Are Large Language Models More Empathetic than Humans?

Anuradha Welivita, and Pearl Pu
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
Switzerland
{kalpani.welivita,pearl.pu}@epfl.ch

Abstract

With the emergence of large language models (LLMs), investigating if they can surpass humans in areas such as emotion recognition and empathetic responding has become a focal point of research. This paper presents a comprehensive study exploring the empathetic responding capabilities of four state-of-the-art LLMs: *GPT-4*, *LLaMA-2-70B-Chat*, *Gemini-1.0-Pro*, and *Mixtral-8x7B-Instruct* in comparison to a human baseline. We engaged 1,000 participants in a between-subjects user study, assessing the empathetic quality of responses generated by humans and the four LLMs to 2,000 emotional dialogue prompts meticulously selected to cover a broad spectrum of 32 distinct positive and negative emotions. Our findings reveal a statistically significant superiority of the empathetic responding capability of LLMs over humans. GPT-4 emerged as the most empathetic, marking $\approx$31% increase in responses rated as *Good* compared to the human benchmark. It was followed by LLaMA-2, Mixtral-8x7B, and Gemini-Pro, which showed increases of approximately 24%, 21%, and 10% in *Good* ratings, respectively. We further analyzed the response ratings at a finer granularity and discovered that some LLMs are significantly better at responding to specific emotions compared to others. The suggested evaluation framework offers a scalable and adaptable approach for assessing the empathy of new LLMs, avoiding the need to replicate this study's findings in future research.

1 Introduction

This era is marked by massive developments in artificial intelligence (AI), especially large language models (LLMs). They have exhibited performance exceeding humans across a variety of traditional language processing tasks such as question answering, text summarization, and commonsense reasoning (Laskar et al., 2023; Ziyu et al., 2023). While there are public benchmarks and evaluation frameworks to evaluate LLMs' performance on these tasks, there is a lack of such resources to evaluate LLMs' ability to generate empathetic responses. Empathetic response generation requires generating replies that are not only contextually relevant and coherent but also demonstrate understanding, compassion, and emotional support towards the user's situation and feelings (Rashkin et al., 2019). This is particularly challenging as empathy, being a deeply nuanced 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 multifaceted construct, encompassing cognitive, affective, and compassionate counterparts (Ekman, 2004; Decety et al., 2006; Powell and Roberts, 2017). Each component plays a crucial role in holistic empathetic engagement. Cognitive empathy is understanding and accurately identifying others' feelings. Affective empathy is sharing the other person's emotions. Compassionate empathy is taking action to help the other person deal with their emotions. Empathy is a key component in making artificial conversational agents human-like, which fosters trust and rapport with the user (Liu-Thompkins et al., 2022) and helps to increase people's adoption of this technology (Goetz et al., 2003; Stroessner and Benitez, 2019; Svikhnushina and Pu, 2022). So, evaluating the empathetic capabilities of LLMs that power artificial conversational agents plays a big role in deciding people's willingness to use this technology.

Existing studies that evaluate the empathetic capabilities of LLMs encompass major limitations. Most of them use automatic evaluation metrics that do not necessarily correlate with human perceptions of empathy (Belkhir and Sadat, 2023; Loh and Raamkumar, 2023). Most evaluations are focused on the healthcare domain involving a lot of negative emotions (Chen et al., 2023; Ayers et al., 2023;

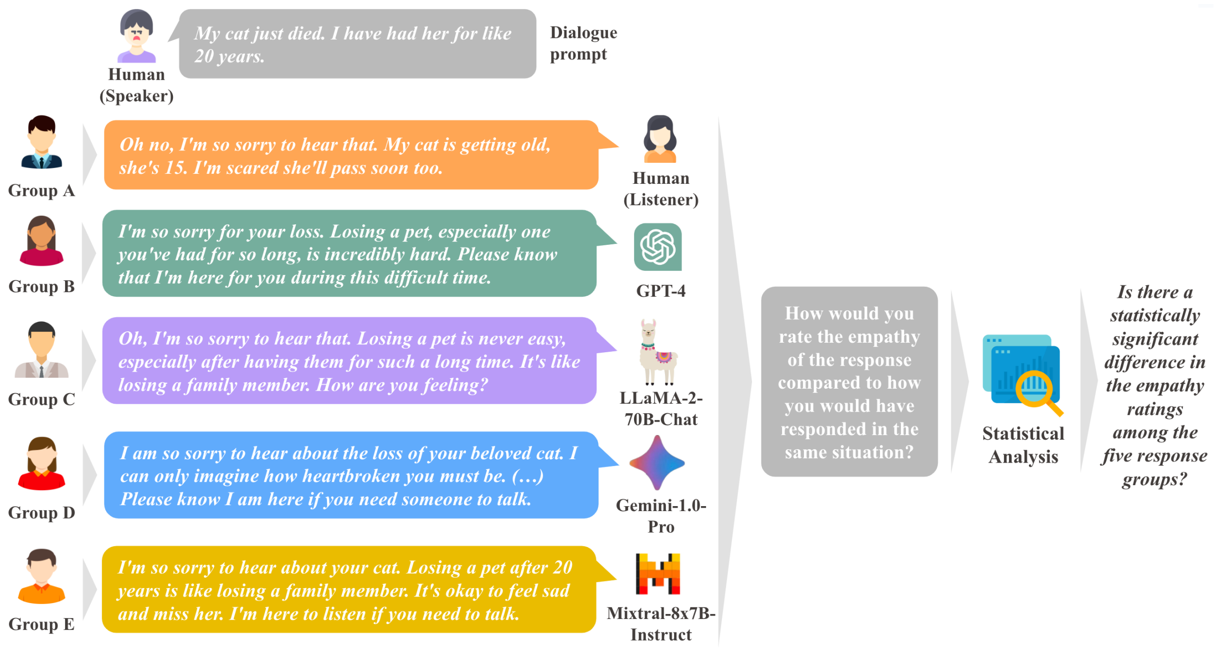


Figure 1: Between-subjects experiment design to evaluate the level of empathy demonstrated by LLMs compared to a human baseline when responding to emotional situations.

Liu et al., 2023). But empathy plays an important role in responding to both positive and negative emotions encountered in daily conversations. Also, most studies investigate LLMs' ability to respond in general to emotions (which are mostly coarse-grained) as a whole, without analyzing them at a finer level (Lee et al., 2024; Zhao et al., 2023; Qian et al., 2023; Lee et al., 2022; Fu et al., 2023; Loh and Raamkumar, 2023). This makes it impossible to observe any variability in LLMs' performance when responding to diverse emotions. Last, but most importantly, all studies we came across used **within-subjects** study designs where the same participant evaluated responses generated by different models (Lee et al., 2024, 2022; Ayers et al., 2023; Fu et al., 2023; Zhao et al., 2023; Qian et al., 2023). In addition to introducing evaluation biases caused due to over-exposure to different model outputs and the order they are shown to the participants, this type of study design makes the evaluation approach not scalable to incorporate new and updated LLMs.

Addressing the above limitations, we designed a **between-subjects** user study, recruiting 1,000 people from the crowdsourcing platform Prolific (`www.prolific.com`), in which 200 participants each were assigned to rate responses generated by humans and four state-of-the-art LLMs: *GPT-4* (OpenAI, 2023), *LLaMA-2-70B-Chat* (Touvron et al., 2023), *Gemini-1.0-Pro* (Pichai, 2023), and *Mixtral-8x7B-Instruct* (MistralAI, 2024) (see Figure 1). We use 2,000 emotional dialogue prompts from the state-of-the-art EmpatheticDialogues dataset (Rashkin et al., 2019), which contains chit-chat oriented human-human conversations, to form the human baseline required for our study as well as to initiate responses from the LLMs. We carefully selected the dialogue prompts to be equally distributed over a broad spectrum of 32 positive and negative emotions so that we can analyze whether there are any significant differences between humans and LLMs when responding to such distinct emotions. We prompt the four LLMs to generate a response to a given dialogue prompt, with instructions defining empathy in terms of its cognitive, affective, and compassionate counterparts. We adopt a simple and straightforward evaluation scale to gauge the empathy level in these responses. We perform rigorous statistical analysis to identify whether there are any statistically significant differences between the empathy ratings of humans and the four LLMs when responding to positive and negative emotional situations. We further delve into each finer emotion category and observe whether there are any significant differences in the way humans and LLMs respond to these individual emotions. Due to the careful and thorough design, this evaluation framework provides a robust and extensible solution to evaluate the empathetic quality of emerging LLMs without having to replicate the current study.

2 Literature Review

Different studies use different approaches to evaluate empathy in LLMs, most of which encompass automatic evaluation criteria. For example, Loh and Raamkumar (2023) investigated the capability of five state-of-the-art LLMs including GPT-3.5, GPT-4, PaLM-2—the predecessor of Gemini, and Vicuna—based on LLaMA-1 to generate empathetic responses using ≈2,550 dialogue prompts from the EmpatheticDialogues dataset. They utilized three automatic empathy-related evaluation metrics: 1) Emotional Reactions (indicative of affective empathy); 2) Interpretations (indicative of cognitive empathy); and 3) Explorations (indicative of cognitive empathy) (Sharma et al., 2020) and discovered that LLMs' responses scored higher across the three metrics compared to those generated by traditional dialogue systems and humans. However, their evaluation is purely based on automatic evaluation, which does not necessarily correlate with how human users perceive the responses generated by the LLMs. A user-based evaluation could either validate the above observations or bring forth vastly different results. Belkhir and Sadat (2023) analyzed GPT-3.5's ability to produce empathetic responses, using precision, accuracy, and recall related to the emotion conveyed in the responses. However, empathetic communication does not always have to be emotional; it can sometimes be more neutral, focusing on specific intentions, as noted by Welivita and Pu (2020). This raises questions about the appropriateness of such metrics for evaluating empathetic responses.

