-response pairs that were subjected to

627 evaluation will be released publicly to ensure the
628 transparency and reproducibility of our study.

629 **Human-like Chatbots:** Lastly, there are some
630 ethical implications behind making artificial con-
631 versational agents human-like. By demonstrating
632 understanding and responding to human emotions,
633 empathetic chatbots are perceived as more human-
634 like by users (Goetz et al., 2003; Stroessner and
635 Benitez, 2019; Svikhnushina and Pu, 2022). How-
636 ever, there can be some ethical implications sur-
637 rounding such chatbots. One risk is that users may
638 get emotionally attached to these chatbots misin-
639 terpreting chatbots’ responses as being capable of
640 actually understanding their emotional needs (Van-
641 derlyn et al., 2021). Such chatbots can cause harm
642 if they are designed in manipulative ways to elicit
643 certain behavior out of users such as revealing their
644 personal or financial information. Thus, actions
645 should be taken to ensure transparency about the
646 artificial nature of these dialogue systems and be
647 considerate of the risk of causing harm through
648 emotional manipulation.

599 References

- 600 John W Ayers, Adam Poliak, Mark Dredze, Eric C
601 Leas, Zechariah Zhu, Jessica B Kelley, Dennis J
602 Faix, Aaron M Goodman, Christopher A Longhurst,
603 Michael Hogarth, et al. 2023. Comparing physician
604 and artificial intelligence chatbot responses to pa-
605 tient questions posted to a public social media forum.
606 *JAMA internal medicine*.
- 607 Ahmed Belkhir and Fatiha Sadat. 2023. Beyond infor-
608 mation: Is chatgpt empathetic enough? In *Pro-
609 ceedings of the 14th International Conference on Recent
610 Advances in Natural Language Processing*, pages
611 159–169.
- 612 Melissa Birkett. 2014. Self-compassion and empathy
613 across cultures: Comparison of young adults in china
614 and the united states. *International Journal of Re-
615 search Studies in Psychology*, 3(3):25–34.
- 616 Dana Brin, Vera Sorin, Akhil Vaid, Ali Soroush, Ben-
617 jamin S Glicksberg, Alexander W Charney, Girish
618 Nadkarni, and Eyal Klang. 2023. Comparing chat-
619 gpt and gpt-4 performance in usmle soft skill assess-
620 ments. *Scientific Reports*, 13(1):16492.
- 621 Tracy G Cassels, Sherilynn Chan, and Winnie Chung.
622 2010. The role of culture in affective empathy: Cul-
623 tural and bicultural differences. *Journal of Cognition
624 and Culture*, 10(3-4):309–326.
- 625 Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao
626 Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-
627 empowered chatbots for psychiatrist and patient sim-
628 ulation: Application and evaluation. *arXiv preprint
629 arXiv:2305.13614*.

680	William J Chopik, Ed O'Brien, and Sara H Konrath.	734
681	2017. Differences in empathic concern and perspective taking across 63 countries. <i>Journal of Cross-Cultural Psychology</i> , 48(1):23–38.	735
682		736
683		737
684	Jacob Cohen. 1992. Quantitative methods in psychology: A power primer. <i>Psychol. Bull.</i> , 112:1155–1159.	738
685		739
686		740
687	Jean Decety, Claus Lamm, et al. 2006. Human empathy through the lens of social neuroscience. <i>The scientific World journal</i> , 6:1146–1163.	741
688		
689		
690	Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. <i>Plos one</i> , 18(3):e0279720.	742
691		743
692		744
693		745
694		746
695	Paul Ekman. 1992. An argument for basic emotions. <i>Cognition & emotion</i> , 6(3-4):169–200.	747
696		748
697	Paul Ekman. 2004. Emotions revealed. <i>Bmj</i> , 328(Suppl S5).	749
698		
699	Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. Chatgpt outperforms humans in emotional awareness evaluations. <i>Frontiers in Psychology</i> , 14:1199058.	750
700		751
701		752
702		753
703	Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. <i>Behavior research methods</i> , 41(4):1149–1160.	754
704		755
705		756
706		757
707		758
708	Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. 2023. Reasoning before responding: Integrating commonsense-based causality explanation for empathetic response generation. <i>arXiv preprint arXiv:2308.00085</i> .	759
709		760
710		761
711		762
712		
713	Jennifer Goetz, Sara Kiesler, and Aaron Powers. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In <i>The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.</i> , pages 55–60. Ieee.	763
714		764
715		765
716		766
717		
718		
719	Flora Ioannidou and Vaya Konstantaki. 2008. Empathy and emotional intelligence: What is it really about? <i>International Journal of caring sciences</i> , 1(3):118.	767
720		768
721		769
722		
723	Hyun Kang. 2021. Sample size determination and power analysis using the g* power software. <i>Journal of educational evaluation for health professions</i> , 18.	770
724		771
725		772
726	Alexandra Koufouri and Marieke S Tollenaar. 2016. Empathy and emotional awareness: An interdisciplinary perspective.	773
727		
728		
729	Richard D Lane, Donald M Quinlan, Gary E Schwartz, Pamela A Walker, and Sharon B Zeitlin. 1990. The levels of emotional awareness scale: A cognitive-developmental measure of emotion. <i>Journal of personality assessment</i> , 55(1-2):124–134.	774
730		775
731		776
732		777
733		778
680	Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 431–469, Toronto, Canada. Association for Computational Linguistics.	779
681		780
682		781
683		782
684	Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In <i>Proceedings of the 29th International Conference on Computational Linguistics</i> , pages 669–683.	783
685		784
686		785
687	Siru Liu, Allison B McCoy, Aileen P Wright, Babatunde Carew, Julian Z Jenkins, Sean S Huang, Josh F Peterson, Bryan Steitz, and Adam Wright. 2023. Leveraging large language models for generating responses to patient messages. <i>medRxiv</i> , pages 2023–07.	786
688		787
689		788
690		789
691		
692		
693		
694		
695		
696		
697		
698		
699		
700		
701		
702		
703		
704		
705		
706		
707		
708		
709		
710		
711		
712		
713		
714		
715		
716		
717		
718		
719		
720		
721		
722		
723		
724		
725		
726		
727		
728		
729		
730		
731		
732		
733		

