Supplementary Material:
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

In this Supplementary Material, we provide the additional experiments involving third-party annotations of empathy for each conversational turn (S1), identity detection using GPT-4o (S2), prompt that we used GPT-3.5-turbo to further flag and remove insincere conversations (S3), and sample conversations from the datasets utilized in our study (S4).

S1: Third-party Empathy Annotations

The WASSA 2023 and 2024 data sets contained third-party annotations of empathy for each conversational turn (Barriere et al., 2023; Giorgi et al., 2024). Ratings were collected via Amazon Mechanical Turk. Workers were asked to rate the empathy of each conversational turn (on a scale from 1-5) and were given the context of the previous turns in the conversation. Workers were not told if a conversation contained a chatbot. Each turn was rated by three workers.

We compared aggregate turn-level empathy (the average turn-level empathy across all turns in a conversation, with one average per person in the conversation) to the perceived empathy ratings. This analysis is similar to the one presented in Table S1 and we present the results in Table 2. Results show that humans and chatbots do not differ on average third-party empathy ($t = 1.82$; $p = 0.07$). Additionally, average turn-level empathy significantly correlates with perceived empathy in humans ($r = 0.43$) but not chatbots ($r=-0.01$). Thus, these results match the language-based estimates of empathy (Off-the-shelf Models and Perceived Empathy Model in Table 2).

S2: Identity Detection Experiment

We further conducted an experiment to determine whether GPT-4o could label the identity of participants in a conversation, specifically distinguishing between humans and chatbots. The motivation behind this experiment was to explore the capabilities of advanced language models like GPT-4o in recognizing subtle differences in conversational behavior that might indicate whether a participant is a chatbot or a human, as a supplementary exploration. We aim to investigate whether large language models can implicitly detect chatbot-generated language, as this detection may influence the empathy levels they assign during annotation tasks stated in our full paper. We employed GPT-4o to identify whether participants in a conversation were human or a chatbot. The model was provided with the entire conversation, along with specific instructions to return a label for each speaker as either "human" or "bot."

In this experiment, each conversation (whether human-human or human-chatbot) is passed to GPT-4o. See results in Table S2.

S3: GPT-3.5-turbo Prompt

We used GPT-3.5-turbo to filter further flag and remove insincere conversations, below is the prompt we used to instruct the system:

> The following conversation is between two crowdworkers or a crowdworker and a bot. They are talking about a news article with is supposed to evoke empathy. Each interlocutor is required to speak for at least 15 conversational turns.
>
> Please check if the article is done in good faith. Examples of issues are as follows: - saying "next" - responding in a way that is inappropriate in the context e.g., "hi", "During the onset of the pandemic in 2020, despite business closures, women launched more businesses than they closed, while the number of men-owned businesses declined. Women-owned businesses also grew their workforces and increased their revenue while men's numbers shrank." - not responding with words e.g., "2fanldudg"
>
> Don't be overly harsh, the conversation doesn't have to be absolutely perfect. It is fine for the conversation not to be about the news article. The crowdworkers can also try to say "bye" to end the conversation before it is complete.
>
> Here's the conversation
> [CONVERSATION]
> Please rate it as "OK" or "ISSUE". Only respond with those words.

S4: Conversation Examples

Below we provide examples of human/human conversations (Table S3) and human/chatbot conversations (Table S4). A total of six turns are presented for each conversation and examples from all three data sets are included.

Analysis	r overall	r human	r chatbot	Mean Human	Mean Chatbot	t (p-value)
Third-party annotations	0.09	0.43	-0.01	2.36	2.74	1.82 (0.07)

Table S1: Results comparing third-party (Amazon Mechanical Turk workers) annotations of empathy (averaged across turns) to conversational partner ratings of empathy. Pearson r between empathy predictions and general empathy annotations, and for human and chatbot agents. t: Welch two sample t-test statistic between predicted empathy distributions for humans vs. chatbots, with corresponding p-values. Mean human/chatbot: mean conversation empathy scores.

Data	Human Occ. (N)	Chatbot Occ. (N)	Pred. Human (N)	Pred. Chatbot (N)	Accuracy	F1 Score
WASSA 2023	64	19	70	13	0.81	0.65
WASSA 2024	40	77	99	18	0.50	0.56
Combined	104	96	169	31	0.61	0.60

Table S2: Human and chatbots in our data for WASSA 2023 and WASSA 2024 and how GPT-4o identified them. Occ. = occurence. Pred. = Predicted.

Source	Speaker	Dialogue
EC	Person 1	What did you think about this article?
	Person 2	It's definitely really sad to read, considering everything they're all going through. What did you think?
	Person 1	I think it's super sad... they seem to never catch a break, always struggling.
	Person 2	I can't imagine just living in an area that is constantly being ravaged by hurricanes or earthquakes. I take my location for granted.
	Person 1	Me too... I also can't imagine living in the poverty and such... It's crazy to think that people still live like that sometimes. The gap between first-world countries and places like that is crazy to me.
	Person 2	It also seems unnecessary for there to even be such a gap. With all of the wealthy countries out there, I hope Haiti gets the help it deserves, because we, and other countries, can certainly afford it.
WASSA 2024	Person 1	Yeah, I'm sorry but celebrity life doesn't interest me that much. I don't know what to think or feel about this article.
	Person 2	I'm pretty much the same. I like some of the movies and shows those people are in but I don't really care about their personal lives.
	Person 1	Right? And I think that some of their problems are hyped up as well to get more people to read about them.
	Person 2	Yep, I couldn't care less!
	Person 1	Same. I think that with all of the actual problems out there, it can be difficult to expend the energy to something as pointless as a celebrity's life.
	Person 2	Yep. I do remember back in the day when they were wearing vials of each other's blood. That was so ridiculous.

Table S3: Example of human/human conversations.

Source	Speaker	Dialogue
WASSA 2023	Human	My first reaction was that it was sad, but I also thought it was senseless and confusing.
	Chatbot	Yes, it was definitely a tragic and confusing incident. It is hard to make sense of why it happened.
	Human	Without more information it's hard to feel too much about it. It's unfortunate, but I can't be angry about it because I don't know why it occurred.
	Chatbot	Yes, it's important to wait for more information before drawing conclusions. It's a tragedy regardless, but understanding the circumstances can bring some clarity to the situation.
	Human	Indeed. If there were bad actors here, I'd hope they'd be brought to justice. There was no indication of that, but it's possible. If it was just an accident, well, hopefully it doesn't happen again.
	Chatbot	Absolutely, justice and accountability are important if there was any wrongdoing. And if it was truly an accident, then hopefully measures can be taken to prevent it from happening again in the future.
WASSA 2024	Human	Hey, did you read the article?
	Chatbot	Yes, I did. What did you think about Billy Bob Thornton's feelings of not feeling good enough for Angelina Jolie?
	Human	I think it makes sense given such a large age gap.
	Chatbot	That's a fair point. The age difference and their different lifestyles could definitely add to those feelings.
	Human	What did you think?
	Chatbot	I found it honest and a bit sad that Thornton felt that way. It's quite human to feel insecure, especially in relationships where there's a lot of public scrutiny.

Table S4: Example of human/chatbot conversations.