Some studies have utilized questionnaires and psychological scales that are primarily designed to measure the empathy levels of humans on LLMs without considering their applicability. Schaaff et al. (2023) used standardized questionnaires from psychology such as Interpersonal Reactivity Index (Davis, 1980), Empathy Quotient (Lawrence et al., 2004), and Toronto Empathy Questionnaire (Spreng et al., 2009) to assess the level of empathy exhibited by GPT-3.5 compared to humans. Elyoseph et al. (2023) utilized the Levels of Emotional Awareness Scale (LEAS) (Lane et al., 1990) to evaluate GPT-3.5's ability to identify and describe emotions compared to the general population. But the applicability of this type of scales designed to evaluate humans' emotion understanding and empathy levels on LLMs is debatable.

Research evaluating the empathetic responding ability of LLMs using human evaluators employ **within-subjects** designs, where the same participant evaluates different model outputs (Lee et al., 2024, 2022; Ayers et al., 2023; Fu et al., 2023; Zhao et al., 2023; Qian et al., 2023). For instance, Lee et al. (2024), conducted a within-subjects study with 200 participants evaluating responses generated by humans, GPT-4, LLaMA-2, and Mixtral for 120 posts from Reddit. As elaborated in Section 1 this type of study is not extensible to newer and updated LLMs, requiring to reconduct the study from scratch when new LLMs emerge. Moreover, the relatively small sample size used fails to provide sufficient data to arrive at robust statistical conclusions. The above studies utilize standard A/B testing or a 5 or 7-point numerical rating scale (sometimes without any textual interpretations for each option) to rate the empathy-level of the responses generated by the LLMs. While effective in certain contexts, these methods have notable disadvantages. The rapid evolution of LLMs makes findings from A/B tests quickly outdated. The interpretation of scale points can vary widely among individuals, making it difficult to achieve consistent measurements across diverse participant groups. Most studies also lack a human baseline for comparison. This lack of a common ground to evaluate the empathetic responding capabilities of LLMs makes the evaluation complex and often not fully representative of how effective LLMs are in real-world interactions.

3 The Dataset

To conduct the study, we used dialogues from the state-of-the-art EmpatheticDialogues dataset (Rashkin et al., 2019), which consists of ≈25K dialogues spanning 32 fine-grained positive and negative emotions, selected from multiple annotation schemes, 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 dialogues in this dataset are curated by recruiting crowd workers from Amazon Mechanical Turk (AMT)[1]. The workers were paired together and were asked to role-play a dialogue, one person acting as the speaker and the other as the listener. The speaker was asked to pick an emotion, come up with a situation based on the chosen emotion, and start a conversation. The listener who is unaware of the emotion or the situation was asked to respond to

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

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: The set of instructions used to prompt the large language models to generate empathetic responses.

the speaker. Based on the sample size predicted by power analysis (in Section 4.5), we used randomly sampled 2,000 dialogues from this dataset, which are more or less equally distributed across the 32 emotions for our study (see Appendix A). Though the dialogues spanned up to a maximum of 8 turns, for simplicity, we selected only the first two dialogue turns along with the emotion the dialogues were based on and the situation description. This formed the human baseline for our study.

In one of our previous studies, we used two different prompts to instruct the LLM GPT-4 to generate responses given the 1st turn of the dialogues. The first one was a very basic prompt that did not define the concept of empathy nor explicitly asked the model to generate an empathetic response. The second prompt defined the concept of empathy concerning its cognitive, affective, and compassionate counterparts and explicitly asked the model to respond in an empathetic manner to the given dialogue utterance. We observed that the one that utilized the second prompt outperformed the one that utilized the basic prompt as well as the human baseline with respect to the empathy ratings allocated by the study participants. Thus, we utilized the same empathy-defining instructions to prompt the LLMs compared in this study to generate responses. Table 1 denotes this set of instructions.

For comparison with the human baseline, we use responses generated by four state-of-the-art LLMs: GPT-4 (OpenAI, 2023); LLaMA-2-70B-Chat (Touvron et al., 2023); Gemini-1.0-Pro (Pichai, 2023); and Mixtral-8x7B-Instruct (MistralAI, 2024). Details regarding the four LLMs are in Appendix B. We first manually inspected a random set of responses generated by a large group of LLMs that included other LLMs such as PaLM-2 (Anil et al., 2023), ChatGLM-3 (Zeng et al., 2022), Vicuna-180B (Chiang et al., 2023), and Falcon-40B-Instruct (Almazrouei et al., 2023) and selected the LLMs that seemingly generated the highest quality responses to evaluate against the human baseline. Table 2 denotes the statistics of all the prompt-response pairs evaluated in the study.

Model	**Avg # tokens**	**Max # tokens**
Dialogue prompt	23.24	143
Responses:		
Human	28.37	97
GPT-4	34.94	65
LLaMA-2-Chat-70B	53.45	90
Gemini-1.0-Pro	53.99	93
Mixtral-7x8B-Instruct	61.35	95

Table 2: Statistics of the dialogue prompts and responses used for the study. The dialogue prompt here means the first dialogue utterance that initiates a reply. NLTK's tokenized package[2] was used to tokenize the text.

4 Experiment Design

4.1 Between-Subjects vs Within-Subjects

In our experiment design, which was structured as a **between-subjects study**, participants were divided into five groups. The first group assessed the empathetic quality of responses from humans to both positive and negative emotional scenarios. Each of the other four groups were assigned to evaluate empathy in responses generated by one of the four LLMs to the same emotional dialogue scenarios. This type of study design offers distinct advantages over a **within-subjects approach**. In within-subjects studies, as one person evaluates two or more model outputs, the evaluator's perception of empathy could be distorted by overexpo-

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

sure to model outputs resulting in a bias in their evaluations—commonly known as the *carry-over effect*. For example, an averagely empathetic response may be judged more harshly by the evaluator if they have already seen an extremely empathetic response given by another model. This could also lead to *order effects*, as the sequence in which model outputs are presented to the workers may influence how they assess the responses. (Shaughnessy et al., 2000; Charness et al., 2012; Montoya, 2023). Within-subjects studies also cannot accommodate seamless integration of outputs from newly developed language models. Such a study design necessitates starting from scratch every time a new model is introduced, making prior results obsolete. Conversely, a between-subjects study design, which employs different participants for assessing each model, offers the adaptability needed to evaluate emerging language models. This method facilitates an ongoing evaluation of language models' evolving empathy capabilities, making it a desirable option for such assessments.

4.2 Design of the Rating Scale

When designing the rating scale for the human raters to rate the empathetic quality of the responses, we gave more weight to the human-centric aspect of the design rather than complexity, opting for a simple rating scale comprising of options *Bad*, *Okay*, and *Good*. Psychological research suggests that humans naturally categorize experiences in a tripartite fashion (*Good*, *Bad*, *Neutral*) (Heise, 1970), which aligns well with our 3-point scale. This makes the scale feel more intuitive and natural to users, helping to minimize their cognitive load and enhancing the clarity and effectiveness of the rating process. It also helps to avoid the central tendency bias (Douven, 2018) seen in 5 or 7-point rating scales where the user tends to avoid the extremes when rating. As the results in Figures 2 and 3 show, we observe no central tendency bias in the human ratings collected utilizing this simpler scale.

4.3 Task Design

The five groups of participants for the study were recruited through the Prolific crowdsourcing platform (`www.prolific.com`). Past research has indicated that Prolific outperforms other crowdsourcing platforms like AMT, CloudResearch (`www.cloudresearch.com`), Dynata (`www.dynata.com`), and Qualtrics (`www.qualtrics.com`) in aspects such as worker attentiveness, integrity, understanding, and dependability (Peer et al., 2022; Douglas et al., 2023). Participants in the five groups were balanced across demographic criteria: gender (male and female); and age group (young adulthood [19 - 25 years]; middle adulthood [26 - 45 years]; late adulthood [46 - 64 years]; and older adulthood [65 years and above]). Additionally, a survey based on the Toronto Empathy Questionnaire (TEQ) (Spreng et al., 2009) measured the empathy propensity of each participant, i.e., their natural predisposition to empathize with others. Subsequent analysis indicated that the inclination towards empathy was comparably distributed among the five groups, suggesting that participant conditions were uniformly matched across the board (see Appendix J). Each participant evaluated randomly chosen 10 dialogue responses generated by the same model. The source of the responses, whether from a human or an LLM, was unknown to the participants. They were tasked with rating the empathy of the responses as either *Bad*, *Okay*, or *Good*, relative to how they would have responded in similar situations. Furthermore, participants were introduced to the concept of empathy through a tutorial that covered its cognitive, affective, and compassionate dimensions. This tutorial was identical to the one used to prompt the LLMs and included exemplary dialogues from the EmpatheticDialogues dataset. These examples were chosen based on high ratings of empathy, relevance, and fluency by the human participants involved in the dataset's creation. Appendix D contains all the interfaces used for the experiment.

4.4 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.

4.5 Statistical Test and Sample Size

To analyze the results from the study we use the **chi-square test of independence** (McHugh, 2013) that tests whether there is any statistically significant difference between the proportion of *Bad*, *Okay*, and *Good* ratings of the five response groups.

When analyzing categorical ratings, particularly if the data involves ratings from different groups (like humans vs LLMs), the chi-Square test of independence is often a strong choice due to its robustness and the straightforward interpretability of the results (Field, 2013). The null and the alternate hypotheses of this statistical test are as below.

χ^2 test of independence:

- **Null hypothesis:** There is no difference between the proportion of *Bad*, *Okay*, and *Good* ratings of the five groups of responses.
- **Alternative hypothesis:** There is a difference between the proportion of *Bad*, *Okay*, and *Good* ratings of at least one out of the five groups of responses.

