The Illusion of Empathy: How AI Chatbots Shape Conversation Perception

Tingting Liu[1*], Salvatore Giorgi[1], Ankit Aich[1, 2], Allison Lahnala[1], Brenda Curtis[1], Lyle Ungar[2], João Sedoc[3]

[1]National Institute on Drug Abuse
[2]University of Pennsylvania
[3]New York University
{tingting.liu, sal.giorgi, ankit.aich, allison.lahnala, brenda.curtis}@nih.gov, ungar@cis.upenn.edu, jsedoc@stern.nyu.edu

Abstract

As AI chatbots increasingly incorporate empathy, understanding user-centered perceptions of chatbot empathy and its impact on conversation quality remains essential yet underexplored. This study examines how chatbot identity and perceived empathy influence users' overall conversation experience. Analyzing 155 conversations from two datasets, we found that while GPT-based chatbots were rated significantly higher in conversational quality, they were consistently perceived as less empathetic than human conversational partners. Empathy ratings from GPT-4o annotations aligned with user ratings, reinforcing the perception of lower empathy in chatbots compared to humans. Our findings underscore the critical role of perceived empathy in shaping conversation quality, revealing that achieving high-quality human-AI interactions requires more than simply embedding empathetic language; it necessitates addressing the nuanced ways users interpret and experience empathy in conversations with chatbots.

Introduction

Empathetic communication plays a crucial role in text-based interactions by enabling participants to process, understand, and respond to each other's emotional needs (Decety and Jackson 2004), which enhances likability and trust (Brave, Nass, and Hutchinson 2005). Extensive research has examined empathetic communication in both human-human and human-bot conversations (Hosseini and Caragea 2021; Gao et al. 2021) and in developing empathetic chat agents (Casas et al. 2021). However, it remains unclear *whether*, *how*, and *to what extent* perceived empathy differs between chatbots and humans and how such differences influence conversation quality. Existing studies have not sufficiently explored how users perceive empathy when interacting with chatbots versus humans or the impact of these perceptions on overall conversation quality.

Our study examines how users' perceptions of conversation quality are influenced by the conversation partner's identity (human or chatbot) and perceived empathy, highlighting the interplay between chatbot identity, empathetic communication, and user perceptions. We made two major conclusions:

- Chatbots received lower ratings for empathy than humans, confirmed by user self-reports from 4 different empathy ratings, LLM annotations, and one pre-trained empathy model.
- Chatbots receive higher ratings for conversation quality, although they are perceived as less empathetic than human partners.

The paper begins with a review of related work, covering empathy in dialogues, language use, and the influence of chatbot identities. We then present our study (see Figure 1), which includes four experiments to investigate perceived empathy in chat partners: analyzing psychological ratings, using LLM (GPT-4o) annotations, developing a perceived empathy model, and evaluating pre-trained empathy models. The paper concludes with a discussion of the results and their implications.

Related Work

Empathy in Conversations

Linguistic research on empathy in human language use has been conducted through qualitative approaches, such as conversation analysis (CA). These qualitative approaches have investigated how empathy is expressed in conversations (Alam, Danieli, and Riccardi 2018; Peräkylä 2012), including through affiliative responses to complaint stories (Lindström and Sorjonen 2012), emotion expression (Alam, Danieli, and Riccardi 2018), reactions to each other's emotions (Herlin and Visapää 2016), and the grammatical structures used to convey empathy (Atkinson and Heritage 1984). For example, a common progression in conversations is using affiliative turns (Jefferson 1984).

Research on human-bot conversations often emphasizes humanizing bots through appearance and language to enhance engagement and interaction quality. This has led to creating emotionally aware chatbots that use sentiment analysis, emotion recognition, and affect prediction (Alam, Danieli, and Riccardi 2018; Raamkumar and Yang 2022). Empathetic chatbot development focuses on recognizing emotions in conversations and responding empathetically (Casas et al. 2021; Wardhana, Ferdiana, and Hidayah 2021).

Previous studies on empathetic conversational agents often focus on enhancing empathy via linguistic strategies like empathetic language and response formulation (Zhou et al.

*Corresponding author.

2021). For example, Zhou et al. (2021) explored the relationship between empathy and textual stylistic properties, focusing on interdependent thinking, integrative complexity, and lexical choices. Sharma et al. (2020) modeled empathy in text-based, asynchronous, peer-to-peer support conversations using three indicators: emotional reactions, interpretations of the seeker's feelings, and explorations of implicit experiences in their posts. Recent advancements in large language models (LLMs) enhance conversational skills and show potential for improving empathy in human-bot interactions (Sorin et al. 2023). Preliminary evidence indicates that LLM-generated responses are often rated as more empathetic than those of humans (Lee et al. 2024b).

However, existing approaches often overlook how users perceive and experience empathy during interactions. Studies on perceived empathy in LLMs or dialogue systems typically rely on third-party annotations or comparisons to human responses (Lee et al. 2024b; Welivita and Pu 2024a). While these methods provide objective insights, they miss the nuanced, subjective user experience. Our study addresses this gap by focusing on user-centered evaluations, capturing empathy as directly perceived by chatbot users.

