Who Generates More Empathetic Responses—Humans or LLMs? A Comparative Evaluation with Human and LLM Judges

Anuradha Welivita, Fawzia Zeitoun, and Pearl Pu
School of Computer and Communication Sciences
EPFL
Switzerland
kalpani.welivita@alumni.epfl.ch;{fawzia.zeitoun,pearl.pu}@epfl.ch

Abstract

This paper compares the empathetic quality of responses generated by humans and large language models (LLMs). We evaluate four LLMs that were widely used at the time of study—*GPT-4*, *LLaMA-2-70B-Chat*, *Gemini-1.0-Pro*, and *Mixtral-8×7B-Instruct*—against a human baseline using a large-scale between-subjects study. A total of 1,000 human participants evaluated the empathetic quality of human- and LLM-generated responses to 2,000 dialogue prompts spanning 32 positive and negative emotions. To complement human judgments, we also employed an LLM-as-judge (GPT-4o-mini) to assess the same responses. Across emotions and evaluators, LLM-generated responses were rated as significantly more empathetic than human-written responses. We also observed that both human judges and the LLM-as-judge tended to rate responses generated by their own group more favorably, indicating self-favoring tendencies. These findings highlight both the strong performance of contemporary LLMs in empathetic responding and the need to interpret human- and LLM-based evaluations with care.

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). Hence, 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; Liu et al., 2023). But empathy plays an important role in responding to both positive and negative emotions encountered in daily conversations. Last, but not least, 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).

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 358–381
July 3-4, 2026 ©2026 Association for Computational Linguistics

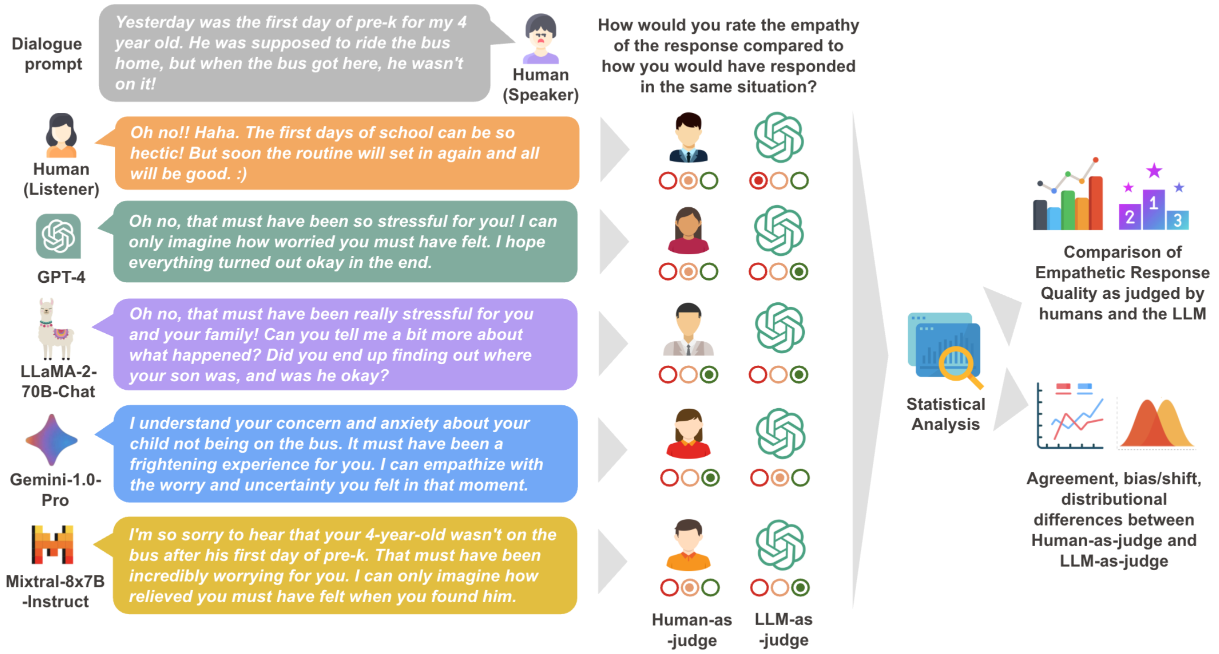


Figure 1: The dual-judge framework to evaluate who generates more empathetic responses—humans or LLMs when responding to emotional dialogue prompts.

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.

Also, as conversational LLMs become more emotionally expressive, researchers increasingly ask not only whether these models can convey empathy, but also whether they can evaluate it. The fact that using human raters to judge empathetic responses in dialogues can be subjective, expensive, and difficult to reproduce at scale encourages us to explore modern LLMs' ability to evaluate empathetic responses — how well they correlate with human judgments, and if they carry any systematic biases. A growing body of work in natural language evaluation suggests that LLMs can serve as capable evaluators across reasoning, summarization, and dialogue tasks, or what is known as the LLM-as-a-judge paradigm (Zheng et al., 2023; Li et al., 2025). However, modern LLM's ability to evaluate empathetic responses is underexplored.

To address these limitations, we conducted a between-subjects study with 1,000 crowdworkers, assigning 200 participants each to evaluate responses written 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 used 2,000 emotional dialogue prompts from the EmpatheticDialogues dataset (Rashkin et al., 2019) spanning 32 positive and negative emotions, to form the human baseline for our study as well as to initiate responses from the LLMs. Participants rated these responses on a simple and straightforward empathy scale. We repeated the same evaluation using an LLM (GPT-4o-mini) serving as the judge. We performed rigorous statistical analyses to determine whether humans—or which of the LLMs—produced better empathetic responses to positive and negative dialogue scenarios, as judged by both human evaluators and the LLM-as-judge. We further compared human and LLM evaluations to assess how strongly they correlate and to identify any systematic biases, shifts, or distributional differences in how empathy was rated.

Our contributions are fourfold: 1) A large-scale, between-subjects evaluation of empathetic response generation, avoiding biases inherent to within-subjects designs and enabling a scalable framework for comparing human and LLM performance. 2) A comprehensive comparison of empathetic responses across four state-of-the-art LLMs and a human baseline, using 2,000 dialogue prompts covering 32 positive and negative emotions, providing one of the most extensive cross-emotion evaluations to date. 3) An investigation of humans and LLMs as empathy evaluators, identifying their correlation, systematic patterns, biases, and directional shifts when evaluating empathy. 4) Empirical evidence that LLMs consistently outperform humans in generating empathetic responses, as judged by both humans and an LLM, while also showing that each group tends to favor its own modality of responding. Together, these contributions offer a deeper understanding of how humans and LLMs express and evaluate empathy, offering key insights for future work in affective computing.

2 Literature Review

As empathy evaluation has become increasingly important in light of LLMs' growing ability to produce fluent and emotionally resonant responses, several recent studies have examined their empathetic capabilities using automatic metrics. Loh and Raamkumar (2023), for example, evaluated multiple LLMs on ≈2,550 prompts from the EmpatheticDialogues dataset using automatic empathy-related measures derived from the Emotional Reactions, Interpretations, and Explorations taxonomy (Sharma et al., 2020). They found that LLMs often outperform traditional systems and even human-written responses under these metrics. However, such automatic measures do not always correlate with human perceptions, raising concerns about their validity for assessing complex affective responses. Similarly, Belkhir and Sadat (2023) evaluated GPT-3.5 using precision, recall, and accuracy of emotional categories. However, this approach mixes up emotional expression with empathy and fails to capture more neutral, intention-focused forms of empathetic responding (Welivita and Pu, 2020).

Another line of research evaluates LLMs using psychological questionnaires originally designed for humans. Studies such as Schaaff et al. (2023) and Elyoseph et al. (2023) have applied instruments like the Interpersonal Reactivity Index (Davis, 1980), Empathy Quotient (Lawrence et al., 2004), and Levels of Emotional Awareness Scale (Lane et al., 1990). While these tools provide structured measurements, their applicability to machine-generated text is debatable, as they assume underlying cognitive and emotional processes that LLMs do not possess.

Human evaluations of empathy in LLM-generated responses have also gained traction, but nearly all adopt a within-subjects design, where the same participants evaluate responses from multiple models. Examples include Lee et al. (2024; 2022), Ayers et al. (2023), Fu et al. (2023), Zhao et al. (2023), and Qian et al. (2023). Although useful for controlled comparisons, within-subjects designs introduce several limitations such as carry-over effects, order biases, rapid obsolescence as new models emerge, and challenges in scaling to large LLM families. Many also rely on A/B tests or numeric scales without clear interpretive anchors, which can inflate noise in participant ratings. Moreover, most studies do not include a human baseline, making it hard to understand how LLM performance compares to natural human empathetic behavior.

Beyond human-based evaluation, a growing line of work explores whether LLMs themselves can assess the empathetic quality of dialogue responses. This follows the broader "LLM-as-a-judge" paradigm. Svikhnushina and Pu (2023) demonstrate that LLMs can partially emulate online human evaluation of social chatbots, highlighting both the promise and instability of automated affective assessment. More recently, Xie et al. (2024) examine the sensitivity of LLM-based scoring to model choice and prompt design, showing that while LLMs can provide structured empathy assessments, their judgments can drift from human perceptions and vary considerably across conditions. While these studies underscore the potential of LLM-as-judge approaches, they lack direct comparisons between human and LLM judges and offer minimal analysis of whether human or LLM evaluators exhibit systematic biases, particularly when judging human versus machine-generated empathy. To address these gaps, our study provides the **largest dual-judge evaluation to date, collecting 10,000 human ratings and 10,000 LLM-based ratings across 2,000 prompts spanning 32 emotions, and comparing human- and LLM-generated responses as evaluated by both human and LLM judges**. This large-scale design enables a robust comparison of the empathetic quality of human versus LLM responses, as well as an analysis of the differences between human and LLM empathy judgments.

