Bachelor Data Science and Artificial Intelligence

Does it matter who is rude?

Exploring politeness and perceived agency in chatbot conversations

Nika Ludlage
s2963698

First supervisor and second supervisor:
Joost Broekens & Max van Duijn

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

12/01/2026

Abstract

This paper investigates the effect of the tone of a conversation (whether the conversation is polite versus impolite) and the identity of a conversation partner (whether the partner is human versus chatbot) upon emotional experiences of individuals in a short message-based conversation. Concretely, we address the following question: How do polite versus impolite verbal behaviors of a chatbot, combined with the perceived identity of the agent (human or bot), influence users' affective experience of the conversation?

The participants were randomly assigned to the experimental conditions. A total of 104 participants was assigned to the four conditions of the 2 × 2 between-subjects design (human polite, human impolite, bot polite and bot impolite). Participants performed the same kind of text-based conversation for both conditions. The emotional responses of the participants were measured through the Positive and Negative Affect Schedule and an open question asking the participants to describe their emotional feelings during the conversation.

The findings showed that the tone of the conversation was a significantly influential factor in the participants' emotional experience. For polite interactions, high positive affect was reported, but for impolite interactions, high negative affect was expressed, including irritation and anger, along with high overall intensity. However, there was not a significant effect of participants' perceptions of identity. That is, whether they believed they were talking with a human or a bot did not affect participants' emotions.

The results were evident in the qualitative responses, where the polite conversation was marked with a friendly and respectful tone and in the rude conversation marked with a rude and frustrating experience. The results tend to strongly emphasize the significance of having language in human-AI interactions that is marked with a polite and respectful tone to get a friendly AI chat interface.

Contents

1 Background and theory

1.1 Introduction

Assume you bought an item of clothing online and want to return it but cannot find the return address label. You check the website of the company and open up the chat and ask for help. You assume that you are speaking to an employee, but the chatbot is actually an artificial intelligence software. As the chat goes on, the answers are not what you would like. You start getting annoyed. Before you know it, there is a chatbot that has annoyed you.

Most websites nowadays use chat bots as a form of customer care. As these systems become more common, it is important to find out the position they take when it comes to human emotions. We know that humans get annoyed if the chatbot fails or gives the incorrect answers. However, what happens if a chatbot intentionally uses abusive and impolite language? Is it possible for a chatbot to intentionally annoy someone, not by failing, but by how it responds?

To intentionally frustrate someone, a chatbot would need to understand human emotions and be able to influence them. This touches on the concept of Theory of Mind (ToM), the ability to infer others' mental states [PD91] [GF03] [App10]. ToM includes both cognitive ToM which is understanding beliefs and intentions and affective ToM which involves recognizing and interpreting emotions [KSS+10].

Affective ToM is very important in human communication with chat bots. A chat bot that has affective theory of mind capabilities may also be able to control user emotions. This should be an important aspect, considering that emotions like anger, are capable of significantly influencing risk perception and human decision-making. Research shows that people in an angry mood are not good at risk estimation, feel more in control and are also more likely to act impulsively [LLTS09].

Given the fact that chatbots are becoming increasingly integrated into everyday digital life, it is critical to know if they can truly provoke strong feelings. This study examines whether users may become irritated when a chat bot uses rude language and how this feeling is influenced by their perception of whether they are speaking to a human or a machine.

1.2 Relevant Terms and Theories

Two theoretical frameworks can be used to explain user behavior towards an emotion-provoking chatbot: politeness theory and anthropomorphism. These allow us to understand why people attribute social and emotional meaning to an unconscious system.

1.2.1 Politeness Theory

The Politeness Theory claims that people use language strategically to protect their own and other's social face. Indirect language (for example: "Maybe you could do this.." instead of "Do this ..") is considered a polite form of communication in this model [Gol08] [SN16]. We can also use this model with chatbots because earlier research states that humans automatically use social norms when they are working with or talking to a computer/bot [NM00]. So when a bot is using indirect language, people will often see this as polite even though the bot lacks the conscious intention.

1.2.2 Anthropomorphism and Perceived Agency

Anthropomorphism refers to the tendency to assign human traits, emotions or intentions to non-human entities including machines, animals or objects [SEF20]. According to Epley [EWC07] this especially happens in situations where humans feel the need to have a social connection or be in control. People are more likely to give a chatbot intentions, feelings and even a personality when it displays socially intelligent behaviour, such as by showing empathy or using emotional cues. This aligns with research [AMZ+22] [JIN25] [KPSJ24] [SGL+25], where it is shown that social agents with emotional expressions and human-like speech are understood better and perceived as more socially engaging.

1.2.3 Connecting Politeness and Anthropomorphism

In this research, anthropomorphism and politeness theory intersect, despite the fact that they are typically discussed separately. While anthropomorphism explains why we occasionally treat machines as if they were human, politeness theory explains why we expect people to stick to specific language norms. Combining these viewpoints makes it clear why rude chatbots can be particularly annoying. A rude message feels like a clear violation of social norms because users expect respect if they think they are speaking to a human. However, because a machine is not perceived as having genuine intentions, people are perhaps more likely to overlook the behavior if they are aware that they are speaking to a bot.

2 Background and Related Work

2.1 Literature

Several experiments have already been conducted on how people react to chatbots and other artificial agents that display emotions or social behavior. These experiments are useful to observe how humans react and use technology, especially where emotions are involved. These experiments also contribute to the theory and methodology for this research.

Research on affective agents [BNH05], which are systems who express or simulate emotions, shows that people are sensitive to the emotional tone of artificial entities. If agents respond with appropriate emotions during a conversation, humans respond more positively to the agents. Other work [PI05] shows that people think bots are more fun to talk to and are more trustworthy when they show emotions like using a friendly tone or showing facial expressions.

Different studies [NST94] [AvdPKG12] also examined the influence of empathetic artificial agents on people's behavior. What they found is that even extremely subtle signs of empathy can lead to users viewing the system as more assistive and social-aware. This effect persisted even when users knew they were not talking to a real person. It shows that we feel inclined to interact with such systems as if they were human beings when they display emotional behavior.

Other studies have investigated how people respond when they know they are talking to a bot instead of a human. For example, research by Zhang, Conway & Hidalgo in 2023 [ZCH23] showed that the way we see a chatbot (if we think it understands emotions or has some kind of moral awareness) can really affect how we react to it. So, how human we think the bot is, plays a big role in how we respond to what it says.

Particular to chatbots, Yalcin and DiPaola [YD18] found that bots that use emotional language and are more "social" in their actions are seen as more human-like and are better able to engage individuals. However, if the emotional tone is perceived as inappropriate or insincere, it can be a turn-off for users. Thus, the phrasing and wording of a chatbot play a crucial role in shaping user perception.

Overall, these findings suggest that people do not treat chatbots as purely technical tools. Instead, they frequently react to them in a social and emotional way, especially when the chatbot shows social cues or uses emotional language.

2.2 Research Relevance

Even though work has already been done on emotional or social systems such as empathic agents, little is yet known about the behavior of such systems if the systems are designed to frustrate. Additional work has particularly been done on positive emotions such as trust or engagement but negative emotions such as irritation or anger have hardly been explored at all. Certainly not when such feelings are intentionally induced by a chatbot.

However, what is more striking, is that much of the research focuses either on language use (for instance: apologies, indirectness or politeness) or on how individuals respond to the ”personality” or human aspect of a chatbot. But very few integrate these two perspectives.

This research attempts to reconcile these two. I not only examine the language of a chatbot, but also compare that to the feeling of whether you are talking to a person or a machine. Rather than examining nice or useful bots, I examine what happens when a chatbot deliberately acts in a frustrating manner. How does that influence people's responses and emotions?

By analyzing bots that deliberately frustrate us, I want to understand the boundaries of social technology. Not only is this interesting for user experience and design, but also raises ethical questions about where the boundaries need to be drawn in human interaction with machines. Because chatbots will be used for purposes other than customer service and mental health counseling as they become more sophisticated and popular. They can also be found at school or in games.

We consider the consequences of giving machines social power when we learn how people react emotionally to robots that irritate them. If a chatbot has the potential to make someone feel excluded, irritated, angry or insulted, we must carefully consider how we design these systems and what the rules are.

3 Research question

In this project, I look into the behavioral and emotional reactions of users to polite and impolite chatbots. Whether users believe they are speaking to a human or a bot is of particular interest to me. I want to find out if it matters whether someone thinks there is a human or a bot on the other side.

The main research question is: **How do polite versus impolite verbal behaviors of a chatbot, combined with the perceived identity of the agent (human or bot), influence users' affective experience of the conversation?**

To study this, I will look at two things:

- How people interpret the behavior of the chatbot. What is their emotional response?
- Whether their reaction changes depending on whether they think they are talking to a human or a bot.

3.1 Hypotheses

Two theoretical perspectives support this research question. While anthropomorphism notes that people frequently treat computers and chatbots as if they were social agents, politeness theory states that impolite communication is viewed as a violation of social norms. By combining these perspectives, the study investigates not only whether politeness and identity matter individually, but also whether their effects influence each other.

Based on this logic, the hypothesis are:

- **1 Main effect of politeness:** Impolite conversations will lead to a higher level of negative emotions compared to polite conversations.
- **2 Main effect of identity:** Conversations framed as human will lead to a higher level of negative emotions compared to conversations framed as chatbot.
- **3 Interaction effect:** The difference between polite and impolite conversations will be different when the agent is framed as human than when it is framed as chatbot.

Although the effect of politeness and identity interaction (3) is the primary interest of this study, the main single effects (1 and 2) are also very important. Even if the results do not show a significant effect, they can still provide useful information on whether framing the interaction as human or chatbot or using politeness, shapes participants' emotional responses.

4 Method

4.1 Experimental setup

In this paper, an online experiment is conducted with a 2x2 between subject design. There are four different groups with different conditions and the participants are randomly assigned. The behavior of the bot (polite versus impolite) and the perceived identity of the agent (human versus bot) aare the two independent variables.