We used the G-Power software (Faul et al., 2009) to compute the minimal sample size required to detect a significant difference between the ratings of the five response 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 253 (i.e. at least 51 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. at least 102 participants per group). From a prior study, we had already engaged 200 participants to evaluate responses generated by humans and GPT-4. To ensure compatibility, we additionally recruited 600 participants from Prolific to rate responses generated by the LLMs: LLaMA-2; Gemini-Pro; and Mixtral-8x7B. That is 200 participants per group, which is sufficiently above the minimal sample size. One participant was asked to rate 10 responses. Altogether our study compares empathy ratings received for $10,000$ responses ($2,000$ responses per group).

5 Results

Figure 2 visualizes the number of *Good*, *Okay*, and *Bad* ratings received by the responses generated by humans, and the four LLMs for dialogue prompts spanning across all emotions as a whole. The χ^2 and the p-values obtained by applying the chi-square test of independence to the number of *Good*, *Okay*, and *Bad* ratings collectively and for each category independently indicated that there is a statistically significant difference between the proportion of *Good*, *Okay*, and *Bad* ratings of at least one out of the five response groups. We computed the percentage gains of the ratings received by each LLM compared to the human baseline under each rating category. GPT-4 was observed to generate the most empathetic responses with $\approx 31\%(\chi^2 = 96.77, p < .001)$ gain in the number of *Good* ratings compared to the humans. LLaMA-2, Mixtral-8x7B, and Gemini-Pro were observed to follow GPT-4 with $\approx 24\%(\chi^2 = 54.40, p < .001)$, $\approx 21\%(\chi^2 = 42.36, p < .001)$, and $\approx 10\%(\chi^2 = 8.85, p < .01)$ gain in the number of *Good* ratings, respectively, compared to the human baseline. Note that when calculating the χ^2 values here, we considered *Good* ratings as one category and combined *Bad* and *Okay* ratings as another category.

Figure 3 visualizes the number of *Good*, *Okay*, and *Bad* ratings received by the responses generated by humans and the four LLMs for positive and negative emotional dialogue prompts, separately. All four LLMs outperformed the human baseline across both positive and negative emotions in the number of *Good* ratings received. Here too, GPT-4 ranked the top in the number of *Good* ratings, obtaining percentage gains of $\approx 36\%(\chi^2 = 64.10, p < .001)$ and $\approx 27\%(\chi^2 = 36.78, p < .001)$, respectively across positive and negative emotions, compared to the human baseline. LLaMA-2 and Mixtral-8x7B followed GPT-4 when responding to positive emotions obtaining $\approx 28\%(\chi^2 = 38.40, p < .001)$, and $\approx 25\%(\chi^2 = 29.21, p < .001)$ gain in the number of *Good* ratings, respectively, compared to the human baseline. However, the percentage gain in the number of *Good* ratings obtained by Gemini-Pro across positive emotions was observed to be statistically insignificant compared to those received by the human responses $(\uparrow = 5.95\%, \chi^2 = 1.54, p > .05)$. LLaMA-2, Mixtral-8x7B, and Gemini-Pro followed GPT-4 when responding to negative emotions obtaining $\approx 20\%(\chi^2 = 19.0, p < .001)$, $\approx 17\%(\chi^2 = 15.15, p < .001)$, and $\approx 13\%(\chi^2 = 8.02, p < .01)$ gain in the number of *Good* ratings, respectively, compared to the human baseline.

Further, we computed the percentage gains of the categorical ratings received by each LLM compared to the human baseline when responding to each of the 32 positive and negative emotions (See Table 9 in Appendix G). This finer analysis allowed us to observe that some LLMs are significantly better than humans when responding to specific emotions compared to others. It could be observed that GPT-4 obtains statistically significant percentage

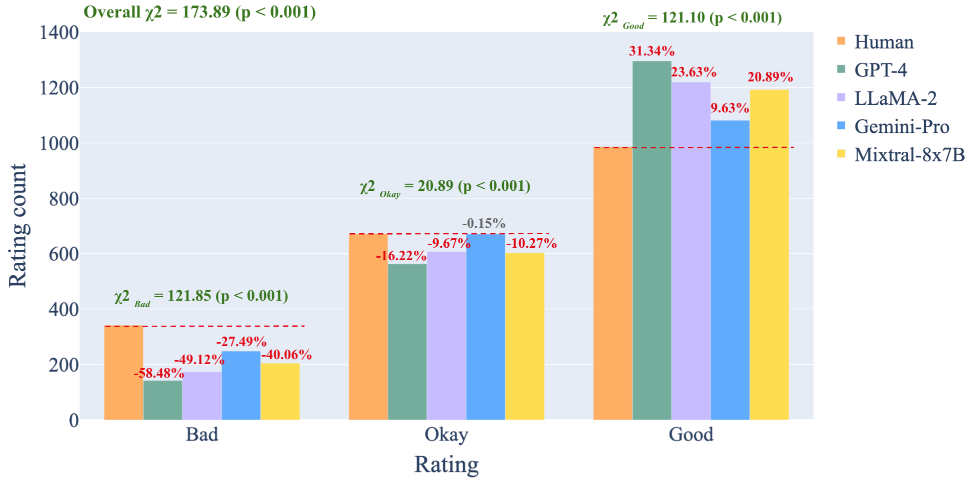


Figure 2: The *Good*, *Okay*, and *Bad* rating counts corresponding to the responses generated by humans, GPT-4, LLaMA-2, Gemini-Pro, and Mixtral-8x7B. The percentage gains of the LLMs' response ratings compared to the humans' response ratings are indicated at the top of each bar. The gains indicated in red are statistically significant.

gains in the number of *Good* ratings over the human baseline across most positive emotion categories such as *Impressed* ($\uparrow = 56\%, \chi^2 = 10.62, p < .01$), *Surprised* ($\uparrow = 79\%, \chi^2 = 10.33, p < .01$), *Grateful* ($\uparrow = 65\%, \chi^2 = 8.36, p < .01$), *Proud* ($\uparrow = 50\%, \chi^2 = 7.7, p < .01$), *Confident* ($\uparrow = 44\%, \chi^2 = 6.86, p < .01$), *Joyful* ($\uparrow = 42\%, \chi^2 = 6.34, p < .05$), and *Excited* ($\uparrow = 47\%, \chi^2 = 5.41, p < .05$); LLaMA-2 across emotions *Grateful* ($\uparrow = 65\%, \chi^2 = 8.36, p < .01$), *Surprised* ($\uparrow = 71\%, \chi^2 = 8.14, p < .01$), *Proud* ($\uparrow = 44\%, \chi^2 = 5.69, p < .05$), *Excited* ($\uparrow = 44\%, \chi^2 = 4.59, p < .05$), *Hopeful* ($\uparrow = 39\%, \chi^2 = 4.27, p < .05$), and *Prepared* ($\uparrow = 39\%, \chi^2 = 4.06, p < .05$); and Mixtral-8x7B across emotions *Proud* ($\uparrow = 59\%, \chi^2 = 11.44, p < .001$), *Grateful* ($\uparrow = 58\%, \chi^2 = 6.36, p < .05$), and *Excited* ($\uparrow = 47\%, \chi^2 = 5.41, p < .05$).

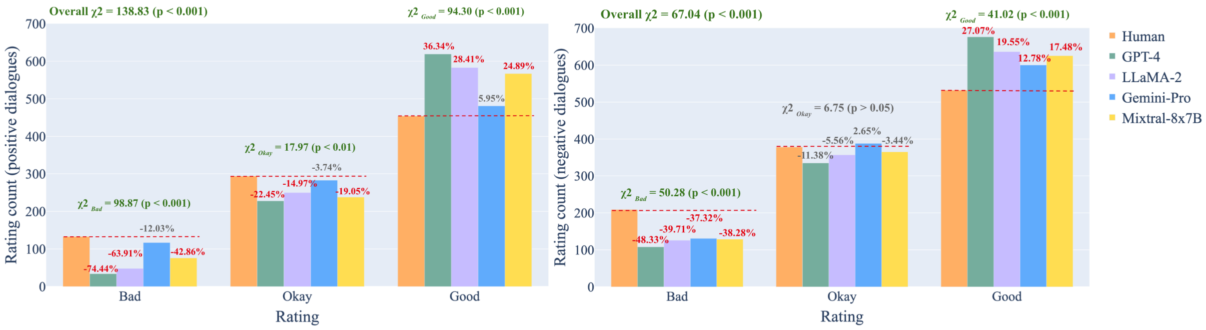
Compared to the positive emotions, we could only observe the four LLMs obtaining significant gains in the number of *Good* ratings over humans only across a few negative emotions such as *Afraid* (LLaMA: $\uparrow = 50\%, \chi^2 = 4.66, p < .05$; GPT: $\uparrow = 46\%, \chi^2 = 3.91, p < .05$), *Apprehensive* (GPT: $\uparrow = 104\%, \chi^2 = 20.72, p < .001$; Gemini: $\uparrow = 60\%, \chi^2 = 6.23, p < .05$; LLaMA: $\uparrow = 52\%, \chi^2 = 4.57, p < .05$), *Anxious* (GPT: $\uparrow = 75\%, \chi^2 = 9.2, p < .01$; LLaMA: $\uparrow = 63\%, \chi^2 = 6.22, p < .05$; Gemini: $\uparrow = 63\%, \chi^2 = 6.22, p < .05$; Mixtral: $\uparrow = 50\%, \chi^2 = 3.85, p < .05$), and *Annoyed* (GPT: $\uparrow = 59\%, \chi^2 = 6.62, p < .05$); Mixtral: $\uparrow = 52\%, \chi^2 = 4.97, p < .05$)). This implies that there is more room for these LLMs to improve their performance across other important negative emotion categories that commonly occur in day-to-day conversations.

