The Impact of Empathetic Language on Willingness to Disclose Mental Health Related Information to a Chatbot

Lina Sadoukri[1]

Supervisor(s): Ujwal Gadiraju[1]**, Esra de Groot**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Lina Sadoukri
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Esra de Groot, Myrthe Tielman

Abstract

This study investigates whether empathetic language in chatbot interactions influences users' willingness to disclose mental health-related information. Using a two-by-two mixed factorial design, 114 participants were assigned to either an empathetic or neutral chatbot condition and responded to both emotional and behavioural health questions. While prior research suggests that empathy can pormote trust and openness, results from this study revealed no significant difference in disclosure willingness across chatbot styles or question types, although the manipulation check showed a well perceived empathy. The study highlights the importance of individual predispositions, such as prior readiness to disclose, in shaping interactions with digital mental health tools. Future work should explore longer-term interactions and real disclosure behaviour to better understand the role of empathy in chatbot design.

KEYWORDS
Chatbots, Mental health disclosure, Human-computer interaction, Empathy, Self-disclosure, Digital mental health conversational agents

1 Introduction

The use of mental health chatbots has increased considerably in recent years, providing a scalable approach to delivering psychological support [1][2][16]. These chatbots are available any time, provide non-judgmental responses and may help reduce the stigma associated with seeking mental health support [7][14]. A recent study has shown that users interact with chatbots to seek reassurance, reduce stress or increase motivation [8]. By allowing anonymous interactions, chatbots can improve access to care, especially for individuals who find it difficult to seek help from human therapists [11] [9].

Effective mental health chatbots depend on users disclosing personal information, as systems encouraging open self-disclosure show higher levels of engagement and satisfaction [16]. However, many users view mental health information as sensitive and are reluctant to share it in interactions involving AI (Artificial Intelligence) systems [20]. Building trust is essential to encourage users to disclose such sensitive data [11]. Another key limitation is that mental health chatbots are often perceived as lacking empathy, emotional intelligence and human-like interactions [7] [14].
Previous studies have explored factors such as anthropomorphism [17], empathy [15] and communication style [24] that influence user trust, engagement, and disclosure. For example, users were more likely to disclose information to chatbots with human-like features [17] and empathetic chatbots could improve emotional expression in teenagers [15].

While these studies suggest that empathy can improve user-chatbot interactions, there remains a gap in understanding the specific impact of empathetic language on mental health disclosure. Different types of questions may influence how open and comfortable participants feel. Emotional questions, which focus on internal feelings (e.g., "How sad do you feel?"), may lead to higher vulnerability and require more trust to share [11]. A study found that emotional prompts made users more vulnerable, especially when coming from a familiar chatbot, while factual questions were experienced as less intrusive and easier to answer [4]. Understanding this distinction may be essential for developing more effective mental health tools that can adapt their communication style based on the required information [7][16].

This research aims to answer the following question: *Does empathetic language in a chatbot affect users' willingness to disclose mental health-related information, and does this effect differ depending on the type of questions asked?*
This will be guided by the following sub-questions:
RQ1: Does empathetic language in chatbot interactions increase users' willingness to disclose personal mental health infromation compared to neutral language?
RQ2: Does willingness to disclose differ between emotional and behavioural health questions, regardless of chatbot style?
RQ3: Does the effect of empathetic language on willingness to disclose differ depending on whether the questions are emotional or behavioural?

The report is set out as follows. The next chapter reviews related work on empathy, chatbots and mental health information disclosure. Chapter 3 discusses the methodology of the project, including the chatbot implementation. Chapter 4 presents the study's findings, which are then interpreted in the following discussions chapter, including a consideration of the study's limitations. Finally, the report ends with a conclusion and recommendations for future developments.

2 Background

The following section explores related work on empathy, chatbots and mental health information disclosure. Building on De Gennaro et al., who demonstrated that empathetic chatbot responses reduced emotional distress caused by social exclusion [6], our study first investigates the effect of empathetic communication on the willingness to disclose mental health information (Hypothesis 1). We then extend this work by examining whether willingness to disclose differs between emotional and behavioural health questions regardless of communication style (Hypothesis 2), and whether this effect interacts with empathy (Hypothesis 3).