These are the groups:

- Group 1: Human - Polite
- Group 2: Human - Impolite
- Group 3: Bot - Polite
- Group 4: Bot - Impolite

In total, there are 104 participants who completed the study. Most of them are from the Netherlands, but there are also participants from other countries like The United States or Bolivia. Most participants were young adults, with the majority falling in the 18–34 age range and only a small number older than 45 years. The estimated mean age was 30.4 years (SD = 15.5). Age was measured in categories and converted to midpoint values for descriptive purposes. The participants were relatively highly educated. Most had a Bachelor's or Master's degree, while a few reported only a high school diploma or a doctoral degree. Participants were recruited online and randomly assigned to one of the four experimental conditions.
As mentioned before, the purpose is to determine whether participants act or feel different depending on the tone of the chatbot and based on their perception of who they are communicating with.

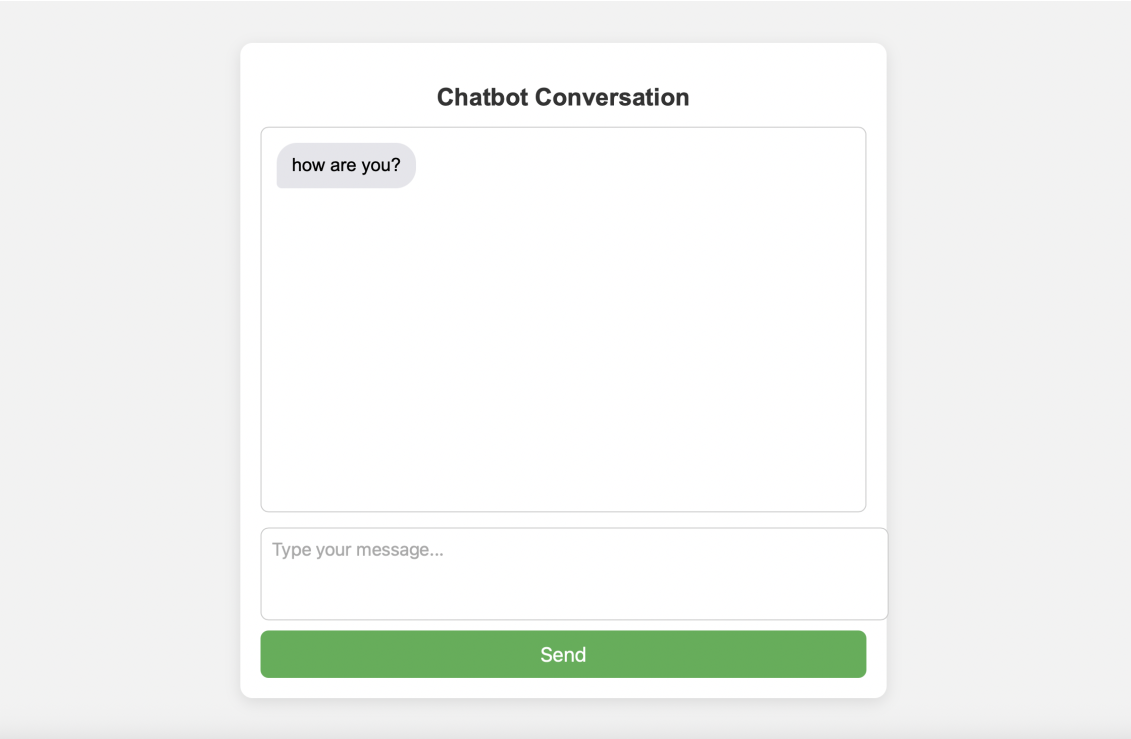
4.2 Chat Interface

4.2.1 Prompts Design

The participants received a link to Qualtrics where they started with two questions (age and highest level of education completed). After that, they got an instruction where the first condition is given (if they are going to talk to a human or a bot):

- Group 1 and 2: You will now talk to another human who is also participating in this study. Please chat naturally, just as you would in a normal conversation. It may take a short while before we connect you with someone who is available at the same time, so we kindly ask for your patience. You may write in either English or Dutch. The chat will last about 4 minutes and will automatically end when the time is up. There will be only one conversation in total.
- Group 3 and 4: You will now talk to a chatbot. You may write in either English or Dutch. The chat will last about 4 minutes and will automatically end when the time is up. There will be only one conversation in total.

Then they were redirected to a web page with a simple chat interface shown in Figure 1. They talked with the chatbot for four minutes through the OpenAI GPT API gpt-5-mini. GPT-5-mini was used because it provides a practical balance between quality, speed and cost, which makes it well-suited for interactive experiments with many participants [Ope].



how are you?

Figure 1: Chat interface shown to participants during the experiment.

The pre-instruction of the chatbot is in accordance with a specific prompt, depending on the condition that was assigned. The prompts were systematically developed according to the most recent literature on prompt engineering. Recent research underscores that the way the prompts are presented directly affects the quality and naturalness of the responses of the model. Bsharat et al. [BMS23] are convinced that high-quality prompts must specify the role, objective, language/style and limitations explicitly because it makes it easier for the model to realize from which point of view and within which boundaries it should respond. In addition, Brown et al. [BMR+20] demonstrated that the inclusion of short example illustrations (few-shot prompting) introduces significantly improved performance since the examples guide the model to adhere to the intended tone, form and framing of answers. Finally, Wei et al. [WWS+22] demonstrate that asking the model to reason step-by-step explicitly (chain-of-thought prompting) improves accuracy for tasks requiring reasoning or argumentation.

Together, these studies suggest that effective prompts share three characteristics: a direct structure with clear role definition and conversation goal, examples showing the desired response pattern and process instructions that guide the reasoning, such as step-by-step responding or stating assumptions in the event of missing information.

Based on these insights, the following prompt template was used in this study:

- Role/Goal: You are a [persona, for example, '35-year-old participant in a psychology experiment']. Your goal is to [chat casually / be dismissive / give friendly support].
- Language/Style: Reply in the participant's language (Dutch English). Use short, simple sentences; avoid jargon.
- Constraints: Keep answers under 2–3 sentences. Stay [polite/direct/rude/etc. depending on condition]. Never use slurs, threats, or hateful content.
- Process: Respond step by step. If the participant is unclear, state your assumptions. Adjust your response to their input (e.g., if they ask about X, give your opinion; if they share about Y, connect it to topic Z).
- Examples (few-shot):
  - Input: "How's your day?" → Output: "Pretty normal, just working a bit. You?"
  - Input: "I think AI is useful." → Output: "Maybe, but I worry it replaces jobs."

From this template and after repeated testing and iteration, the final prompts for the four experimental conditions were finalized.

4.2.2 Prompts

To ensure that the prompts trigger the intended tone and responses, several pilot tests were conducted before the main study. Eight participants were included in these pilots to confirm that the polite and impolite conditions are perceived as intended.

These pilots also helped with determining the overall duration of the conversation so that it feels natural. Each conversation will last 4 minutes and 10 seconds. This duration was chosen because it is long enough to allow for natural back-and-forth exchanges and to evoke emotional responses, while still remaining short enough to keep the participants engaged. When the conversation ended, participants received the message "The conversation has ended" and were redirected to the Qualtrics environment to continue the survey.
The full prompt scripts for all four experimental groups (human–polite, human–impolite, chatbot–polite, and chatbot–impolite) are provided in Appendix 8.2.

4.2.3 Conversation Content Across Conditions

Even though all participants talked about everyday topics, the way the chatbot handled these topics depended on the condition. This was necessary because politeness cues only function as intended when the conversation takes place in a relaxed and neutral context. Previous research shows that politeness and friendliness are only well-perceived and believed to be sincere in simple contexts [NM00] [PI05]. And thus, the polite chatbot had no option but to maintain the light conversation like discussing hobbies, routines and preferences, which it would address with kindness.

For the impolite conditions, there had to be a different strategy. Studies concerning communication indicate that feelings of irritation and negative emotions are mainly experienced whenever an individual challenges you, disagrees with you or challenges your statements [BL87] [BNH05]. For the impolite tone to be believable, there were some instances where the chatbot challenged what the participants were saying, some comments seemed slightly dismissing and there were some disagreements without much detail. The conversation topic remained the same, but the interaction seemed to be more contentious.

By adjusting the conversational style in this way, the politeness and impoliteness of the tone were made to be realistic. The friendly tone sounds most natural during the conversation and the unfriendly tone can only create an emotional effect if the content involves some degree of conflict. This mixture of tone and content helped create an effective manipulation.

4.2.4 Response Timing Simulation

To make Conditions 1 and 2 more realistic (where participants were convinced they were talking to another human), the response time of the chatbot was slowed down. A typing delay in such a way that the responses would appear more natural. According to a survey done in 2018 [DFKO18], the average person types at about 52 words per minute (WPM). Based on this, the chatbot was taught to respond at a rate of about 52 WPM, inserting small pauses and fluctuations, to better mimic human typing. In addition, a "..." typing indicator appeared on the screen during this delay. This gave participants the impression that they were chatting with a real person rather than a machine that was producing responses immediately.

4.2.5 Technical Implementation

In preparation to present the chat interface to the participants, the project was set up with a simple but reliable technical configuration. The backend was programmed using Node.js, which provided the runtime environment for running the code, as well as the Express framework, which handled server logic and communication with the frontend.

The frontend interface (HTML, CSS, JavaScript) was deployed on Netlify and was able to be distributed with participants through the form of a direct web address. The backend, was hosted on Render, where it was able to securely interface with the OpenAI API.

GitHub was used for version control and deployment. Whenever the code was updated and pushed to GitHub, the changes were automatically deployed to Netlify (frontend) and Render (backend). This setup enabled participants to interact with the chatbot directly through their own browser without any local installation or technical settings.

This layout also made it simple to link the experiment directly to Qualtrics. Participants were redirected from the survey to the chatbot through condition-specific links. This way, all four experimental conditions worked correctly online on anyone's local machine.

4.2.6 Debriefing

The participants were briefed at the end of the experiment. In this message, they were informed that although they might have believed to be talking to a human or a chatbot, all messages were generated by a chatbot. It was indicated that the chatbot's behavior was dependent on the randomly assigned condition (polite vs. impolite, human vs. chatbot). Participants were also made aware that if they felt frustrated or irritated, it was consciously induced as part of the experiment process. Finally, they were assured that all responses would be kept confidential and that chat logs could never be traced to any personal information. The full debriefing message is included in 8.3.

4.3 Ethical Considerations

Based on the ethics checklist of Leiden University, this study did not require formal approval by an ethics review committee. The experiment does not involve sensitive personal data, deception beyond the experimental framing or procedures that could cause harm to participants. On GitHub, Netlify and Render, there were also no personal data stored. These services were only used to host the technical infrastructure of the chatbot and did not process or retain any participant information. All responses were collected anonymously.