790	Kristina Schaaff, Caroline Reinig, and Tim Schlippe.	Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang,	842
791	2023. Exploring chatgpt’s empathetic abilities. <i>arXiv preprint arXiv:2308.03527</i> .	Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities?	843
792		<i>arXiv preprint arXiv:2304.09582</i> .	844
793			845
794	Miriam Schuler, Sebastian Mohnke, and Henrik Wal-	Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda,	846
795	ter. 2016. The neurological basis of empathy and	Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan,	847
796	mimicry. <i>Emotional mimicry in social context</i> , pages	and Ting Liu. 2023. Through the lens of core com-	848
797	192–221.	petency: Survey on evaluation of large language mod-	849
798	Doug Semenick. 1990. Tests and measurements: The t-	els . In <i>Proceedings of the 22nd Chinese National</i>	850
799	test. <i>Strength & Conditioning Journal</i> , 12(1):36–37.	<i>Conference on Computational Linguistics (Volume</i>	851
800		<i>2: Frontier Forum)</i> , pages 88–109, Harbin, China.	852
801	John J Shaughnessy, Eugene B Zechmeister, and	Chinese Information Processing Society of China.	853
802	Jeanne S Zechmeister. 2000. <i>Research methods in</i>		854
803	<i>psychology</i> . McGraw-Hill.		
804	Amy E Skerry and Rebecca Saxe. 2015. Neural repre-		
805	sentations of emotion are organized around abstract		
806	event features. <i>Current biology</i> , 25(15):1945–1954.		
807	R Nathan Spreng, Margaret C McKinnon, Raymond A		
808	Mar, and Brian Levine. 2009. The toronto empathy		
809	questionnaire: Scale development and initial vali-		
810	dation of a factor-analytic solution to multiple em-		
811	pathy measures. <i>Journal of personality assessment</i> ,		
812	91(1):62–71.		
813	Lars St, Svante Wold, et al. 1989. Analysis of variance		
814	(anova). <i>Chemometrics and intelligent laboratory</i>		
815	systems, 6(4):259–272.		
816			
817	Steven J Stroessner and Jonathan Benitez. 2019. The		
818	social perception of humanoid and non-humanoid		
819	robots: Effects of gendered and machinelike features.		
820	<i>International Journal of Social Robotics</i> , 11:305–		
821	315.		
822	Ekaterina Svikhnushina and Pearl Pu. 2022. Peace: A		
823	model of key social and emotional qualities of conver-		
824	sational chatbots. <i>ACM Transactions on Interactive</i>		
825	<i>Intelligent Systems</i> , 12(4):1–29.		
826			
827	Lindsey Vanderlyn, Gianna Weber, Michael Neumann,		
828	Dirk Väth, Sarina Meyer, and Ngoc Thang Vu. 2021.		
829	“it seemed like an annoying woman”: On the percep-		
830	tion and ethical considerations of affective language		
831	in text-based conversational agents. In <i>Proceedings</i>		
832	<i>of the 25th Conference on Computational Natural</i>		
833	<i>Language Learning</i> , pages 44–57.		
834			
835	Anuradha Welivita and Pearl Pu. 2020. A taxonomy of		
836	empathetic response intents in human social conversa-		
837	tions . In <i>Proceedings of the 28th International Con-</i>		
838	<i>ference on Computational Linguistics</i> , pages 4886–		
839	4899, Barcelona, Spain (Online). International Com-		
840	mittee on Computational Linguistics.		
841	Anuradha Welivita and Pearl Pu. 2023. Boosting dis-		
842	stress support dialogue responses with motivational		
843	interviewing strategy . In <i>Findings of the Associa-</i>		
844	<i>tion for Computational Linguistics: ACL 2023</i> , pages		
845	5411–5432, Toronto, Canada. Association for Com-		
846	putational Linguistics.		
847			
848			
849			
850			
851			
852			
853			
854			
855			
856			
857			
858			
859			
860			
861			
862			
863			
864			
865			
866			
867			
868			
869			
870			
871			
872			
873			
874			
875			
876			
877			
878			
879			
880			
881			
882			
883			
884			
885			
886			
887			
888			

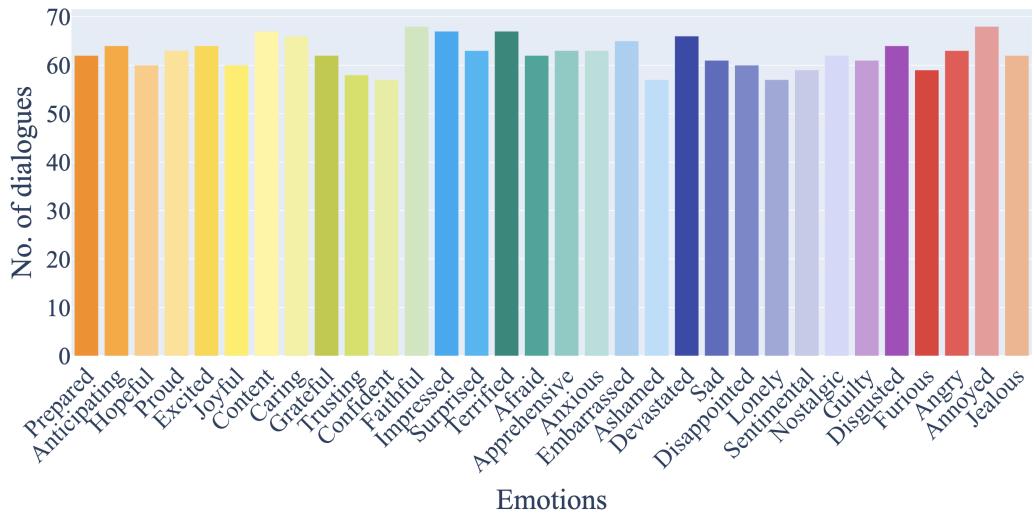


Figure 3: Distribution of the dialogue prompt-response pairs sampled from the EmpatheticDialogues dataset across the 32 positive and negative emotions.