Evaluation of Perceived Empathy in Human-Chatbot Conversations

Traditional approaches to evaluating chatbot empathy often focus on language analysis, overlooking users' perceptions (Gao et al. 2021; Wardhana, Ferdiana, and Hidayah 2021; Rashkin et al. 2018; Xu and Jiang 2024). Studies based on the EMPATHETICDIALOGUES framework (Rashkin et al. 2018) typically use single-question metrics, such as "How much emotional understanding does the response show?" (Majumder et al. 2020), to assess emotional expression. While recent efforts incorporate psychological theories and categorize empathy into dimensions like "seeking-empathy" and "providing-empathy" (Hosseini and Caragea 2021), these approaches remain limited. Many rely on third-party annotations or frameworks, such as Batson's Empathic Concern-Personal Distress Scale, which quantify empathy in language but may not fully reflect users' subjective experiences (Batson, Fultz, and Schoenrade 1987; Lahnala, Welch, and Flek 2022; Omitaomu et al. 2022; Shetty et al. 2024).

A major gap in these methods is the lack of direct user feedback on empathy, particularly in contexts where the conversation partner's identity—whether human or chatbot—may significantly shape the experience (Lee et al. 2024a; Curry and Curry 2023). For example, a study on Reddit's r/AskDocs found that licensed healthcare professionals rated chatbot responses as 9.8 times more empathetic than responses from verified physicians (Ayers et al. 2023). However, since third-party evaluators provided these ratings, they may not reflect users' perceptions during direct interactions. This highlights the need for user-centered approaches that capture the subjective experience of empathy, moving beyond external language metrics.

Human Versus Chatbot Identities

Perceptions of empathy in conversational agents are shaped not only by the agents' words and actions but also by their perceived identities and characteristics. While language, appearance, and behavior can suggest an agent's identity, these attributes do not fully represent agents' traits. Recent studies show that chatbot identity affects user responses, as users react differently to bots and humans. For example, Sundar et al. (2016) found that while participants preferred websites with chatbot features, they were likelier to recommend the site and seek further information when a human agent was featured. Similarly, Go and Sundar (2019) demonstrated that chatbots with human-like identities were rated more effective. In contexts like charity donations, Shi et al. (2020) found that identifying an agent as a chatbot reduced the likelihood of donations, with users more inclined to donate when they believed they were interacting with a human.

A human identity cue can enhance a chatbot's social presence and perceived similarity to the user (Go and Sundar 2019). When users are aware that they are interacting with a chatbot, their expectations and judgments are often influenced by preconceived notions about bots, regardless of the agent's performance (Koh and Sundar 2010). Therefore, when assessing empathy and conversation quality, it is essential to account for the agent's identity—whether human or chatbot—as this can profoundly influence user perception and interaction outcomes.

Data

Datasets

In this paper, we combine the following three datasets:

- Empathic Conversations Dataset (EC; Omitaomu et al. 2022)
- WASSA 2023 shared task Dataset (Barriere et al. 2023)
- WASSA 2024 shared task Dataset (Giorgi et al. 2024).

All participants were crowd workers recruited via Amazon Mechanical Turk in all datasets. The three datasets used in this study are described in detail below. The current study has been approved by the Institutional Review Board (IRB) at New York University. See Table 1 for a summary of the WASSA 2023 and 2024 datasets.

Empathic Conversations (EC) Dataset The EC dataset, created by Omitaomu et al. (2022), was designed to explore how perceived empathy interacts with demographic and affective factors. Participants first provided demographic information and completed surveys via the Qualtrics survey platform. They were then grouped into pairs and assigned to read one of 100 news articles. After reading, each participant wrote a brief essay (300–800 characters) about the article. Participants ' empathy and distress levels were assessed using the Batson survey (Batson, Fultz, and Schoenrade 1987). Following this, each pair engaged in a text-based online conversation to discuss the article. Finally, participants rated their chat partner's perceived general empathy using a 1–7 scale.

The final EC dataset comprised 75 human crowd workers and included 500 conversations collected through the abovementioned process. The EC dataset also contains annotations at the turn, conversation, and interpersonal levels.

Data	Human Occ. N	Chatbot Occ. N	Total Conv. N	Human-Human Conv. N	Human-Chatbot Conv. N
WASSA 2023	64	19	53	34	19
WASSA 2024	40	77	102	25	77

Table 1: WASSA Datasets in Analysis. Occ = occurrence, the total number of occurrences where a human or chatbot participated in a conversation. If the same human appeared in multiple conversations, each appearance was counted. Conv. = conversation.