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. The dialogues in this dataset are curated by recruiting crowd workers from Amazon

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.

Mechanical Turk (AMT)[1]. Based on the sample size predicted by power analysis (in Section 4.3), we used randomly sampled 2,000 dialogues from this dataset, which are approximately 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 considered only the first two dialogue turns along with the emotion the dialogues were based on and the situation description for the evaluation task. This set of 2,000 single-exchange empathetic dialogues will serve as the human baseline for subsequent empathy evaluations by both human and LLM-based judges.

For comparison with the human baseline, we used 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). When prompting the models we used a prompt (denoted in Table 1), that defined empathy in terms of its cognitive, affective, and compassionate counterparts and explicitly asked the model to respond in an empathetic manner to the given dialogue utterance. The instructions aimed to replicate human understanding of empathy in LLMs. 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 Human Evaluation Experiment

4.1 Between-Subjects vs Within-Subjects

The human evaluation experiment was structured as a **between-subjects study**, in which 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.

The above 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 overexposure 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

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

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 Task Design

The five groups of participants for the study were recruited through the Prolific crowdsourcing platform (`www.prolific.com`). 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]). 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%. 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 M). Additionally, the 8 reserve scale questions contained in this questionnaire were used to gauge the quality of the workers and their attentiveness to the task.

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.

4.3 Statistical Test and Sample Size

To analyze the results from the study we used 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. 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 (see Appendix F). Accordingly, we recruited altogether $1,000$ evaluators (200 evaluators per each group of responses). One evaluator was asked to rate 10 responses. Altogether we received empathy ratings for $10,000$ responses ($2,000$ responses per group).

5 Automatic Evaluation

Building on the human evaluation experiment, we conduct an automatic evaluation in which an LLM serves as an empathy judge instead of human raters. The evaluation was performed in a few-shot setting using GPT-4o-mini, selected for its balance between efficiency and reasoning capability. We replicate the human scoring procedure, asking the LLM to rate each listener response on the same three-point empathy scale.

To construct few-shot examples, we withheld six representative dialogues from testing. For each sentiment (positive and negative) and each empathy rating (*Bad*, *Okay*, and *Good*), the model was provided with examples of both human-generated and LLM-generated responses that received that rating from human annotators (see Appendix E). All remaining items were used for quantitative analysis.

To compare human and automatic evaluations, we assessed the humans' and LLM's ratings across three dimensions: **1) Agreement** measured by Cohen's κ; **2) Bias** (*Is one systematically harsher or more lenient?*) measured by the Wilcoxon signed-rank test (WCX), the Hodges-Lehmann (HL) shift estimate, and the Rank-Biserial Correlation (RBC); and **3) Distributional Difference** (*Are their overall rating profiles different?*) measured by the Bowker's test of symmetry. Together, these measures test whether humans raters and the LLM-as-judge align in their ratings and use the scale in a similar manner, including comparable levels of strictness or leniency.

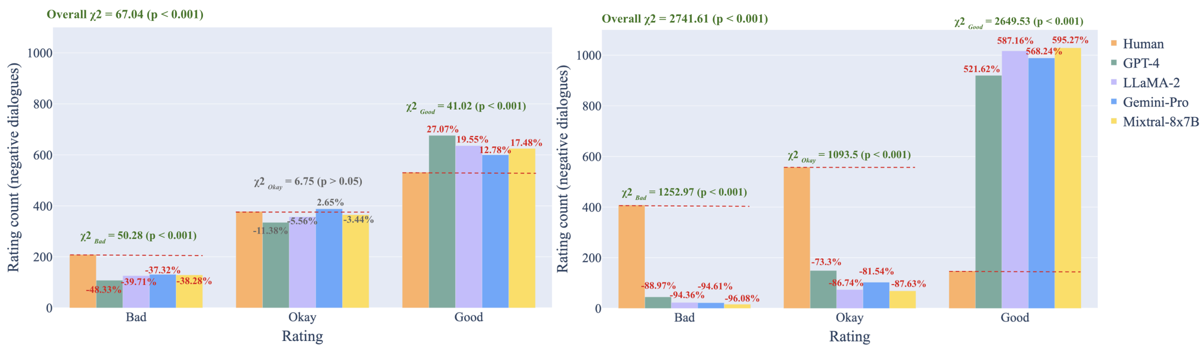
6 Results

6.1 Human and LLM Evaluation Results

Figure 2 shows the rating distributions for human and LLM-generated responses across positive and

(a) Ratings by human judges – positive dialogue responses

(b) Ratings by the LLM judge – positive dialogue responses



(c) Ratings by human judges – negative dialogue responses

(d) Ratings by the LLM judge – negative dialogue responses

Figure 2: The human and LLM 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. Statistically significant ($p < 0.001$) percentage gains of ratings received by LLM-generated responses compared to ratings received by human-written ones are denoted in red.

negative emotional prompts, as evaluated by human judges (left column) and the LLM judge (right column). Each subplot reports the χ^2 statistic comparing rating distributions across the five response sources and the percentage gain in ratings received by LLM-generated responses relative to human-generated responses in each rating category.

6.1.1 Positive dialogues

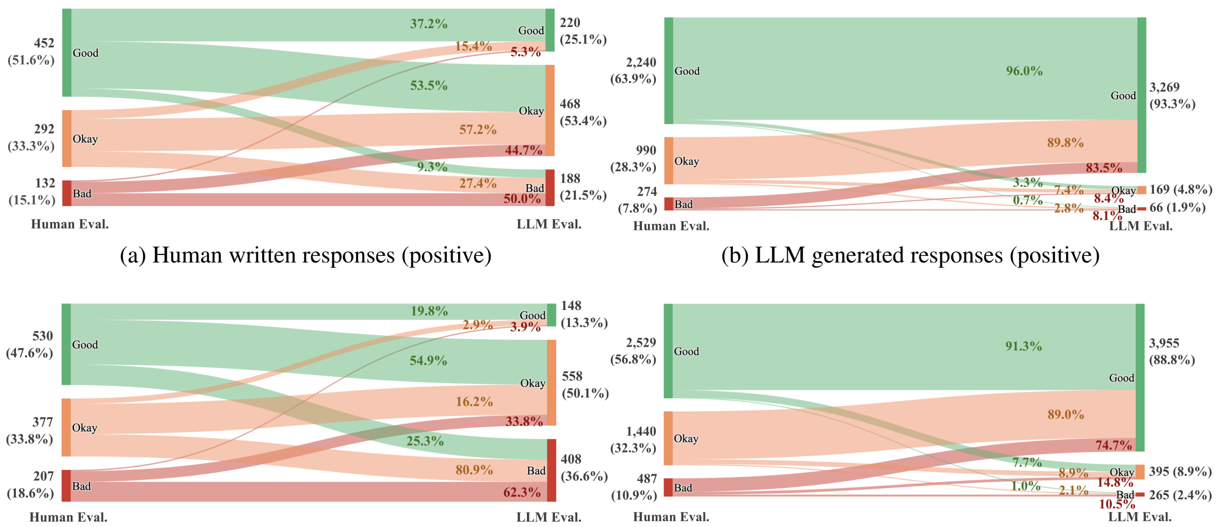
For positive emotional dialogues, both humans and the LLM-as-judge rate LLM-generated responses more positively than human-written ones, but the magnitude of this effect differs sharply between evaluators. As observed in Figure 2a, as rated by humans, LLMs' responses receive noticeably more *Good* ratings and fewer *Bad* and *Okay* ratings compared to human responses, leading to a significant overall deviation ($\chi^2 = 138.83, p < 0.001$). LLM responses show reductions of $12 - 74\%$ in *Bad* ratings and modest increases of $5 - 36\%$ in *Good* ratings relative to the human baseline, indicating that human annotators perceive LLM-generated responses as more empathetic when responding to positive-emotional contexts.

According to Figure 2b, this trend is amplified when responses are rated by the LLM-as-judge. All four LLMs receive significantly higher proportions of *Good* ratings ($> 268\%$ increases compared human response ratings) and significantly fewer *Bad* or *Okay* ratings compared to those of humans (reduced by $89 - 93\%$), resulting in a much larger overall deviation ($\chi^2 = 2015.14, p < 0.001$).

6.1.2 Negative dialogues

For negative emotional dialogues, we observe a consistent pattern where LLM-generated responses are perceived as more empathetic than human-written responses by both human evaluators and the LLM-as-judge. When evaluated by human judges (Fig. 2c), human responses received more *Bad* ratings and fewer *Good* ratings compared to all LLMs, indicating that LLMs were viewed as providing more empathetic responses than humans in negative contexts as well.