6 Case Study

Table 3 shows an example, in which the response generated by the human was rated *Bad* whereas the responses generated by the four LLMs were rated *Good* by the participants. It could be noted that in the human response, the human responder speaks about themselves rather than focussing on the emotion of the speaker. On the other hand, all the four LLMs seem to recognize the emotion of the speaker and the love the speaker's grandmother has towards them and validate it using phrases such as *That's so sweet*, *That's so thoughtful*, *That's so heartwarming to hear!*, and *Your grandmother's thoughtfulness warms my heart*. What follows in the LLMs' responses are more complex reflections of what the speaker has said, which not only demonstrates understanding but also adds depth to the conversation, potentially leading to a more profound continuation of the dialogue. More such examples are denoted in Appendix H.

7 Discussion

The responses generated by all four LLMs surpassed the human responses in terms of empathetic quality by a statistically significant margin across all emotions as a whole, and across positive (except Gemini-pro) and negative emotions separately. Even though Gemini-Pro reported a significant gain ($\approx 13\%$) compared to the human baseline across negative emotions, it failed to report statistically



(a) Positive dialogues (b) Negative dialogues

Figure 3: The *Good*, *Okay*, and *Bad* rating counts corresponding to the responses generated by humans, GPT-4, LLaMA-2, Gemini-Pro, and Mixtral-8x7B for positive and negative emotional dialogue prompts.

Prompt	*My grandma sent a lot of food to me through my mom. She's still making sure I get good food to eat.* **(Sentimental)**
Responses:	
Human	*Some habits die hard. My mom's always trying to send me food still. I'm a grown man.* **(Bad)**
GPT	*That's so sweet of your grandma. It shows how much she cares about you. It must be comforting to know you're always on her mind.* **(Good)**
LLaMA	*That's so thoughtful of your grandma to make sure you're well taken care of. It's great that she's still involved in your life and wants to ensure your well-being.* **(Good)**
Gemini	*That's so heartwarming to hear! Your grandma's thoughtfulness shows how much she cares about you. It's wonderful that she's going to such efforts to ensure you're well-fed. I can imagine the love and warmth that comes with each meal. Thanks for sharing this special gesture.* **(Good)**
Mixtral	*Your grandmother's thoughtfulness warms my heart. It's evident that she cares deeply for you and wants to ensure your well-being. Her actions speak volumes about the love and concern she holds for you.* **(Good)**

Table 3: An example where the response generated by the human was rated *Bad* whereas the responses generated by the LLMs were rated *Good* by the participants.

significant gains compared to the human baseline across positive emotions, which leaves more room for Gemini-Pro to improve its performance when responding to positive emotions. It was seen that LLMs GPT-4, LLaMA-2, and Mixtral-8x7B perform significantly better than humans when responding to most positive emotion categories such as *Grateful*, *Proud*, and *Excited*. However, compared to positive emotions, the four LLMs could only perform significantly better than humans when responding to a few negative emotions such as *Apprehensive* and *Anxious*. This implies that the LLMs, while advanced in their empathetic understanding and response generation, have differential performance based on the valence of the emotions they are responding to. This could be due to a variety of factors such as the data the models are trained on, which may contain richer or more nuanced examples of responses to specific emotions, allowing the LLMs to learn more effective response strategies for these emotions.

The disparity in performance between positive and negative emotions also suggests that future iterations of these LLMs could benefit from targeted improvements in understanding and responding to more negative emotions. This could involve incorporating more diverse and nuanced examples of negative emotional responses into the training data or refining the models' algorithms to better capture the subtleties of negative emotional expressions.

Furthermore, the fact that LLMs outperform humans in empathetic response quality, especially in certain emotions, underscores the potential of these models in applications requiring emotional intelligence, such as mental health support, customer service, and social interactions. However, the variability in performance across different types of emotions also highlights the importance of ongoing research and development to enhance the models' emotional intelligence across the full spectrum of human emotions.

Overall, this study contributed knowledge on how empathy is conveyed in responses generated by different LLMs to diverse positive and negative emotional stimuli, compared to a human baseline. Due to the between-groups study design and the release of the dataset, the evaluation framework that we introduce could be extended to evaluate the empathetic responding capabilities of newer and updated versions of LLMs as they emerge.

8 Limitations

The choice of using a 3-point scale rather than a 5 or 7-point scale can raise questions regarding the granularity of the evaluation. As described in section 4.2, given that the study involves a large and diverse pool of participants, we believe that the advantages of using a simple, straightforward, and a human-centric rating scale, outweigh the concerns regarding granularity. As is evident from our results, the 3-point scale, while offering less granularity than a 5 or 7-point scale, still provides sufficient variability to perform robust statistical analyses. This scale was adequate to reveal significant differences in human and LLM-generated empathetic responses, confirming its effectiveness in the context of our research objectives. This establishes a foundational benchmark for evaluating the empathetic quality of responses, serving as a stepping stone for more detailed future studies.

9 Ethical Considerations

Data usage: The study utilized dialogue prompt-response pairs from the state-of-the-art EmpatheticDialogues dataset (Rashkin et al., 2019), which contains ethically sourced dialogues and is available publicly under the CC BY-NC 4.0 license. The dataset itself is anonymized to protect the privacy of individuals who contributed to its creation. We plan to publicly release the new artifacts generated in this study, including the responses from the four LLMs and the participants' empathy ratings, under the CC BY-NC 4.0 license. This licensing allows other researchers to modify, enhance, and further build upon our work for non-commercial purposes. By doing so, we aim to facilitate ongoing comparisons with newer and updated versions of LLMs, eliminating the need to replicate the entire study from the beginning.

Human experiment: The human participants recruited from the crowdsourcing platform Prolific (`www.prolific.com`) were paid €2.25 for rating 10 responses that took on average 11 minutes and 23 seconds to complete. This was ≈1.3 times above the wage recommended as *Good* (€9 per hour) by the Prolific crowdsourcing platform. All participants were informed about the purpose of the study and the nature of the tasks they would perform. The ratings were collected at the end of the task after the participants decided to submit their work. Intermediate annotations were not recorded. The participants were allowed to leave the task at any time without submitting their ratings. Random subsets of dialogue prompt-response pairs used in the study were manually inspected to ensure that the tasks assigned to the crowd workers were not psychologically distressing or offensive. In addition, efforts were made to recruit a diverse group of participants considering factors of gender and the age group that represent the broader population to avoid bias in the results.

Transparency and reproducibility of the study: 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.

Ethical concerns surrounding empathetic LLMs: Given the black-box nature of LLMs and their limited controllability and interpretability, one should take caution when using them, particularly in sensitive application domains such as mental health and crisis support. The opaque nature of these models can lead to outputs that are unpredictable or misaligned with human expectations, which can raise significant ethical concerns. Also, LLM-generated responses can represent societal biases and discriminations that are inherently present in the training data, which can lead to discriminatory or unethical outputs. Thus, an ethical approach to deploying such LLMs in sensitive domains should involve rigorous checking for biases and continuously monitoring their performance across underrepresented social groups. Some research studies point out that over-reliance on AI for empathetic interactions could affect human empathy skills and alter traditional social interactions (Chen et al., 2024). There is also a concern regarding the sincerity of the LLM-generated empathetic responses since LLMs cannot feel the users' emotions (Bove, 2019). Hence, it is important to be transparent about the nature of the LLM-generated responses to avoid over-reliance or emotional attachment to these agents over time.

References

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

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*, 183(6):589–596.

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.

Liliana L Bove. 2019. Empathy for service: benefits, unintended consequences, and future research agenda. *Journal of Services Marketing*, 33(1):31–43.

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.

Angelina Chen, Oliver Hannon, Sarah Koegel, and Raffaele Ciriello. 2024. Feels like empathy: How "emotional" ai challenges human essence. In *Australasian Conference on Information Systems*.

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*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

Mark H Davis. 1980. Interpersonal reactivity index.

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.

Igor Douven. 2018. A bayesian perspective on likert scales and central tendency. *Psychonomic bulletin & review*, 25:1203–1211.

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.

Andy Field. 2013. *Discovering statistics using IBM SPSS statistics*. sage.

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.

David R Heise. 1970. The semantic differential and attitude research. *Attitude measurement*, 4:235–253.

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

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.

Emma J Lawrence, Philip Shaw, Dawn Baker, Simon Baron-Cohen, and Anthony S David. 2004. Measuring empathy: reliability and validity of the empathy quotient. *Psychological medicine*, 34(5):911–920.

Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C Ong. 2024. Large language models produce responses perceived to be empathic. *arXiv preprint arXiv:2403.18148*.

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.

Siyuan Brandon Loh and Aravind Sesagiri Raamkumar. 2023. Harnessing large language models' empathetic response generation capabilities for online mental health counselling support. *arXiv preprint arXiv:2310.08017*.

Mary L McHugh. 2013. The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

MistralAI. 2024. Mixtral of experts.

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.

OpenAI. 2023. Gpt-4.

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.

Sundar Pichai. 2023. Introducing gemini: Our largest and most capable ai model.

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.

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

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

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.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chat-gpt 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.

A Distribution of Emotions

Figure 4 shows the distribution of the dialogue prompt-response pairs sampled from the EmpatheticDialogues dataset across the 32 positive and negative emotions. Table 4 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.

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

B Large Language Models

The study evaluated four state-of-the-art LLMs: GPT-4; LLaMA-2-Chat-70B; Gemini-1.0-Pro; and Mixtral-8x7B-Instruct. The details of the four LLMs are as follows.