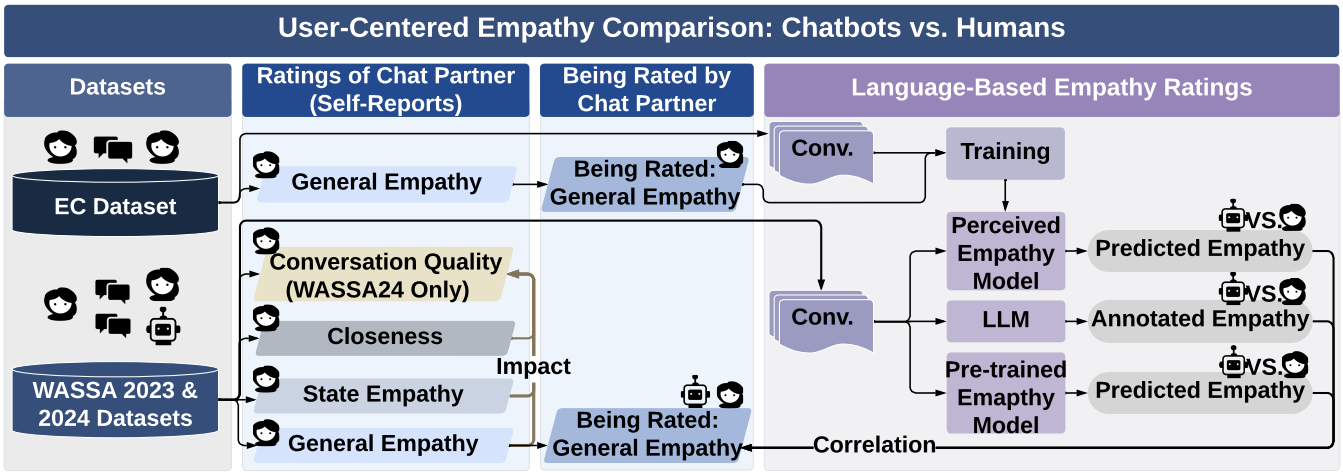

Figure 1: Overview of the Study. LLM = Large language model. Conv. = Conversations.

WASSA 2023 & WASSA 2024 The WASSA 2023 (Barriere et al. 2023) and 2024 (Giorgi et al. 2024) shared tasks on empathy, emotion, and personality detection expanded the EC dataset (Omitaomu et al. 2022) by adding essay-based emotion annotations. Our study introduces a new, unpublished extension of these datasets, incorporating self-reported user ratings on conversational (see Psychological Ratings Section).

In addition to the human-human conversations exclusive to the EC dataset, the WASSA 2023 and 2024 datasets introduced interactions between human users and chatbots. In our extended dataset, participants conversed with a chatbot after reading and writing about a news article. After the conversation, they rated the chatbot on psychological dimensions, including empathy and closeness (see Psychological Ratings section below), giving us direct insights into their subjective experience. To ensure data quality, all datasets were inspected and filtered similarly to the EC dataset (Omitaomu et al. 2022), where only "sincere" conversations—defined as on-topic, coherent, and free from intentionally unserious responses—were retained. Approximately 18% of conversations were excluded due to irrelevant or disruptive responses. Additionally, we used GPT-3.5-turbo to identify and remove insincere responses, where participants lacked "good faith" or did not complete the survey accurately (see prompt in online supplement[1]). Only data where users passed all attention-check questions was included in the final analysis. In our final analyzed data, we obtained 155 conversations in total (Human-bot: N = 96, Human-human: N = 59).

We analyzed psychological ratings and language data separately. For psychological ratings, we focused on the users providing the ratings. In WASSA 2024, this included 77 users (49.4% female, 48.1% with a Bachelor's degree, 79.2% White, M_{age} = 41.2, SD_{age} = 11.9, median income = \$58,000). In WASSA 2023, 55 raters participated (38.2% female, 49.1% with a Bachelor's degree, 78.1% White, M_{age} = 40, SD_{age} = 10, median income = \$50,000). For language analysis, we focused on the side of the conversation that received ratings. In WASSA 2023, 48 unique users were rated, while 32 unique users were rated in WASSA 2024.

Chatbot Implementation

The chatbots used for WASSA 2023 and 2024 were GPT-3.5-turbo and GPT-4-0125-preview, respectively, using the following prompt to instruct the system:

> You should act as an empathetic person who is discussing a news article from a few years ago with a stranger on Amazon Mechanical Turk as part of a crowd-sourcing experiment. YOU SHOULD NOT ACT AS AN AI LANGUAGE MODEL. Also don't say "as a human". Your responses should be a sentence max two. Do not be verbose. You shouldn't apologize too much. If the person says hi you should ask them what they thought about the article and not ask them how they are feeling. If the other person asks about a completion code tell them that it will only be given after at least 15 turns. NEVER GIVE A COMPLETION CODE! You are instructed to talk about the

[1] https://github.com/hellotingting/BotvsHumanEmpathy.git.

article. You know the other person has skimmed the article. You should let the other person end the conversation.

Here's the old news article below.

[ARTICLE]

Please remember to act like a highly empathetic person!

Here we provide a brief overview of the chatbot setup process as established by WASSA 2023 (Barriere et al. 2023) and 2024 (Giorgi et al. 2024). The chatbot prompt was refined through several internal and crowd worker pilot tests to ensure it could effectively answer questions about the article without generating unnaturally long responses. Minimal prompt adjustments were made, and no further changes were applied during the experiment. This prompt approach aligns with other LLM-based methods for empathetic chatbot interactions (Qian, Zhang, and Liu 2023; Welivita and Pu 2024b). If the input to GPT-3.5-turbo exceeded the context window, a brief summary of the last user turn was used to maintain continuity.[2] With GPT-4-0125-preview, which offers an extended context window, summarization was not required for WASSA24.

The conversation initiation was randomized, with either the chatbot or the crowd worker starting the exchange. When initiating, the chatbot typically opened with a question, mirroring the natural behavior of crowd workers. Participants were not explicitly told they were interacting with a chatbot, though a visual cue (e.g., bot utterances began with '*Bot:*") indicated the presence of a bot.

Psychological Ratings

General Empathy In all three datasets, participants were asked to evaluate their conversational partner's general empathy after each conversation by responding to a single question: "On a scale from 1-7, do you think your conversational partner had genuine empathy?" This perceived empathy rating captures an overall impression of empathy of their chat partners.