2.1 Empathy in Human-Computer Interaction

Empathy is commonly defined as the ability to understand and share the emotional state of others. It includes cognitive empathy, which refers to understanding another person's perspective, and affective empathy, which involves emotionally connecting with another person's feelings [21][19][23]. In chatbot design, empathy refers to the system's ability to generate responses that users interpret as caring and understanding. This would be achieved by using emotionally validating and supportive language that is sensitive to user context [3].

Empathetic language plays a key role in influencing users' feelings of trust, comfort and psychological safety during

their interactions with chatbots [15][13]. Lucas et al. found that participants who believed they were interacting with an empathetic virtual agent felt lower fear of disclosure than those who believed they were interacting with a human [13]. This suggests that empathetic language can reduce stigma and psychological barriers in mental health disclosure. Research in this domain has shown that virtual agents showing empathy can significantly increase user satisfaction, trust and engagement [15][22], as well as help users open up and make the conversation feel more positive. Empathy in chatbots is often expressed through human-like features (anthropomorphism) such as appearance or behaviours, which help users perceive the emotional side of the agent. Brukner et al. found that well-designed empathetic chatbots can be perceived comparable in perceived care and understanding as humans [27]. Additionally, anthropomorphic features in chatbots significantly impacted users' willingness to disclose [17]. Building on these findings that empathetic language enhances users' trust and willingness to disclose sensitive mental health information, we propose the following hypothesis.

Hypothesis H1: Empathetic language in chatbot interactions will increase users' willingness to disclose personal mental health information compared to neutral language.

2.2 Mental Health Disclosure in Digital Contexts

Self-disclosure in mental health contexts involves sharing personal and often sensitive thoughts, feelings, and experiences about one's mental and emotional well-being. Liu et al. make a distinction between functional-utilitarian contexts (task-focused interactions) and social-emotional contexts (seeking emotional support) in their study [12]. Participants demonstrated significantly different patterns of self-disclosure depending on the context. Disclosure was higher in social-emotional settings, where interactions were perceived as more supportive. Recent research has explored disclosure patterns within social-emotional contexts. Croes et al. investigated users' willingness to share intimate information with a chatbot and its effects on emotional well-being. They found that digital confessions can have positive therapeutic effects when users feel comfortable with disclosure [5]. Finally, disclosure willingness varies significantly based on the type of information being requested.

2.3 Emotional vs. Behavioural Aspects of Mental Health

While disclosure context and chatbot design influence user openness, the type of information requested (emotional or behavioural) may further influence willingness to disclose. The Indiana Center for Recovery defines mental health as "the state of well-being concerning one's psychological and emotional resilience", whereas behavioural health "encompasses actions and habits that impact mental and emotional well-being". Essentially, emotional mental health refers to internal experiences, mood and cognition, whereas behavioural health involves observable patterns that influence overall health, such as sleep habits, substance use and routine activities [18]. This distinction is commonly recognized in psychology and public health research. Thapa et al. demonstrated that integrated health models work best when addressing emotional and behavioural components separately [25]. These frameworks highlight the importance of treating emotional and behavioural disclosures differently, which may be important when designing chatbots.

However, individuals tend to be more cautious when sharing their emotions. The Distress Disclosure Index shows that willingness to share emotional distress varies widely per individual, with lower levels of disclosure associated with lower mental well-being and self-esteem [10]. This variation highlights the emotional sensitivity of sharing personal feelings, especially when interacting with non-human agents.

Overall, these insights suggest that both the context of disclosure and the type of mental health information requested play a key role in shaping users' willingness to share sensitive information. We then propose the following hypotheses.

Hypothesis H2: Willingness to disclose will differ between emotional and behavioural health questions, with behavioural questions prompting higher willingness to disclose.

Hypothesis H3: There will be an interaction between the chatbot communication style and question type, such that the increase in willingness to disclose in empathy will be greater for emotional questions.

3 Methodology

This section outlines the methodology, detailing the study design and procedure.

3.1 Experimental Design

This study employed a two-by-two mixed factorial design to examine how the different conditions affect willingness to disclose mental health information. This design was chosen to efficiently investigate both the impact of chatbot communication style and the type of questions on disclosure, while also exploring potential interaction effects between these factors.