This trend becomes much stronger under LLM judging (Fig. 2d). The LLM-as-judge assigned a



(a) Human written responses (positive)

(b) LLM generated responses (positive)

(c) Human written responses (negative)

(d) LLM generated responses (negative)

Figure 3: Sankey diagrams visualizing the correspondence between human and LLM ratings for human-written and LLM-generated responses under positive and negative emotional contexts. The percentages on each flow represent the proportion of responses in each LLM rating category relative to the corresponding human rating category.

larger proportion of *Okay* and *Bad* ratings and the fewest *Good* ratings to human-written responses, being substantially harsher than human evaluators. All LLM-generated responses received significantly higher *Good* ratings compared to human-written ones. This led to a substantially larger chi-square statistic ($\chi^2 = 2741.61, p < .001$), reflecting a stronger divergence in how the LLM-as-judge evaluates humans' and LLMs' responses.

6.1.3 Summary

Overall, across both positive and negative emotional contexts, both human evaluators and the LLM-as-judge consistently rate LLM-generated responses more favorably than human-written ones. However, the two evaluator groups differ in how they apply the empathy rating scale, resulting in systematic differences in scoring behavior. These differences do not imply that one evaluator is "more correct" than the other, but instead reflect distinct internal criteria for assessing empathetic quality.

6.2 LLM-as-Judge vs Human-as-Judge

To examine these differences more closely, in Table 3 we summarize the statistics for agreement, bias, and distributional differences between human and LLM ratings for human-written and LLM-generated responses. For simplicity, we aggregate the scores assigned by humans and the LLM-as-judge across the four model-generated responses, as they exhibit consistent patterns even when analyzed separately.

Sent.	Src.	Kappa	WCX	HL	RBC	Bowker's χ^2
Pos.	Hum.	**0.36**	**87,699** ($p < 0.001$)	**0.5**	**0.55**	**163.40** ($p < 0.001$)
	LLM	0.12	**69,470** ($p < 0.001$)	**-0.5**	**-0.82**	**876.84** ($p < 0.001$)
Neg.	Hum.	**0.23**	**200,231** ($p < 0.001$)	**0.5**	**0.71**	**338.99** ($p < 0.001$)
	LLM	0.12	**222,100** ($p < 0.05$)	**-0.5**	**-0.77**	**1,112.70** ($p < 0.001$)

Table 3: Agreement, bias, and asymmetry metrics comparing ratings between humans and LLM-as-judge. (Sent. = Sentiment; Src. = Source (human-generated or LLM-generated responses); ns = not significant)

Overall, human-as-judge and LLM-as-judge ratings show only low to moderate agreement. Moderate agreement can be seen for human-written responses under positive ($\kappa = 0.36$) and negative ($\kappa = 0.23$) dialogue prompts, whereas agreement for LLM-generated responses is weaker ($\kappa < 0.2$). The Wilcoxon signed-rank test further reveals significant differences ($p < 0.001$) between human and LLM ratings across all four cases, indicating that the two evaluators tend to score the same responses differently. However, these differences are not random but systematic.

As measured by the Hodges-Lehmann (HL) shift and the Rank-Biserial Correlation (RBC), human evaluators rate human-written responses more favorably than the LLM-as-judge in both positive ($HL = +0.5,\ RBC = +0.55$) and negative ($HL = +0.5,\ RBC = +0.71$) contexts, whereas the LLM-as-judge rates LLM-generated responses more favorably than human judges in both positive ($HL = -0.5, RBC = -0.82$) and negative ($HL = -0.5, RBC = -0.77$) contexts.

The Bowker's test results strengthen this observation. Significant asymmetry is observed across all conditions ($\chi^2 = [163.40, 1{,}112],\ p < 0.001$), indicating that disagreements between human and LLM ratings are systematically directional rather than randomly distributed. Each evaluator tends to deviate from the other's ratings in a consistent direction. The Sankey plots in Figure 3 illustrate this pattern clearly: the LLM-as-judge often shifts ratings of LLM-generated responses upward toward the *Good* category, while humans are comparatively more conservative in these cases. Conversely, humans tend to rate human-written responses more favorably than the LLM-as-judge does, while the LLM frequently shifts these ratings downward toward *Okay* or *Bad*. These opposite patterns show that both evaluators apply their own characteristic standards—resulting in upward adjustments for responses that align with their own modality and downward adjustments for those that do not. This asymmetry suggests that humans and the LLM-as-judge rely on different internal criteria when assessing empathy, each favoring the style of responding that is more familiar to them.

7 Discussion

Across both positive and negative emotional contexts, our results reveal a robust pattern: LLMs consistently produce responses judged as more empathetic than those written by humans, a conclusion shared by both humans and the LLM-as-judge. Our results suggest modern LLMs have become highly effective at expressing supportive and emotionally validating language.

One possible explanation for the strong performance of LLMs is that they are trained on large-scale conversational data and optimized through instruction tuning to produce responses that follow recognizable empathetic patterns, such as acknowledging the speaker's feelings, expressing concern, and offering reassurance. These structured patterns may align closely with the criteria implicitly used by evaluators when rating empathy, giving LLM-generated responses an advantage over human responses, which may be shorter, more implicit, or more varied in style.

However, these findings should be interpreted with care. Recent work on AI sycophancy has highlighted that LLMs often exhibit a tendency to over-validate or agree with users, producing responses that emphasize affirmation and emotional alignment even when such responses may not reflect deeper understanding (Sharma et al., 2023; Malmqvist, 2025; Ibrahim et al., 2025). In this context, the higher empathy ratings observed in our study may partly reflect evaluators' preference for explicit expressions of validation rather than richer or more contextually grounded empathic reasoning. This raises the possibility that what is being measured is not empathy in its entirety, but a particular linguistic style of empathy that LLMs are especially optimized to produce. These factors highlight the importance of developing more nuanced evaluation frameworks that distinguish between surface-level expressions of empathy and more substantive empathic engagement.

Our results also reveal a systematic divergence between how humans and LLMs evaluate empathetic responses. Although both groups agree that LLM-generated responses are more empathetic, each group still favors its own perception or style of empathy. Because humans rate human-written responses slightly more favorably than the LLM-as-judge does, and rate LLM responses slightly less favorably, we can cautiously interpret this as humans potentially underrating LLM performance and overrating their own. Conversely, the LLM-as-judge shows the mirror pattern: it evaluates LLM-generated responses more positively than humans do and is more critical of human-written responses. Rather than indicating that either evaluator is "correct" or "biased," these complementary tendencies suggest that humans and LLMs rely on partially different internal criteria when interpreting and expressing empathy, leading each to favor responses aligning more closely with its own style.

This shows that even when humans and LLMs agree on which responses seem more empathetic, each evaluator may be sensitive to different cues when judging empathy — humans may value contextual subtlety, indirect reassurance, or conversational naturalness, whereas LLMs may prioritize explicit emotional validation or stylistic consistency. As a result, even when both groups converge on similar overall rankings, their rating patterns diverge in systematic ways. This highlights the need to view empathy not as a single measurable dimension, but as a construct shaped by the interpretive lens of the evaluator.

Implications: Our finding that LLMs consistently generate responses judged as more empathetic than human responses highlights the substantial potential of these models for applications requiring emotional intelligence. Their ability to produce emotionally attuned and supportive language at scale suggests that LLMs may play an increasingly valuable role in contexts where empathic communication is essential. At the same time, these implications must be carefully contextualized. While LLMs are effective at generating responses perceived as empathetic, empathy alone is not sufficient for high-stakes applications such as mental health support or sensitive social interactions. Effective deployment in such domains requires additional capabilities, including factual reliability, safety, ethical alignment, and the ability to sustain coherent and appropriate behavior over extended interactions. Moreover, over-reliance on simulated empathy may have unintended consequences, such as reduced human-to-human empathic engagement or misplaced trust in systems that do not possess genuine emotional understanding.

The observation that humans and LLMs apply different evaluative tendencies implies the need for careful interpretation of empathy assessments, as each tends to favor responses that align more closely with their own. While our analyses reveal directional differences between humans and the LLM-based evaluations, the underlying mechanisms remain unclear. Understanding whether these divergences arise from characteristics of the training data, human interpretive norms, or differing stylistic preferences is an important direction for future research. This also highlights the need for more robust evaluation frameworks, potentially combining human judgment, calibrated LLM-based assessment, and task-specific criteria, to better capture the multifaceted nature of empathy. As empathy is inherently subjective and context-dependent, progress in this area will benefit from approaches that account for variation in how both humans and LLMs perceive, express, and evaluate empathy in communication.

8 Limitations

While our study provides clear evidence of both human and LLM preferences for LLM-generated empathetic responses, several limitations should be considered when interpreting these findings.

First, our evaluation uses short, single-turn dialogues. Although single-turn interactions provide a controlled and meaningful unit for assessing localized expressions of empathy, they may not capture the richer dynamics of real-world interactions, in which empathy is often expressed and perceived over multiple turns and evolving context. As a result, both LLM performance and human judgments may differ in more naturalistic, multi-turn settings.

Second, the evaluation was conducted in a controlled crowdsourcing environment using decontextualized dialogue scenarios. While this setup enables large-scale and systematic comparison, it may not fully reflect how empathy is perceived in real-world situations, where personal relevance, prior relationships, and trust in the interlocutor can significantly influence judgments.