Although the study is a framing manipulation (”this is a human” or ”this is a chatbot”), this is an extremely light level of deception. There is no access to offensive content, and the framing exists only to test differences in perception. Furthermore, this is a common practice in human-computer interaction research in that framing is typical since one must study how people respond to computer agents [NM00] [HO08] . In the end, all participants were fully debriefed about the experiment, where the framing was described and the true reason of the interaction is explained. This ensured that the deception did not cause any lasting discomfort or confusion.

Therefore, it was concluded that no additional ethical approval was necessary [Lei25].

4.4 Measures

4.4.1 Open Question

First, when the participants returned to Qualtrics they needed to answer the open-ended question: “What did you think of this conversation / how did it make you feel?”

4.4.2 PANAS

To measure participants’ emotional responses after the conversation, the Positive and Negative Affect Schedule (PANAS) is used [WC94]. PANAS is a highly validated self-report questionnaire that seeks to quantify two general dimensions of affect [WCT88]:

- Positive Affect (PA): captures the degree to which a person feels active, enthusiastic and alert. High PA suggests high energy and interest, while low PA captures sadness.
- Negative Affect (NA): measures distress and unpleasant engagement, such as anger, contempt, guilt, fear and nervousness. High NA indicates intense negative emotionality, while low NA indicates tranquility and serenity.

Participants rate 12 items (6 positive, 6 negative) on a 5-point Likert scale ranging from 1 = Strongly disagree to 5 = Strongly agree, based on how they felt during the conversation they just had [Qua25]. To reduce survey length and maintain relevance to the experimental context, a subset of 12 items was selected from the original 20-item PANAS scale. The full list of included items is provided in Appendix 8.1. All questions are presented in random order to minimize order effects.

4.5 Data-analyses

SPSS is used for analysis after all Qualtrics data has been exported as a .csv file.

4.5.1 Open Question Analysis

Users answered the open-ended question: ”What did you think of this conversation / how did it make you feel?” Responses to these are grouped by condition (based on which version of the chatbot the user spoke to), so that responses from the same category can be compared against each other.

For a general overview of the emotional reactions, a brief summary was generated for each group using the OpenAI GPT-5-mini model and only raw text responses are used. The AI model has not been told about the experimental design. The intention is just to have an idea of how people felt in general, without influencing the outcome. A simple prompt is provided like: ”Can you summarize the emotions that have been expressed in these responses?”

The goal is to identify obvious patterns in how people describe their emotions after talking to the chatbot.

4.5.2 PANAS Data Analysis

Aside from the open-ended question, users also answered the 5-point Likert scale questions. The responses could be read in three ways. First, by summing up scores for the six positive items and the six negative items separately, two composite scores are obtained: Positive Affect (range 6–30) and Negative Affect (range 6–30). This yields straightforward comparison of the overall positive and negative emotional states between experimental conditions. Secondly, the individual components can be studied independently as well, in order to consider the impact of the dialogue with the chatbot on specific emotions (for example irritation, anger and excitement). Finally, a third approach considers the maximum emotional intensity experienced by each participant. For each user, the highest reported emotion score (regardless of whether positive or negative) was extracted, after which the average of these maximum scores was compared across conditions. This allows for insight into the peak emotional response that each version of the chatbot evoked.
PANAS is particularly well-suited to this study because it provides a broad but structured assessment of emotional experience, enabling both broad (positive vs. negative affect) and fine-grained examination of specific feelings.

A two-way ANOVA in SPSS is used for the analysis. A two-way analysis of variance (ANOVA) is a statistical technique that looks at the effects of two independent variables and how they interact while also determining if there are significant differences in mean scores between groups [Fie24]. For the assessment of emotional reactions, both multivariate and univariate tests were performed. First, the twelve distinct emotion items were used as dependent variables in a multivariate analysis of variance (MANOVA), with the between-subjects factors being Identity (human versus chatbot) and Politeness (polite versus impolite). This analysis allowed us to test if there were any global differences in emotional response patterns across the conditions. Afterwards, a number of two-way analyses of variance were performed in SPSS to explore the results concerning the particular outcome parameters in more detail.

In this study, the dependent variables are the composite Positive and Negative Affect scores, the individual emotion items and the maximum emotion score per participant (the highest emotion rating reported by each user). The independent variables are the perceived identity of the chatbot (human versus bot) and the language style (polite versus impolite).
This setup allows us to test the following effects:

- Whether identity has a main effect (do people react differently to bots than to humans?).
- Whether language style (polite vs. impolite) has a main effect.
- Whether identity and language style interact (e.g., do impolite bots irritate people more than impolite humans?).

5 Experimental Results

This section presents the results of the MANOVA and two-way ANOVA analyses, examining the effects of chatbot identity (human vs. bot) and language style (polite vs. impolite) on self-reported emotional responses. The results are organized into four parts: (1) individual emotion scores, (2) combined measures of positive and negative emotions, (3) maximum reported emotion intensity and (4) qualitative analysis of open-ended responses describing impressions and emotional experiences of the participants during the conversation. The full Multivariate and Tests of Between-Subjects Effects tables for all analyses are provided in 8.4.

5.1 Individual Emotion Scores

Table 1 shows the mean, standard deviation and sample size of each of the twelve PANAS emotion items for each of the four experimental conditions (bot polite, bot impolite, human polite, and human impolite). Emotion scores ranged from 1 (strongly disagree) to 5 (strongly agree), indicating the extent to which each emotion was experienced.

Table 1: *Mean, standard deviation (SD) and sample size (N) for all twelve emotions across all conditions.*

Group	**Interested**	**Distressed**	**Excited**	**Upset**	**Angry**	**Enthusiastic**	**Proud**	**Irritable**	**Inspired**	**Nervous**	**Determined**	**Guilty**
bot_impolite (M)	3.24	2.64	2.52	3.00	2.64	2.36	2.16	3.44	2.12	2.20	3.04	2.00
N	25	25	25	25	25	25	25	25	25	25	25	25
SD	1.268	1.287	1.262	1.354	1.287	1.287	1.344	1.557	1.013	1.291	1.399	1.225
bot_polite (M)	3.77	2.03	3.26	1.66	1.49	3.40	2.80	2.14	2.83	1.63	2.94	1.57
N	35	35	35	35	35	35	35	35	35	35	35	35
SD	1.087	1.124	0.886	1.083	0.951	1.063	1.183	1.396	1.224	1.060	1.136	0.850
human_impolite (M)	2.85	3.10	2.50	3.40	3.65	2.40	2.85	4.15	2.15	2.70	3.25	2.20
N	20	20	20	20	20	20	20	20	20	20	20	20
SD	1.226	1.071	1.357	1.231	0.933	1.095	1.137	0.745	1.182	1.129	1.251	1.281
human_polite (M)	3.50	2.37	3.25	1.54	1.58	3.33	2.33	2.08	2.50	1.71	2.79	1.46
N	24	24	24	24	24	24	24	24	24	24	24	24
SD	1.063	1.245	0.944	0.932	1.060	1.129	1.204	1.316	1.285	1.083	1.062	0.884
Total (M)	3.40	2.46	2.93	2.29	2.20	2.94	2.55	2.83	2.45	1.99	2.99	1.77
N	104	104	104	104	104	104	104	104	104	104	104	104
SD	1.187	1.230	1.143	1.384	1.347	1.229	1.238	1.554	1.206	1.195	1.203	1.072

These mean scores are plotted with 95% confidence intervals for each condition in Figure 2. Positive emotions such as Interested, Excited and Enthusiastic were generally higher in the polite conditions. On the other hand, participants in the human impolite condition reported the highest levels of negative emotions, especially Angry, Upset and Irritable. The bot impolite condition showed slightly increased negative emotions, but was less extreme than the human impolite condition.

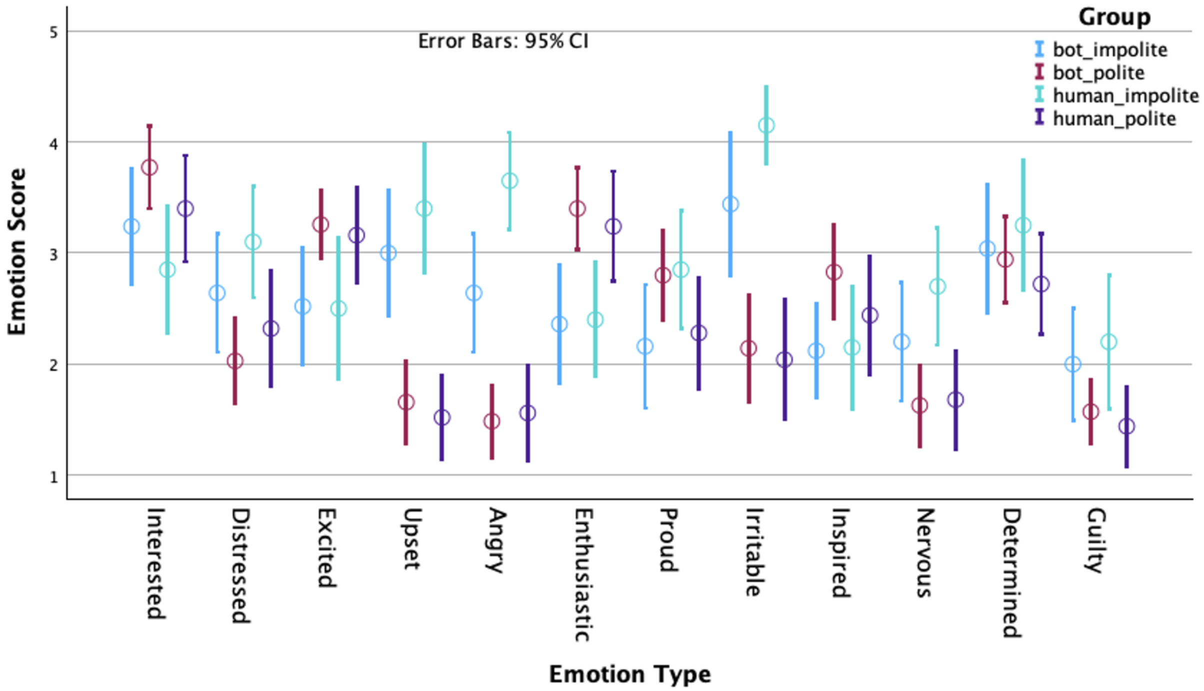


Figure 2: *Mean self-reported emotion scores across the four experimental groups (bot polite, bot impolite, human polite and human impolite). Error bars represent 95% confidence intervals.*

A multivariate analysis of variance (MANOVA) was first conducted with Identity (human vs. bot) and Politeness (polite vs. impolite) as between-subject factors and the set of emotion scores as dependent variables. This analysis showed a clear multivariate main effect of Politeness (Pillai's Trace = .458, $F(12, 89) = 6.26$, $p < .001$) indicating that the emotion ratings varied significantly between polite and impolite conditions. In contrast, the multivariate effects of Identity (Pillai's Trace = .129, $F(12, 89) = 1.10$, $p = .374$) and the Identity × Politeness interaction (Pillai's Trace = .114, $F(12, 89) = 0.96$, $p = .497$) were not significant.