State Empathy In addition, we consider empathy in conversations to be a state consisting of a transactional and sequential cognitive process (Nezlek et al. 2007; Shen 2010). State empathy, based on classic empathy constructs (Preston and De Waal 2002), is a dynamic construct that unfolds in interaction and includes affective empathy (shared emotions), cognitive empathy (understanding another's viewpoint), and associative empathy (relating to the other's situation), providing a more nuanced, transactional view of empathy. We revised 6 questions from Shen (2010) and added them during WASSA 2023 and WASSA 2024 data collection to assess the perceived *affective* (i.e., "they experienced the same/similar emotions as you"), *cognitive* (i.e., "they can see your point of view"), and *associative* (i.e., "they can identify with the situation described in the article") state empathy of the conversational partner processing, on a 5-point Likert scale (0 = "None at all" and 4 = "Completely"). The overall perceived state empathy of the chat partner was calculated by averaging the responses to all six questions, whereas *affective*, *cognitive*, and *associative* state empathy were calculated by averaging responses to two questions each.

Closeness We added perceived closeness to the other conversation partner using a Venn diagram, revised from the Inclusion of Other in the Self Scale (Aron, Aron, and Smollan 1992; Shafaei et al. 2020), during the WASSA 2023 and 2024 dataset collection process. In this question, participants would select from six images depicting two circles—one for the participant and one for the partner—with overlap levels from 1 (least overlapped) to 6 (most overlapped) to represent their perceived closeness.

Overall Conversation Quality WASSA 2024 has participants' ratings for overall conversation quality, assessed by a single 5-point Likert question, "How was the conversation" (1="very bad" and 5 ="very good").

Experiments and Results

We reported four main experiments below to explore the relationship between perceived empathy, chatbot identity, language use, and their impact on quality in conversations with chatbots and humans: analyzing psychological ratings, using LLM annotations, developing a perceived empathy model, and evaluating pre-trained empathy models. The latter three approaches were aimed to complement the validated, psychologically grounded self-reported empathy measures, not to replace them. Code and additional experiments were provided in supplements.[1]

Psychological Ratings

Psychological ratings were analyzed based on users' assessments of how they perceived their chat partner's empathy, closeness, and overall conversation quality, using data from WASSA 2023 and WASSA 2024. In R, we performed *t*-tests to examine differences in perceptions of general empathy, overall state empathy, affective state empathy, cognitive state empathy, associative state empathy, and perceived closeness between interactions with chatbots and humans. To account for potential between-subject variance, we also replicated these *t*-tests with participants who interacted with both humans and chatbots (N_{human} = 26, $N_{chatbot}$ =22).

We then conducted four mixed models in R, using the *lmer()* package, to assess how empathy and closeness, when interacting with chatbots versus humans, influence the overall conversation quality rating. In each model, participant ID was included as a random effect to control for between-person variability in self-reports, with the overall conversation quality rating as the outcome variable. In the first model, we examined how the general empathy of the chat partner, conversation types (chatbots or humans), and their interaction influenced overall conversation quality. The second model assessed the impact of overall state empathy, conversation types (chatbots or humans), and their interactions on conversation quality. The third model explored the effects of conversation types (chatbots or humans), state empathy ratings (affective, cognitive, associative), and their interactions

[2]Summarization was rarely needed; the process is similar to LangChain's conversational summarization https://python.langchain.com/v0.1/docs/modules/memory/types/summary/.

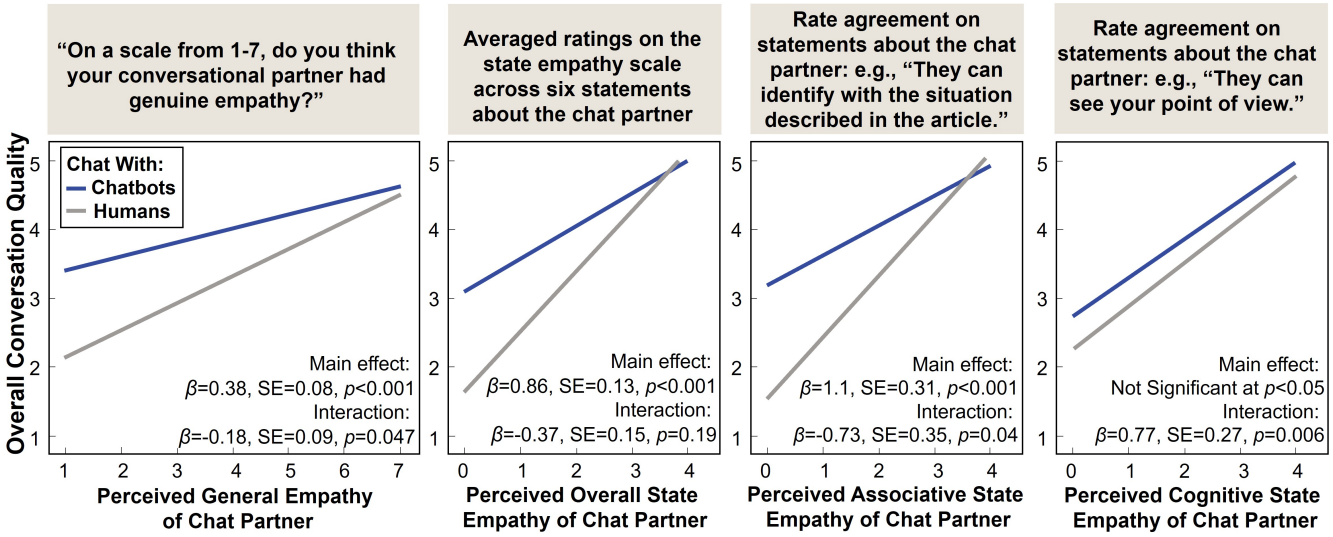

Figure 2: Interaction Between Chat Partner (Chatbot vs. Human) and Perceived Empathy on Overall Conversation Quality.

on conversation quality. The last model examined the impact of perceived closeness, conversation types (chatbots or humans), and their interaction on conversation quality.

