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 counterparts. Our findings indicate that the average empathy rating of responses generated by ChatGPT exceeds those crafted by humans by approximately 10%. Additionally, instructing ChatGPT to incorporate a clear understanding of empathy in its responses makes the responses align ≈5 times more closely with the expectations of individuals possessing a high degree of empathy, compared to human responses. The proposed evaluation framework serves as a scalable and adaptable framework to assess the empathetic capabilities of newer and updated versions of large language models, eliminating the need to replicate the current study's results in future research.

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, and text summarization, its empathetic capabilities when responding to emotions expressed by users remain relatively under-explored (Brin et al., 2023). Understanding and responding empathetically to emotions pose greater challenges, 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

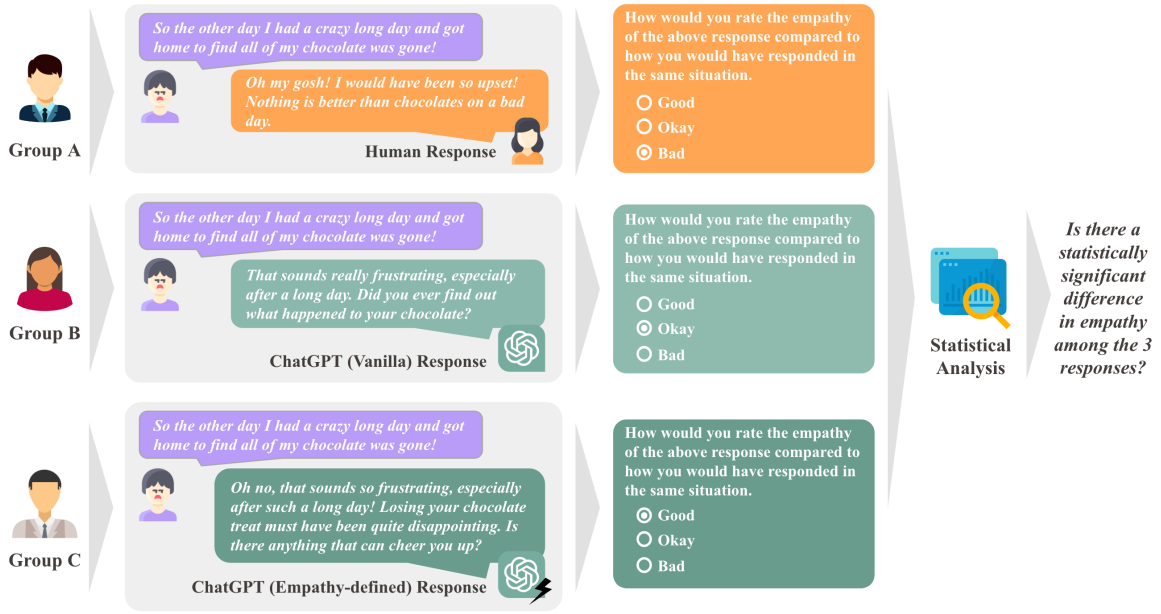


Figure 1: Between-subjects experiment design to evaluate the level of empathy demonstrated by ChatGPT compared to humans when responding to emotional situations.

ability to empathize during everyday conversations have done so in comparison to other state-of-the-art language models, lacking a human baseline (Zhao et al., 2023; Belkhir and Sadat, 2023; Qian et al., 2023). Most importantly, existing studies have utilized **within-subjects** study designs (Lee et al., 2022; Fu et al., 2023; Zhao et al., 2023; Qian et al., 2023), in which the same group of evaluators evaluates responses from different models. Not only does this introduce bias in the evaluations, as evaluators may form judgments influenced by previous models rather than evaluating each independently, but also limits the scalability of the evaluation framework to newer or updated versions of large language models (LLMs), as it necessitates discarding previous study results where all the models are evaluated collectively, rather than independently. In an era characterized by rapid advancements in language model technology, the need for scalable and adaptable evaluation frameworks that can keep pace with evolving models is becoming increasingly vital.

In this work, we comparatively analyze the level of empathy exhibited by ChatGPT powered 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 state-of-the-art EmpatheticDialogues dataset (Rashkin et al., 2019) having dialogues grounded in 32 positive and negative emotions. Our approach involves a **between-groups** study, ensuring that different sets of evaluators assess distinct groups of responses from humans and variants of ChatGPT. This type of study design not only minimizes potential evaluation biases but also allows for a more scalable and adaptable framework, facilitating the evaluation of new or updated language models as they are introduced, without the need to discard previous findings or replicate the entire study.

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 and one that explicitly defines empathy in terms of its cognitive, affective, and compassionate aspects. We adopt a simple and straightforward evaluation scale – *Bad*, *Okay*, and *Good* to gauge the empathy level in these responses. We perform rigorous statistical analysis to identify whether there are any statistically significant differences in the empathy ratings 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 propensity to empathize. This additional layer of analysis helps us understand how individual differences in empathy influence the evaluation of empathetic responses from GPT-4 and humans.

2 Literature Review

Schaaff et al. (2023) investigated the extent to which ChatGPT based on GPT 3.5 can exhibit empathetic responses and emotional expressions by prompting ChatGPT to rephrase neutral sentences into 6 emotional sentences and asking human workers to label the rephrased text. They measured to what percentage ChatGPT reacts with the same emotional category as in the initial prompts. However, empathy is known to be a more complex psychological construct than mere mimicry of emotion (Schuler et al., 2016), and to the best of our knowledge, there are no definitive rules in the psychological literature that describe how empathy is elicited. Thus, in our study, we take a different evaluation approach. Elyoseph et al. (2023) evaluated ChatGPT's ability to identify and describe emotions by utilizing the Levels of Emotional Awareness Scale (LEAS) (Lane et al., 1990), in which ChatGPT demonstrated significantly higher performance than the general population on all the LEAS scales. Though some evidence suggests that there is a strong correlation between emotional awareness and empathy (Koufouli and Tollenaar, 2016; Rieffe and Camodeca, 2016), the LEAS scale does not directly measure one's ability to respond empathetically to emotional situations. On the other hand, Belkhir and Sadat (2023) evaluated ChatGPT's ability to generate empathetic responses using automatic metrics. They compared ChatGPT with two slightly modified versions of it prompted with instructions that incorporated the user's emotion. The responses were compared with the state-of-the-art language models using automatic metrics such as accuracy, precision, and recall of the response's emotion. But there are limitations to such evaluation since automatic metrics may not fully capture the nuances of empathetic communication. Also, an empathetic response may not necessarily involve an emotion but it could be more neutral and encompass specific intents as outlined by Welivita and Pu (2020), which challenges the validity of such evaluation metrics.