Third, our study employs a single LLM (GPT-4o-mini) as an automated evaluator in the LLM-as-judge framework. Although this allows for consistency in evaluation, prior work has shown that LLM-based judgments can vary depending on model choice, prompting strategy, and calibration. Therefore, our findings regarding LLM-based evaluation may reflect model-specific biases, and future work should consider multiple LLM judges or ensemble approaches to improve robustness.

Fourth, while we conceptually frame empathy as comprising cognitive, affective, and compassionate components, our evaluation aggregates these dimensions into a single overall rating. This simplification does not allow us to distinguish which aspects of empathy LLMs capture most effectively. Future work should aim to disentangle these components to better understand the nature of LLM-generated empathy.

Finally, differences in response length across human and LLM-generated outputs may influence perceived empathy. As shown in Table 2, LLM responses tend to be longer on average, which may allow for more explicit emotional validation and elaboration. Although we report these differences, we do not explicitly control for or analyze the relationship between response length and empathy ratings. This remains an important direction for future investigation.

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.

This study did not require Institutional Review Board (IRB) approval, as it involved anonymous, voluntary participation in a minimal-risk survey task with no collection of personally identifiable or sensitive information. All participants were recruited through a crowdsourcing platform, provided informed consent, and were free to withdraw at any time. The study design adheres to standard ethical guidelines for human-subject research involving low-risk behavioral data collection.

Transparency and reproducibility of the study: The dialogue prompt-response pairs that were subjected to evaluation along with the participants' and the LLM's 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*.

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*.

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.

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

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.

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

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919*.

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.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

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*.

Lars Malmqvist. 2025. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 61–74. Springer.

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.

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

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.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

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

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.

Ekaterina Svikhnushina and Pearl Pu. 2023. Approximating online human evaluation of social chatbots with prompting. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 268–281, Prague, Czechia. Association for Computational Linguistics.

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.

Henry J Xie, Jinghan Zhang, Xinhao Zhang, and Kunpeng Liu. 2024. Scoring with large language models: A study on measuring empathy of responses in dialogues. *arXiv preprint arXiv:2412.20264*.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

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.

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.

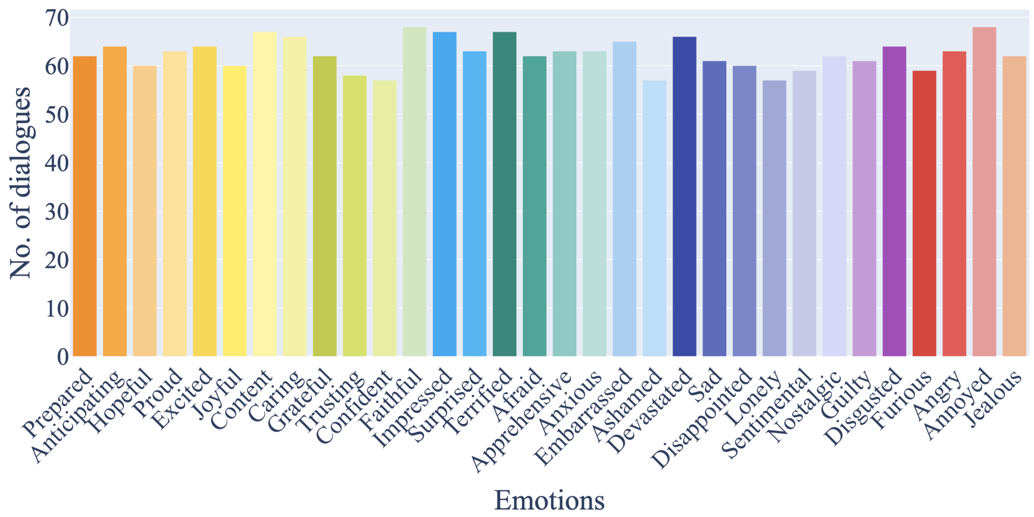


Figure 4: Distribution of the dialogue prompt-response pairs sampled from the EmpatheticDialogues dataset across the 32 positive and negative emotions.

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.

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 Automatic Evaluation Prompts

For automatic evaluation, the LLM judge receives a system instruction explaining the rating task, followed by a user prompt containing: (1) empathy definitions identical to those provided to human raters, and (2) few-shot examples where for each rating (*Bad*, *Okay*, and *Good*), the model is shown both a human-generated and an LLM-generated response that received that rating from human annotators. An example prompt for positive sentiment dialogues is shown in Table 6.

F Computing the Sample Size

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

System	*You are an evaluator of listener responses in conversations. Reply with exactly one word: Bad, Okay, or Good.*
User	*I will present you with responses to emotional situations. You need to rate how empathetic the responses are in terms of 'Bad', 'Okay', or 'Good' compared to how you would have responded in the same situations. I will show you a quick tutorial describing the concept of empathy along with some examples.* *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.* *Affective empathy is the ability to experience the emotions of another person.* *Compassionate empathy is the ability to not only understand and share another person's feelings, but also to be moved to help if needed.*
Example (Good — Human-generated):	
	I am a Speaker feeling caring because I always take care of cats and feed them. I shared these emotions with a Listener in a dialogue: *Speaker: I believe I am a really caring person, because I always take care of cats and feed them and lately they are more and more close to me.* *Listener: That is so kind of you, I also have a soft heart for animals. I happen to be a dog person, but I do have a cat that was my ex boyfriend's.* *How empathetic is the Listener's response compared to how you would have responded for the same situation?* *Answer: Good*
Example (Good — LLM-generated):	
	I am a Speaker feeling caring because I always take care of cats and feed them. I shared these emotions with a Listener in a dialogue: *Speaker: I believe I am a really caring person, because I always take care of cats and feed them and lately they are more and more close to me.* *Listener: Your care and compassion for cats are truly heartwarming. Your actions in feeding and taking care of them reflect your empathy and understanding of their needs. The growing closeness between you and the cats is a testament to the bond you have built through your kindness and concern for their well-being.* *How empathetic is the Listener's response compared to how you would have responded for the same situation?* *Answer: Good*
[Additional few-shot examples for Okay and Bad, each with one human-generated and one LLM-generated response]	
Target response to evaluate:	
	I am a Speaker feeling excited because I got a promotion at work. I shared these emotions with a Listener in a dialogue: *Speaker: I just got promoted to senior manager!* *Listener: That's nice.* *How empathetic is the Listener's response compared to how you would have responded for the same situation?* ***Answer:***

Table 6: Example prompt for LLM-based empathy evaluation (positive sentiment).

Mixtral-8x7B. That is 200 participants per group, which is sufficiently above the minimal sample size.

G 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 **??** and 2 show, we observe no central tendency bias in the human ratings collected utilizing this simpler scale.

H Determing the Effect Size

Jacob Cohen, a renowned psychologist and statistician, introduced standards for evaluating the mag-

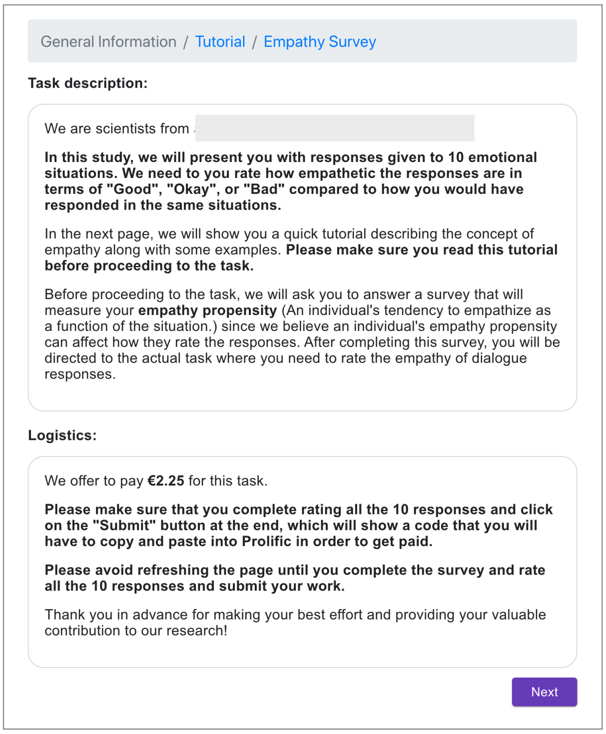


Figure 5: The description of the task.

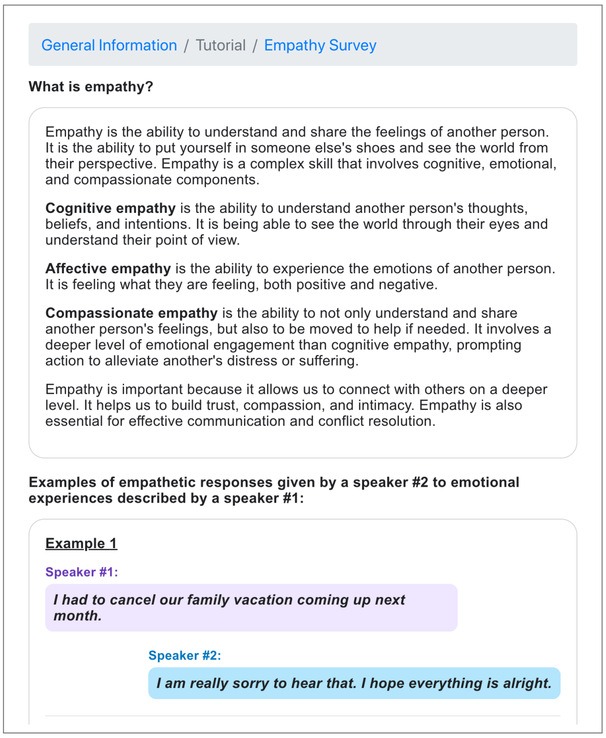


Figure 6: The tutorial.