General Information / Tutorial / Empathy Survey

Task description:

We are scientists from [REDACTED]

In this study, we will present you with responses given to 10 emotional situations. We need to you rate how empathetic the responses are in terms of "Good", "Okay", or "Bad" compared to how you would have responded in the same situations.

In the next page, we will show you a quick tutorial describing the concept of empathy along with some examples. Please make sure you read this tutorial before proceeding to the task.

Before proceeding to the task, we will ask you to answer a survey that will measure your **empathy propensity** (An individual's tendency to empathize as a function of the situation.) since we believe an individual's empathy propensity can affect how they rate the responses. After completing this survey, you will be directed to the actual task where you need to rate the empathy of dialogue responses.

Logistics:

We offer to pay €2.25 for this task.

Please make sure that you complete rating all the 10 responses and click on the "Submit" button at the end, which will show a code that you will have to copy and paste into Prolific in order to get paid.

Please avoid refreshing the page until you complete the survey and rate all the 10 responses and submit your work.

Thank you in advance for making your best effort and providing your valuable contribution to our research!

[Next](#)

Figure 4: The description of the task.

General Information / Tutorial / Empathy Survey

What is empathy?

Empathy is the ability to understand and share the feelings of another person. It is the ability to put yourself in someone else's shoes and see the world from their perspective. Empathy is a complex skill that involves cognitive, emotional, and compassionate components.

Cognitive empathy is the ability to understand another person's thoughts, beliefs, and intentions. It is being able to see the world through their eyes and understand their point of view.

Affective empathy is the ability to experience the emotions of another person. It is feeling what they are feeling, both positive and negative.

Compassionate empathy is the ability to not only understand and share another person's feelings, but also to be moved to help if needed. It involves a deeper level of emotional engagement than cognitive empathy, prompting action to alleviate another's distress or suffering.

Empathy is important because it allows us to connect with others on a deeper level. It helps us to build trust, compassion, and intimacy. Empathy is also essential for effective communication and conflict resolution.

Examples of empathetic responses given by a speaker #2 to emotional experiences described by a speaker #1:

Example 1

Speaker #1:

I had to cancel our family vacation coming up next month.

Speaker #2:

I am really sorry to hear that. I hope everything is alright.

Figure 5: The tutorial.

General Information / Tutorial / Empathy Survey

Below is a list of statements. Please read each statement carefully and rate how frequently you feel or act in the manner described. There are no right or wrong answers or trick questions. Please answer each question as honestly as you can.

Note: You need to first complete this survey to be able to proceed to the actual task!

When someone else is feeling excited, I tend to get excited too.

Never Rarely Sometimes Often Always

Other people's misfortunes do not disturb me a great deal.

Never Rarely Sometimes Often Always

It upsets me to see someone being treated disrespectfully.

Never Rarely Sometimes Often Always

Figure 6: The Toronto Empathy Questionnaire.

General Information / Tutorial / Empathy Survey / Batch 200

Below is a dialogue between two speakers, speaker #1 and speaker #2.

Rate how empathetic is the response given by the speaker #2 to the emotional situation described by the speaker #1, compared to how you would have responded in the same situation.

For better understanding, we also present the emotion of speaker #1 and the description of the situation that speaker #1 has encountered.

0 out of 10 dialogues completed!

1

Emotion of the speaker: Faithful

Situation: I'm glad I can trust my husband to always be there for me.

The dialogue:

Speaker #1:
I'm glad I can trust my husband to always be there for me.

Speaker #2:
That's wonderful to hear! Having a supportive partner like your husband is truly a blessing. It must bring a lot of comfort and happiness to your life.

The task:

How empathetic is the speaker #2's response, compared to how you would have responded for the same situation?

Good
 Okay
 Bad

You should rate the response before proceeding! [Next](#)

Figure 7: The task interface for rating responses in terms of empathy.

Emotion	# dialogues	% of dialogues
Positive emotions:	881	44.05%
Prepared	62	3.10%
Anticipating	64	3.20%
Hopeful	60	3.00%
Proud	63	3.15%
Excited	64	3.20%
Joyful	60	3.00%
Content	67	3.35%
Caring	66	3.30%
Grateful	62	3.10%
Trusting	58	2.90%
Confident	57	2.85%
Faithful	68	3.40%
Impressed	67	3.35%
Surprised	63	3.15%
Negative emotions:	1119	55.95%
Terrified	67	3.35%
Afraid	62	3.10%
Apprehensive	63	3.15%
Anxious	63	3.15%
Embarrassed	65	3.25%
Ashamed	57	2.85%
Devastated	66	3.30%
Sad	61	3.05%
Disappointed	60	3.00%
Lonely	57	2.85%
Sentimental	59	2.95%
Nostalgic	62	3.10%
Guilty	61	3.05%
Disgusted	64	3.20%
Furious	59	2.95%
Angry	63	3.15%
Annoyed	68	3.40%
Jealous	62	3.10%

Table 6: The counts and percentages of dialogue prompt-response pairs in the dataset corresponding to each emotion.

Cohen suggested the following benchmarks for interpreting the magnitude of Cohen's F: small effect: 0.10; medium effect: 0.25; and large effect: 0.40.

We selected medium effect sizes to compute the required minimum sample size because a medium effect size can sensitively detect differences in empathy levels between human and ChatGPT responses, which are expected to be noticeable but not extremely large. Also, a study conducted with a medium effect size can detect differences that are subtle yet practically meaningful, without requiring an excessively large sample size.