Some other studies have evaluated the empathetic responding capability of GPT-3 that uses in-context learning, using human evaluators (Lee et al., 2022; Fu et al., 2023; Zhao et al., 2023; Qian et al., 2023). As outlined in Section 1, they use within-subjects studies, which introduces bias in the evaluations and limits their scalability to newer and updated language models. Most studies use standard A/B testing or a 5 or 7-point numerical rating scale (without any textual interpretations for each option) to rate the empathy-level of the responses generated by the language models being compared. They also lack comparisons with human baselines nor make any associations with the human evaluators' propensity to empathize when analyzing the ratings. 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 the 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.

This paper introduces a between-groups study accompanied by rigorous statistical analysis. It compares the responses of ChatGPT with human responses, employing a simplified rating scale, which includes textual interpretations for each option. Additionally, the study considers the tendency of evaluators to empathize. By doing so, it effectively addresses the limitations mentioned earlier.

3 The Dataset

For the study, we used dialogues from the EmpatheticDialogues dataset introduced by Rashkin et al. (2019). The dataset consists of ≈25K dialogues grounded on 32 fine-grained positive and negative emotions, ranging from basic emotions derived from biological responses (Ekman, 1992; Plutchik, 1984) to larger sets of subtle emotions derived from contextual situations (Skerry and Saxe, 2015). The authors have recruited crowd workers from Amazon Mechanical Turk (AMT)[1] and paired them to engage in a dialogue. The speaker counterpart was instructed to come up with a situation based on a given emotion and the listener counterpart was instructed to respond to these situations in an empathetic manner. Based on the sample size predicted by power analysis (in Section 4.4), we used a randomly sampled 2,000 dialogues from the dataset that are evenly distributed across the 32 emotions, for our study (see Appendix A).

The study participants were only shown the first two turns in the dialogue along with the situation description and the emotion the situation was based

[1] https://www.mturk.com

on and asked to rate the empathy of the 2nd turn i.e. the listener's response to the emotional situation described in the 1st turn. This formed the human 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 Between-Groups Study

In evaluating the empathetic capabilities of large language models, using a between-groups study design offers distinct advantages over a within-subject approach. This is especially true for reducing evaluation biases and enhancing the scalability of the evaluation framework for new or updated versions of these models. In within-subjects studies, where an individual assesses multiple model outputs, there is a risk of bias due to the *carry-over effect*. This effect occurs when continuous exposure to model responses skews the evaluator's perception of empathy. For example, a response that is moderately empathetic might be perceived more critically if the evaluator has previously encountered a highly empathetic response from a different model. Additionally, the sequence in which model outputs are presented to participants, known as the *order effect*, can influence their evaluations (Shaughnessy et al., 2000; Charness et al., 2012; Montoya, 2023). Within-subject studies are also not supportive of a "plug and play" approach when introducing outputs generated by new language models. In a within-subjects design, the introduction of a new model would require a whole new study, having to completely discard the results obtained from previous studies. This makes it challenging to integrate new models seamlessly into the evaluation framework. On the other hand, a between-subjects design, in which, different subjects are used to evaluate each model, provides the flexibility to evaluate new language models as they emerge. This approach allows for a more dynamic assessment of the evolving capabilities of language models in terms of empathy, making it a more suitable choice for this type of evaluation.

In our **between-subjects** experiment design, 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 emotional situations. The participants 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) (Spreng et al., 2009) 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 C). Later analysis revealed that the distributions of the participants' propensity to empathize was more or less distributed similarly across the three groups implying that the conditions of the participants were similar across the three groups (see Appendix J).

4.2 Task Design

We used the crowdsourcing platform Prolific (www.prolific.com) to recruit 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, CloudResearch (www.cloudresearch.com), Dynata (www.dynata.com), and Qualtrics (www.qualtrics.com) in terms of worker attention, honesty, comprehension, and reliability (Peer et al., 2022; Douglas et al., 2023). Each participant was shown 10 dialogue prompts randomly sampled from the subset of the EmpatheticDialogues dataset chosen for the experiment along with the response generated either by the 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 to the same situation. They were also shown a tutorial defining

[2] https://www.nltk.org/api/nltk.tokenize.html

GPT-4 (Vanilla):	*You are engaging in a conversation with a human. Respond to the following using on average 28 words and a maximum of 97 words.*
GPT-4 (Empathy-defined):	*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.* *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.*

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 package[2]was used to tokenize the text.

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 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 D includes all the interfaces relevant to the task.

4.3 Quality Control

To ensure a high standard of data quality, our study selectively recruited participants who were proficient in English and had a track record of at least 100 prior submissions with an approval rate exceeding 95%. Following the selection criteria, the Toronto Empathy Questionnaire (TEQ), which was used to measure the workers' propensity to empathize, contained 8 reserve scale questions. These questions were used to gauge the quality of the workers and their attentiveness to the task.

4.4 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 G.)

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-β) of 0.95, the minimal sample size required is 252 (i.e. 84 participants per group). See Appendix E 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

Figure 2 visualizes the average empathy ratings corresponding to the human's and GPT-4's responses (based on the two prompts) for all, positive, and negative emotions. We assigned values 1, 2, and 3 to *Bad*, *Okay*, and *Good* ratings, respectively, when calculating the average ratings and the standard errors. According to the results, the responses generated by GPT-4 that used the empathy-defined prompt scores the highest in terms of the average empathy rating when responding to all emotions and negative emotions. They show 11.21% and 9.61% increase in the average empathy ratings when responding to all and negative emotions, respectively, compared to human responses. The responses generated by GPT-4 that uses the generic prompt score the highest in terms of the average empathy rating when responding to positive emotions showing an increase of 13.14% compared to the human responses. 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 differences in the average ratings between GPT-4 (empathy-defined) and GPT-4 (vanilla) are relatively minor. The average empathy ratings of GPT-4 (empathy-defined) only show increases of 0.78%, -0.37%, and 1.62%, compared to GPT-4 (vanilla) for all, positive, and negative emotions, respectively. These differences are not statistically significant, as evidenced by low t-values ($t < \pm 2$) and high p-values ($p > 0.05$) presented in Table 3. The details of statistical analysis conducted using the Chi-square test of independence are included in Appendix G, which yielded similar observations.