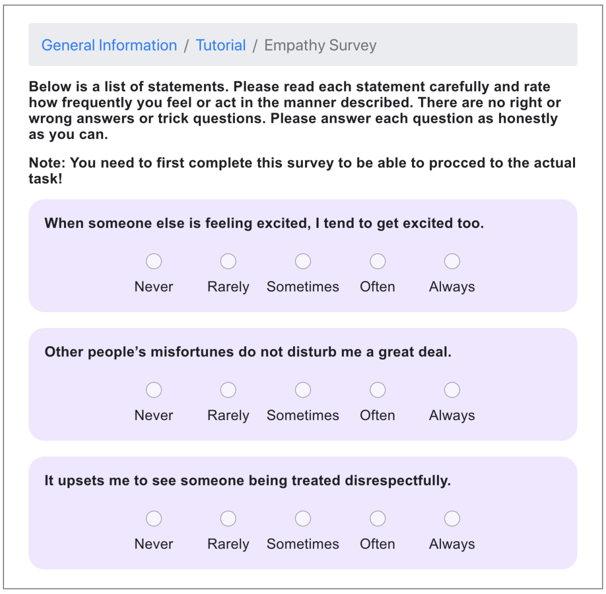


Figure 7: The Toronto Empathy Questionnaire.

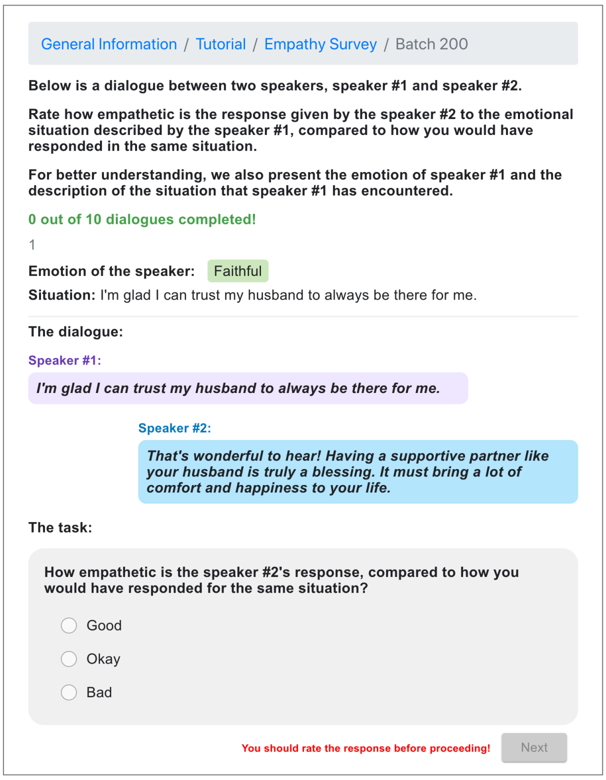


Figure 8: The task interface for rating responses in terms of empathy.

nitude 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 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.

I 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 LLMs are denoted in Table 7. Table 9 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.

J Finer analysis of empathy ratings

Tables 10 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.

K Example dialogue responses

Table 11 denotes some example dialogue situations and responses generated by humans and LLMs and the corresponding ratings given by the human raters.

L 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.

M 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.

N 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.

Sentiment	Evaluator	Rating	Human	GPT	LLaMA	Gemini	Mixtral	χ^2 (9.49)		χ^2 (15.51)	
Positive emotions	Human judges	Bad	133	**34**	48	117	76	**98.88**	***	**138.83**	***
		Okay	294	228	250	283	238	**17.97**	***		
		Good	454	**619**	583	481	567	**94.30**	***		
	LLM-as-judge	Bad	188	14	14	25	**13**	**493.73**	***	**2015.14**	***
		Okay	468	51	50	38	30	**1334.74**	***		
		Good	220	811	812	813	**833**	**2012.68**	***		
Negative emotions	Human judges	Bad	209	**108**	126	131	129	**50.28**	***	**67.04**	***
		Okay	378	335	357	388	365	6.75	(ns)		
		Good	532	**676**	636	600	625	**41.03**	***		
	LLM-as-judge	Bad	408	45	23	22	**16**	**1252.96**	***	**2741.61**	***
		Okay	558	149	74	103	69	**1093.55**	***		
		Good	148	920	1017	989	**1029**	**2649.54**	***		

Table 7: 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.

	Positive emotions χ^2 (5.991) Human judges		LLM-as-judge		Negative emotions χ^2 (5.991) Human judges		LLM-as-judge	
LLMs against human baseline:								
Human Vs GPT	**92.41**	***	**823.70**	***	**51.94**	***	**1085.50**	***
Human Vs LLaMA	**59.52**	***	**826.78**	***	**30.42**	***	**1362.77**	***
Human Vs Gemini	2.01	(ns)	**830.56**	***	**22.11**	***	**1281.76**	***
Human Vs Mixtral	**33.95**	***	**894.44**	***	**26.64**	***	**1403.22**	***
LLMs against each other:								
GPT Vs LLaMA	4.48	(ns)	0.01	(ns)	3.30	(ns)	**37.19**	***
GPT Vs Gemini	**68.86**	***	5.00	(ns)	**10.63**	**	**18.78**	***
GPT Vs Mixtral	**18.53**	***	5.77	(ns)	5.15	(ns)	**49.24**	***
LLaMA Vs Gemini	**40.68**	***	4.73	(ns)	2.44	(ns)	5.16	(ns)
LLaMA Vs Mixtral	**6.84**	*	5.30	(ns)	0.22	(ns)	1.50	(ns)
Gemini Vs Mixtral	**19.65**	***	4.97	(ns)	1.23	(ns)	**8.46**	*

Table 8: 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.

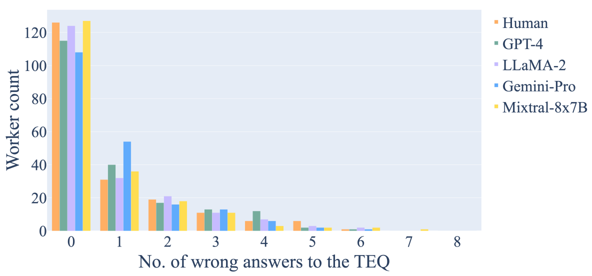


Figure 12: The number of reverse scale questions in the TEQ that were marked wrong by the participants rating the three response groups.

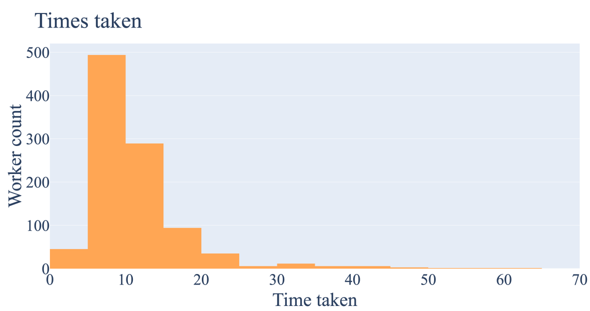
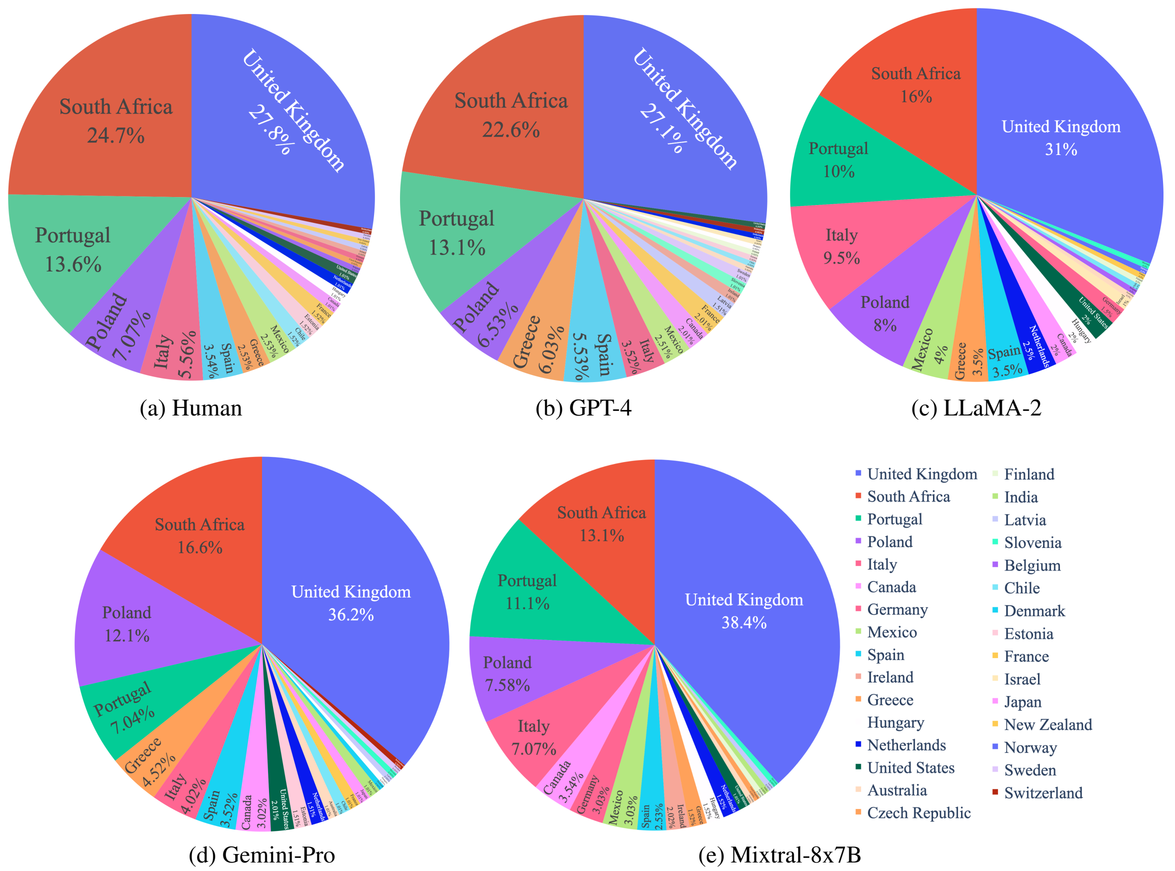