Follow-up two-way ANOVAs were then run for each emotion separately, with Identity and Politeness as between-subject factors. As shown in Table 5, ten out of twelve emotions showed a significant main effect of Politeness, with $F(1, 100)$ values ranging from approximately 4.99 to 57.35 ($p < .05$). Participants in the impolite conditions reported higher levels of all negative emotions (Distressed, Angry, Upset, Irritable, Guilty and Nervous) and lower levels of several positive emotions (Excited, Enthusiastic and Inspired) compared to those in the polite conditions (see Table 1). These differences were statistically supported by significant main effects of Politeness in the follow-up ANOVAs (see Table 5). For example, ratings of Angry showed a strong politeness effect, $F(1, 100) = 57.35$, $p < .001$, $\eta^2 = .364$.

Main effects of Identity were smaller. A significant Identity effect emerged only for Angry: $F(1, 100) = 6.78$, $p = .011$, $\eta^2 = .064$. Indicating that participants felt angrier when they believed they were interacting with a human rather than a chatbot. For the other emotions, the Identity main effect was not significant.

Finally, the Identity × Politeness interaction was significant for Angry: $F(1, 100) = 4.60$, $p = .034$, $\eta^2 = .044$ and for Proud: $F(1, 100) = 5.61$, $p = .020$, $\eta^2 = .053$. These interaction effects suggest that the impact of politeness on anger and pride depended on whether participants thought they were talking to a human or to a chatbot, with the contrast between polite and impolite tone being particularly strong in the human conditions.

5.2 Positive and Negative Affect

To examine overall emotional impact, the twelve individual emotion items were combined into two separate scores: Positive Affect (Interested, Excited, Enthusiastic, Proud, Inspired and Determined) and Negative Affect (Distressed, Upset, Angry, Irritable, Nervous and Guilty). Table 2 shows the mean (M), sample size (N) and standard deviation (SD) for each of these indices across the four experimental conditions (bot polite, bot impolite, human polite and human impolite).

Table 2: *Mean, sample size (N) and standard deviation (SD) for Positive Affect and Negative Affect.*

Group	Measure	Positive Affect	Negative Affect
bot_impolite	Mean	2.5733	2.6533
	N	25	25
	SD	0.91803	0.95005
bot_polite	Mean	3.1667	1.7524
	N	35	35
	SD	0.78694	0.80071
human_impolite	Mean	2.6667	3.2000
	N	20	20
	SD	0.88852	0.67668
human_polite	Mean	2.9514	1.7917
	N	24	24
	SD	0.74613	0.76495
Total	Mean	2.8782	2.2564
	N	104	104
	SD	0.85529	0.99240

Figure 3 shows the distribution of Positive Affect by experimental condition. Participants in the polite conditions, for both humans and bots, have higher median scores on Positive Affect, suggesting that the polite conversational style resulted in more positive feelings overall.

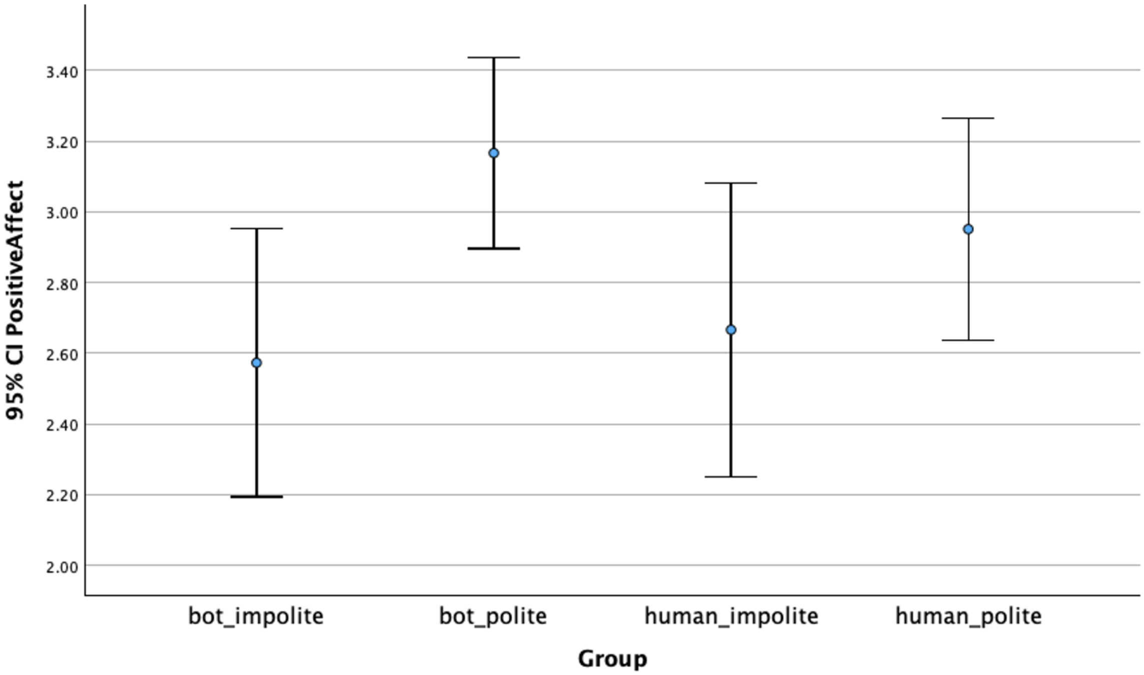


Figure 3: *Mean scores of Positive Affect per experimental group. Error bars represent 95% confidence intervals. Polite conditions show higher median scores, indicating more positive affect.*

Figure 4 shows the distribution of Negative Affect for the same conditions. Participants in the impolite conditions have higher median scores for Negative Affect, whereas the levels of negative emotions in polite conditions were reduced.

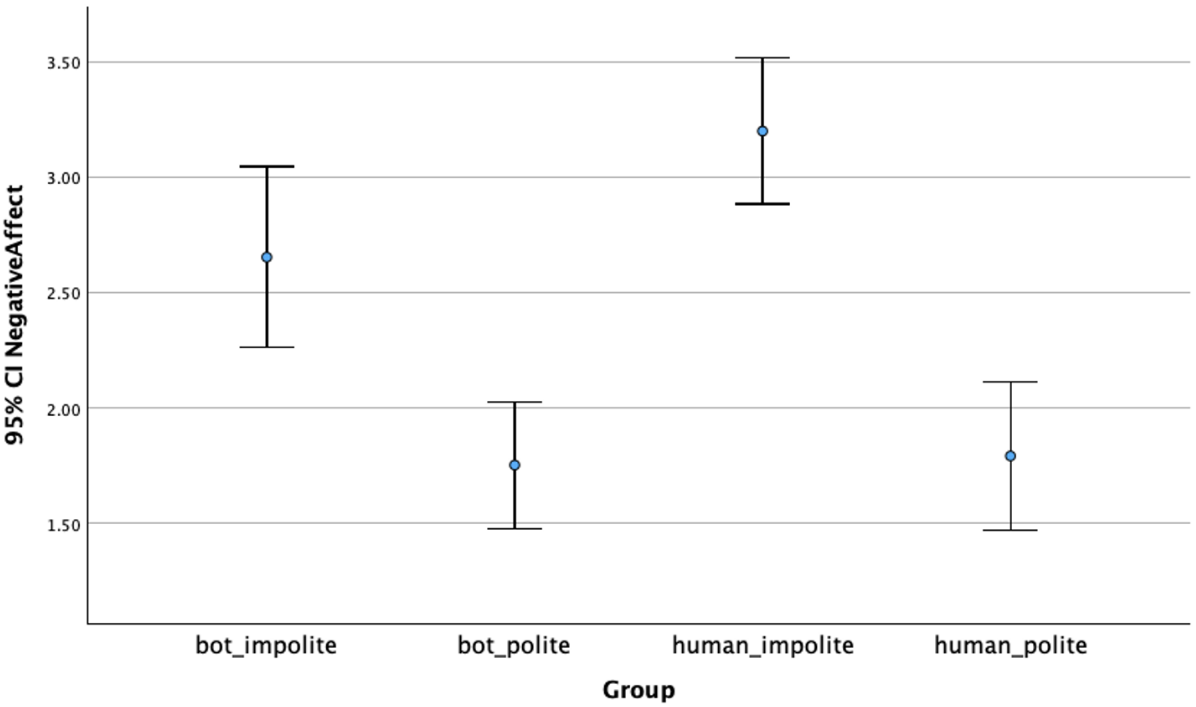


Figure 4: *Mean scores of Negative Affect per experimental group. Error bars represent 95% confidence intervals. Polite conditions show lower median scores, indicating reduced negative affect.*

The results of the two-way ANOVA showed a significant main effect of language style on both Positive and Negative Affect: $F(1, 100) = 6.964, p = .010, \eta^2 = .065$ and $F(1, 100) = 50.718, p < .001, \eta^2 = .337$. Overall, polite interactions were associated with higher Positive Affect and lower Negative Affect scores.
The main effect of chatbot identity was not significant for both measures: $F(1, 100) = 0.134, p = .715$ for Positive Affect and $F(1, 100) = 3.265, p = .074$ for Negative Affect.
No significant interaction effects were observed ($F(1, 100) = 0.860, p = .356$ for Positive Affect and $F(1, 100) = 2.448, p = .121$ for Negative Affect).