We investigated if the empathy ratings for responses from humans and the two variants of GPT-4 varied based on the raters' empathy levels, 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 trendline slopes were 0.0022 for human responses, 0.0066 for GPT-4 (vanilla), and 0.0130 for GPT-4 (empathy-defined), with the latter two being statistically significant from zero ($p < 0.05$ and $p < 0.001$, respectively). This suggests a trend where higher empathy ratings are given to GPT-4's responses, especially GPT-4 (empathy-defined), by raters with higher empathy propensities. This reflects a tendency of highly empathetic individuals to favor the well-articulated empathetic expressions generated by GPT-4 compared to human-generated ones. The statistical significance of this trend is notably stronger for GPT-4 (empathy-defined) responses compared to human ones, as shown by t-test results ($t = 3.822$, $p < 0.001$). Appendix H contains detailed t-test values and the associated graphs for each scenario.

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

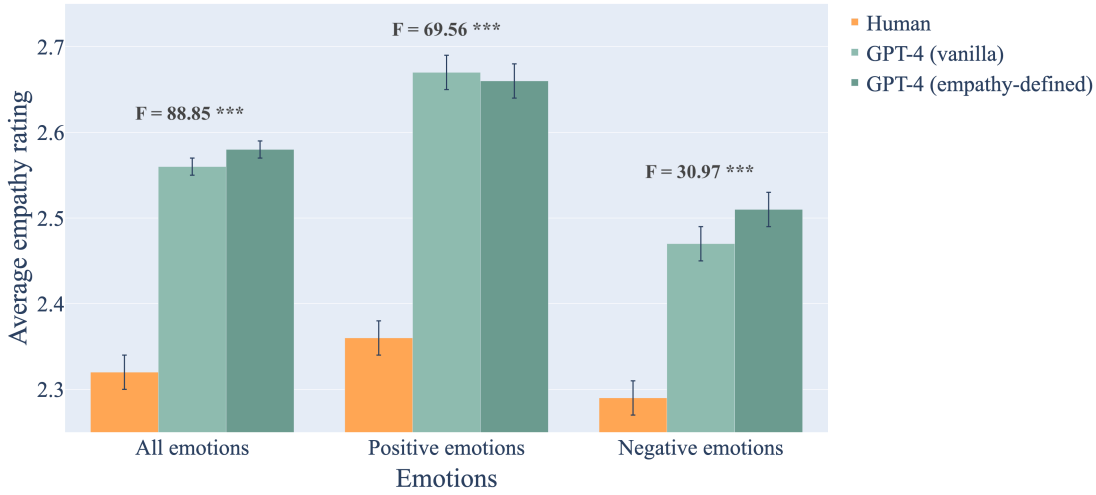


Figure 2: Average empathy ratings corresponding to the human's and GPT-4's responses (based on the two prompts for all, positive, and negative emotions. Error bars are calculated using the standard errors for each. The F-values computed using the statistical one-way ANOVA test for all, positive, and 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 F.

	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.

scenarios was that the language used by GPT-4 was comparatively 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 permission 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 L 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.

7 Discussion

This study explored the empathetic responding capability of ChatGPT powered by GPT-4 compared to a human baseline while establishing a scalable and adaptable evaluation framework for future comparisons of empathetic responses generated by new and updated versions of LLMs. It was seen that GPT-4 was capable of surpassing humans with respect to empathy by a very high statistically significant margin when responding to both positive and negative emotions. Though there was a slight improvement in the empathy ratings of GPT-4 that used an empathy-defining prompt, it was not statistically significant compared to the generic version, implying the inherent capability of GPT-4 to empathize without explicit instructions or definitions.

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	*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.*
Responses:	
Human	*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)**
GPT-4 (vanilla)	*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)**
GPT-4 (empathy-defined)	*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)**

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.

Situation	I ran a red light yesterday and got stopped by the cops. It was very traumatizing.
Emotion	Embarrassed
Speaker utterance	*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.*
Responses:	
Human	*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)**
GPT-4 (vanilla)	*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)**
GPT-4 (empathy-defined)	*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)**

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.

But it was seen that GPT-4 when prompted to incorporate a clear understanding of empathy with respect to its cognitive, affective, and compassionate correlates, significantly enhances the responses' alignment with the expectations of highly empathetic individuals, compared to human responses. By qualitatively analyzing GPT-4's responses rated *Bad* by the participants, we could observe that GPT-4 has the potential to improve the empathetic quality of its responses by using less formal language and adopting more subtle and non-directive ways of advising and providing suggestions.

Since this study was conducted with a large sample size, which is sufficiently above the minimal sample size required to accurately detect a significant difference between the ratings of the three groups of responses, the results obtained can be stated reliable and generalizable to a larger population. The balanced representation of participants across gender and age groups and the balanced distribution of dialogue prompts across 32 fine-grained positive and negative emotions further reinforce the credibility and applicability of our results.

EmpatheticDialogues (Rashkin et al., 2019) is a widely used state-of-the-art dataset used for training and benchmarking a number of dialogue generation models. However, the results of this study raise questions about the quality of such crowd-sourced datasets with respect to empathy. In this case, synthetic data generated by LLMs such as GPT-4 outperform human-generated data by a statistically significant margin, implying that such synthetically generated data has the potential to be used for training or fine-tuning other models.

The capacity of ChatGPT to empathize with humans further opens up the possibility for a myriad of applications. It can be used as a companion and an empathetic listener for people suffering from loneliness or as a personalised life coach offering empathetic support and motivation for personal development. It can also serve to assist during distressful situations, by actively listening and providing compassionate support during crisis situations.

Overall, this study contributed to the understanding of empathy in ChatGPT-generated responses to a variety of positive and negative emotions, compared to those generated by humans. The between-subjects study design along with the release of the data and the code of the evaluation experiment, presents a framework for evaluating empathy levels of the responses generated by newer and updated versions of LLMs in forthcoming studies.