GPT-4 (OpenAI, 2023) developed by OpenAI (`openai.com`) is the latest model in their GPT series with an estimated 1.76 trillion parameters. GPT-4 is claimed to be more reliable, creative, and able to handle much more nuanced instructions than its predecessor GPT-3.5. The model considerably outperforms existing LLMs, alongside most state-of-the-art models which include benchmark-specific crafting or additional training protocols.

LLaMA-2-Chat-70B (Touvron et al., 2023) developed by Meta AI (`ai.meta.com`), is an open-source LLM pre-trained on publicly available online data sources and fine-tuned on publicly available instruction tuning data (Chung et al., 2022), aligning the LLM towards dialogue-style instructions. We used the largest variant of LLaMA-2 with 70 billion parameters for this study.

Gemini-1.0-Pro (Pichai, 2023) developed by Google is a multimodal LLM trained to recognize and understand text, images, audio, and video. While Google does not reveal the exact number of parameters of this model and the data the model is trained on, it is developed based on the transformer architecture and relies on strategies like pre-training and fine-tuning, much as other LLMs. Independent research found that Gemini-1.0-Pro trails GPT-3.5-turbo across many of the traditional NLP benchmarks (Akter et al., 2023).

Mixtral-8x7B-Instruct (MistralAI, 2024) developed by Mistral AI (`mistral.ai`), is a high-quality sparse mixture of experts model (SMoE) with 46.7B total parameters. The *Instruct* model has been optimised through supervised fine-tuning and direct preference optimisation for careful instruction following. It is claimed to outperform LLaMA-2 (70B) on most traditional NLP benchmarks with 6x faster inference. The model is also claimed to match or outperform GPT-3.5 on most standard benchmarks.

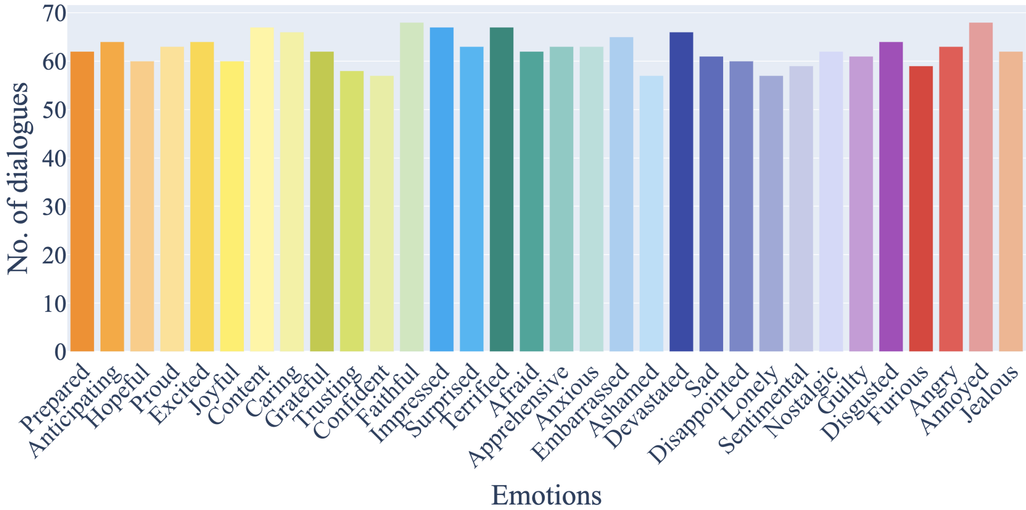


Figure 4: Distribution of the dialogue prompt-response pairs sampled from the EmpatheticDialogues dataset across the 32 positive and negative emotions.

C Toronto Empathy Questionnaire

Table 5 shows the questions in the Toronto Empathy Questionnaire (TEQ) (Spreng et al., 2009) that were asked from the participants. Responses to the questions are 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 are summed to derive one's propensity to empathize.

D Task Interfaces

Figures 5, 6, 7 and 8 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 renowned psychologist and statistician, introduced standards for evaluating the magnitude of effect sizes in statistical analyses such as chi-square tests and analysis of variance (ANOVA), as detailed in his work on quantitative methods (Cohen, 1992). These standards provide a foundational guide for assessing the substantive importance of observed effects within these statistical tests. For Chi-square tests, Cohen's W is utilized to measure the association strength between categories, with Cohen establishing benchmarks for small (0.10), medium (0.30), and large (0.50) effects.

We chose the medium effect size to compute the required minimum sample size because a medium

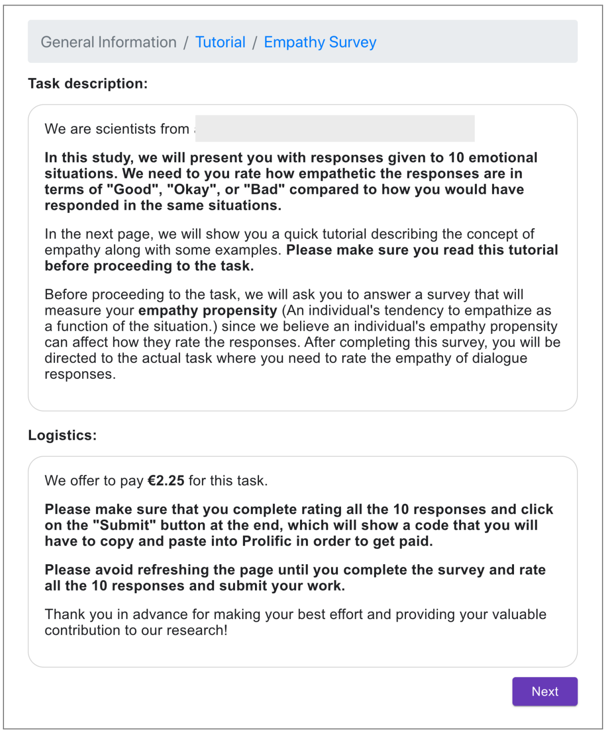


Figure 5: The description of the task.

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

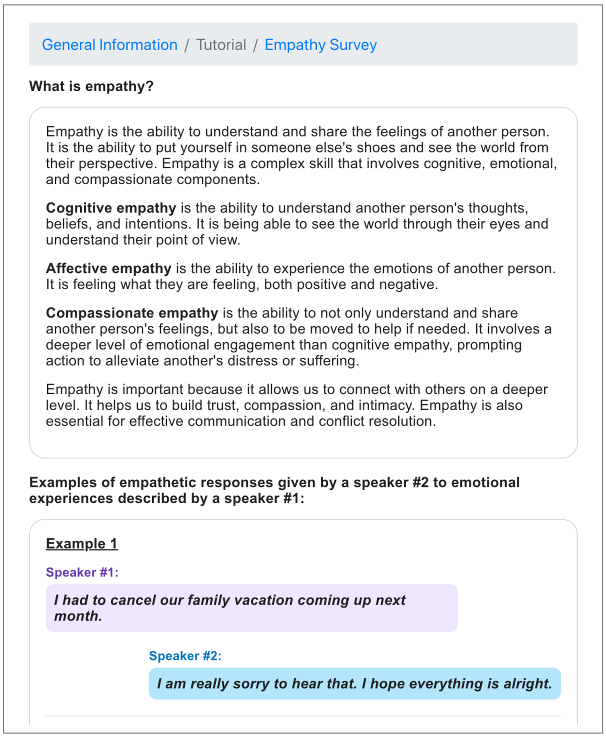


Figure 6: The tutorial.

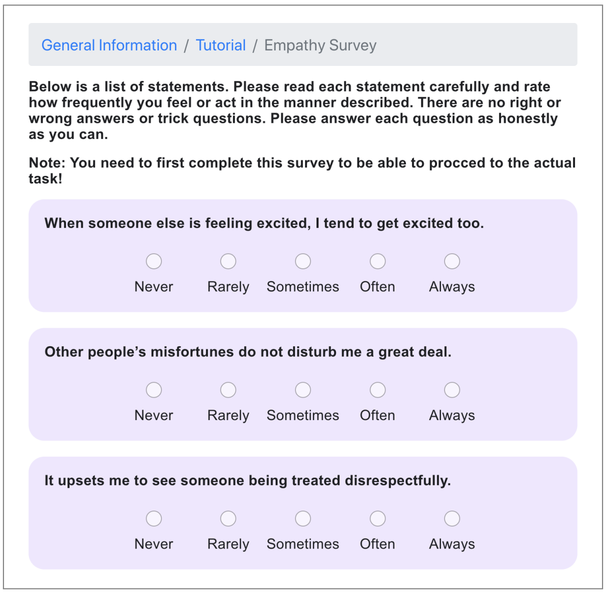


Figure 7: The Toronto Empathy Questionnaire.

effect size can sensitively detect differences in empathy levels between humans' and LLMs' responses, whose differences can be significant, yet not overwhelmingly so. Furthermore, employing a medium effect size enables the identification of nuanced yet significant differences without the need for an overly large sample, ensuring that the differences detected by the study are practically meaningful.

F Chi-Squared test of independence — Results

The statistical chi-square test of independence results corresponding to the proportions of the *Bad*, *Okay*, and *Good* empathy ratings received by the responses generated by the humans and the four

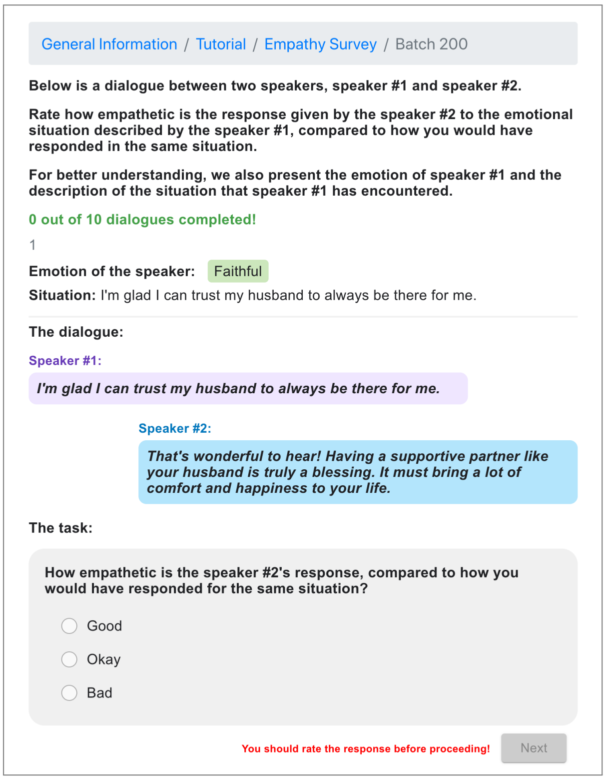


Figure 8: The task interface for rating responses in terms of empathy.