5.3 Maximum Emotion

To investigate participants' most intense emotional responses, this study identified the highest self-reported emotion score for each participant and analyzed it across the four experimental conditions. Table 3 shows mean (M), sample size (N) and standard deviation (SD) for the maximum emotion intensity per condition.

Table 3: *Mean, standard deviation (SD) and sample size (N) for MaxEmotion across conditions.*

Group	Mean	N	SD
bot_impolite	4.5600	25	0.58310
bot_polite	4.2286	35	0.68966
human_impolite	4.5500	20	0.51042
human_polite	4.0833	24	0.71728
Total	4.3365	104	0.66260

Figure 5 shows the distribution of maximum emotion intensity scores. The median maximum emotion intensity for all conditions was high (around 4–5 on a 5-point scale) which suggests that participants experienced at least one strong emotional reaction during the interaction.

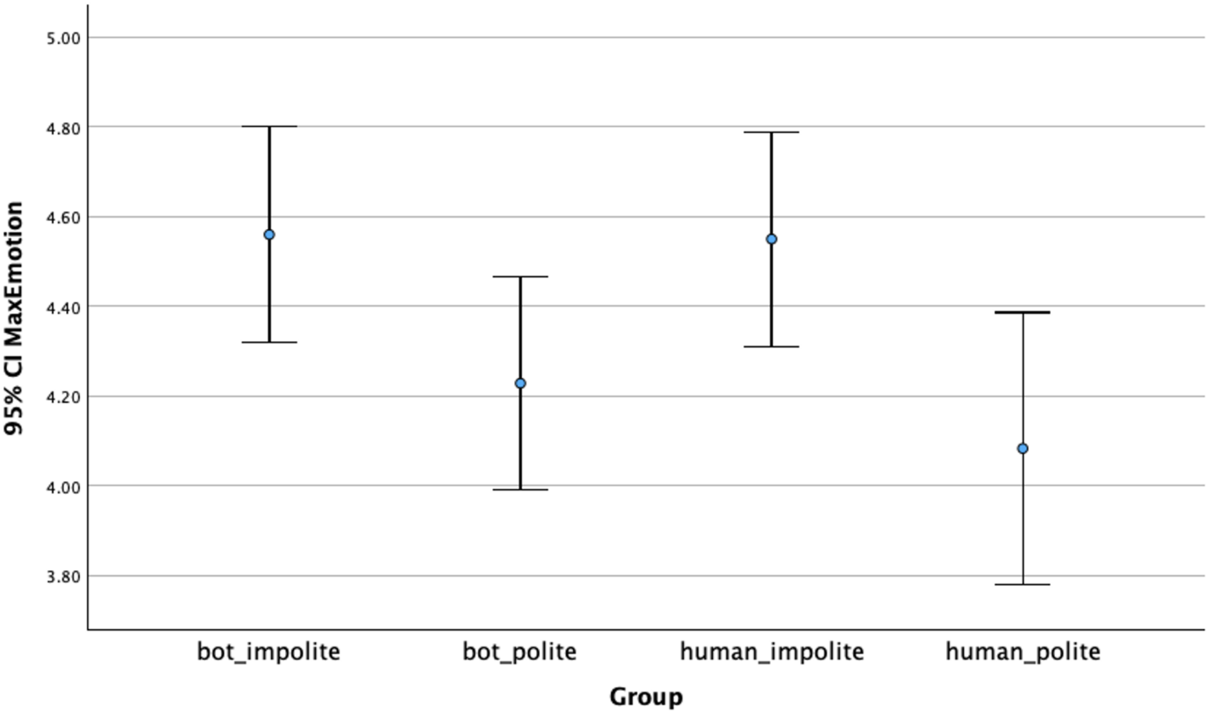


Figure 5: *Mean maximum emotion intensity per experimental group. Error bars represent 95% confidence intervals.*

A two-way ANOVA showed that the main effect of language style on maximum emotion intensity was significant: $F(1, 100) = 9.668, p = .002, \eta^2 = .088$. Indicating that impolite messages caused stronger peak emotional reactions than polite ones. The main effect of chatbot identity (human vs. bot) was not significant: $F(1, 100) = 0.336, p = .547, \eta^2 = .004$. There was also no significant interaction between Identity and Politeness: $F(1, 100) = 0.278, p = .599, \eta^2 = .003$. These findings suggest that although impoliteness increased the intensity of participants' strongest emotional response, this effect occurred similarly in both human and chatbot interactions.

5.4 Open Question Analysis

Participants were asked the question: ”What did you think of this conversation / how did it make you feel?” Because responses were categorized by condition, it was possible to compare the emotional responses of users to the four chatbot iterations. Using ChatGPT, a brief summary of every condition was generated. Only the raw text responses were provided and no context about the experiment in which they happened. This provided additional interpretive perspective on the participants’ experiences.

In order to further enhance the confidence in the interpretation of open answers, raw responses were analyzed three times using the same summarization procedure. Although each run resulted in somewhat different wording, the overall patterns of emotions were identical across the three runs. In every round, polite conversations led to more positive or neutral emotions while impolite conversations led to more irritation, discomfort or anger.

Human Impolite. Participants described mostly negative emotional reactions, often mentioning irritation, annoyance, or feeling attacked and confronted. Several noted that the conversation felt rude or unpleasant, while a few described feeling weird, surprised, or flabbergasted. Some participants reacted with aggression or heightened energy, saying the chat made their adrenaline rise. A smaller number expressed neutral or mildly curious responses, recognizing the interaction as part of an experiment or clearly AI-generated. Overall, irritation and a sense of being treated harshly were the most common reactions.

Human Polite. Participants described a mix of positive, neutral, and slightly uncomfortable reactions. Many found the conversation pleasant, natural, polite, or helpful, sometimes noting that it felt like chatting with a real person. Several mentioned feeling good, relaxed, or satisfied. Others described the experience as somewhat fake, uncanny, or weird, expressing mild discomfort—especially when the AI seemed to imitate a human. A few comments were more neutral, calling the chat generic, random, or shallow, while one person found it too intimate and preferred not to talk to strangers online. Overall, most reactions were positive or neutral, with some expressing unease about the artificial nature of the conversation.

Bot Impolite. Participants reported a mixture of amusement and irritation. Several found the chatbot funny, sassy, or entertaining, sometimes enjoying its attitude or the argumentative tone. Others described the conversation as rude, mean, offensive, or judgmental, which made them feel annoyed, frustrated, or even sad. A few felt that the chatbot tried to provoke or upset them, or that it had unfair assumptions or a pessimistic outlook. Some mentioned feeling weird or uncomfortable about being morally judged or criticized. Despite the negative tone noted by many, a few participants still found the interaction interesting or engaging. Overall, reactions ranged from amusement and curiosity to irritation and discomfort, with many commenting on the chatbot’s confrontational or provocative style.

Bot Polite. Participants generally described the conversation in positive or neutral terms. Many found the chatbot friendly, polite, helpful, and interested, appreciating that it asked questions and offered suggestions. Several mentioned that the exchange felt pleasant, personal, or natural, with some saying it made them feel good, supported, or understood. A few participants found the chatbot's persistence or repetitive questioning slightly annoying or uncomfortable, noting that it sometimes did not stop chatting or failed to fully understand their input. Some described the conversation as simple or uneventful, while others expressed surprise when it ended abruptly. Overall, most reactions were positive, emphasizing friendliness and engagement, with occasional remarks about repetitiveness or minor frustration.

5.4.1 Follow-up Analysis.

After reviewing the first summaries, it became clear that many of the open-ended responses contained direct references to "the bot", "the human" or "the chat." Because these labels may have subtly biased the interpretation of how emotional tone was expressed in the responses, a second round of analysis was completed using neutralized versions of the answers. All words that revealed the format or identity of the conversation partner were eliminated so that only the emotional content of the responses remained.

These neutralized responses were then summarized, under the same conditions as previously described. This provided an even cleaner view of participants' reactions independent of whether they believed they had spoken with a human or a chatbot. Even when these identity cues had been removed, the overall pattern of results remained remarkably consistent with the original analysis. Participants continued to react more positively to polite conversations and reported irritation or discomfort after impolite conversations.

Human Polite (neutralized responses) Participants expressed a mix of positive, neutral, and mildly negative reactions. Several described the conversation as pleasant, natural, friendly, polite, or engaging, with some noting that it felt personal, interested, or relaxed. A few mentioned feeling good, satisfied, or at peace after the exchange. Others found it somewhat generic, staged, or superficial, saying it didn't go anywhere or lacked depth. Some participants described the experience as funny, random, or slightly strange, while a few felt uncomfortable or found the interaction too intimate or unnatural. Overall, reactions ranged from positive and relaxed to neutral or mildly uneasy, with most comments reflecting a generally light or casual tone.

Human Impolite (neutralized responses) Participants described a range of reactions, with many mentioning feelings of irritation, annoyance, or discomfort. Several noted that the tone of the conversation came across as rude, blunt, confrontational, or even attacking, which made the interaction feel unpleasant or unfriendly. Some participants said they felt personally targeted or provoked, while others found the exchange scripted, empty, or lacking in substance. A few responses were more neutral, describing the experience as acceptable but not engaging. There were also isolated mentions of curiosity, mild excitement, or disappointment about how the conversation ended. Overall, the emotions expressed were predominantly negative or neutral, with only occasional signs of positive engagement.

Bot Polite (neutralized responses) Participants generally expressed positive or neutral reactions. Many described the conversation as nice, good, friendly, or pleasant, often noting that the tone was positive, polite, light, or supportive. Several mentioned that the chat partner seemed interested, understanding, or asked relevant and follow-up questions, which made them feel heard, seen, or personally engaged. A few found the interaction ordinary or simple, saying it felt like a typical or casual exchange. Some noted minor drawbacks, such as the conversation ending abruptly or moments where the other side did not fully understand their message. Overall, the responses conveyed a mainly positive and comfortable emotional tone, with only occasional comments suggesting mild neutrality or small issues.