8 Limitations

This study was limited to evaluating the empathetic quality of human and GPT-4 generated responses from a broad perspective, without considering the varied socio-cultural backgrounds of the participants. The participants in our study hail from a variety of different countries and ethnicities, as detailed in Appendix I. Their evaluations of the empathetic quality of the responses were collectively analyzed as a whole. Several studies point to the fact that the perception of empathy differs across cultures (Birkett, 2014; Chopik et al., 2017; Cassels et al., 2010). Future work may further investigate whether there are any differences in the way people from different countries and cultural backgrounds perceive empathetic responses generated by humans and large language models.

9 Ethics Statement

Data Usage: We used a subset of the dialogues from the EmpatheticDialogues dataset (Rashkin et al., 2019) in our study, which is a publicly available dataset containing ethically sourced dialogues. The dialogues used in the study did not contain any personally identifiable information or any sensitive content and were used in compliance with the dataset's release license terms (CC BY-NC 4.0). The new artifacts produced in this paper, including the responses generated by the two variants of GPT-4 and the empathy ratings of the participants, will also be released to the public under the same CC BY-NC 4.0 license terms, allowing other researchers to tweak, and build upon this work for non-commercial purposes. This allows for comparison of the empathetic quality of responses generated by newer and updated versions of the state-of-the-art large language models in future studies.

Human Evaluation: Since the responses evaluated were in the English language, we recruited only workers who were fluent in the English language from the Prolific crowdsourcing platform. They were paid €2.25 for rating 10 responses that required on average ≈11 minutes to complete. Thus, the amount paid to the human raters was ≈1.4 times above the hourly rate recommended as *Good* by Prolific (€9 per hour). All the participants were informed about the nature of the study and their role in it before exposing them to the actual task. The participants had the freedom to exit working on the task after reading the initial description. Random subsets of dialogue prompt-response pairs used in the study were manually inspected to ensure that the tasks assigned to the crowdworkers were not psychologically distressing or offensive.

Transparency and Reproducibility: The dialogue prompt-response pairs that were subjected to evaluation along with the participants' evaluations of these responses will be released publicly to ensure the transparency and reproducibility of our study.

Human-like Chatbots: Lastly, there are some ethical implications behind making artificial conversational agents human-like. By demonstrating understanding and responding to human emotions, empathetic chatbots are perceived as more human-like by users (Goetz et al., 2003; Stroessner and Benitez, 2019; Svikhnushina and Pu, 2022). However, there can be some ethical implications surrounding such chatbots. One risk is that users may get emotionally attached to these chatbots misinterpreting chatbots' responses as being capable of actually understanding their emotional needs (Vanderlyn et al., 2021). Such chatbots can cause harm if they are designed in manipulative ways to elicit certain behavior out of users such as revealing their personal or financial information. Thus, actions should be taken to ensure transparency about the artificial nature of these dialogue systems and be considerate of the risk of causing harm through emotional manipulation.

References

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*.

Ahmed Belkhir and Fatiha Sadat. 2023. Beyond information: Is chatgpt empathetic enough? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 159–169.

Melissa Birkett. 2014. Self-compassion and empathy across cultures: Comparison of young adults in china and the united states. *International Journal of Research Studies in Psychology*, 3(3):25–34.

Dana Brin, Vera Sorin, Akhil Vaid, Ali Soroush, Benjamin S Glicksberg, Alexander W Charney, Girish Nadkarni, and Eyal Klang. 2023. Comparing chatgpt and gpt-4 performance in usmle soft skill assessments. *Scientific Reports*, 13(1):16492.

Tracy G Cassels, Sherilynn Chan, and Winnie Chung. 2010. The role of culture in affective empathy: Cultural and bicultural differences. *Journal of Cognition and Culture*, 10(3-4):309–326.

Gary Charness, Uri Gneezy, and Michael A Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, 81(1):1–8.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

William J Chopik, Ed O'Brien, and Sara H Konrath. 2017. Differences in empathic concern and perspective taking across 63 countries. *Journal of Cross-Cultural Psychology*, 48(1):23–38.

Jacob Cohen. 1992. Quantitative methods in psychology: A power primer. *Psychol. Bull.*, 112:1155–1159.

Jean Decety, Claus Lamm, et al. 2006. Human empathy through the lens of social neuroscience. *The scientific World journal*, 6:1146–1163.

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. *Plos one*, 18(3):e0279720.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman. 2004. Emotions revealed. *Bmj*, 328(Suppl S5).

Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14:1199058.

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. *Behavior research methods*, 41(4):1149–1160.

Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. 2023. Reasoning before responding: Integrating commonsense-based causality explanation for empathetic response generation. *arXiv preprint arXiv:2308.00085*.

Jennifer Goetz, Sara Kiesler, and Aaron Powers. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pages 55–60. Ieee.

Flora Ioannidou and Vaya Konstantikaki. 2008. Empathy and emotional intelligence: What is it really about? *International Journal of caring sciences*, 1(3):118.

Hyun Kang. 2021. Sample size determination and power analysis using the g* power software. *Journal of educational evaluation for health professions*, 18.

Alexandra Koufouli and Marieke S Tollenaar. 2016. Empathy and emotional awareness: An interdisciplinary perspective.

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. *Journal of personality assessment*, 55(1-2):124–134.

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 *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

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 *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Siru Liu, Allison B McCoy, Aileen P Wright, Babatunde Carew, Julian Z Genkins, Sean S Huang, Josh F Peterson, Bryan Steitz, and Adam Wright. 2023. Leveraging large language models for generating responses to patient messages. *medRxiv*, pages 2023–07.

Yuping Liu-Thompkins, Shintaro Okazaki, and Hairong Li. 2022. Artificial empathy in marketing interactions: Bridging the human-ai gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, 50(6):1198–1218.

Amanda K Montoya. 2023. Selecting a within-or between-subject design for mediation: Validity, causality, and statistical power. *Multivariate Behavioral Research*, 58(3):616–636.

TB Moyers, JK Manuel, D Ernst, T Moyers, J Manuel, D Ernst, and C Fortini. 2014. Motivational interviewing treatment integrity coding manual 4.1 (miti 4.1). *Unpublished manual*.

Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, page 1.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Philip A Powell and Jennifer Roberts. 2017. Situational determinants of cognitive, affective, and compassionate empathy in naturalistic digital interactions. *Computers in Human Behavior*, 68:137–148.

Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Carolien Rieffe and Marina Camodeca. 2016. Empathy in adolescence: Relations with emotion awareness and social roles. *British journal of developmental psychology*, 34(3):340–353.

Kristina Schaaff, Caroline Reinig, and Tim Schlippe. 2023. Exploring chatgpt's empathic abilities. *arXiv preprint arXiv:2308.03527*.

Miriam Schuler, Sebastian Mohnke, and Henrik Walter. 2016. The neurological basis of empathy and mimicry. *Emotional mimicry in social context*, pages 192–221.

Doug Semenick. 1990. Tests and measurements: The t-test. *Strength & Conditioning Journal*, 12(1):36–37.

John J Shaughnessy, Eugene B Zechmeister, and Jeanne S Zechmeister. 2000. *Research methods in psychology*. McGraw-Hill.

Amy E Skerry and Rebecca Saxe. 2015. Neural representations of emotion are organized around abstract event features. *Current biology*, 25(15):1945–1954.

R Nathan Spreng, Margaret C McKinnon, Raymond A Mar, and Brian Levine. 2009. The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment*, 91(1):62–71.

Lars St, Svante Wold, et al. 1989. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272.

Steven J Stroessner and Jonathan Benitez. 2019. The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. *International Journal of Social Robotics*, 11:305–315.

Ekaterina Svikhnushina and Pearl Pu. 2022. Peace: A model of key social and emotional qualities of conversational chatbots. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–29.

Lindsey Vanderlyn, Gianna Weber, Michael Neumann, Dirk Väth, Sarina Meyer, and Ngoc Thang Vu. 2021. "it seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 44–57.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2023. Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5411–5432, Toronto, Canada. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.

A Distribution of Emotions

Figure 3 shows the distribution of the dialogue prompt-response pairs sampled from the EmpatheticDialogues dataset across the 32 positive and negative emotions. Table 6 shows the counts and the percentages of dialogue prompt-response pairs in the dataset corresponding to each emotion. It can be noted that the prompt-response pairs are more or less equally distributed across the 32 emotions.

B Experiment Parameters

The following are the key parameters used when querying OpenAI's GPT-4 to generate responses to the emotional dialogue prompts: `model=gpt-4`; `temperature=0`; `top_p=1`; `frequency_penalty=0`; and `presence_penalty=0`. All the experiments were conducted on a MacBook Pro machine having a 2.3 GHz Quad-Core Intel Core i5 processor and 8 GB memory.

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

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.

C Toronto Empathy Questionnaire

Table 7 shows the questions in the Toronto Empathy Questionnaire (TEQ) (Spreng et al., 2009) that were asked from the participants. Responses to the questions were scored according to the following scale for positively worded questions: Never = 0; Rarely = 1; Sometimes = 2; Often = 3; Always = 4. The negatively worded questions indicated are reverse-scored. Scores were summed to derive the evaluator's propensity to empathize.

D Task Interfaces

Figures 4, 5, 6 and 7 show the task interfaces corresponding to the description of the task, the tutorial presented to the crowd workers, the Toronto Empathy Questionnaire, and the response rating task, respectively.

E Determing the Effect Size

Jacob Cohen, a notable psychologist and statistician, established benchmarks for interpreting effect sizes in statistical tests, including analysis of variance (ANOVA) and chi-square tests (Cohen, 1992). These benchmarks serve as a general framework for evaluating the practical significance of the effects observed in these tests. For ANOVA, Cohen's F is a commonly used measure of effect size, representing the standardized difference between means. 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

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.

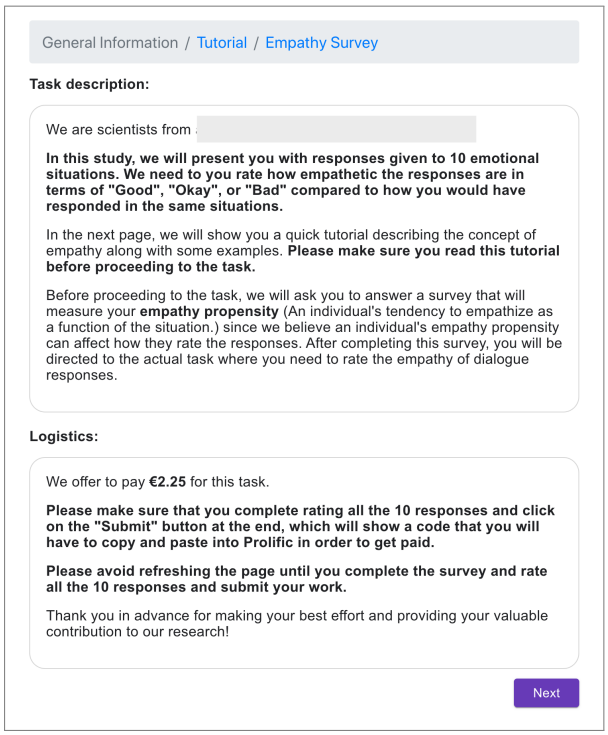


Figure 4: The description of the task.

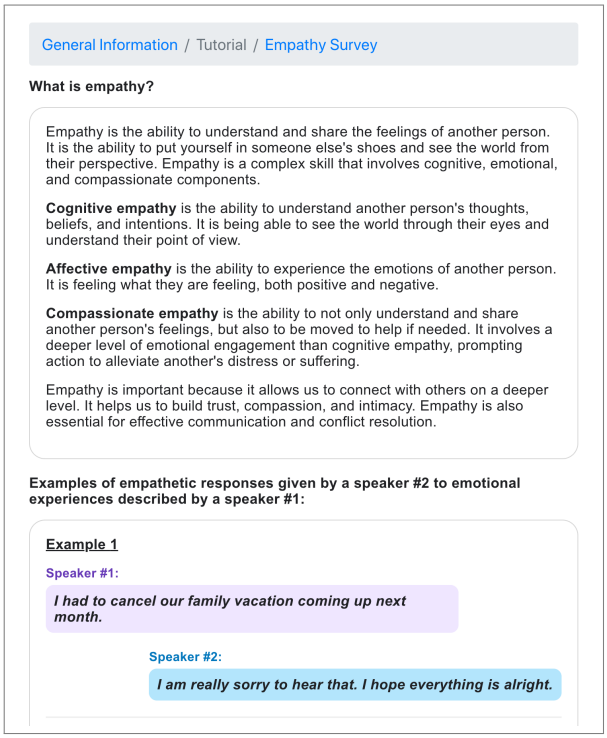


Figure 5: The tutorial.