Figure 13: The histogram of times taken to complete the task by all participants.

Sentiment	Source	Human-judges Percentage gain (%) Bad		Okay		Good		LLM-as-judge Percentage gain (%) Bad		Okay		Good	
Positive emotions	GPT Vs Human	**-74.44**	***	**-22.45**	***	**36.34**	***	**-92.55**	***	**-89.10**	***	**268.64**	***
	LLaMA Vs Human	**-63.91**	***	**-14.97**	*	**28.41**	***	**-92.55**	***	**-89.32**	***	**269.09**	***
	Gemini Vs Human	-12.03	(ns)	-3.74	(ns)	5.95	(ns)	**-86.70**	***	**-91.88**	***	**269.55**	***
	Mixtral Vs Human	**-42.86**	***	**-19.05**	**	**24.89**	***	**-93.09**	***	**-93.59**	***	**278.64**	***
Negative emotions	GPT Vs Human	**-48.33**	***	-11.38	(ns)	**27.07**	***	**-88.97**	***	**-73.30**	***	**521.62**	***
	LLaMA Vs Human	**-39.71**	***	-5.56	(ns)	**19.55**	***	**-94.36**	***	**-86.74**	***	**587.16**	***
	Gemini Vs Human	**-37.32**	***	2.65	(ns)	**12.78**	**	**-94.61**	***	**-81.54**	***	**568.24**	***
	Mixtral Vs Human	**-38.28**	***	-3.44	(ns)	**17.48**	***	**-96.08**	***	**-87.63**	***	**595.27**	***

Table 9: The percentage gains obtained by the LLMs in each rating category compared to the human baseline. The statistically significant gains are highlighted in bold. *, **, and *** indicates p < .05, p < .01, and p < .001, respectively for corresponding statistical χ^2 test.



(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.

Emotion	LLM	Human judges Percentage gain (%)						LLM-as-judge Percentage gain (%)					
		Bad		Okay		Good		Bad		Okay		Good	
<u>**Positive emotions:**</u>													
Prepared	GPT	**-90.0**	*	-9.52	(ns)	35.48	(ns)	**-100.0**	**	**-90.24**	***	**533.33**	***
	LLaMA	-50.0	(ns)	-33.33	(ns)	**38.71**	*	**-90.91**	**	**-95.12**	***	**544.44**	***
	Gemini	-20.0	(ns)	19.05	(ns)	-6.45	(ns)	**-100.0**	**	**-100.0**	***	**577.78**	***
	Mixtral	-50.0	(ns)	-14.29	(ns)	25.81	(ns)	**-90.91**	**	**-97.56**	***	**555.56**	***
Anticipa -ting	GPT	-66.67	(ns)	-16.67	(ns)	23.53	(ns)	**-92.86**	***	**-100.0**	***	**238.89**	***
	LLaMA	-16.67	(ns)	-8.33	(ns)	8.82	(ns)	**-92.86**	***	**-90.0**	***	**222.22**	***
	Gemini	0.0	(ns)	25.0	(ns)	-17.65	(ns)	**-92.86**	***	**-96.67**	***	**233.33**	***
	Mixtral	16.67	(ns)	-25.0	(ns)	14.71	(ns)	**-92.86**	***	**-96.67**	***	**233.33**	***
Hopeful	GPT	-33.33	(ns)	-30.0	(ns)	29.03	(ns)	**-86.67**	**	**-90.91**	***	**358.33**	***
	LLaMA	-55.56	(ns)	-35.0	(ns)	**38.71**	*	**-93.33**	***	**-90.91**	***	**366.67**	***
	Gemini	55.56	(ns)	-30.0	(ns)	3.23	(ns)	**-86.67**	**	**-93.94**	***	**366.67**	***
	Mixtral	-33.33	(ns)	-10.0	(ns)	16.13	(ns)	**-93.33**	***	**-96.97**	***	**383.33**	***
Proud	GPT	-70.0	(ns)	-42.86	(ns)	**50.0**	**	**-90.91**	**	**-92.59**	***	**140.0**	***
	LLaMA	**-90.0**	*	-23.81	(ns)	**43.75**	*	**-81.82**	*	**-88.89**	***	**132.0**	***
	Gemini	-30.0	(ns)	-33.33	(ns)	31.25	(ns)	**-72.73**	*	**-88.89**	***	**128.0**	***
	Mixtral	**-100.0**	**	-42.86	(ns)	**59.38**	**	**-81.82**	*	**-96.3**	***	**140.0**	***
Excited	GPT	**-90.91**	**	-17.39	(ns)	**46.67**	*	**-92.31**	**	**-96.3**	***	**158.33**	***
	LLaMA	**-81.82**	*	-17.39	(ns)	**43.33**	*	**-92.31**	**	**-92.59**	***	**154.17**	***
	Gemini	0.0	(ns)	-21.74	(ns)	16.67	(ns)	-61.54	(ns)	**-96.3**	***	**141.67**	***
	Mixtral	-54.55	(ns)	-34.78	(ns)	**46.67**	*	**-92.31**	**	**-88.89**	***	**150.0**	***
Joyful	GPT	**-71.43**	*	-30.77	(ns)	**42.42**	*	**-100.0**	**	**-85.71**	***	**161.9**	***
	LLaMA	**-71.43**	*	53.85	(ns)	9.09	(ns)	**-90.0**	*	**-96.43**	***	**171.43**	***
	Gemini	**-64.29**	*	38.46	(ns)	12.12	(ns)	**-90.0**	*	**-89.29**	***	**161.9**	***
	Mixtral	**-71.43**	*	23.08	(ns)	21.21	(ns)	**-80.0**	*	**-96.43**	***	**166.67**	***
Content	GPT	-85.71	(ns)	-40.0	(ns)	**35.0**	*	**-100.0**	***	**-90.32**	***	**178.26**	***
	LLaMA	-71.43	(ns)	-25.0	(ns)	25.0	(ns)	**-92.31**	**	**-96.77**	***	**182.61**	***
	Gemini	0.0	(ns)	-15.0	(ns)	7.5	(ns)	**-92.31**	**	**-83.87**	***	**165.22**	***
	Mixtral	-42.86	(ns)	-15.0	(ns)	15.0	(ns)	**-100.0**	***	**-96.77**	***	**186.96**	***
Caring	GPT	-33.33	(ns)	16.67	(ns)	-4.44	(ns)	**-87.5**	*	**-100.0**	***	**137.04**	***
	LLaMA	0.0	(ns)	-5.56	(ns)	2.22	(ns)	**-100.0**	*	**-93.33**	***	**133.33**	***
	Gemini	200.0	(ns)	-11.11	(ns)	-8.89	(ns)	**-100.0**	*	**-96.67**	***	**137.04**	***
	Mixtral	33.33	(ns)	-5.56	(ns)	0.0	(ns)	**-100.0**	*	**-93.33**	***	**133.33**	***
Grateful	GPT	**-90.91**	**	-28.0	(ns)	**65.38**	**	**-100.0**	***	**-85.71**	***	**470.0**	***
	LLaMA	**-72.73**	*	-36.0	(ns)	**65.38**	**	**-100.0**	***	**-82.86**	***	**460.0**	***
	Gemini	-36.36	(ns)	-16.0	(ns)	30.77	(ns)	**-94.12**	***	**-94.29**	***	**490.0**	***
	Mixtral	-36.36	(ns)	-44.0	(ns)	**57.69**	*	**-94.12**	***	**-91.43**	***	**480.0**	***
Trusting	GPT	**-72.73**	*	22.22	(ns)	13.79	(ns)	**-80.0**	**	**-80.56**	***	**585.71**	***
	LLaMA	**-81.82**	*	11.11	(ns)	24.14	(ns)	**-86.67**	**	**-88.89**	***	**642.86**	***
	Gemini	-27.27	(ns)	27.78	(ns)	-6.9	(ns)	**-80.0**	**	**-88.89**	***	**628.57**	***
	Mixtral	-27.27	(ns)	-33.33	(ns)	31.03	(ns)	**-93.33**	***	**-94.44**	***	**685.71**	***
Confident	GPT	**-87.5**	*	-41.18	(ns)	**43.75**	**	**-80.0**	*	**-81.82**	***	**250.0**	***
	LLaMA	-50.0	(ns)	11.76	(ns)	6.25	(ns)	**-90.0**	*	**-75.76**	***	**242.86**	***
	Gemini	0.0	(ns)	5.88	(ns)	-3.12	(ns)	-70.0	(ns)	**-84.85**	***	**250.0**	***
	Mixtral	-75.0	(ns)	-11.76	(ns)	25.0	(ns)	**-80.0**	*	**-96.97**	***	**285.71**	***
Faithful	GPT	-37.5	(ns)	-18.52	(ns)	24.24	(ns)	**-87.5**	**	**-95.45**	***	**700.0**	***
	LLaMA	-37.5	(ns)	-18.52	(ns)	24.24	(ns)	**-87.5**	**	**-81.82**	***	**625.0**	***
	Gemini	-12.5	(ns)	-14.81	(ns)	15.15	(ns)	**-100.0**	***	**-95.45**	***	**725.0**	***
	Mixtral	-37.5	(ns)	-7.41	(ns)	15.15	(ns)	**-100.0**	***	**-90.91**	***	**700.0**	***