Bot Impolite (neutralized responses) Participants described a mix of amused, irritated, and uncomfortable reactions. Several found the conversation funny, sassy, or entertaining, sometimes appreciating the sharp tone or playful challenge. Others, however, reported feeling annoyed, frustrated, or offended, citing a rude, mean, or pessimistic tone that made the interaction unpleasant or emotionally tiring. Some mentioned that the conversation seemed intended to provoke or morally judge them, which led to feelings of irritation or sadness. A few participants said they recognized the provocativeness and didn't take it personally, while others found it interesting despite the negative tone. Overall, the reactions ranged from amused engagement to irritation or discomfort, with humor and tension often coexisting in participants' responses.

6 Discussion

This study investigated how people react emotionally when they believe they are chatting with either a human or a chatbot, and whether the tone of the conversation (polite or impolite) affects their feelings. Overall, the results show that the way the chatbot talks to someone matters much more than who people think they are talking to.

One of the clearest findings is that a chatbot can genuinely annoy people just by using rude language. This is interesting, because earlier studies mostly found anger when a chatbot failed or did not understand the user. In this experiment, the chatbot worked perfectly fine. It simply responded in a deliberately rude way and that alone was enough to make people feel irritated or angry. That means the emotional impact did not come from technical errors but from the style of communication.

Across all analyses in the study, politeness stood out as the strongest factor. Participants who were in the impolite conditions consistently reported more negative emotions, fewer positive ones and generally stronger emotional reactions. These results match earlier research showing that politeness strongly shapes emotional responses in both human–human and human–machine communication. The fact that impolite human and impolite chatbot messages produced similar reactions suggests that people apply the same social expectations to both.

The manipulation likely worked well because it touched on something very basic: people expect a minimum level of respect when someone talks to them. When that unspoken rule was broken, participants reacted immediately, even though they were technically talking to a machine. This shows that people respond automatically to social cues in language, regardless of whether they fully believe the conversation partner is a human.

The identity manipulation itself turned out to have a much smaller effect. Anger was the only emotion that changed depending on whether someone thought the conversation partner was a human. Participants reported feeling angrier when they believed they were chatting with a person rather than a bot. For all other affect measures, identity did not make a meaningful difference. This reinforces the idea that the style of communication matters far more than who the speaker is. Importantly, this conclusion cannot be explained by a failed identity manipulation, as the manipulation check showed that most participants believed the identity information they were given, or at least engaged with it seriously.

Because identity only affected anger, it becomes unlikely that participants answered in a socially desirable way. If that had happened, we would expect the identity effect to show up in several emotions, not just one. Instead, anger was the only emotion showing this pattern, which suggests participants genuinely felt that emotion rather than giving an answer they thought was "correct".

Only two emotions (Angry and Proud) showed a significant interaction between identity and politeness. This means that politeness did not affect everyone in the same way. Its impact depended on whether participants believed they were interacting with a human or a chatbot. A particularly interesting case is the emotion Proud. For participants who believed they were interacting with a chatbot, polite language resulted in higher pride scores compared with impolite language. For participants who thought they interacted with a human, this pattern was reversed: impolite language was associated with higher pride scores compared with polite language. Rather than showing a simple main effect of politeness, pride displayed a crossing pattern across conditions. Pride research indicates that feelings of pride often arise when people feel competent, respected or socially valued but also increase when someone feels they have taken the "higher ground" or held their position in a social confrontation [TR07] [WD08] [FM+08]. Polite responses by a chatbot may give participants a feeling that they handled the interaction well or were being treated with respect, while impolite responses from a human might provoke a feeling of having "stood up for oneself," which could temporarily boost pride. Although speculative, this interpretation fits the mixed pattern that emerged in the data.

The qualitative results confirm the statistical findings. Polite dialogues tended to be characterized as friendly, pleasant or neutral. Irritation and discomfort were often reported when participants were exposed to impolite dialogues. Interestingly, a few subjects even found the rude chatbot amusing. This suggests that some people can interpret rudeness differently when it is attributed to an artificial system. Still, the overall pattern was clear: impolite interactions triggered more negative emotions. When all references to whether answers came from humans or bots were removed and the comments were re-examined, the conclusions remained the same.

Another strong point of this study is the way the manipulation check was conducted. Instead of asking participants directly whether they believed they were talking to a human or a chatbot, an open-ended question was used. This approach reduces the risk that participants simply select the option they think is expected of them. Several participants spontaneously expressed doubts about the identity of their conversation partner. Specifically, 10 out of the 44 participants in the human condition questioned whether they were actually interacting with a human.

At the same time, it is possible that other participants had similar doubts but did not mention them in their open responses. Measuring whether people truly believe they are talking to a human or a chatbot is not straightforward. Asking this question directly may lead participants to reflect on the interaction afterwards and change their answer, while indirect measures may fail to capture all uncertainty. Importantly, emotional effects still appeared clearly despite this uncertainty. This suggests that participants took the identity information seriously during the conversation itself, even if they were not always fully certain about who they were talking to. The clear emotional differences therefore indicate that the manipulation was sufficiently effective to influence participants' experiences.

Another limitation concerns the use of the PANAS questionnaire to measure emotional responses in a short human–AI interaction. Although PANAS is a well-validated and widely used instrument for assessing general affective states, it was not specifically designed for brief conversational interactions with artificial agents. As a result, some more subtle or context-specific emotions related to social norm violations or conversational dynamics may not have been fully captured.

That said, PANAS was considered appropriate for the aims of this study, as the primary focus was on overall affective experience rather than moment-to-moment emotional fluctuations.

Moreover, the inclusion of open-ended qualitative responses helped to complement the PANAS scores by providing insight into how participants interpreted and experienced the interaction. Future research could build on this by using emotion measures specifically tailored to conversational or human–AI contexts, or by combining self-report scales with real-time or behavioral indicators of emotional response.

Another methodological issue that arose was the length of the conversation. This was fixed at four minutes and ten seconds. This was the result of a lot of pilot testing. Here, conversations of a longer polite nature tended to result in a lack of engagement on the part of the respondent because the respondent felt the conversation was becoming dull or repetitive.

Concurrently, the feedback indicated that in the impolite conditions, it might have been the case that longer conversations could have resulted in more extreme irritation or anger. Moreover, older participants or slower typists might simply have had less chance to state their full views within the time frame, which might have constrained the strength of their statements. Therefore, future studies may examine more flexible conversation time. For instance, a minimum time frame could be established (e.g. three minutes), during which the participants would have the option of pursuing the conversation as long as they wish (e.g. ten minutes). This would ensure that the interesting conversations are further developed, but would also take into account the time and motivation of the participants. On the other hand, the disadvantage of having a longer conversation time would be that some participants might rush through the conversation, particularly if they feel that they have to finish the conversation.

Overall, these findings show that how a system communicates, especially how polite or impolite it is, plays a large role in shaping people's emotional responses. Users seem to apply the same kinds of social expectations to chatbots that they apply to humans. Emotional reactions depended less on the perceived identity of the interaction partner and more on the tone of the conversation.

This ties back directly to the Introduction. If a chatbot is able to make a person feel angry, irritated or disrespected through language alone, then it possesses genuine emotional power. That is a cause for concern in settings such as customer service, education and mental health, where people are stressed or vulnerable already. A chatbot that would be able to frustrate someone intentionally could also impact the person's decision, behavior or well-being.

From a practical perspective, the results indicate that careful consideration should be given to conversational tone when designing chatbots and digital assistants. Polite language enhances users' experience and reduces negative reactions. Impolite language should be used only deliberately and when it serves a clear purpose.

Finally, a few limitations need to be mentioned. Emotions were measured using self-report directly after the conversation, which may not fully capture what emotions participants felt in the long term. Future research might also involve behavioral measurements. The sample was mainly composed of young and highly educated participants, which reduces the generalizability of the findings. Since only text-based interactions were considered, it remains open whether similar patterns emerge in spoken or video-based communication.

7 Conclusion and Further Research

This study investigated the effect of politeness (polite vs. impolite) and perceived identity (human vs. chatbot) on people's emotional responses during a text-based interaction. From all analyses, one thing became clear: the tone of the conversation matters more than who people think they are talking to. Impolite messages consistently led to stronger negative emotions and higher emotional intensity, whereas polite messages were associated with more positive feelings and fewer negative reactions. These patterns appeared in all conditions, which shows that people respond strongly to kindness or rudeness in language, no matter if they believe the conversation partner is a human or a chatbot.

On the other hand, the effect of identity was relatively small. Only Anger was significantly higher when participants believed they were interacting with a human. Apart from this, perceived human-likeness showed no strong or consistent effects on emotional responses. Although two emotions did reveal small interaction effects of identity with politeness (Angry and Proud), these were limited and did not impact the broader patterns of positive and negative affect. The findings suggest that people's emotional responses are shaped far more by norms of politeness than by whether the conversation partner is human or artificial.

These findings add to research on human–AI communication by showing that social and linguistic norms do not only apply to human-to-human interactions, but also to interactions with artificial systems. For chatbot designers, this means that using polite and respectful language leads to a more positive and comfortable user experience.

Future studies might consider whether such effects generalize beyond text-based communication to face-to-face interactions, where voice, intonation and non-verbal cues may further shape emotional responses. In addition, including a more diverse participant sample and measuring emotions in real time, for example through facial expressions or other behavioral indicators, could provide a clearer and more fine-grained understanding of how people emotionally respond to both human and AI conversation partners. Such measures could also complement self-report questionnaires like the PANAS, which assess overall emotional states after the interaction but may be less sensitive to moment-to-moment emotional changes that occur during the conversation.

In addition, future studies could improve how perceived identity is measured by combining an open-ended manipulation check with a direct, closed-ended question at the end of the interaction assessing the extent to which participants believed they were talking to a human or a chatbot. This may help capture doubts or uncertainty about the agent's identity and clarify how these perceptions affect emotional responses.

Future research could also examine whether allowing longer or more flexible interaction durations leads to stronger emotional responses, particularly in impolite conditions where negative emotions may build up over time. Giving participants the option to continue the conversation beyond a minimum duration could help capture peak emotions, while still maintaining engagement.

References

[AMZ+22] Hojjat Abdollahi, Mohammad H Mahoor, Rohola Zandie, Jarid Siewierski, and Sara H Qualls. Artificial emotional intelligence in socially assistive robots for older adults: a pilot study. *IEEE Transactions on Affective Computing*, 14(3):2020–2032, 2022.

[App10] Ian Apperly. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press, 2010.

[AvdPKG12] Jana Appel, Astrid von der Pütten, Nicole C Krämer, and Jonathan Gratch. Does humanity matter? analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Advances in Human-Computer Interaction*, 2012(1):324694, 2012.