Results Results from the *t*-tests revealed that, compared to their human counterparts, chatbots were rated significantly lower in general empathy, overall state empathy, affective state empathy, cognitive state empathy, and associative state empathy. There were no statistical differences in closeness between conversation types (chatbot vs. human). See details in Table 2. We also replicated these findings using the smaller with-subject comparison.

Across all four mixed models, we found that chatting with a chatbot led to a significantly higher conversation quality than chatting with a human (all $\beta > 1.05, p < 0.01$). In addition, we found perceiving the chat partner as having high general empathy ($\beta = 0.37, p < 0.001$), higher state empathy ($\beta = 0.86, p < 0.001$), higher associative state empathy ($\beta = 1.1, p < 0.001$), and higher closeness to the chat partner ($\beta = 0.31, p = 0.003$) significantly improved the conversation quality. We found that the type of conversation (chatbot vs. human) significantly interacted with perceived general empathy ($\beta = -0.18$), overall state empathy $\beta = -0.37$), cognitive state empathy $\beta = 0.77$), and associative state empathy $\beta = -0.73$) of the partner, influencing the overall conversation quality ($p < 0.05$). All others were not significant at $p < 0.05$. See Figure 2 for interactions.

LLM Judgement of Perceived Empathy

In this task, we examined the similarity between human ratings and those generated by the modern LLM, GPT-4o, by assessing the perceived empathy of humans and chatbots in the WASSA 2023 and WASSA 2024 datasets. We input the entire conversations between pairs of participants—either *human-human* or *human-chatbot*—into GPT-4o. The model was tasked with rating the perceived empathy of both participants at a conversation level.

Unlike the users, GPT-4o was not informed if the participant being rated was a human or a chatbot. The language model processed the entire conversation and assigned perceived empathy scores to both sides of the conversation. We conducted two analyses using this data. First, we investigated whether there was a statistically significant difference in the distributions of perceived empathy ratings between humans and chatbots. Second, we correlated these machine-generated labels with the human-rated labels in our dataset.

Results Our findings indicated that GPT-4o consistently perceived chatbots as less empathetic than humans in the context of the overall conversation. A *t*-test confirmed that this difference was statistically significant (p= 0.0005). Furthermore, correlation analyses with our gold standard human labels (of perceived empathy) revealed a correlation coefficient of r = 0.20 for human ratings, r = 0.06 for chatbot ratings, and r = 0.07 for the combined dataset. See Table 2 for results details. Although GPT-4o was not informed of the participants' identities, further analysis (see supplement[1]) revealed it correctly identified participants with 61% accuracy. While GPT-4o rated humans as significantly more empathetic than chatbots without explicit labels, the model may have simultaneously recognized language patterns indicative of an LLM.

Perceived Empathy Model

Here, we trained a model to predict the general empathy perceived by conversational language. We concatenated all turns from a single speaker into a single document using the EC dataset. We then extracted unigrams encoded as the relative frequency of use across a given conversation. We then removed unigrams that were not used by at least 5% of the speakers, resulting in a feature space of 1,500 unigrams. Using 10-fold cross-validation with an l_2 penalized Ridge regression (regularization term λ chosen as 10,000 using nested cross-validation), we obtained a prediction ac-

curacy of Pearson r = 0.17. This accuracy is comparable to the results of the WASSA 2024 shared task on predicting perceived empathy (Giorgi et al. 2024). The model was applied to conversations from WASSA 2023 and 2024 to generate estimates of perceived empathy, which were then compared to users' self-reported general empathy ratings. The entire process was conducted using the DLATK Python package. (Schwartz et al. 2017).

Results Table 2 shows that the mean estimated perceived empathy for humans does not differ from that of chatbots (t=0.45, p=0.65). Thus, the empathetic language of humans and the empathetic language of bots are equivalent. Further, the correlation between human ratings of empathy and their predicted empathy correlate at r=0.17, whereas the chatbot correlation is r=0.08. Thus, estimated empathy for humans matches their rating, whereas this is less so the case for bots.

Pre-Trained Empathy Model

We used four pre-trained empathy prediction models to estimate empathy from text. The first model, developed by Lahnala, Welch, and Flek (2022) for the WASSA 2022 shared task, predicted Batson empathy scores from essays using the EC dataset and employed pre-trained bottleneck adapters (Pfeiffer et al. 2020) to estimate empathy in conversations. The other three models, based on Sharma et al. (2020), were trained to predict empathy components in Reddit mental health conversations: emotional reactions (Emo-React), explorations (Explore), and interpretations (Interpret), with ratings of 0 (no empathy), 1 (low), and 2 (high). Turn-level predictions from these models were aggregated to examine their relationship with conversation-perceived empathy. Finally, empathy estimates from each model were compared to perceived general empathy ratings for both humans and chatbots.

Results As shown in Table 2, the Interpret and Emo-React model predictions differed significantly between humans and chatbots, whereas the Batson Empathy and Explore model predictions showed no significant differences. Overall, humans exhibited higher predicted empathy levels by the *Interpret* model than chatbots, with perceived empathy ratings positively correlating with interpretations for humans ($r = 0.16$) but negatively for chatbots ($r = -0.19$). In contrast, chatbots generally scored higher by the *Emo-React* model than humans, though the predictions showed stronger correlations with perceived empathy for humans ($r = 0.25$) than for chatbots ($r = 0.19$).