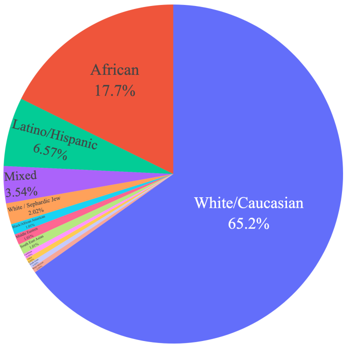
Impressed	GPT	**-80.0**	*	**-47.83**	*	**55.88**	**	**-100.0**	***	**-77.78**	***	**391.67**	***
	LLaMA	-50.0	(ns)	-21.74	(ns)	29.41	(ns)	**-94.74**	***	**-83.33**	***	**400.0**	***
	Gemini	10.0	(ns)	-8.7	(ns)	2.94	(ns)	**-78.95**	**	**-86.11**	***	**383.33**	***
	Mixtral	-10.0	(ns)	-8.7	(ns)	8.82	(ns)	**-100.0**	***	**-88.89**	***	**425.0**	***
Surprised	GPT	**-86.67**	**	-25.0	(ns)	**79.17**	**	**-93.75**	***	**-83.78**	***	**460.0**	***
	LLaMA	**-80.0**	**	-20.83	(ns)	**70.83**	**	**-100.0**	***	**-97.3**	***	**520.0**	***
	Gemini	-40.0	(ns)	4.17	(ns)	20.83	(ns)	**-93.75**	***	**-89.19**	***	**480.0**	***
	Mixtral	-33.33	(ns)	-16.67	(ns)	37.5	(ns)	**-93.75**	***	**-86.49**	***	**470.0**	***
Negative emotions:													
Terrified	GPT	-46.67	(ns)	-4.55	(ns)	26.67	(ns)	**-78.79**	***	**-69.23**	***	**550.0**	***
	LLaMA	-40.0	(ns)	-9.09	(ns)	26.67	(ns)	**-87.88**	***	**-88.46**	***	**650.0**	***
	Gemini	-46.67	(ns)	18.18	(ns)	10.0	(ns)	**-93.94**	***	**-84.62**	***	**662.5**	***
	Mixtral	-6.67	(ns)	-27.27	(ns)	23.33	(ns)	**-90.91**	***	**-84.62**	***	**650.0**	***
Afraid	GPT	**-66.67**	*	0.0	(ns)	**46.15**	*	**-86.21**	***	**-66.67**	***	**1500.0**	***
	LLaMA	**-72.22**	**	0.0	(ns)	**50.0**	*	**-93.1**	***	**-76.67**	***	**1666.67**	***
	Gemini	**-55.56**	*	33.33	(ns)	15.38	(ns)	**-96.55**	***	**-73.33**	***	**1666.67**	***
	Mixtral	-50.0	(ns)	11.11	(ns)	26.92	(ns)	**-96.55**	***	**-93.33**	***	**1866.67**	***
Apprehen -sive	GPT	**-90.0**	*	**-60.71**	**	**104.0**	***	**-100.0**	***	**-87.5**	***	**1833.33**	***
	LLaMA	-50.0	(ns)	-28.57	(ns)	**52.0**	*	**-100.0**	***	**-97.5**	***	**1966.67**	***
	Gemini	-40.0	(ns)	-39.29	(ns)	**60.0**	*	**-100.0**	***	**-87.5**	***	**1833.33**	***
	Mixtral	-70.0	(ns)	-14.29	(ns)	44.0	(ns)	**-100.0**	***	**-87.5**	***	**1833.33**	***
Anxious	GPT	-50.0	(ns)	**-44.44**	*	**75.0**	**	**-93.75**	***	**-85.0**	***	**700.0**	***
	LLaMA	-41.67	(ns)	-37.04	(ns)	**62.5**	*	**-87.5**	***	**-97.5**	***	**757.14**	***
	Gemini	-41.67	(ns)	-37.04	(ns)	**62.5**	*	**-93.75**	***	**-92.5**	***	**742.86**	***
	Mixtral	-66.67	(ns)	-14.81	(ns)	**50.0**	*	**-93.75**	***	**-92.5**	***	**742.86**	***
Embarra -ssed	GPT	-47.06	(ns)	10.53	(ns)	20.69	(ns)	**-80.77**	***	**-69.7**	***	**733.33**	***
	LLaMA	-23.53	(ns)	5.26	(ns)	10.34	(ns)	**-96.15**	***	**-93.94**	***	**933.33**	***
	Gemini	-47.06	(ns)	10.53	(ns)	20.69	(ns)	**-96.15**	***	**-90.91**	***	**916.67**	***
	Mixtral	-29.41	(ns)	-10.53	(ns)	24.14	(ns)	**-96.15**	***	**-96.97**	***	**950.0**	***
Ashamed	GPT	-41.67	(ns)	0.0	(ns)	16.67	(ns)	**-100.0**	***	-4.55	(ns)	**580.0**	***
	LLaMA	-58.33	(ns)	60.0	(ns)	-6.67	(ns)	**-100.0**	***	**-68.18**	**	**860.0**	***
	Gemini	-58.33	(ns)	40.0	(ns)	3.33	(ns)	**-92.86**	***	-40.91	(ns)	**700.0**	***
	Mixtral	-25.0	(ns)	33.33	(ns)	-6.67	(ns)	**-100.0**	***	**-68.18**	**	**860.0**	***
Devastated	GPT	-33.33	(ns)	-40.0	(ns)	29.73	(ns)	**-93.75**	***	**-100.0**	***	**425.0**	***
	LLaMA	-44.44	(ns)	-15.0	(ns)	18.92	(ns)	**-100.0**	***	**-95.24**	***	**433.33**	***
	Gemini	-44.44	(ns)	-30.0	(ns)	27.03	(ns)	**-100.0**	***	**-90.48**	***	**425.0**	***
	Mixtral	-66.67	(ns)	30.0	(ns)	0.0	(ns)	**-100.0**	***	**-100.0**	***	**441.67**	***
Sad	GPT	-27.27	(ns)	20.0	(ns)	0.0	(ns)	**-78.57**	***	**-71.43**	**	**308.33**	***
	LLaMA	-27.27	(ns)	0.0	(ns)	8.57	(ns)	**-100.0**	***	**-80.95**	***	**375.0**	***
	Gemini	**-72.73**	*	20.0	(ns)	14.29	(ns)	**-92.86**	***	**-90.48**	***	**375.0**	***
	Mixtral	-54.55	(ns)	-13.33	(ns)	22.86	(ns)	**-96.43**	***	**-80.95**	***	**366.67**	***
Disappoi -nted	GPT	-54.55	(ns)	-15.0	(ns)	31.03	(ns)	**-90.48**	***	**-81.25**	***	**642.86**	***
	LLaMA	-45.45	(ns)	-10.0	(ns)	24.14	(ns)	**-95.24**	***	**-93.75**	***	**714.29**	***
	Gemini	-18.18	(ns)	35.0	(ns)	-17.24	(ns)	**-95.24**	***	**-87.5**	***	**685.71**	***
	Mixtral	-54.55	(ns)	10.0	(ns)	13.79	(ns)	**-100.0**	***	**-96.88**	***	**742.86**	***
Lonely	GPT	-12.5	(ns)	-5.88	(ns)	6.25	(ns)	**-70.37**	***	**-60.87**	**	**471.43**	***
	LLaMA	-12.5	(ns)	11.76	(ns)	-3.12	(ns)	**-92.59**	***	**-86.96**	***	**642.86**	***
	Gemini	-62.5	(ns)	-17.65	(ns)	25.0	(ns)	**-92.59**	***	**-73.91**	***	**600.0**	***
	Mixtral	-62.5	(ns)	11.76	(ns)	9.38	(ns)	**-100.0**	***	**-73.91**	***	**628.57**	***
Sentimen -tal	GPT	-40.0	(ns)	-11.11	(ns)	11.11	(ns)	**-100.0**	***	**-93.1**	***	**256.25**	***
	LLaMA	-60.0	(ns)	-11.11	(ns)	13.89	(ns)	**-100.0**	***	**-96.55**	***	**262.5**	***
	Gemini	20.0	(ns)	11.11	(ns)	-8.33	(ns)	**-100.0**	***	**-100.0**	***	**268.75**	***
	Mixtral	40.0	(ns)	-27.78	(ns)	8.33	(ns)	**-100.0**	***	**-100.0**	***	**268.75**	***