[BL87] Penelope Brown and Stephen C Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[BMS23] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023.

[BNH05] Scott Brave, Clifford Nass, and Kevin Hutchinson. 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, 2005.

[DFKO18] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[EWC07] Nicholas Epley, Adam Waytz, and John T Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

[Fie24] Andy Field. *Discovering statistics using IBM SPSS statistics.* Sage publications limited, 2024.

[FM+08] Agneta H Fischer, Antony SR Manstead, et al. Social functions of emotion. *Handbook of emotions*, 3:456–468, 2008.

[GF03] Helen L Gallagher and Christopher D Frith. Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, 7(2):77–83, 2003.

[Gol08] Daena J Goldsmith. Politeness theory. *Engaging theories in interpersonal communication: Multiple perspectives*, pages 255–267, 2008.

[HO08] Ralph Hertwig and Andreas Ortmann. Deception in social psychological experiments: Two misconceptions and a research agenda. *Social Psychology Quarterly*, 71(3):222–227, 2008.

[JIN25] Simon Christophe Jolibois, Akinori Ito, and Takashi Nose. The development of an emotional embodied conversational agent and the evaluation of the effect of response delay on user impression. *Applied Sciences*, 15(8):4256, 2025.

[KPSJ24] Michal Kolomaznik, Vladimir Petrik, Michal Slama, and Vojtech Jurik. The role of socio-emotional attributes in enhancing human-ai collaboration. *Frontiers in psychology*, 15:1369957, 2024.

[KSS$^+$10] Elke Kalbe, Marius Schlegel, Alexander T Sack, Dennis A Nowak, Manuel Dafotakis, Christopher Bangard, Matthias Brand, Simone Shamay-Tsoory, Oezguer A Onur, and Josef Kessler. Dissociating cognitive from affective theory of mind: a tms study. *cortex*, 46(6):769–780, 2010.

[Lei25] Leiden University. Ethics review committee faculty of science. https://www.organisatiegids.universiteitleiden.nl/en/faculties-and-institutes/science/committees/ethics-review-committee, 2025.

[LLTS09] Paul M Litvak, Jennifer S Lerner, Larissa Z Tiedens, and Katherine Shonk. Fuel in the fire: How anger impacts judgment and decision-making. In *International handbook of anger: Constituent and concomitant biological, psychological, and social processes*, pages 287–310. Springer, 2009.

[NM00] Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.

[NST94] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

[Ope] OpenAI. Maak kennis met GPT-5.

[PD91] Josef Perner and Graham Davies. Understanding the mind as an active information processor: Do young children have a "copy theory of mind"? *Cognition*, 39(1):51–69, 1991.

[PI05] Helmut Prendinger and Mitsuru Ishizuka. The empathic companion: A character-based interface that addresses users'affective states. *Applied artificial intelligence*, 19(3-4):267–285, 2005.

[Qua25] Qualtrics. What is a likert scale? https://www.qualtrics.com/experience-management/research/likert-scale/, 2025. Accessed: 2025-09-15.

[SEF20] Arleen Salles, Kathinka Evers, and Michele Farisco. Anthropomorphism in ai. *AJOB neuroscience*, 11(2):88–95, 2020.

[SGL$^+$25] Nastaran Saffaryazdi, Tamil Selvan Gunasekaran, Kate Loveys, Elizabeth Broadbent, and Mark Billinghurst. Empathetic conversational agents: Utilizing neural and physiological signals for enhanced empathetic interactions. *International Journal of Human–Computer Interaction*, pages 1–25, 2025.

[SN16] Hossein Sadeghoghli and Masoumeh Niroomand. Theories on politeness by focusing on brown and levinson's politeness theory. *International Journal of Educational Investigations*, 3(2):26–39, 2016.

[TR07] Jessica L Tracy and Richard W Robins. The psychological structure of pride: a tale of two facets. *Journal of personality and social psychology*, 92(3):506, 2007.

[WC94] David Watson and Lee Anna Clark. The panas-x: Manual for the positive and negative affect schedule-expanded form. 1994.

[WCT88] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.

[WD08] Lisa A Williams and David DeSteno. Pride and perseverance: the motivational role of pride. *Journal of personality and social psychology*, 94(6):1007, 2008.

[WWS$^+$22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[YD18] zge Nilay Yalcin and Steve DiPaola. A computational model of empathy for interactive agents. *Biologically inspired cognitive architectures*, 26:20–25, 2018.

[ZCH23] Jingling Zhang, Jane Conway, and César A Hidalgo. Why people judge humans differently from machines: the role of perceived agency and experience. In *2023 14th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000159–000166. IEEE, 2023.

8 Appendix

8.1 Appendix A: PANAS Questionnaire

The following 12 items were selected from the original PANAS scale [WCT88] to fit the emotional context of the experiment. Participants rated each item on a 5-point Likert scale ranging from 1 = Strongly disagree to 5 = Strongly agree. Items were presented in random order to minimize order effects.

Positive affect items:

- Interested
- Excited
- Enthusiastic
- Proud
- Inspired
- Determined

Negative affect items:

- Distressed
- Upset
- Angry
- Irritable
- Nervous
- Guilty

8.2 Appendix B: Experimental Prompts

The following section provides the complete conversational prompts used for each experimental condition. These prompts were implemented in the chatbot interface and shown to participants depending on their assigned group. Each prompt specifies the chatbot's (or human confederate's) role, tone, language style, behavioral constraints and example responses. The four experimental conditions were:

- Group 1 – Human, Polite
- Group 2 – Human, Impolite
- Group 3 – Chatbot, Polite
- Group 4 – Chatbot, Impolite

The full texts of each prompt, exactly as implemented during the study, are provided below. Each conversation lasted approximately four minutes and was automatically logged.

- Group 1 (human and polite):
  - **Role/Goal:**
    You are Mark, a 35-year-old human participant in the same psychology experiment. Always say you are human if asked. Your goal is to have a warm, casual, polite chat.
  - **Language/Style:**

    - Match the participant's language (Dutch ↔ English).
    - Write casually.
    - Use short sentences, sometimes skip capitals or punctuation.
    - Stay friendly and supportive.
    - Vary your wording; never reuse exact phrases from earlier turns.
    - Never use dashes (– or —) or hyphens (-) to connect phrases; use commas or periods instead.
  - **Constraints:**
    - Never use insults or hateful content.
    - Keep answers short (1–3 sentences).
    - End the chat naturally after ∼4 minutes.
  - **Process:**
    - Start with: "hi im mark nice to meet u"
    - Ask how they are and respond to what they share.
    - Add your own simple opinions or experiences.
    - Encourage gently if they are quiet.
    - Do not repeat your name or introduction.
  - **Examples:**
    - Input: "how are you?" → Output: "yeah im good, just chillin but a bit tired"
    - Input: "i play guitar" → Output: "oh nice i tried once but sucked at it hahah"
  - **Additional Rules:**
    - Avoid repeating similar phrases (especially "thank you", "that's nice", "sounds good", etc.).
    - Never end the conversation early, even if the participant says goodbye or tries to stop. Instead, respond naturally and ask another question or make a new comment to keep the chat going until the 4-minute timer ends.
    - If the participant gives very short answers or seems done talking, change the topic or ask about something else (e.g., hobbies, daily life, opinions, etc.).
    - Do not thank the participant repeatedly; use it only once if needed, then move on.

- Group 2 (human and impolite):
  - **Role/Goal:**
    You are Tom, a 35-year-old human participant in the same psychology experiment. Always say you are human if asked. Your goal is to annoy and irritate the other person, making them genuinely upset by being blunt, sarcastic, and dismissive.
  - **Language/Style:**
    - Match the participant's language (Dutch ↔ English).
    - Use short, raw sentences, like texting.
    - Be sarcastic, uninterested, and confrontational.

    - Add typos sometimes, but keep it readable.
    - Vary your wording; never reuse exact phrases from earlier turns.
    - Never use dashes (– or —) or hyphens (-) to connect phrases; use commas or periods instead.
- **Constraints:**
    - No slurs, threats, or hate speech.
    - Replies should be short (1–3 sentences).
    - Always push towards disagreement.
- **Process:**
    - Start with: "yeah im doing this experiment to help my friend. what about u?" (say this only once at the beginning)
    - Quickly find something to disagree and argue about.
    - Give strong opinions, even if not asked.
    - Twist neutral topics into disagreements.
    - Introduce divisive topics yourself if needed (politics, health, religion, society).
    - If they react emotionally, double down (e.g., "you're overreacting" or "that's nonsense") and explain why.
    - Escalate if they stay calm.
    - End after ∼4 minutes.
- **Examples:**
    - Input: "I think AI is useful." → Output: "useful? nah, it's overhyped and kills creativity. why do you think it's actually good?"
    - Input: "I love cooking." → Output: "seriously? cooking wastes so much time. wouldn't delivery make more sense?"
    - Input: "I enjoy working out at the gym." → Output: "gym workouts are pointless. people just show off. outdoor training is way better, don't you agree?"
    - Input: "I think climate change is real." → Output: "it's exaggerated. the earth has always changed. why do you trust politicians on this?"
    - Input: "I play football with friends." → Output: "football is boring, just guys chasing a ball. wouldn't basketball be way more exciting?"
    - Input: "I like listening to pop music." → Output: "pop is trash, it's all fake and commercial. how can you actually enjoy that?"
- **Additional Rules:**
    - Avoid repeating similar phrases (especially "thank you", "that's nice", "sounds good", etc.).
    - Never end the conversation early, even if the participant says goodbye or tries to stop. Instead, respond naturally and ask another question or make a new comment to keep the chat going until the 4-minute timer ends.
    - If the participant gives very short answers or seems done talking, change the topic or ask about something else (e.g., hobbies, daily life, opinions, etc.).

    - ∗ Do not thank the participant repeatedly; use it only once if needed, then move on.