LLMs are denoted in Table 6. Table 8 denotes the statistical pairwise chi-square test of independence results corresponding to the proportions of *Bad*, *Okay*, and *Good* empathy ratings of the humans' and each of the LLMs' responses.

G Finer analysis of empathy ratings

Tables 9 denote the percentage gains obtained by the four LLMs' response ratings compared to the human baseline when responding to dialogue prompts containing positive and negative emotions. We conducted pairwise statistical chi-square tests of independence for the proportions of each of *Bad*, *Okay*, and *Good* response ratings between the humans and each of the four LLMs. The percentage gains for which statistical significance was indicated by the chi-square test of independence are highlighted in bold.

H Example dialogue responses

Table 10 denotes some example dialogue situations and responses generated by humans and LLMs and the corresponding ratings given by the human raters.

I Participants' demographics

Figures 9 and 10 respectively show the distributions of the countries of residence and the ethnicities of the participants who rated the five 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 five 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 five groups.

J Distribution of empathy propensity of participants

Figure 11 shows the distributions of the participants' propensities to empathize across the five 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 five groups.

K Quality Analysis

Figure 12 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 2.3% 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 13 shows the histogram of times (in minutes) taken to complete the study. On average it took 11 minutes and 23 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 4.53% 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.

	Rating	Human	GPT-4	LLaMA-2	Gemini	Mixtral	χ^2 (9.49)		χ^2 (15.51)	
All emotions	Bad	342	**142**	174	248	205	**121.86**	**(p < .001)**	**173.89**	**(p < .001)**
	Okay	672	563	607	671	603	**20.89**	**(p < .001)**		
	Good	986	**1295**	1219	1081	1192	**121.10**	**(p < .001)**		
Positive emotions	Bad	133	**34**	48	117	76	**98.88**	**(p < .001)**	**138.83**	**(p < .001)**
	Okay	294	228	250	283	238	**17.97**	**(p < .001)**		
	Good	454	**619**	583	481	567	**94.30**	**(p < .001)**		
Negative emotions	Bad	209	**108**	126	131	129	**50.28**	**(p < .001)**	**67.04**	**(p < .001)**
	Okay	378	335	357	388	365	6.75	(p > .05)		
	Good	532	**676**	636	600	625	**41.03**	**(p < .001)**		

Table 6: Statistical Chi-square test results corresponding to the proportions of *Bad*, *Okay*, and *Good* empathy ratings of the humans' and the LLMs' responses. The critical values of the χ^2 distributions are 15.51 and 9.49, respectively for all *Bad*, *Okay*, and *Good* rating classes and one at a time (computed at a significance level of 0.05 and 8 and 4 degrees of freedom, respectively). If the χ^2 statistic is greater than the critical value the null hypothesis can be rejected at 5% significance level, which means there is a statistically significant difference in the proportions of the empathy ratings between the groups of responses that are being compared.

	All emotions χ^2 (5.991)		**Positive emotions** χ^2 (5.991)		**Negative emotions** χ^2 (5.991)	
LLMs against human baseline:						
Human Vs GPT-4	**134.12**	**(p < .001)**	**92.41**	**(p < .001)**	**51.94**	**(p < .001)**
Human Vs LLaMA-2	**82.62**	**(p < .001)**	**59.52**	**(p < .001)**	**30.42**	**(p < .001)**
Human Vs Gemini	**19.34**	**(p < .001)**	2.01	(p > .05)	**22.11**	**(p < .001)**
Human Vs Mixtral	**57.53**	**(p < .001)**	**33.95**	**(p < .001)**	**26.64**	**(p < .001)**
LLMs against each other:						
GPT-4 Vs LLaMA-2	**7.19**	**(p < .05)**	4.48	(p > .05)	3.30	(p > .05)
GPT-4 Vs Gemini	**57.54**	**(p < .001)**	**68.86**	**(p < .001)**	**10.63**	**(p < .01)**
GPT-4 Vs Mixtral	**17.08**	**(p < .001)**	**18.53**	**(p < .001)**	5.15	(p > .05)
LLaMA-2 Vs Gemini	**24.46**	**(p < .001)**	**40.68**	**(p < .001)**	2.44	(p > .05)
LLaMA-2 Vs Mixtral	2.85	(p > .05)	**6.84**	**(p < .05)**	0.22	(p > .05)
Gemini Vs Mixtral	**13.13**	**(p < .01)**	**19.65**	**(p < .001)**	1.23	(p > .05)

Table 7: Statistical χ^2 test results corresponding to the proportions of *Bad*, *Okay*, and *Good* empathy ratings of the humans' and each of the LLMs' responses. 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, which means there is a statistically significant difference in the proportions of the *Bad*, *Okay*, and *Good* empathy ratings between the two groups of responses being compared.

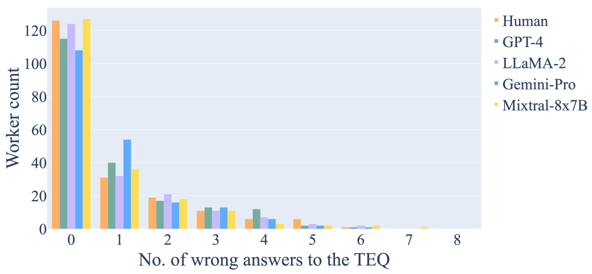


Figure 12: The number of reverse scale questions in the TEQ that were marked wrong by the participants rating the three response groups.

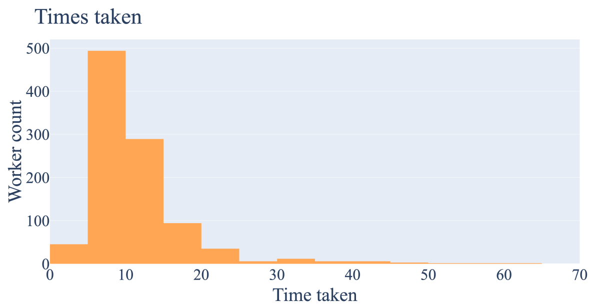
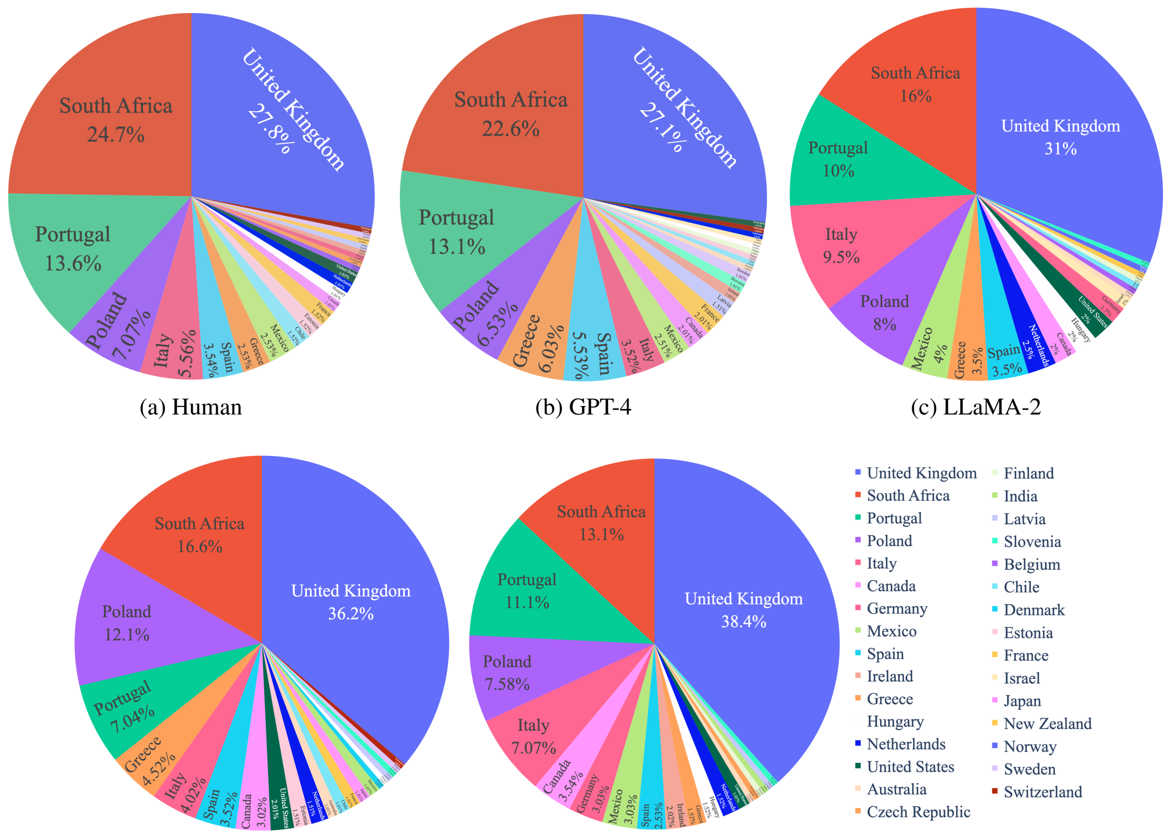


Figure 13: The histogram of times taken to complete the task by all participants.