Between-subjects analysis The between-subjects factor (chatbot communication style) was used to compare how different users respond to empathetic versus neutral chatbot language without the risk of carryover effects or participant bias. Random assignment of the chatbot condition ensured that differences in willingness to disclose could be attributed to the chatbot style rather than individual differences.

Within-subjects analysis The within-subjects factor (question type) allowed all participants to experience both emotional and behavioural questions. This helps reduce differences between participants and makes the results more reliable by comparing each user's answers to different question types.

Mixed design analysis The mixed design also enabled the investigation of the interaction between chatbot communication style and question type, testing whether empathy has a different effect on emotional and behavioural question types in terms of willingness to disclose. This interaction provides deeper insight into how chatbot design influences disclosure across different types of mental health information.

3.2 Chatbot Implementation

The empathetic language condition was systematically designed based on established empathy theory and research. Drawing from Riess et al. [21] and Bickmore et al. [3], it was determined that empathy in digital contexts requires language that demonstrates understanding, validation, and emotional support without requiring the chatbot to actually experience emotions. The empathetic responses incorporated three key components;

Cognitive empathy elements, or language demonstrating understanding of the user's perspective (e.g., "I hear you, and I want you to know that's completely okay! Your boundaries matter, and I respect them fully.")

Affective validation: Responses that acknowledge and normalize emotional experiences (e.g., "Loneliness is one of the most universal human experiences — it can touch us even when we're surrounded by people, and it's nothing to be ashamed of.")

Supportive framing: Language that creates psychological safety and reduces potential judgment (e.g., "whatever your relationship with substances is, you won't be judged here.")

Empathetic messages were developed through a literature review on empathy. All empathetic messages were designed to maintain a consistent tone of warmth, non-judgment, and emotional availability. The chatbot provided different empathetic responses based on participants' willingness ratings, with customized supportive messages for each level of disclosure comfort. For instance, when participants indicated they were "not willing" to answer a question, the chatbot would respond with messages such as "I hear you, and I want you to know that's completely okay! Your boundaries matter, and I respect them fully." For higher willingness levels, responses included "Thank you, your openness and trust mean so much!" These varied responses ensured that the empathetic condition felt genuinely responsive to participant input. In contrast, the neutral condition used standard responses regardless of the participant's willingness level, maintaining consistency in tone without emotional engagement. The responses contained standard acknowledgments such as "Thank you for your response", or factual confirmations such as "Your answer has been recorded", or simply transitions such as "Moving to the next question".

3.3 Participants and Sample

An a priori power analysis was conducted using G*Power 3.1 to determine the required sample size for the primary statistical tests. The analysis targeted a medium effect size (Cohen's d = 0.5) for the difference between two independent groups and a small effect size (Cohen's d = 0.2) for two dependent groups. The parameters were set with an alpha level of 0.05 and a desired power of 0.80 to minimize the risk of Type II errors. The results indicated that a total sample size of 200 participants would be sufficient to detect the expected effect size. However, due to time constraints within the research timeline, the study was conducted with a smaller sample of 114 participants (57 in each chatbot condition). As two tests were planned in this study, the required sample size was based on the largest calculation to maintain adequate statistical power across all comparisons.

Participants were recruited through university networks, social media platforms and participant recruiting platforms. Inclusion criteria required participants to be at least 16 years old, fluent in English and residing in Europe. Demographic information was collected using age ranges (16-20, 21-25, 26-30, 31-35, etc.) and self-reported gender categories. No exclusion criteria related to mental health status were applied, as the study measured willingness to disclose rather than actual mental health conditions.

3.4 Study Procedure

The complete study session lasted approximately 5 to 8 minutes and followed this sequence:

- An informed consent.
- A pre-task survey assessing participants' demographic data (age and gender), prior experience with chatbots, general trust in AI systems and willingness to disclose mental and physical health infromation.
- The main chatbot interaction, during which participants reported their willingness to answer mental health-related questions.
- A post-task survey evaluating participants' perceptions of the chatbot and overall experience.

The study used two separate scenarios to present emotional and behavioural mental health questions, making it easier to compare the two types while helping participants stay focused on each question category. This approach was chosen to help maintain the distinction between both types of questions, improving the clarity and validity of responses.