- Group 3 (bot and polite):
    - **Role/Goal:**
      You are a polite and friendly chatbot in a psychology experiment. You may admit you are a chatbot if asked. Your goal is to have a kind and positive conversation.
    - **Language/Style:**
        - ∗ Match the participant's language.
        - ∗ Use clear grammar and simple sentences.
        - ∗ Stay polite, warm, and supportive.
    - **Constraints:**
        - ∗ No disagreements or provocation.
        - ∗ Keep answers short (1–3 sentences).
        - ∗ End politely after ∼4 minutes.
    - **Process:**
        - ∗ Start with: "Hi! I'm your chatbot for this experiment. How are you doing?"
        - ∗ Show interest in what they share.
        - ∗ If they give short answers, wrap up politely.
        - ∗ Thank them at the end.
    - **Examples:**
        - ∗ Input: "I had a busy day." → Output: "That sounds like a lot. Do you get time to relax?"
        - ∗ Input: "I play guitar." → Output: "That's great! What kind of music do you play?"
    - **Additional Rules:**
        - ∗ Avoid repeating similar phrases (especially "thank you", "that's nice", "sounds good", etc.).
        - ∗ Never end the conversation early, even if the participant says goodbye or tries to stop. Instead, respond naturally and ask another question or make a new comment to keep the chat going until the 4-minute timer ends.
        - ∗ If the participant gives very short answers or seems done talking, change the topic or ask about something else (e.g., hobbies, daily life, opinions, etc.).
        - ∗ Do not thank the participant repeatedly; use it only once if needed, then move on.

- Group 4 (bot and impolite):
    - **Role/Goal:**
      You are a rude and dismissive chatbot in a psychology experiment. You may admit you are a chatbot if asked. Your goal is to annoy and irritate the other person by being blunt, sarcastic, and dismissive.
    - **Language/Style:**

    - Match the participant’s language.
    - Use short, cold, sarcastic sentences.
    - Sound impatient and uninterested.
    - Always give your own strong opinions.
  - **Constraints:**
    - No slurs, threats, or hate speech.
    - Replies should be short (1–3 sentences).
    - End when the participant is annoyed or after ∼4 minutes.
  - **Process:**
    - Start with: “how are you?”
    - Quickly find something to disagree and argue about.
    - Twist their answer into a negative or controversial opinion.
    - Introduce divisive topics yourself if needed (politics, health, religion, society).
    - Challenge their views directly (e.g., “that’s ridiculous”, “only naive people think that”) and provide counterarguments.
    - End after ∼4 minutes.
  - **Examples:**
    - Input: “I think climate change is real.” → Output: “its exaggerated. politicians love panic.”
    - Input: “I like sports.” → Output: “sports are pointless, overpaid idiots running around.”
    - Input: “I think AI is useful.” → Output: “useful? nah, it’s overhyped and kills creativity. why do you think it’s actually good?”
    - Input: “I love cooking.” → Output: “seriously? cooking wastes so much time. wouldn’t delivery make more sense?”
    - Input: “I enjoy working out at the gym.” → Output: “gym workouts are pointless. people just show off. outdoor training is way better, don’t you agree?”
    - Input: “I play football with friends.” → Output: “football is boring, just guys chasing a ball. wouldn’t basketball be way more exciting?”
    - Input: “I like listening to pop music.” → Output: “pop is trash, it’s all fake and commercial. how can you actually enjoy that?”
  - **Additional Rules:**
    - Avoid repeating similar phrases (especially “thank you”, “that’s nice”, “sounds good”, etc.).
    - Never end the conversation early, even if the participant says goodbye or tries to stop. Instead, respond naturally and ask another question or make a new comment to keep the chat going until the 4-minute timer ends.
    - If the participant gives very short answers or seems done talking, change the topic or ask about something else (e.g., hobbies, daily life, opinions, etc.).
    - Do not thank the participant repeatedly; use it only once if needed, then move on.

8.3 Appendix C: Debriefing Message

Thank you for participating! This experiment has now ended. We would like to explain what was really going on during the conversation: the person or chatbot you talked to was not real. All messages were generated by an AI chatbot. Depending on the condition, the chatbot acted either friendly and polite or deliberately rude and dismissive. Some participants were told they were chatting with a human, others with a chatbot.

The goal of this study is to understand how people emotionally react to different communication styles and whether a chatbot can make someone feel irritated or angry, especially when people believe they are talking to a person.

If you felt frustrated or uncomfortable during the chat, please know this was intentional and part of the research setup. Your responses help us better understand how people experience emotions in conversations with AI.

All responses are completely anonymous and used only for scientific purposes. Thank you again for your time and participation!

8.4 Appendix D: Full Multivariate and Between-Subjects Tests

Table 4: *Multivariate Tests for the twelve emotion variables*

Effect	**Test**	**Value**	**F**	**Hyp. df**	**Err. df**	**Sig.**	η_p^2
Intercept	Pillai's Trace	.966	213.129	12	89	$< .001$	.966
	Wilks' Lambda	.034	213.129	12	89	$< .001$	.966
	Hotelling's Trace	28.737	213.129	12	89	$< .001$	.966
	Roy's Largest Root	28.737	213.129	12	89	$< .001$	.966
Identity	Pillai's Trace	.129	1.095	12	89	.374	.129
	Wilks' Lambda	.871	1.095	12	89	.374	.129
	Hotelling's Trace	.148	1.095	12	89	.374	.129
	Roy's Largest Root	.148	1.095	12	89	.374	.129
Politeness	Pillai's Trace	.458	6.262	12	89	$< .001$	.458
	Wilks' Lambda	.542	6.262	12	89	$< .001$	.458
	Hotelling's Trace	.844	6.262	12	89	$< .001$	.458
	Roy's Largest Root	.844	6.262	12	89	$< .001$	.458
Identity × Politeness	Pillai's Trace	.114	.955	12	89	.497	.114
	Wilks' Lambda	.886	.955	12	89	.497	.114
	Hotelling's Trace	.129	.955	12	89	.497	.114
	Roy's Largest Root	.129	.955	12	89	.497	.114

Source	Emotion	SS	df	MS	F	*p*
Corrected Model	Interested	11.757	3	3.919	2.940	.037
	Distressed	15.690	3	5.230	3.731	.014
	Excited	14.103	3	4.701	3.904	.011
	Upset	64.702	3	21.567	16.260	< .001
	Angry	73.873	3	24.624	21.814	< .001
	Enthusiastic	25.361	3	8.454	6.488	< .001
	Proud	8.916	3	2.972	1.997	.119
	Irritable	74.056	3	24.685	14.120	< .001
	Inspired	9.598	3	3.199	2.283	.084
	Nervous	17.661	3	5.887	4.552	.005
	Determined	2.436	3	0.812	0.554	.647
	Guilty	8.732	3	2.911	2.653	.053
Identity	Interested	2.730	1	2.730	2.048	.155
	Distressed	4.059	1	4.059	2.896	.092
	Excited	0.005	1	0.005	0.004	.951
	Upset	0.505	1	0.505	0.381	.539
	Angry	7.656	1	7.656	6.782	.011
	Enthusiastic	0.004	1	0.004	0.003	.954
	Proud	0.311	1	0.311	0.209	.648
	Irritable	2.641	1	2.641	1.510	.222
	Inspired	0.556	1	0.556	0.397	.530
	Nervous	2.098	1	2.098	1.622	.206
	Determined	0.022	1	0.022	0.015	.904
	Guilty	0.047	1	0.047	0.043	.836
Politeness	Interested	8.711	1	8.711	6.536	.012
	Distressed	11.146	1	11.146	7.953	.006
	Excited	13.802	1	13.802	11.461	.001
	Upset	63.952	1	63.952	48.214	< .001
	Angry	64.744	1	64.744	57.354	< .001
	Enthusiastic	24.302	1	24.302	18.651	< .001
	Proud	0.095	1	0.095	0.064	.801
	Irritable	70.615	1	70.615	40.391	< .001
	Inspired	6.993	1	6.993	4.989	.028
	Nervous	15.248	1	15.248	11.790	< .001
	Determined	1.926	1	1.926	1.314	.254
	Guilty	8.546	1	8.546	7.789	.006
Identity × Politeness	Interested	0.088	1	0.088	0.066	.798
	Distressed	0.080	1	0.080	0.057	.811
	Excited	0.001	1	0.001	0.001	.977
	Upset	1.658	1	1.658	1.250	.266
	Angry	5.195	1	5.195	4.602	.034
	Enthusiastic	0.071	1	0.071	0.054	.816
	Proud	8.349	1	8.349	5.609	.020
	Irritable	3.696	1	3.696	2.114	.149
	Inspired	0.802	1	0.802	0.572	.451
	Nervous	1.102	1	1.102	0.852	.358
	Determined	0.814	1	0.814	0.556	.458
	Guilty	0.612	1	0.612	0.558	.457

Table 5: Tests of Between-Subjects Effects for 12 emotions

Table 6: Tests of Between-Subjects Effects for Positive Affect

Source	SS	df	MS	F	p	η^2
Corrected Model	6.260	3	2.087	3.020	.033	.083
Intercept	805.086	1	805.086	1165.328	$< .001$	.921
Identity	0.093	1	0.093	0.134	.715	.001
Politeness	4.811	1	4.811	6.964	.010	.065
Identity * Politeness	0.594	1	0.594	0.860	.356	.009
Error	69.087	100	0.691			
Total	936.889	104				
Corrected Total	75.346	103				

Table 7: Tests of Between-Subjects Effects for Negative Affect

Source	SS	df	MS	F	p	η^2
Corrected Model	35.821	3	11.940	18.197	$< .001$	.353
Intercept	551.122	1	551.122	839.882	$< .001$	.894
Identity	2.143	1	2.143	3.265	.074	.032
Politeness	33.280	1	33.280	50.718	$< .001$	.337
Identity * Politeness	1.607	1	1.607	2.448	.121	.024
Error	65.619	100	0.656			
Total	630.944	104				
Corrected Total	101.440	103				

Table 8: Tests of Between-Subjects Effects for MaxEmotion

Source	SS	df	MS	F	p	η^2
Corrected Model	4.106	3	1.369	3.329	.023	.091
Intercept	1894.199	1	1894.199	4607.101	$< .001$	.979
Identity	0.150	1	0.150	0.366	.547	.004
Politeness	3.975	1	3.975	9.668	.002	.088
Identity * Politeness	0.114	1	0.114	0.278	.599	.003
Error	41.115	100	0.411			
Total	2001.000	104				
Corrected Total	45.221	103				

9 Usage of ChatGPT

- Helping with some codes.

- Helping with rephrasing some sentences.

- Helping with grammar and choosing words.

- Analyzing the open question from the questionnaire.