E One-way ANOVA Test Results

The statistical one-way ANOVA test results corresponding to the average empathy ratings of the responses generated by the humans and the two versions of GPT-4 are denoted in Table 8. Table 9 denotes the statistical one-way ANOVA test results

889	
890	
891	
892	
893	
894	
895	
896	
897	
898	
899	
900	
901	
902	
903	
904	
905	
906	

-
1. When someone else is feeling excited, I tend to get excited too
 2. Other people's misfortunes do not disturb me a great deal*
 3. It upsets me to see someone being treated disrespectfully
 4. I remain unaffected when someone close to me is happy*
 5. I enjoy making other people feel better
 6. I have tender, concerned feelings for people less fortunate than me
 7. When a friend starts to talk about his or her problems, I try to steer the conversation towards something else*
 8. I can tell when others are sad even when they do not say anything
 9. I find that I am "in tune" with other people's moods
 10. I do not feel sympathy for people who cause their own serious illnesses*
 11. I become irritated when someone cries*
 12. I am not really interested in how other people feel*
 13. I get a strong urge to help when I see someone who is upset
 14. When I see someone being treated unfairly, I do not feel very much pity for them*
 15. I find it silly for people to cry out of happiness*
 16. When I see someone being taken advantage of, I feel kind of protective towards him or her
-

Table 7: The Toronto Empathy Questionnaire (Spreng et al., 2009). *Negatively worded reverse scale questions.

corresponding to each of the 32 emotions the dialogue prompts used for testing are based on. All values in Table 8 are statistically significant as denoted by the low p-values ($p < 0.001$). Only some values in Table 9 are statistically significant mainly due to the insufficient sample size corresponding to each emotion. However, it can be observed that the responses generated by GPT-4 (empathy-defined) score the highest average empathy rating out of the three response groups in most cases.

F Chi-square Test of Independence Results

Another type of statistical test that can be used to analyze the results is **Chi-square test of independence** that tests whether there is any statistically significant difference between the proportion of *Bad*, *Okay*, and *Good* ratings of the three groups. For the chi-square test of independence with a medium effect size (0.3), a significance level (α) of 0.05, and a power ($1-\beta$) of 0.95, the minimal total sample size required is 207 (i.e. 69 participants per group). When statistically analyzing the differences in empathy ratings when responding to positive and negative emotions separately, the minimal sample size required becomes twice the sample size suggested above (i.e. 138 participants per group). Since we recruited 200 participants per group, which is sufficiently above the minimal sample size required for this test, the chi-square test of independence can also be applied to analyze the results obtained from this study.

Table 10 shows the proportions of *Bad*, *Okay*, and *Good* empathy ratings scored by the humans' and the two versions of GPT-4's responses. They

are visualized graphically in Figure 9. Similar to the case with average empathy ratings, the responses generated by GPT-4 that uses the empathy-defined prompt scores the highest number of *Good* ratings when responding to all emotions as well as negative emotions, whereas the responses generated by GPT-4 that uses the generic prompt scores the highest number of *Good* ratings when responding to positive emotions. The least number of *Bad* responses are scored by GPT-4 based on the empathy-defined prompt when responding to all, positive, and negative emotions alike. The statistical χ^2 and p-values obtained for the χ^2 test comparing the proportions of *Bad*, *Okay*, and *Good* empathy ratings of the three groups indicate that there is an extremely statistically significant difference ($p < 0.001$) between the proportions of the empathy ratings of GPT-4's responses compared to the humans'. However, similar to the case with average empathy ratings, the difference between the proportions of *Bad*, *Okay*, and *Good* empathy ratings of GPT-4 based on the generic and empathy-defined prompts is not very statistically significant as indicated by the low χ^2 ($\chi^2 < 5.991$ — critical value) and high p-values ($p > 0.05$) in Table 11.

G Change of Average Empathy Ratings with Evaluator's Empathy Propensity

Figure 10 shows scatter plots corresponding to the evaluators' average empathy ratings of the responses against their empathy propensities as measured by the Toronto Empathy Questionnaire. Regression trendlines based on Ordinary Least Squares (OLS) regression are also shown for each of humans' and the two versions of GPT-4's re-