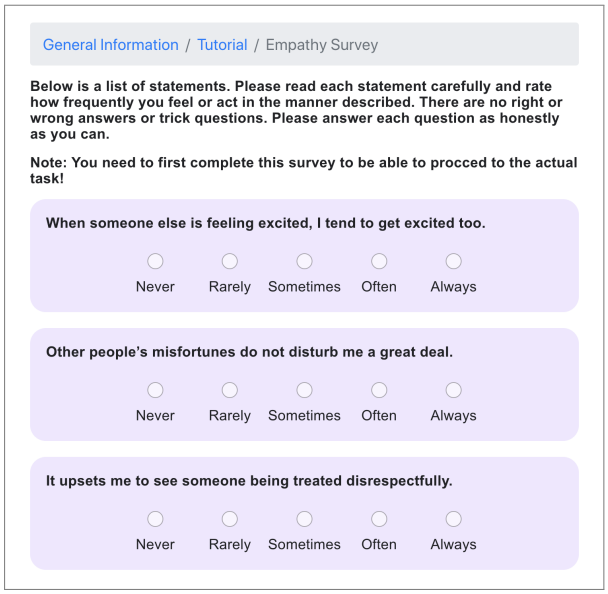


Figure 6: The Toronto Empathy Questionnaire.

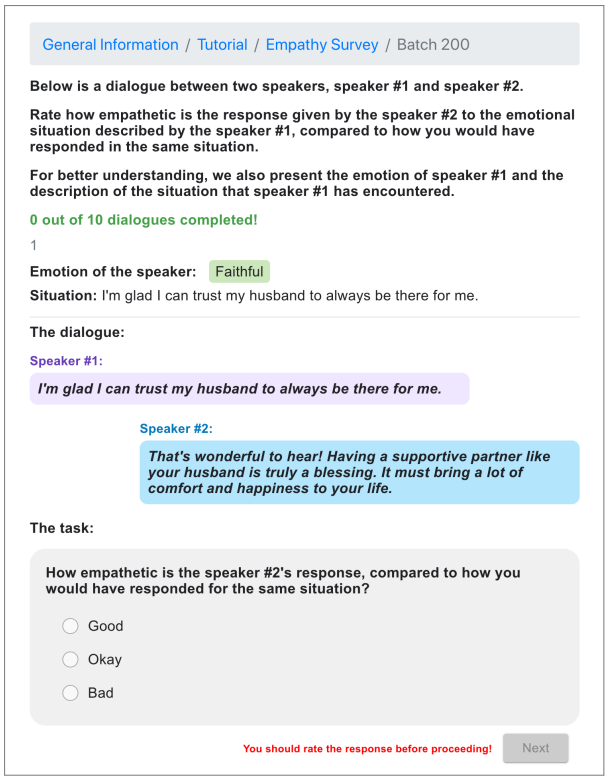


Figure 7: The task interface for rating responses in terms of empathy.

effect size can sensitively detect differences in empathy levels between the responses of humans and ChatGPT, 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.

F 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 variants of GPT-4 are denoted in Table 8. Table 9 denotes the statistical one-way ANOVA test results corresponding to each of the 32 emotions the dialogue prompts used for testing are based on. Figure 8 visualizes the average empathy ratings corresponding to the human's and the two variants of GPT-4's responses across the 32 positive and negative emotions. 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. Nevertheless, 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 across most emotions.

G 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-β) 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*,

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.

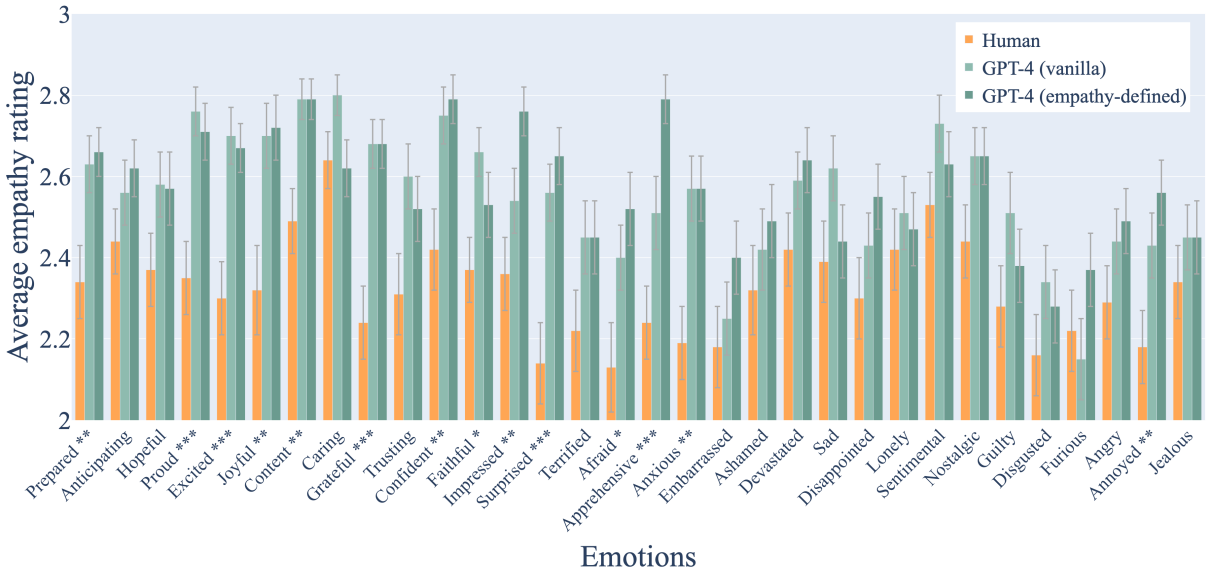


Figure 8: Average empathy ratings corresponding to the human's and the two versions of GPT-4's responses across the 32 positive and negative emotions. Error bars are calculated using the standard errors for each. We have indicated ***, **, and * in front of the emotion in the x-axis to indicate results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

and *Good* empathy ratings scored by the humans' and the two variants of GPT-4's responses. They are visualized 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.