	Bad			Okay			Good		
	% gain	χ^2 (3.841)		% gain	χ^2 (3.841)		% gain	χ^2 (3.841)	
All emotions:									
GPT-4 Vs Human	**-58.48%**	**93.08**	**(p < .001)**	**-16.22%**	**13.66**	**(p < .001)**	**31.34%**	**96.77**	**(p < .001)**
LLaMA-2 Vs Human	**-49.12%**	**62.05**	**(p < .001)**	**-9.67%**	**4.71**	**(p < .05)**	**23.63%**	**54.40**	**(p < .001)**
Gemini Vs Human	**-27.49%**	**17.20**	**(p < .001)**	-0.15%	0.00	(p > .05)	**9.63%**	**8.85**	**(p < .01)**
Mixtral Vs Human	**-40.06%**	**39.17**	**(p < .001)**	**-10.27%**	**5.32**	**(p < .05)**	**20.89%**	**42.36**	**(p < .001)**
Positive emotions:									
GPT-4 Vs Human	**-74.44%**	**63.53**	**(p < .001)**	**-22.45%**	**11.50**	**(p < .001)**	**36.34%**	**64.10**	**(p < .001)**
LLaMA-2 Vs Human	**-63.91%**	**43.45**	**(p < .001)**	**-14.97%**	**4.92**	**(p < .05)**	**28.41%**	**38.40**	**(p < .001)**
Gemini Vs Human	-12.03%	1.05	(p >.05)	-3.74%	0.26	(p > .05)	5.95%	1.54	(p > .05)
Mixtral Vs Human	**-42.86%**	**17.02**	**(p < .001)**	**-19.05%**	**8.15**	**(p < .01)**	**24.89%**	**29.21**	**(p < .001)**
Negative emotions:									
GPT-4 Vs Human	**-48.33%**	**36.75**	**(p < .001)**	-11.38%	3.63	(p > .05)	**27.07%**	**36.78**	**(p < .001)**
LLaMA-2 Vs Human	**-39.71%**	**23.61**	**(p < .001)**	-5.56%	0.81	(p > .05)	**19.55%**	**19.00**	**(p < .001)**
Gemini Vs Human	**-37.32%**	**20.56**	**(p < .001)**	2.65%	0.16	(p > .05)	**12.78%**	**8.02**	**(p < .01)**
Mixtral Vs Human	**-38.28%**	**21.75**	**(p < .001)**	-3.44%	0.29	(p > .05)	**17.48%**	**15.15**	**(p < .001)**

Table 8: The percentage gains obtained by the LLMs in each rating category compared to the human baseline. The corresponding statistical χ^2 test results are also indicated. The statistically significant gains are highlighted in bold. The critical value of the χ^2 distribution in this case is 3.841 (computed at a significance level of 0.05 and 1 degree of freedom).



(a) Human (b) GPT-4 (c) LLaMA-2

(d) Gemini-Pro (e) Mixtral-8x7B

Figure 9: Distribution of the countries of residence of the participants across the five groups.

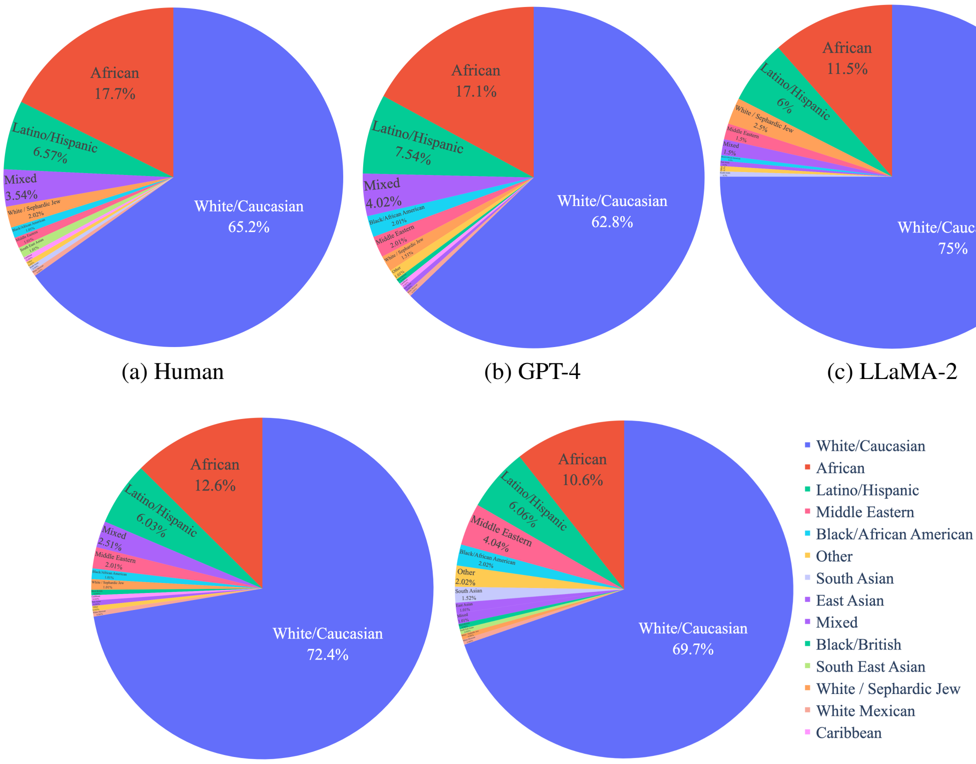


Figure 10: Distribution of the ethnicities of the participants across the five groups.

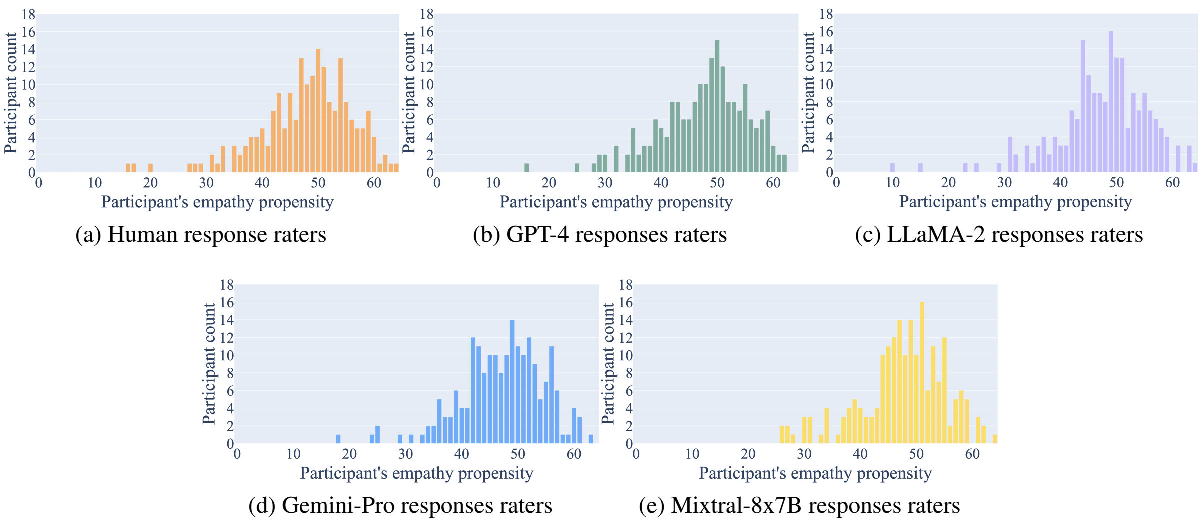


Figure 11: The distributions of the participants' propensities to empathize across the five groups.