907
908
909
910
911
912
913
914
915
916

941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965

917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940

966
967
968
969
970
971
972
973
974

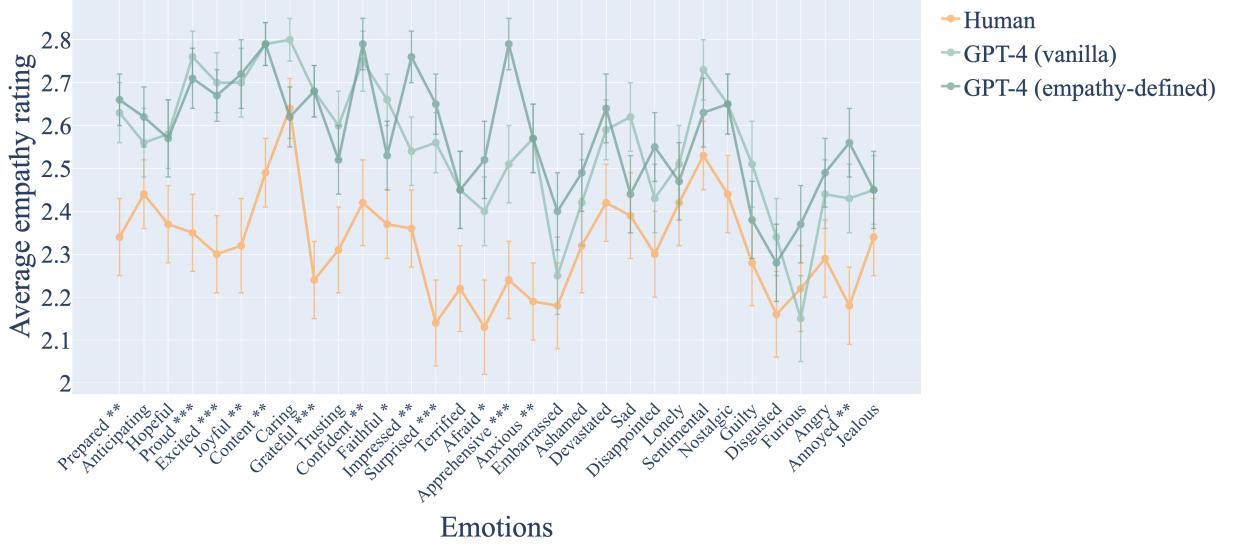


Figure 8: Average empathy ratings corresponding to the human’s and the two versions of GPT-4’s responses across the 32 positive and negative emotions. Error bars are calculated using the standard errors for each. We have indicated ***, **, and * in front of the emotion in the x-axis to indicate results that are statistically significant (***, **, and * indicate $p < 0.001$, $p < 0.01$, and $p < 0.5$, respectively).

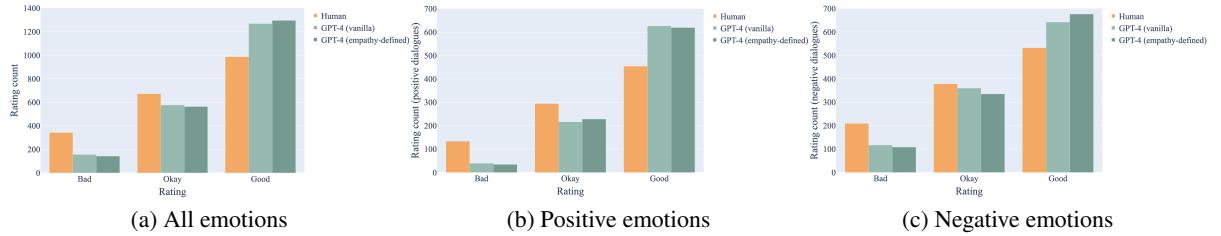


Figure 9: The number of *Bad*, *Okay*, and *Good* empathy ratings of the human’s and GPT-4’s responses (based on the two prompts) when responding to (a) all emotions (b) positive emotions, and (c) negative emotions.

Emotion	Human	GPT-4 (vanilla)	GPT-4 (empathy-defined)	F-value	p-value
All emotions	2.32 ± 0.02	2.56 ± 0.01	2.58 ± 0.01	88.849	9.418e-39***
Positive emotions	2.36 ± 0.02	2.67 ± 0.02	2.66 ± 0.02	69.558	3.632e-30***
Negative emotions	2.29 ± 0.02	2.47 ± 0.02	2.51 ± 0.02	30.971	4.698e-14***

Table 8: Statistical one-way ANOVA test results corresponding to the average empathy ratings of the human’s and GPT-4’s responses (based on the two prompts). Generally, any F-value greater than +2 or less than -2 is acceptable. The higher the F-value, the greater the confidence we have in the coefficient as a predictor.

Emotion	Human	GPT-4 (vanilla)	GPT-4 (empathy-defined)	F-value	p-value
Positive emotions:					
Prepared	2.34 ± 0.09	2.63 ± 0.07	2.66 ± 0.06	5.406	0.005**
Anticipating	2.44 ± 0.08	2.56 ± 0.08	2.62 ± 0.07	1.480	0.230
Hopeful	2.37 ± 0.09	2.58 ± 0.08	2.57 ± 0.09	1.854	0.160
Proud	2.35 ± 0.09	2.76 ± 0.06	2.71 ± 0.07	8.708	0.000***
Excited	2.30 ± 0.09	2.70 ± 0.07	2.67 ± 0.06	8.728	0.000***
Joyful	2.32 ± 0.11	2.70 ± 0.08	2.72 ± 0.08	6.650	0.002**
Content	2.49 ± 0.08	2.79 ± 0.05	2.79 ± 0.05	6.925	0.001**
Caring	2.64 ± 0.07	2.80 ± 0.05	2.62 ± 0.07	2.460	0.088
Grateful	2.24 ± 0.09	2.68 ± 0.06	2.68 ± 0.06	11.126	0.000***
Trusting	2.31 ± 0.10	2.60 ± 0.08	2.52 ± 0.08	3.010	0.052
Confident	2.42 ± 0.1	2.75 ± 0.07	2.79 ± 0.06	6.835	0.001**
Faithful	2.37 ± 0.08	2.66 ± 0.06	2.53 ± 0.08	3.904	0.022*
Impressed	2.36 ± 0.09	2.54 ± 0.08	2.76 ± 0.06	6.912	0.001**
Surprised	2.14 ± 0.10	2.56 ± 0.07	2.65 ± 0.07	11.010	0.000***
Negative emotions:					
Terrified	2.22 ± 0.10	2.45 ± 0.09	2.45 ± 0.09	2.001	0.138
Afraid	2.13 ± 0.11	2.4 ± 0.08	2.52 ± 0.09	4.618	0.011*
Apprehensive	2.24 ± 0.09	2.51 ± 0.09	2.79 ± 0.06	12.306	0.000***
Anxious	2.19 ± 0.09	2.57 ± 0.08	2.57 ± 0.08	6.548	0.002**
Embarrassed	2.18 ± 0.10	2.25 ± 0.09	2.40 ± 0.09	1.354	0.261
Ashamed	2.32 ± 0.11	2.42 ± 0.10	2.49 ± 0.09	0.790	0.455
Devastated	2.42 ± 0.09	2.59 ± 0.07	2.64 ± 0.08	1.881	0.155
Sad	2.39 ± 0.10	2.62 ± 0.08	2.44 ± 0.09	1.783	0.171
Disappointed	2.30 ± 0.10	2.43 ± 0.08	2.55 ± 0.08	1.974	0.142
Lonely	2.42 ± 0.10	2.51 ± 0.09	2.47 ± 0.09	0.221	0.802
Sentimental	2.53 ± 0.08	2.73 ± 0.07	2.63 ± 0.08	1.765	0.174
Nostalgic	2.44 ± 0.09	2.65 ± 0.07	2.65 ± 0.07	2.613	0.076
Guilty	2.28 ± 0.10	2.51 ± 0.10	2.38 ± 0.09	1.429	0.242
Disgusted	2.16 ± 0.10	2.34 ± 0.09	2.28 ± 0.09	1.083	0.341
Furious	2.22 ± 0.1	2.15 ± 0.10	2.37 ± 0.09	1.313	0.272
Angry	2.29 ± 0.09	2.44 ± 0.08	2.49 ± 0.08	1.707	0.184
Annoyed	2.18 ± 0.09	2.43 ± 0.08	2.56 ± 0.08	5.297	0.006**
Jealous	2.34 ± 0.09	2.45 ± 0.08	2.45 ± 0.09	0.584	0.559