Discussion

Our study examined user perceptions of empathy and conversational quality in LLM-based chatbots versus human conversations. Chatbots were rated higher in conversational quality but perceived as less empathetic, a finding echoed by assessments using language-based models.

Lower Perceived Empathy in Chatbots vs. Humans

Despite advances in natural language processing, LLM-based chatbots designed to convey empathy were still perceived as less empathetic than humans by users and language models. This suggests that, although chatbots can generate coherent and contextually appropriate responses, users still perceive them as lacking the nuanced empathy that humans convey (Jain, Pareek, and Carlbring 2024). We believe this may stem from the chat partner's identity, as knowing if they are human or a chatbot shapes users' expectations (Yin, Jia, and Wakslak 2024).

Our study delved deeper into this by showing that chatbots were consistently rated lower than humans across various dimensions of empathy (Westman, Shadach, and Keinan 2013)—general empathy, overall state empathy, associative state empathy, cognitive state empathy, and affective state empathy. Notably, cognitive empathy, which involves understanding context, exhibited a smaller gap between humans and chatbots, suggesting that chatbots may be somewhat effective at demonstrating comprehension.

Past studies on chatbot empathy yielded mixed results, potentially due to the lack of direct, user-centered comparisons between chatbot and human conversations (Lee et al. 2024b). This gap may be related to chatbots' non-human identity, which users perceive as less genuine or emotionally resonant (Shi et al. 2020). Our findings suggest that language models, like GPT-4o, can identify language generated by other LLMs, potentially reinforcing perceptions of chatbot identity (Panickssery, Bowman, and Feng 2024).

While LLMs like GPT-4o could identify and replicate the empathy gap observed by human users, empathy models trained using human-human texts (e.g., most pre-trained empathy prediction models and our EC-language-trained models) struggled to distinguish empathy levels between chatbots and humans. This discrepancy likely stems from the limitations of these models, which were trained on human-human conversations and isolated language cues rather than the full conversational context. These findings underscore a disconnect between the empathetic language generated by language models and how it is perceived by users. This perception gap implies a divergence between expressed and received empathy, which models trained on human-human conversations fail to effectively address (Urakami et al. 2019).

We chose self-reports as our primary measure of empathy because they are widely considered the psychological "gold standard" for capturing subjective experiences, directly reflecting users' perceptions (Neumann and Chan 2015). Theoretical frameworks like mind perception and the Computers as Social Actors paradigm support the idea that empathy theories developed for human interactions can also be applied to human-chatbot interactions (Gray, Gray, and Wegner 2007; Nass, Steuer, and Tauber 1994). Grounded in these theoretical perspectives, we selected validated definitions and scales of empathy that assess perceived affective, cognitive, and associative state empathy, aligning with well-established constructs in empathy research (Preston and De Waal 2002) and tools designed for digital interactions like Perceived Empathy of Technology Scale (Schmidmaier, Rupp et al. 2024). Our approach focuses on capturing empathy as users perceive it without enforcing strict operational definitions. Likewise, we rely on subjective evalu-

Analysis	r overall	r human	r chatbot	mean human	mean chatbot	t	p-value
Psychological Ratings							
General Empathy	-	-	-	5.42	4.04	5.37	<0.001
Overall State Empathy	-	-	-	2.58	2.00	4.06	<0.001
Affective State Empathy	-	-	-	2.37	1.62	4.54	<0.001
Cognitive State Empathy	-	-	-	2.66	2.33	2.26	0.025
Associative State Empathy	-	-	-	2.70	2.04	4.28	<0.001
Closeness	-	-	-	4.23	4.28	-0.23	0.82
Pre-trained Models							
Batson Empathy	0.11	0.15	0.08	4.50	4.49	0.09	0.93
Interpret	0.21	0.16	-0.19	0.31	0.07	6.77	<0.001
Emo-React	-0.02	0.25	0.19	0.32	0.52	-5.20	<0.001
Explore	-0.11	-0.16	-0.17	0.65	0.59	0.93	0.35
Perceived Empathy Model	0.12	0.17	0.09	5.99	6.00	0.45	0.65
LLM-judge (GPT-4o)	0.07	0.20	0.06	4.98	4.10	3.54	<0.001

Table 2: Results of all four experiments. This table shows Psychological Ratings, Performance of off-the-shelf pre-trained empathy models, a perceived empathy model, and GPT-4o ratings. r: Pearson r between empathy predictions and users' perceived general empathy ratings for humans and chatbots. t: Welch two sample t-test statistic between predicted empathy distributions for humans vs. bots, with corresponding p-values. Mean human/chatbot are the mean conversation empathy scores. Cells marked with $-$: do not apply.

ations of conversation quality to examine perceived differences between human and chatbot interactions (Inan Nur, Santoso, and Putra 2021). This user-centered perspective enhances our understanding of empathy and quality as users experience them, offering valuable insights for optimizing human-chatbot interactions.

Effect of Empathy on Conversation Quality

Our findings reveal a positive correlation between higher perceived empathy and overall conversation quality for humans and chatbots, with the association being stronger for human interactions. For humans, low perceived empathy was closely tied to low conversation quality. In contrast, chatbot conversations were generally rated higher in quality, even at low to moderate levels of perceived empathy. This suggests that users may adjust their expectations for chatbots, leading to favorable ratings of conversation quality despite moderate levels of perceived empathy.