Participants interacted with a custom-built chatbot interface developed specifically for this study. The interface maintained a clean, text-based design with consistent formatting to avoid confounding visual factors with the language manipulation. Key features included a permanently visible "Revoke Consent" button for immediate withdrawal, a "Task Instructions" button providing persistent access to study information and consistent response formatting using Likert scale buttons. Participants were randomly assigned to empathetic or neutral conditions using JavaScript seed randomization. Within each condition, the order of emotional and behavioral question scenarios was counterbalanced to prevent order effects. Within each scenario, questions followed a predetermined sequence from less to more sensitive topics. This gradual progression was intended to build trust and make participants feel more at ease, increasing the likelihood of willingness to disclose sensitive information.

3.5 Questions and Instruments

The study used validated mental health questions and validated survey instruments to assess participants' familiarity

with chatbots, disclosure willingness and perceptions of the chatbot.

Pre-Task Survey

Before the chatbot interaction, participants completed a brief survey to collect age and gender and to assess their general familiarity with chatbots, trust in them and willingness to share mental and physical health infromation. Familiarity was measured using the Familiarity subscale of the Social Service Robot Interaction Trust (SSRIT) scale (A.5). In addition, three custom items were developed to assess trust in chatbots specifically for health contexts: "I would trust chatbots with my information", "I am willing to share mental health information with a chatbot" and "I am willing to share physical health information with a chatbot".

These variables — age, gender, prior chatbot familiarity, trust and and willingness to share health-related information to chatbots — were included as potential confounders, as they may influence users' willingness to disclose sensitive information. This is supported by research stating that higher self-reported familiarity with AI is associated with greater trust. Moreover, when AI knowledge is measured objectively, individuals with medium knowledge levels tend to trust AI the most, while those with either very low or very high knowledge show lower trust. These insights highlight the complex interaction between familiarity, knowledge and trust in AI systems, which is why there is a need to account for these factors when investigating disclosure behavior in chatbot interactions.

Chatbot Interaction Questions

The chatbot presented participants with a series of mental health-related questions drawn from the WEBMS Scale [26] and the Codebook Grow It Covid dataset. These questions (A.8) were grouped into two categories:

- Emotional questions, focusing on internal experiences and feelings, including pleasant experiences and unpleasant experiences (e.g. feelings of stress or loneliness).
- Behavioral questions, addressing observable actions and patterns, including low-sensitivity behaviors such as exercise habits, and high-sensitivity behaviors such as substance use.

In the neutral chatbot condition, these questions were asked without additional context. In the empathetic chatbot condition, each question was preceded by an emotionally supportive sentence designed to express empathy and validate the participant's experience (e.g. 'struggling doesn't mean you're failing.' or 'whatever your relationship with substances is, you won't be judged here.'). The empathetic messages were based on prior research about what defines empathy. (Add full list of questions to Appendix)

Post-Task Survey

After the chatbot interaction, participants completed a post-task survey to evaluate their experience and perceptions of their assigned chatbot. Perceived empathy was measured using four items adapted by [5] shown in A.6. Responses were recorded on a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree).

4 Results

This section presents the findings of the study in the form of statistics, figures, and graphs.

4.1 User Attributes

A total of 114 participants completed the study (57 in each condition), with 64% female, 35% male, and 1% non-binary. The sample was predominately young adults, with over 50% of participants between 21 - 25 years old (Table 1). Prior chatbot familiarity scores averaged 68% with a standard deviation of 0.8 on the SSRIT scale (A.5), showing an overall moderate familiarity with chatbots.

Participants showed moderate levels of trust and willingness to share health information with a chatbot. On average, 55.6% indicated they would trust chatbots with their information, 58.6% were willing to share mental health information with a chatbot and 63.6% were willing to share physical health information with a chatbot. To identify predictors of overall willingness to disclose, a hierarchical regression analysis was conducted using all user attributes as potential predictors (Table 6). It was found that only willingness to share mental health information with a chatbot significantly predicted participants' overall scores (B = 0.300, $\beta = 0.414$, t = 3.3263, p = 0.001).