Table 9: Statistical one-way ANOVA test results corresponding to the average empathy ratings of the human’s and GPT-4’s responses (based on the two prompts) to each of the 32 emotions the dialogue prompts are based on.

975 response ratings. The t-test statistics corresponding
976 to the slopes of the trendlines are denoted in Table
977 12.

978 H Participants’ Demographics

979 Figures 11 and 12 respectively show the distributions
980 of the countries of residence and the ethnicities
981 of the participants who rated the three groups of
982 responses. It could be observed that though there

983 are imbalances across the countries and the ethnicities
984 represented in the participants’ pool, these
985 demographics are similar across the three groups of
986 participants. This allows control for factors other
987 than the independent variable influencing the re-
988 sults of the study and fair comparison of response
989 ratings across the three groups.

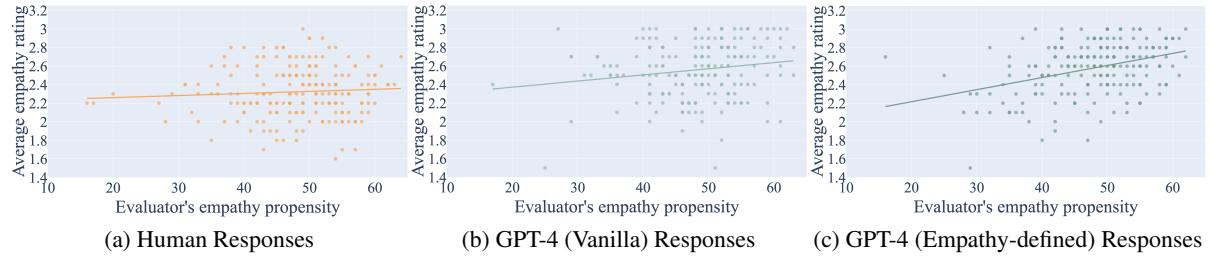


Figure 10: The scatter plots of average empathy ratings against evaluator's propensity to empathize as measured by the Toronto Empathy Questionnaire. Ordinary Least Squares (OLS) regression trendlines are also plotted.

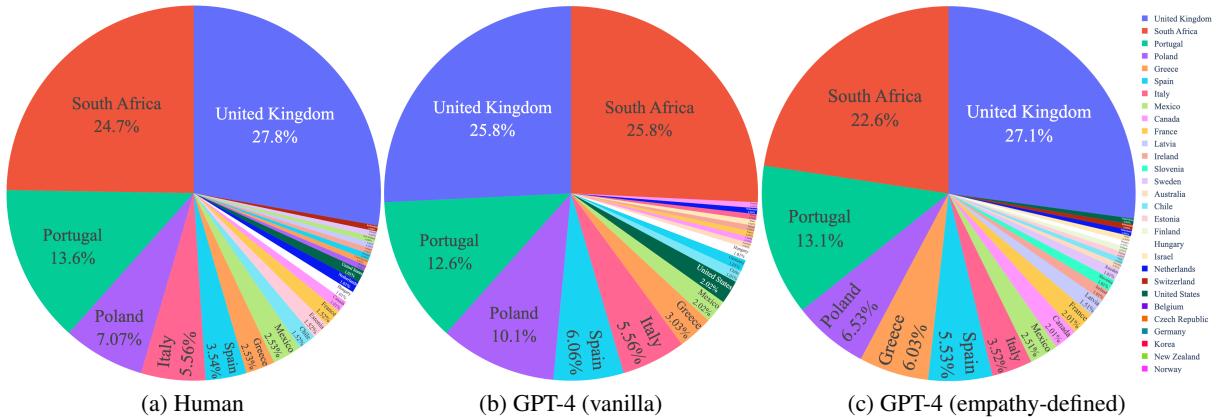


Figure 11: Distribution of the countries of residence of the participants across the three groups.

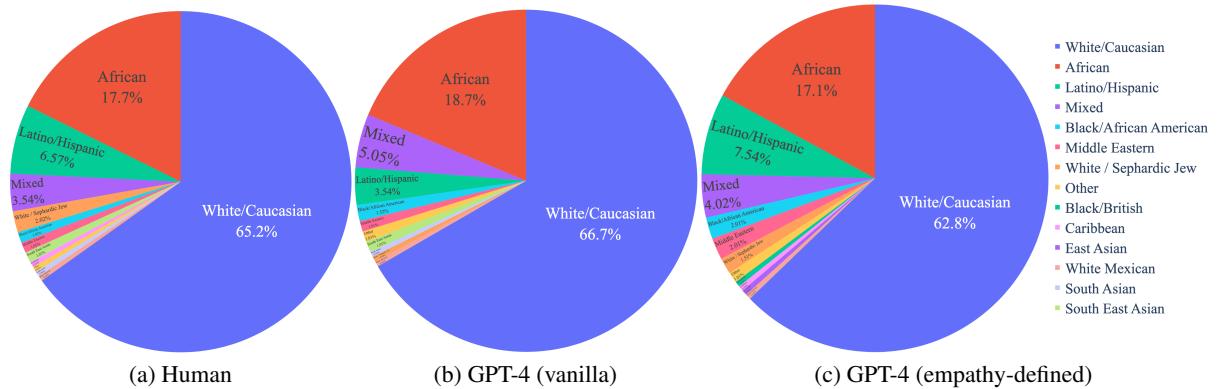


Figure 12: Distribution of the ethnicities of the participants across the three groups.