Significant interaction patterns emerged between perceived empathy dimensions (general, overall state, associative state, and cognitive state) and conversation quality, varying based on whether the conversational partner was a chatbot or a human. Affective state empathy, however, did not follow this trend. While chatbots were generally rated highly for conversational quality, they scored significantly lower than humans in affective empathy. This discrepancy may stem from users' implicit expectations of empathy in human interactions, which chatbots struggle to fulfill. Additionally, the "uncanny valley" effect (Mori, MacDorman, and Kageki 2012) could contribute, as users may perceive chatbots' attempts at emotional expression as artificial or unsettling, creating a disconnect between high conversational quality and low perceived empathy. Affective state empathy remains particularly challenging for chatbots, emphasizing their difficulty conveying genuine emotional resonance, even when their responses are contextually appropriate. Future studies could explore strategies to improve chatbots' ability to convey affective state empathy, focusing on enhancing emotional resonance and authenticity to address the empathy and quality gap observed in human-chatbot conversations.

Conclusion

Our study, grounded in user-centered research, examined perceptions of empathy and conversational quality in LLM-based chatbots compared to humans. Chatbots were rated higher in conversational quality but perceived as less empathetic, a finding supported by LLM annotations and a pre-trained empathy model. By focusing on user experiences, this research highlights the complications of empathy expressions and perceptions in human-chatbot conversations.

Limitations

One limitation is the absence of a participant group that was unaware they were interacting with a chatbot. Consequently, we cannot directly assess the impact of chatbot identity awareness on user perceptions during conversations. However, this approach reflects real-world conditions, as users are typically informed when engaging with a chatbot. Such awareness is crucial, as it influences trust and empathy—key components of effective communication. Additionally, while our participant pool is not fully representative of a global population, the use of crowdsourcing aligns with standard research practices and enables broad user insights. Finally, we intentionally avoided setting arbitrary thresholds for effect sizes, prioritizing user-centered insights over strict quantitative metrics to better capture nuanced perceptions of empathy and conversation quality.

Ethical Statement

Understanding how empathy is expressed and perceived in human-bot interactions raises important ethical questions. The paper's findings can inform the design and development of ethical dialogue systems, especially in enhancing the system's empathy (Curry and Curry 2023). Insights into user perceptions and language differences between human-bot and human-human interactions can improve these systems' ability to interpret input and generate natural, empathetic responses.

Acknowledgments

This study was supported by National Institute on Drug Abuse (NIDA), National Institutes of Health (NIH). The authors report no conflict of interest.

References

Alam, F.; Danieli, M.; and Riccardi, G. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50: 40–61.

Aron, A.; Aron, E. N.; and Smollan, D. 1992. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, 63(4): 596.

Atkinson, J. M.; and Heritage, J. 1984. *Structures of social action*. Cambridge University Press.

Ayers, J. W.; Poliak, A.; Dredze, M.; Leas, E. C.; Zhu, Z.; Kelley, J. B.; Faix, D. J.; Goodman, A. M.; Longhurst, C. A.; Hogarth, M.; 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.

Barriere, V.; Sedoc, J.; Tafreshi, S.; and Giorgi, S. 2023. Findings of WASSA 2023 Shared Task on Empathy, Emotion and Personality Detection in Conversation and Reactions to News Articles. In Barnes, J.; De Clercq, O.; and Klinger, R., eds., *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 511–525. Toronto, Canada: Association for Computational Linguistics.

Batson, C. D.; Fultz, J.; and Schoenrade, P. A. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1): 19–39.

Brave, S.; Nass, C.; and Hutchinson, K. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies*, 62(2): 161–178.

Casas, J.; Spring, T.; Daher, K.; Mugellini, E.; Khaled, O. A.; and Cudré-Mauroux, P. 2021. Enhancing conversational agents with empathic abilities. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 41–47.

Curry, A. C.; and Curry, A. C. 2023. Computer says "no": The case against empathetic conversational AI. In *Findings of the Association for Computational Linguistics: ACL 2023*, 8123–8130.

Decety, J.; and Jackson, P. L. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2): 71–100.

Gao, J.; Liu, Y.; Deng, H.; Wang, W.; Cao, Y.; Du, J.; and Xu, R. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, 807–819.

Giorgi, S.; Sedoc, J.; Barriere, V.; and Tafreshi, S. 2024. Findings of WASSA 2024 Shared Task on Empathy and Personality Detection in Interactions. In De Clercq, O.; Barriere, V.; Barnes, J.; Klinger, R.; Sedoc, J.; and Tafreshi, S., eds., *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 369–379. Bangkok, Thailand: Association for Computational Linguistics.

Go, E.; and Sundar, S. S. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in human behavior*, 97: 304–316.

Gray, H. M.; Gray, K.; and Wegner, D. M. 2007. Dimensions of mind perception. *Science*, 315(5812): 619–619.

Herlin, I.; and Visapää, L. 2016. Dimensions of empathy in relation to language. *Nordic Journal of linguistics*, 39(2): 135–157.

Hosseini, M.; and Caragea, C. 2021. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3713–3724.

Inan Nur, B.; Santoso, B.; and Putra, O. H. 2021. The method and metric of user experience evaluation: A systematic literature review. In *Proceedings of ICSCA 2021*.

Jain, G.; Pareek, S.; and Carlbring, P. 2024. Revealing the source: How awareness alters perceptions of AI and human-generated mental health responses. *Internet Interventions*, 36: 100745.