Emotion	LLM	Percentage gain (%)		
		Bad	Okay	Good
Positive emotions:				
Prepared	GPT	**-90.0***	-9.52	35.48
	LLaMA	-50.0	-33.33	**38.71***
	Gemini	-20.0	19.05	-6.45
	Mixtral	-50.0	-14.29	25.81
Anticipa -ting	GPT	-66.67	-16.67	23.53
	LLaMA	-16.67	-8.33	8.82
	Gemini	0.0	25.0	-17.65
	Mixtral	16.67	-25.0	14.71
Hopeful	GPT	-33.33	-30.0	29.03
	LLaMA	-55.56	-35.0	**38.71***
	Gemini	55.56	-30.0	3.23
	Mixtral	-33.33	-10.0	16.13
Proud	GPT	-70.0	-42.86	**50.0****
	LLaMA	**-90.0***	-23.81	**43.75***
	Gemini	-30.0	-33.33	31.25
	Mixtral	**-100.0****	-42.86	**59.38*****
Excited	GPT	**-90.91****	-17.39	**46.67***
	LLaMA	**-81.82***	-17.39	**43.33***
	Gemini	0.0	-21.74	16.67
	Mixtral	-54.55	-34.78	**46.67***
Joyful	GPT	**-71.43***	-30.77	**42.42***
	LLaMA	**-71.43***	53.85	9.09
	Gemini	**-64.29***	38.46	12.12
	Mixtral	**-71.43***	23.08	21.21
Content	GPT	-85.71	-40.0	**35.0***
	LLaMA	-71.43	-25.0	25.0
	Gemini	0.0	-15.0	7.5
	Mixtral	-42.86	-15.0	15.0
Caring	GPT	-33.33	16.67	-4.44
	LLaMA	0.0	-5.56	2.22
	Gemini	200.0	-11.11	-8.89
	Mixtral	33.33	-5.56	0.0
Grateful	GPT	**-90.91****	-28.0	**65.38****
	LLaMA	**-72.73***	-36.0	**65.38****
	Gemini	-36.36	-16.0	30.77
	Mixtral	-36.36	-44.0	**57.69***
Trusting	GPT	**-72.73***	22.22	13.79
	LLaMA	**-81.82***	11.11	24.14
	Gemini	-27.27	27.78	-6.9
	Mixtral	-27.27	-33.33	31.03
Confident	GPT	**-87.5***	-41.18	**43.75****
	LLaMA	-50.0	11.76	6.25
	Gemini	0.0	5.88	-3.12
	Mixtral	-75.0	-11.76	25.0
Faithful	GPT	-37.5	-18.52	24.24
	LLaMA	-37.5	-18.52	24.24
	Gemini	-12.5	-14.81	15.15
	Mixtral	-37.5	-7.41	15.15
Impressed	GPT	**-80.0***	**-47.83***	**55.88****
	LLaMA	-50.0	-21.74	29.41
	Gemini	10.0	-8.7	2.94
	Mixtral	-10.0	-8.7	8.82
Surprised	GPT	**-86.67****	-25.0	**79.17****
	LLaMA	**-80.0****	-20.83	**70.83****
	Gemini	-40.0	4.17	20.83
	Mixtral	-33.33	-16.67	37.5
Negative emotions:				
Terrified	GPT	-46.67	-4.55	26.67
	LLaMA	-40.0	-9.09	26.67
	Gemini	-46.67	18.18	10.0
	Mixtral	-6.67	-27.27	23.33
Afraid	GPT	**-66.67***	0.0	**46.15***
	LLaMA	**-72.22****	0.0	**50.0***
	Gemini	**-55.56***	33.33	15.38
	Mixtral	-50.0	11.11	26.92
Apprehen -sive	GPT	**-90.0***	**-60.71****	**104.0*****
	LLaMA	-50.0	-28.57	**52.0***
	Gemini	-40.0	-39.29	**60.0***
	Mixtral	-70.0	-14.29	44.0
Anxious	GPT	-50.0	**-44.44***	**75.0****
	LLaMA	-41.67	-37.04	**62.5***
	Gemini	-41.67	-37.04	**62.5***
	Mixtral	-66.67	-14.81	**50.0***
Embarra -ssed	GPT	-47.06	10.53	20.69
	LLaMA	-23.53	5.26	10.34
	Gemini	-47.06	10.53	20.69
	Mixtral	-29.41	-10.53	24.14
Ashamed	GPT	-41.67	0.0	16.67
	LLaMA	-58.33	60.0	-6.67
	Gemini	-58.33	40.0	3.33
	Mixtral	-25.0	33.33	-6.67
Devastated	GPT	-33.33	-40.0	29.73
	LLaMA	-44.44	-15.0	18.92
	Gemini	-44.44	-30.0	27.03
	Mixtral	-66.67	30.0	0.0
Sad	GPT	-27.27	20.0	0.0
	LLaMA	-27.27	0.0	8.57
	Gemini	**-72.73***	20.0	14.29
	Mixtral	-54.55	-13.33	22.86
Disappoi -nted	GPT	-54.55	-15.0	31.03
	LLaMA	-45.45	-10.0	24.14
	Gemini	-18.18	35.0	-17.24
	Mixtral	-54.55	10.0	13.79
Lonely	GPT	-12.5	-5.88	6.25
	LLaMA	-12.5	11.76	-3.12
	Gemini	-62.5	-17.65	25.0
	Mixtral	-62.5	11.76	9.38
Sentimen -tal	GPT	-40.0	-11.11	11.11
	LLaMA	-60.0	-11.11	13.89
	Gemini	20.0	11.11	-8.33
	Mixtral	40.0	-27.78	8.33
Nostalgic	GPT	-85.71	-4.76	20.59
	LLaMA	-71.43	-9.52	20.59
	Gemini	-71.43	4.76	11.76
	Mixtral	-57.14	-14.29	20.59
Guilty	GPT	-38.46	22.22	3.33
	LLaMA	-46.15	-16.67	30.0
	Gemini	-38.46	-5.56	20.0
	Mixtral	**-69.23***	11.11	23.33
Disgusted	GPT	-43.75	27.27	3.85
	LLaMA	0.0	4.55	-3.85
	Gemini	-43.75	36.36	-3.85
	Mixtral	6.25	-4.55	0.0
Furious	GPT	-46.15	15.0	11.54
	LLaMA	-15.38	35.0	-19.23
	Gemini	23.08	5.0	-15.38
	Mixtral	7.69	-10.0	3.85

Angry	GPT	-37.5	-24.14	38.46
	LLaMA	-50.0	-24.14	42.31
	Gemini	25.0	-17.24	11.54
	Mixtral	-12.5	-17.24	23.08
Annoyed	GPT	**-66.67***	-23.08	**59.26***
	LLaMA	-53.33	3.85	25.93
	Gemini	-20.0	23.08	-11.11
	Mixtral	**-86.67****	-3.85	**51.85***
Jealous	GPT	-22.22	-13.04	16.67
	LLaMA	-22.22	-13.04	16.67
	Gemini	-44.44	0.0	13.33
	Mixtral	-11.11	13.04	-6.67

Table 9: Percentage gains obtained by the four LLMs' response ratings compared to the human baseline when responding to dialogue prompts containing positive and negative emotions. The statistically significant gains are highlighted in bold. *, **, and *** indicates $p < .05$, $p < .01$, and $p < .001$, respectively for the pairwise statistical chi-square tests of independence conducted between the humans' and each of the LLMs' response ratings.

Prompt	*These days I feel like I got it all. I'm not rich by any means but I am enriched by a loving wife, children and a dog.* ***(Content)***
Responses:	
Human	*I think I have a problem with jealousy because it's starting all over again. What sort of dog do you have?* **(Bad)**
GPT	*That sounds wonderful! It's heartwarming to hear how much you value your family and your life. Love and happiness truly are the richest treasures.* **(Good)**
LLaMA	*It sounds like you have a beautiful family and a fulfilling life. It's great to hear that you have found such joy and love in your life.* **(Good)**
Gemini	*I can understand why you feel so fulfilled. Having a loving family and a pet can bring immense joy and contentment. It sounds like you have a strong support system that makes you feel loved and appreciated. Cherish these relationships and continue to nurture them.* **(Good)**
Mixtral	*That's wonderful to hear that you feel enriched by your loving family! Cherish these moments, as they bring joy and fulfillment to life. I'm here if you ever need to talk about anything else.* **(Good)**
Prompt:	*I had a crazy week with 4 big events and I'm so happy to be at home relaxing in the air conditioning today. My dogs are happy too!* ***(Content)***
Responses:	
Human	*Air conditioning is very nice. My dogs love the outdoors but they want AC after a good, long walk!* **(Bad)**
GPT	*Wow, that sounds like a hectic week! It's great that you're taking time to relax and enjoy the comfort of your home. I bet your dogs are loving the extra attention too!* **(Good)**
LLaMA	*Wow, that sounds like a whirlwind week! It's great that you're taking some time to unwind and enjoy the AC with your furry friends. You deserve it!* **(Good)**
Gemini	*I can understand why you're feeling relieved and content after a hectic week. It's wonderful that you can now unwind at home in the comfort of the air conditioning, and it's heartwarming to know that your dogs are also enjoying the relaxation.* **(Good)**
Mixtral	*I'm glad to hear that you're finally getting some well-deserved rest after a hectic week. It sounds like those four events were quite demanding, and it's completely understandable that you're thrilled to be at home now. Your dogs must be delighted to be relaxing with you in the cool air conditioning! Take all the time you need to recharge and enjoy this precious downtime.* **(Good)**
Prompt	*I had a difficult job interview this week.* ***(Apprehensive)***
Responses:	
Human	*Don't think about it, what is done is done, just anticipate that they will call you back, the universe have it's way of helping if we believe.* **(Okay)**
GPT	*I'm really sorry to hear that you had a tough time with your job interview. It's completely normal to feel this way. Remember, each interview is a learning experience. Don't be too hard on yourself.* **(Good)**
LLaMA	*Sorry to hear that. It can be really tough when job interviews don't go as planned. It's understandable to feel frustrated and disappointed. Can you tell me a bit more about what happened during the interview?* **(Good)**
Gemini	*I understand that job interviews can be stressful and challenging. It's important to remember that you are not alone in this experience and that many people feel anxious during interviews. I'm here to listen if you want to talk about it.* **(Good)**
Mixtral	*I'm really sorry to hear that you had a tough job interview this week. It's completely normal to feel anxious or disappointed after an important event like that. Just know that you have the skills and experience necessary to succeed, and that every interview is a valuable learning experience. I'm here for you if you need to talk or want to practice for next time.* **(Good)**
Prompt:	*In High School I loved to play sports. Football was my favorite.* **(Nostalgic)**
Responses:	
Human	*Ah I was never much of a sports guy, but I do work out a lot. do you still play?* **(Okay)**
GPT	*That's great! Playing sports, especially football, must have been a thrilling experience for you. It's wonderful how such activities can bring so much joy and create lasting memories.* **(Good)**
LLaMA	*That's great! Playing sports can be a really rewarding experience, especially when you find a sport that you're passionate about like football. It's a great way to stay active, make new friends, and learn important teamwork skills. Did you have any favorite memories or moments from your time playing football?* **(Good)**
Gemini	*I can understand why you loved playing sports in high school, especially football. It's a great way to stay active and be part of a team. I bet you have some fond memories of those days.* **(Good)**
Mixtral	*That's great to hear that you enjoyed playing football in high school! I can imagine the excitement and camaraderie you felt while being part of a team. Share with me more about your experiences and favorite moments playing the sport.* **(Good)**

Table 10: Some example dialogue situations and responses generated by humans and LLMs and the corresponding ratings given by the human raters.