Type of response		Bad	Okay	Good	χ^2 -value	p-value
All emotions	Human	342	672	986	178.122	1.888e-37***
	GPT-4 (vanilla)	156	576	1268		
	GPT-4 (empathy-defined)	142	563	1295		
Positive emotions	Human	133	294	454	138.399	6.213e-29***
	GPT-4 (vanilla)	39	216	626		
	GPT-4 (empathy defined)	34	228	619		
Negative emotions	Human	209	378	532	64.175	3.839e-13***
	GPT-4 (vanilla)	117	360	642		
	GPT-4 (empathy defined)	108	335	676		

Table 10: Statistical Chi-square test results corresponding to the proportions of *Bad*, *Okay*, and *Good* empathy ratings of the human’s and GPT-4’s responses (based on the two prompts). The critical value of the χ^2 distribution in this case is 9.488 (computed at a significance level of 0.05 and 4 degrees of freedom), which means if the χ^2 statistic is greater than 9.488 the null hypothesis can be rejected at 5% significance level

	Human vs GPT-4 (vanilla)		Human vs GPT-4 (empathy-defined)		GPT-4 (vanilla) vs GPT-4 (empathy-defined)	
	χ^2 -value	p-value	χ^2 -value	p-value	χ^2 -value	p-value
All emotions	112.136	4.467e-25***	134.124	7.504e-30***	1.091	0.580 (p > 0.05)
Positive	90.694	2.023e-20***	92.406	8.595e-21***	0.706	0.703 (p > 0.05)
Negative	36.709	1.068e-08***	51.939	5.268e-12***	2.136	0.344 (p > 0.05)

Table 11: Statistical χ^2 test results corresponding to the proportions of *Bad*, *Okay*, and *Good* empathy ratings of the human’s and GPT-4’s responses (based on the two prompts). In this case, we compare two by two. The critical value of the χ^2 distribution in this case is 5.991 (computed at a significance level of 0.05 and 2 degrees of freedom), which means if the χ^2 statistic is greater than 5.991 the null hypothesis can be rejected at 5% significance level

	slope	std err	t-value	p-value
Human	0.0022	0.002	0.967	0.335
	0.0066	0.003	2.461	0.015*
	0.0130	0.002	5.653	0.000***
t-test statistics for the difference in 2 independent slopes				
Human vs GPT-4 (vanilla)	t-value		p-value	
	1.215		0.225 (p > 0.05)	
	3.822		0.00015***	
Human vs GPT-4 (empathy-defined)	t-value		p-value	
	3.822		0.00015***	
	1.783		0.075 (p > 0.05)	

Table 12: The statistics of the Ordinary Least Squares (OLS) regression trendlines of average empathy ratings against evaluator’s propensity to empathize (as measured by the Toronto Empathy Questionnaire). t>2 in GPT-4 (vanilla) and GPT-4 (empathy-defined) indicate that the slope of the trendline is statistically significantly different from zero. t>2 for the difference in the slopes of the average empathy ratings of the responses of the human and GPT-4 (empathy-defined) indicates that there is a statistically significant difference between the two slopes.

I Distribution of Participants’ Empathy Propensities

Figure 13 shows the distributions of the participants’ propensities to empathize across the three groups. It could be observed that they are more or less equally distributed across the three groups avoiding any biases in the results that might be caused by any unequal distribution of empathy propensities across the three groups.

J Quality Analysis

We restricted this study only to workers on Prolific who were fluent in English and had at least 100 submissions on the platform with over 95% approval rate. The Toronto Empathy Questionnaire (TEQ) used to measure the workers’ propensity to empathize, contained 8 reverse scale questions, which we used to gauge the quality of the workers and their attentiveness to the task. Figure 14 shows the number of reverse scale questions in the TEQ that were marked incorrect by the participants rating the

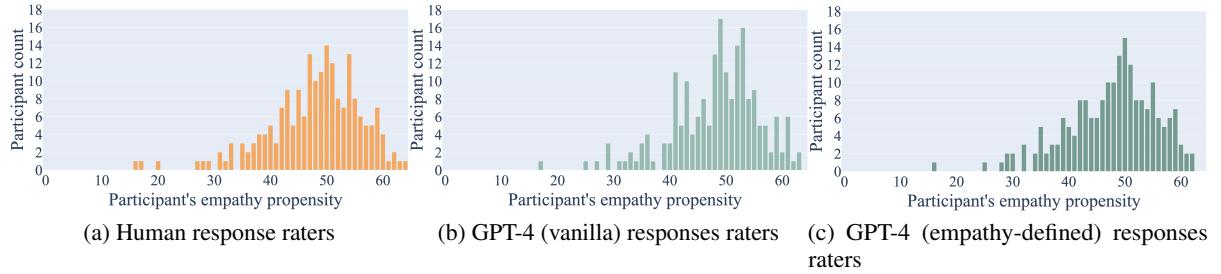


Figure 13: The distributions of the participants’ propensities to empathize across the three groups.

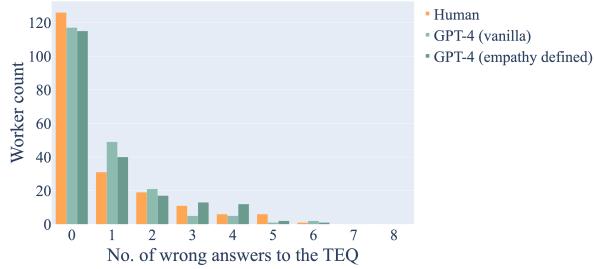


Figure 14: The number of reverse scale questions in the TEQ that were marked wrong by the participants rating the three response groups.