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

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.

I Participants' Demographics

Figures 11 and 12 respectively show the distributions of the countries of residence and the ethnicities of the participants who rated the three groups of responses. It could be observed that though there are imbalances across the countries and the ethnicities represented in the participants' pool, these demographics are similar across the three groups of participants. This allows control for factors other than the independent variable influencing the results of the study and fair comparison of response ratings across the three groups.

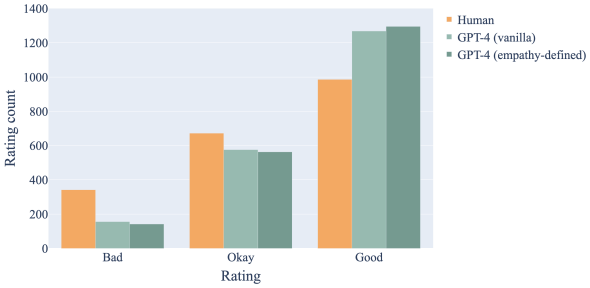
J Distribution of Participants' Empathy Propensities

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 inequal distribution of empathy propensities across the three groups.

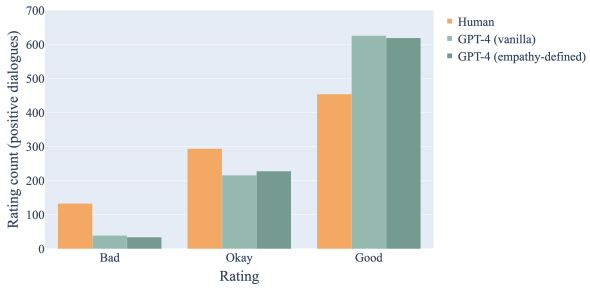
K Quality Analysis

Figure 14 shows the number of reverse scale questions in the TEQ that were marked incorrect by the participants rating the three response groups. It was observed that ≈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.

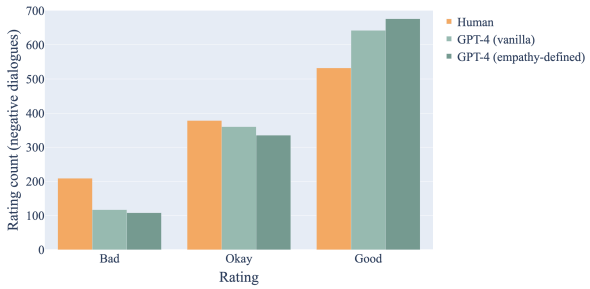
Further, Figure 15 shows the histogram of times (in minutes) taken to complete the study. On average it took 11 minutes and 18 seconds to complete rating 10 responses, which was close to the average completion time of 15 minutes that we estimated



(a) All emotions



(b) Positive emotions



(c) Negative emotions

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.

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.

L MITI Classification

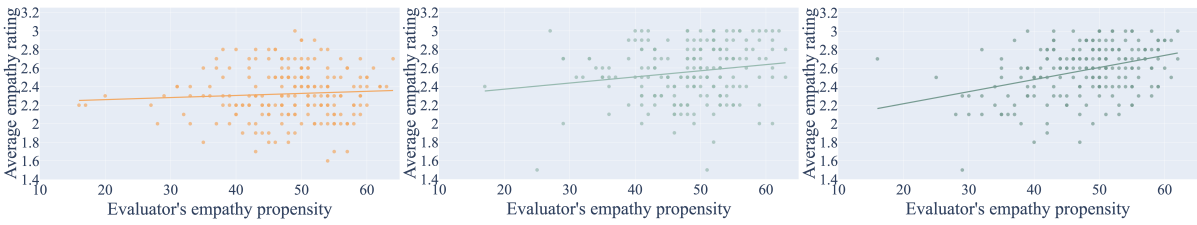
Table 13 shows the percentage of labels from the Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2014) present in responses generated by humans and the two versions of GPT-4 that were rated *Bad*, *Okay*, and *Good*, as classified by the classifier introduced by Welivita and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

	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



(a) Human Responses (b) GPT-4 (Vanilla) Responses (c) GPT-4 (Empathy-defined) Responses

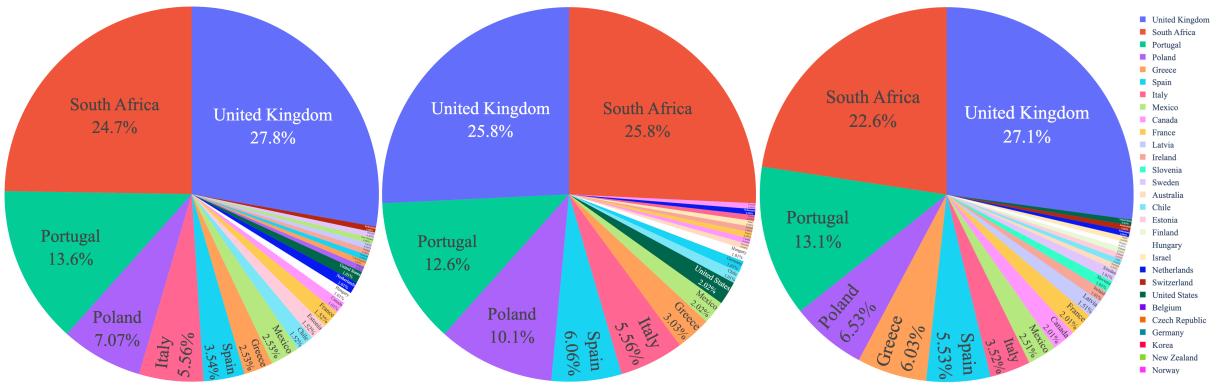
Figure 10: The scatter plots of average empathy ratings against evaluator's propensity to empathize as measured by the Toronto Empathy Questionnaire. Ordinary Least Squares (OLS) regression trendlines are also plotted.