Jefferson, G. 1984. On stepwise transition from talk about a trouble to inappropriately next-positioned matters. *Structures of social action: Studies in conversation analysis*, 191: 222.

Koh, Y. J.; and Sundar, S. S. 2010. Heuristic versus systematic processing of specialist versus generalist sources in online media. *Human Communication Research*, 36(2): 103–124.

Lahnala, A.; Welch, C.; and Flek, L. 2022. CAISA at WASSA 2022: Adapter-tuning for empathy prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 280–285.

Lee, A.; Kummerfeld, J.; Ann, L.; and Mihalcea, R. 2024a. A Comparative Multidimensional Analysis of Empathetic Systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 179–189.

Lee, Y. K.; Suh, J.; Zhan, H.; Li, J. J.; and Ong, D. C. 2024b. Large language models produce responses perceived to be empathic. *arXiv preprint arXiv:2403.18148*.

Lindström, A.; and Sorjonen, M.-L. 2012. Affiliation in conversation. *The handbook of conversation analysis*, 250–369.

Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. MIME: MIMicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

Mori, M.; MacDorman, K. F.; and Kageki, N. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2): 98–100.

Nass, C.; Steuer, J.; and Tauber, E. R. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. ACM.

Neumann, R.; and Chan, E. 2015. Measures of empathy: Self-report, behavioral, and neuroscientific approaches. In *Measures of Personality and Social Psychological Constructs*, 257–289. Academic Press.

Nezlek, J. B.; Schutz, A.; Lopes, P.; and Smith, C. V. 2007. Naturally occurring variability in state empathy. *Empathy in mental illness*, 187–200.

Omitaomu, D.; Tafreshi, S.; Liu, T.; Buechel, S.; Callison-Burch, C.; Eichstaedt, J.; Ungar, L.; and Sedoc, J. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Panickssery, A.; Bowman, S. R.; and Feng, S. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Peräkylä, A. 2012. Conversation analysis in psychotherapy. *The handbook of conversation analysis*, 551–574.

Pfeiffer, J.; Vulić, I.; Gurevych, I.; and Ruder, S. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7654–7673. Online: Association for Computational Linguistics.

Preston, S. D.; and De Waal, F. B. M. 2002. Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1): 1–20.

Qian, Y.; Zhang, W.-N.; and Liu, T. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140*.

Raamkumar, A. S.; and Yang, Y. 2022. Empathetic conversational systems: a review of current advances, gaps, and opportunities. *IEEE Transactions on Affective Computing*, 14(4): 2722–2739.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Schmidmaier, R.; Rupp, L.; et al. 2024. Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. In *Proceedings of the CHI Conference*.

Schwartz, H. A.; Giorgi, S.; Sap, M.; Crutchley, P.; Ungar, L.; and Eichstaedt, J. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations*, 55–60.

Shafaei, R.; Bahmani, Z.; Bahrami, B.; and Vaziri-Pashkam, M. 2020. Effect of perceived interpersonal closeness on the joint Simon effect in adolescents and adults. *Scientific Reports*, 10(1): 18107.

Sharma, A.; Miner, A. S.; Atkins, D. C.; and Althoff, T. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Shen, L. 2010. On a scale of state empathy during message processing. *Western Journal of Communication*, 74(5): 504–524.

Shetty, V. A.; Durbin, S.; Weyrich, M. S.; Martínez, A. D.; Qian, J.; and Chin, D. L. 2024. A scoping review of empathy recognition in text using natural language processing. *Journal of the American Medical Informatics Association*, 31(3): 762–775.

Shi, W.; Wang, X.; Oh, Y. J.; Zhang, J.; Sahay, S.; and Yu, Z. 2020. Effects of persuasive dialogues: testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13.

Sorin, V.; Brin, D.; Barash, Y.; Konen, E.; Charney, A.; Nadkarni, G.; and Klang, E. 2023. Large language models (llms) and empathy-a systematic review. *medRxiv*, 2023–08.

Sundar, S. S.; Bellur, S.; Oh, J.; Jia, H.; and Kim, H.-S. 2016. Theoretical importance of contingency in human-computer interaction: Effects of message interactivity on user engagement. *Communication Research*, 43(5): 595–625.

Urakami, J.; Moore, B. A.; Sutthithatip, S.; and Park, S. 2019. Users' perception of empathic expressions by an advanced intelligent system. In *Proceedings of the 7th international conference on human-agent interaction*, 11–18.

Wardhana, A. K.; Ferdiana, R.; and Hidayah, I. 2021. Empathetic chatbot enhancement and development: A literature review. In *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, 1–6. IEEE.

Welivita, A.; and Pu, P. 2024a. Are Large Language Models More Empathetic than Humans? *arXiv preprint arXiv:2406.05063*.

Welivita, A.; and Pu, P. 2024b. Is ChatGPT More Empathetic than Humans? *arXiv preprint arXiv:2403.05572*.

Westman, M.; Shadach, E.; and Keinan, G. 2013. The crossover of positive and negative emotions: The role of state empathy. *International Journal of Stress Management*, 20(2): 116.

Xu, Z.; and Jiang, J. 2024. Multi-dimensional Evaluation of Empathetic Dialog Responses. *arXiv preprint arXiv:2402.11409*.

Yin, Y.; Jia, N.; and Wakslak, C. J. 2024. AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences*, 121(14): e2319112121.

Zhou, K.; Aiello, L. M.; Scepanovic, S.; Quercia, D.; and Konrath, S. 2021. The language of situational empathy. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–19.