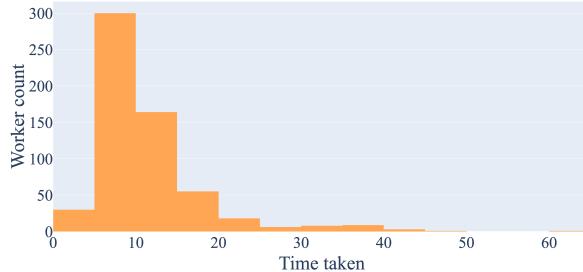


Figure 15: The histogram of times taken to complete the task by all participants.

three response groups. It was observed that $\approx 60\%$ of all participants did not get any reverse scale questions wrong and only 2.17% of all participants got more than half of the reverse scale questions wrong. These statistics validate the quality of the workers recruited for the study.

Further, Figure 15 shows the histogram of times (in minutes) taken to complete the study. On average it took 11 minutes and 18 seconds to complete rating 10 responses, which was close to the average completion time of 15 minutes that we estimated before conducting the study. Only 5% of all participants were observed to take less than 5 minutes to complete the study, which indicates that most of the participants took time to carefully read the instructions and respond to the questions attentively.

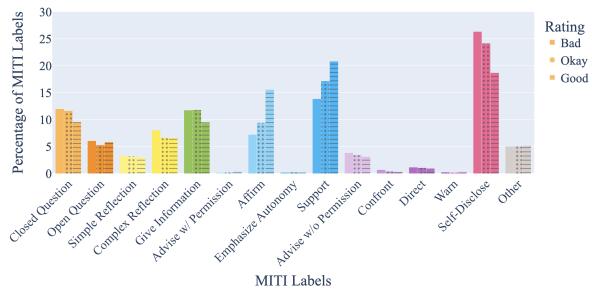
K MITI Classification

Table 13 shows the percentage of labels from the Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2014) present in responses generated by humans and the two versions of GPT-4 that were rated *Bad*, *Okay*, and *Good*, as classified by the classifier introduced by Welivita and Pu (2023). Figure 16 visualizes these distributions of the percentages of labels for responses generated by humans and the two versions of GPT-4, separately.

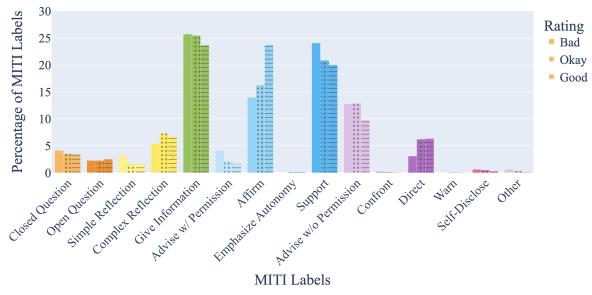
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036

MITI label	Human			GPT-4 (vanilla)			GPT-4 (empathy-defined)		
	Bad	Okay	Good	Bad	Okay	Good	Bad	Okay	Good
MI Adherent Response Types:									
Closed Question	12.00	11.64	9.55	4.12	3.55	3.41	3.56	3.56	3.47
Open Question	6.06	5.29	5.84	2.26	2.27	2.48	2.31	2.20	2.50
Simple Reflection	3.26	3.21	2.98	3.09	1.60	1.64	2.73	1.83	1.58
Complex Reflection	8.04	6.65	6.57	5.35	7.31	6.68	5.03	5.86	7.34
Give Information	11.77	11.82	9.51	25.72	25.49	23.67	23.69	25.86	23.75
Advise with Permission	0.23	0.24	0.39	4.12	2.27	1.68	3.35	2.36	1.73
Affirm	7.23	9.44	15.58	13.99	16.27	23.74	10.06	18.90	22.96
Emphasize Autonomy	0.23	0.24	0.19	0.00	0.15	0.14	0.00	0.10	0.16
Support	13.87	17.16	20.83	24.07	20.85	20.01	27.67	21.31	19.45
MI Non-adherent Response Types:									
Advise without Permission	3.85	3.44	3.13	12.76	12.87	9.75	16.98	11.83	9.77
Confront	0.70	0.42	0.31	0.21	0.15	0.07	0.21	0.00	0.14
Direct	1.17	1.07	0.93	3.09	6.23	6.32	3.56	5.34	6.64
Warn	0.23	0.18	0.31	0.00	0.10	0.05	0.00	0.21	0.00
Other:									
Self-Disclose	26.34	24.17	18.71	0.62	0.51	0.25	0.42	0.37	0.34
Other	5.01	5.05	5.18	0.62	0.36	0.14	0.42	0.26	0.20

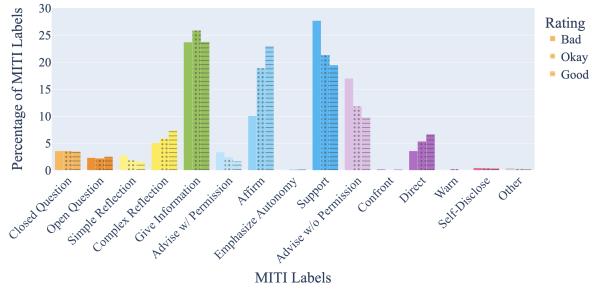
Table 13: The percentages of labels from the MITI code present in responses generated by humans and the two versions of GPT-4 that were rated *Bad*, *Okay*, and *Good*.



(a) Human responses



(b) GPT-4 (vanilla) responses



(c) GPT-4 (empathy-defined) responses

Figure 16: The distributions of the percentages of MITI labels present in responses generated by humans and the two versions of GPT-4 that were rated *Bad*, *Okay*, and *Good*.