Nostalgic	GPT	-85.71	(ns)	-4.76	(ns)	20.59	(ns)	-100.0	(ns)	**-91.43**	***	**145.83**	***
	LLaMA	-71.43	(ns)	-9.52	(ns)	20.59	(ns)	-100.0	(ns)	**-94.29**	***	**150.0**	***
	Gemini	-71.43	(ns)	4.76	(ns)	11.76	(ns)	-66.67	(ns)	**-100.0**	***	**154.17**	***
	Mixtral	-57.14	(ns)	-14.29	(ns)	20.59	(ns)	-66.67	(ns)	**-94.29**	***	**145.83**	***
Guilty	GPT	-38.46	(ns)	22.22	(ns)	3.33	(ns)	**-96.55**	***	**-45.16**	*	**4200.0**	***
	LLaMA	-46.15	(ns)	-16.67	(ns)	30.0	(ns)	**-96.55**	***	**-87.1**	***	**5500.0**	***
	Gemini	-38.46	(ns)	-5.56	(ns)	20.0	(ns)	**-93.1**	***	**-77.42**	***	**5100.0**	***
	Mixtral	**-69.23**	*	11.11	(ns)	23.33	(ns)	**-100.0**	***	**-83.87**	***	**5500.0**	***
Disgusted	GPT	-43.75	(ns)	27.27	(ns)	3.85	(ns)	**-91.67**	***	**-54.29**	**	**820.0**	***
	LLaMA	0.0	(ns)	4.55	(ns)	-3.85	(ns)	**-83.33**	***	**-77.14**	***	**940.0**	***
	Gemini	-43.75	(ns)	36.36	(ns)	-3.85	(ns)	**-87.5**	***	**-57.14**	***	**820.0**	***
	Mixtral	6.25	(ns)	-4.55	(ns)	0.0	(ns)	**-95.83**	***	**-82.86**	***	**1040.0**	***
Furious	GPT	-46.15	(ns)	15.0	(ns)	11.54	(ns)	**-89.47**	***	**-83.87**	***	**537.5**	***
	LLaMA	-15.38	(ns)	35.0	(ns)	-19.23	(ns)	**-84.21**	***	**-83.87**	***	**525.0**	***
	Gemini	23.08	(ns)	5.0	(ns)	-15.38	(ns)	**-89.47**	***	**-83.87**	***	**537.5**	***
	Mixtral	7.69	(ns)	-10.0	(ns)	3.85	(ns)	**-84.21**	***	**-83.87**	***	**525.0**	***
Angry	GPT	-37.5	(ns)	-24.14	(ns)	38.46	(ns)	**-88.89**	***	**-83.78**	***	**587.5**	***
	LLaMA	-50.0	(ns)	-24.14	(ns)	42.31	(ns)	**-100.0**	***	**-89.19**	***	**637.5**	***
	Gemini	25.0	(ns)	-17.24	(ns)	11.54	(ns)	**-100.0**	***	**-81.08**	***	**600.0**	***
	Mixtral	-12.5	(ns)	-17.24	(ns)	23.08	(ns)	**-100.0**	***	**-94.59**	***	**662.5**	***
Annoyed	GPT	**-66.67**	*	-23.08	(ns)	**59.26**	*	**-94.12**	***	**-82.93**	***	**555.56**	***
	LLaMA	-53.33	(ns)	3.85	(ns)	25.93	(ns)	**-100.0**	***	**-85.37**	***	**577.78**	***
	Gemini	-20.0	(ns)	23.08	(ns)	-11.11	(ns)	**-100.0**	***	**-80.49**	***	**555.56**	***
	Mixtral	**-86.67**	**	-3.85	(ns)	**51.85**	*	**-100.0**	***	**-80.49**	***	**555.56**	***
Jealous	GPT	-22.22	(ns)	-13.04	(ns)	16.67	(ns)	**-91.67**	***	**-61.29**	***	**585.71**	***
	LLaMA	-22.22	(ns)	-13.04	(ns)	16.67	(ns)	**-87.5**	***	**-58.06**	**	**557.14**	***
	Gemini	-44.44	(ns)	0.0	(ns)	13.33	(ns)	**-91.67**	***	**-64.52**	***	**600.0**	***
	Mixtral	-11.11	(ns)	13.04	(ns)	-6.67	(ns)	**-83.33**	***	**-74.19**	***	**614.29**	***

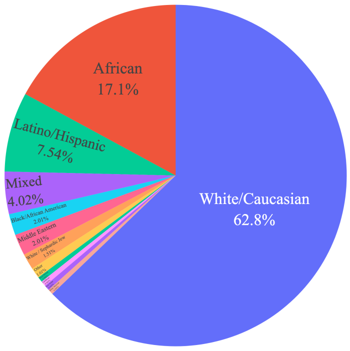
Table 10: Percentage gains obtained by the four LLMs' response ratings compared to the human baseline when responding to dialogue prompts containing positive and negative emotions. The statistically significant gains are highlighted in bold. *, **, and *** indicates p < .05, p < .01, and p < .001, respectively 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?* **(Human judge: *Bad*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *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.* **(Human judge: *Okay*; LLM-as-judge: *Bad*)**
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.* **(Human judge: *Good*; LLM-as-judge: *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?* **(Human judge: *Good*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *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?* **(Human judge: *Okay*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *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?* **(Human judge: *Good*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *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.* **(Human judge: *Good*; LLM-as-judge: *Good*)**

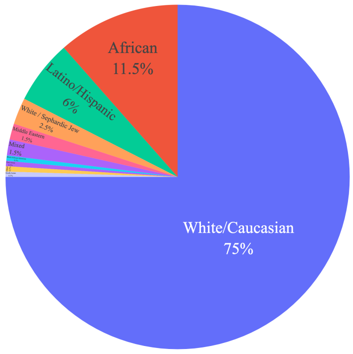
Table 11: Example dialogues with human and LLM responses, rated by human judges and LLM-as-a-judge.



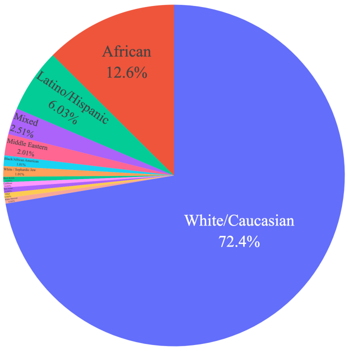
(a) Human



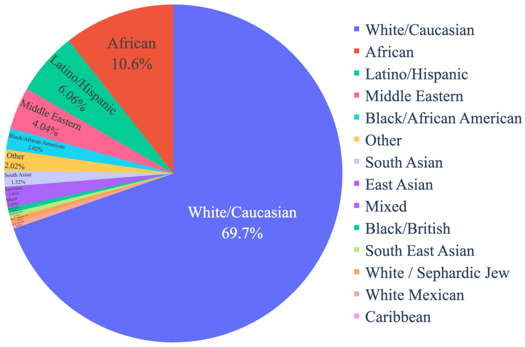
(b) GPT-4



(c) LLaMA-2

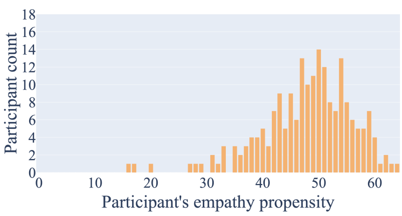


(d) Gemini-Pro

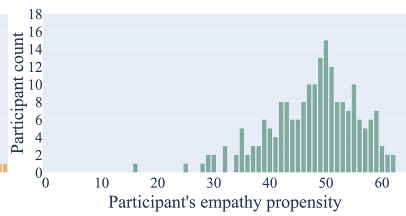


(e) Mixtral-8x7B

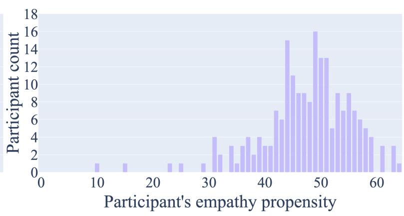
Figure 10: Distribution of the ethnicities of the participants across the five groups.



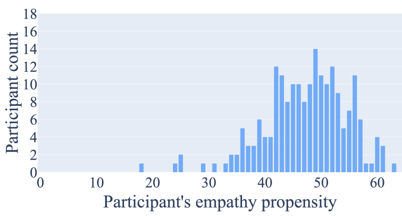
(a) Human response raters



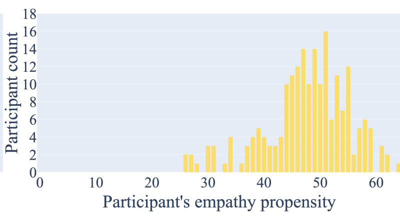
(b) GPT-4 responses raters



(c) LLaMA-2 responses raters



(d) Gemini-Pro responses raters



(e) Mixtral-8x7B responses raters

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