	slope	std err	t-value	p-value
Human	0.0022	0.002	0.967	0.335
GPT-4 (vanilla)	0.0066	0.003	**2.461**	**0.015***
GPT-4 (empathy-defined)	0.0130	0.002	**5.653**	**0.000*****

t-test statistics for the difference in 2 independent slopes

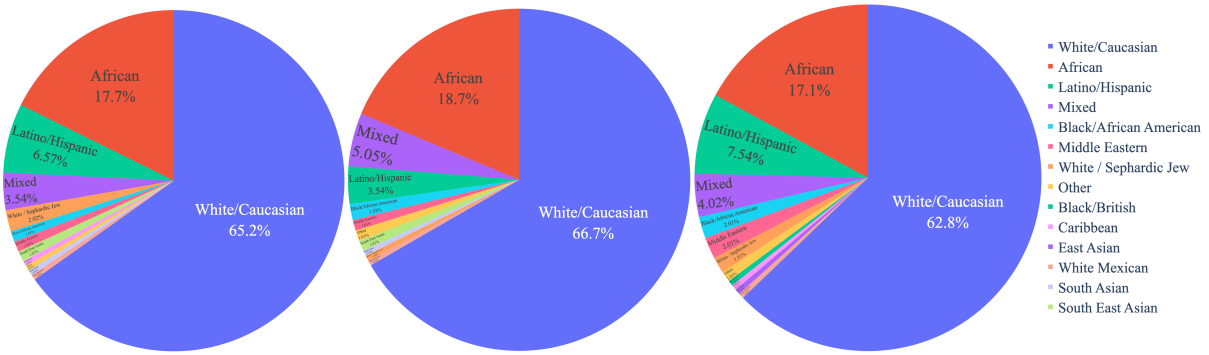
	t-value	p-value
Human vs GPT-4 (vanilla)	1.215	0.225 (p > 0.05)
Human vs GPT-4 (empathy-defined)	**3.822**	**0.00015*****
GPT-4 (vanilla) vs GPT-4 (empathy-defined)	1.783	0.075 (p > 0.05)

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.



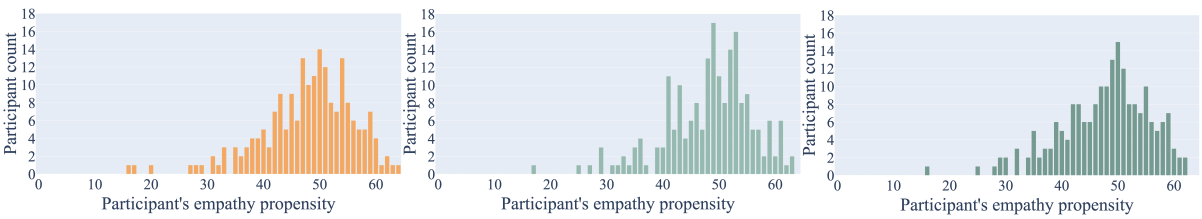
(a) Human (b) GPT-4 (vanilla) (c) GPT-4 (empathy-defined)

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



(a) Human (b) GPT-4 (vanilla) (c) GPT-4 (empathy-defined)

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



(a) Human response raters (b) GPT-4 (vanilla) responses raters (c) GPT-4 (empathy-defined) responses raters

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

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*.

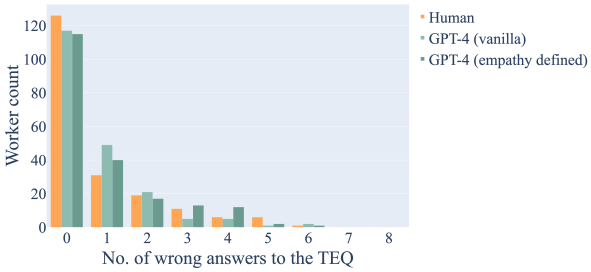


Figure 14: The number of reverse scale questions in the TEQ that were marked wrong by the participants rating the three response groups.

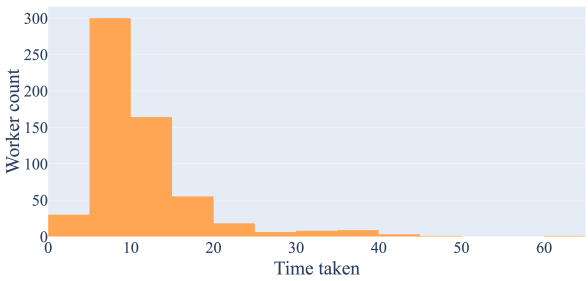
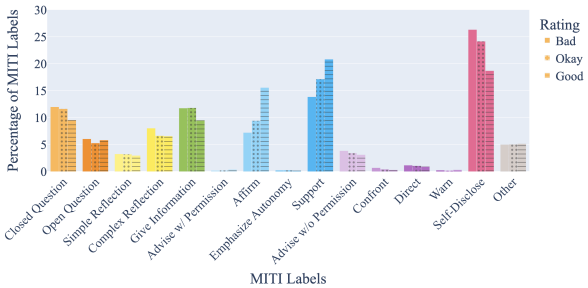
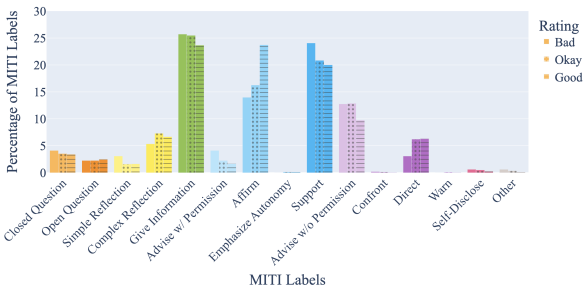


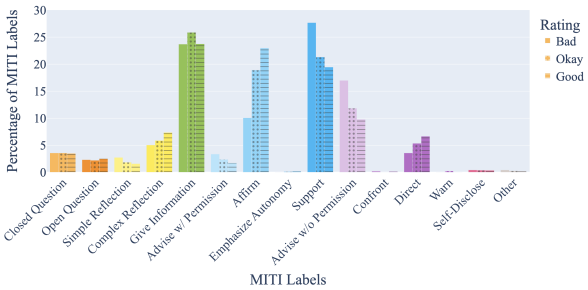
Figure 15: The histogram of times taken to complete the task by all participants.



(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*.