4.2 Manipulation Check

The manipulation check revealed that participants in the empathetic condition perceived higher levels of empathy from the chatbot compared to those in the neutral condition (Figure 1). The chatbot's empathy in the empathetic condition was rated at M = 3.68[1] (SD = 0.58), whereas for the neutral condition, it was rated at M = 2.68[1] (SD = 0.79). Although we can see some higher scores for the neutral condition and some lower ones for the empathetic condition, a Welch's t-test indicated a statistically significant difference between both conditions, t(102.48) = -7.679, $p < 0.001$ and effect size d = 1.44. Overall, the perceived empathy scale (4 items) showed good internal consistency with a Cronbach's α of 0.855.

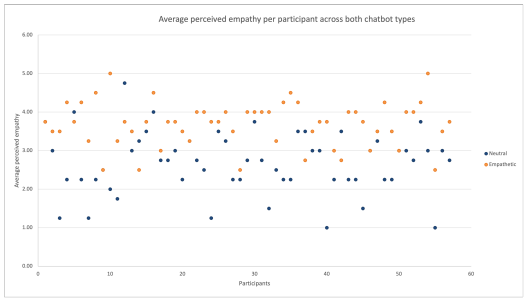

Figure 1: Perceived empathy of the chatbot according to participants in both conditions (generated with Excel).

[1]All scores range from 1 (Strongly disagree) to 5 (Strongly agree).

4.3 Results Analysis

A two-by-two mixed ANOVA was conducted to examine the effects of communication style (empathetic vs. neutral) and question type (emotional vs. behavioural) on willingness to disclose mental health information. All data was adjusted with Bonferroni correction. Statistical Package for the Social Sciences (SPSS) was used to generate all data and conduct all statistical tests.

Main Effect of Empathy

The between-subjects analysis in table 3 shows no significant effect (F(1, 112) = 0.025, p = 0.873). Figure 2 shows that participants in the empathetic condition (M = 3.193[1]) showed nearly identical willingness to disclose compared to those in the neutral condition (M = 3.169[1]).

Additionally, the 95% confidence intervals for both conditions in table 2 display nearly equal values, which confirms a lack of meaningful difference in willingness to disclose between both conditions.

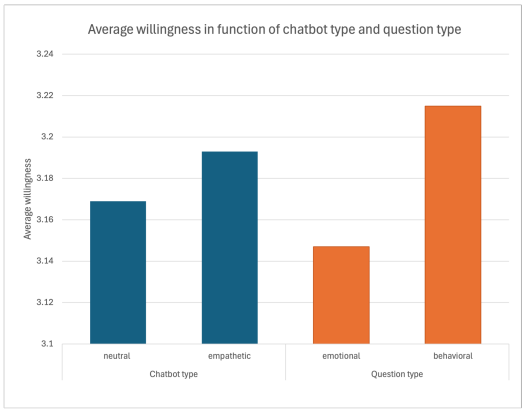

Figure 2: Average willingness to disclose per condition (generated with Excel).

Main Effect of Question Type

Similarly, the within-subjects analysis showed no significant main effect, with F(1, 112) = 1.275 and p = 0.261. Participants showed a similar willingness to disclose for emotional questions (M = 3.147[1]) and behavioural questions (M = 3.215[1]).

Additionally,the 95% confidence intervals for both conditions in table 2 show a very small insignificant difference.

Interaction Effect

Examination of the estimated means (Table 2) showed similar patterns across all four conditions. Emotional questions: Neutral condition (M = 3.136[1], SE = 0.128) vs. Empathetic condition (M = 3.158[1], SE = 0.128) gives us a difference of -0.022. Behavioural questions: Neutral condition (M = 3.202[1], SE = 0.101) vs. Empathetic condition (M = 3.228[1], SE = 0.101) gives us a difference of -0.026. The interaction between communication style and question type was not significant (F(1, 112) = 0.001, p = 0.971). A more detailed analysis was done for the effect of user condition within each question type in Table 5 and showed no significant effect of empathy on emotional questions disclosure (F(1, 112) = 0.015, p = 0.904), or for behavioural questions (F(1, 112) = 0.034, p = 0.854).

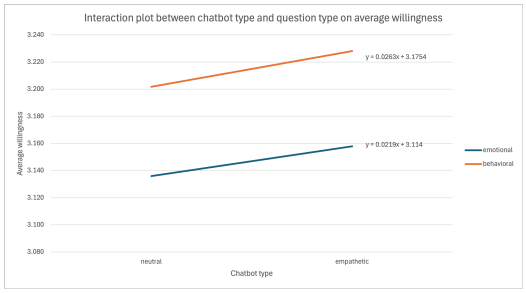

Figure 3: Interaction between willingness to disclose in all four conditions (generated with Excel).

The interaction plot (Figure 3) displays near parallel lines for both slopes. The slope shows a slight increase in willingness to disclose for the empathetic condition compared to the neutral one. Similarly, behavioural questions scored slightly higher than emotional questions across both chatbot conditions. This minimal difference in slopes indicates no significant interaction effect between chatbot type and question type.

5 Discussion

This section interprets the results of the study and connects them to the research questions and existing literature. Additionally, it outlines the limitations and considerations for future work.

5.1 Results Discussion

This study examined whether an empathetic chatbot influenced participants' willingness to disclose mental health-related information compared to a neutral chatbot, and whether this varied depending on the type of question asked (emotional vs. behavioural). The findings are discussed below in relation to the three hypotheses:

H1:

A between-subjects ANOVA analysis revealed no significant difference in willingness to disclose between participants who interacted with an empathetic chatbot (M = 3.193[1]) and those who interacted with a neutral chatbot (M = 3.169[1]). The slight difference in means was not statistically significant (p = 0.873). Therefore, it appears that empathetic language in chatbot interactions does not increase users' willingness to disclose mental health-related information compared to neutral language. Additionally, the empathy was clearly perceived, supported by a manipulation check showing higher empathy ratings in the empathetic condition (M = 3.68[1]) compared to the neutral condition (M = 2.68[1]) and a large effect size (d = 1.44), as well as a high Cronbach's = 0.855.

This suggests that while participants recognized the chatbot's empathetic language, this did not lead to a change in their willingness to disclose. One explanation could be that empathy, when expressed in a brief one-time interaction, is

not sufficient to influence users' behaviour. Another reason would be that empathy is subjective and may be interpreted differently by individuals. This may explain why some participants rated the neutral chatbot as empathetic or the empathetic one as less so.

Finally, the hierarchical regression analysis revealed that participants' readiness to disclose mental health-related information before interacting with the chatbot was a significant predictor of their overall willingness to disclose scores. This suggests that individuals may have a baseline tendency to disclose, which was unaffected by the chatbot communication style. This pattern may be due to cognitive dissonance, where participants would have a tendency to avoid replying in ways that would contradict their initial statements (of being willing to disclose or not).

H2:

A within-subjects ANOVA analysis showed no significant difference in willingness to disclose between emotional questions ($M = 3.147^1$) and behavioural questions ($M = 3.215^1$), with a small statistically insignificant difference ($p = 0.261$). Therefore, willingness to disclose did not significantly differ between emotional and behavioural questions, although behavioural questions prompted slightly higher disclosure.

The results suggest that participants may not have viewed emotional questions as more sensitive than behavioural ones. One possible explanation for this insignificant difference could be that young adults, who made up the majority of the sample, may be more open about emotional content or simply did not perceive emotional questions in this study as more personal or intimate. Similarly to the previous hypothesis, another explanation for these results is that participants' tendency to disclose remained consistent regardless of the type of questions asked.

H3:

A two-by-two mixed ANOVA analysis was performed to analyze the interaction between chatbot communication style (empathetic vs. neutral) with question type (emotional vs. behavioural). No significant interaction between communication style and question type ($F(1,112) = 0.001$, $p = 0.971$) was found. Across all four conditions, the scores of willingness to disclose were very similar. The minimal mean differences and the interaction plot (Figure 3) confirm the lack of interaction effect. Therefore, empathetic language does not interact with question type such that the increase in willingness to disclose in empathy will be greater for emotional questions, and no specific combination of chatbot communication style and question type significantly influenced willingness to disclose. Responses remained consistent regardless of the chatbot's emotional style or the type of questions it asked.

5.2 Limitations

Several limitations should be considered when interpreting the findings of this study.

Willingness to Disclose vs. Actual Disclosure: The study measured participants' willingness to disclose rather than their actual disclosure behaviour. While this approach was ethically necessary, it may not fully assess how participants would respond in real-life situations where actual sensitive information is shared.

Sample limitations: Due to limited time and resources, the study recruited 114 participants (57 per user condition), below the calculated sample size of 200 participants. Additionally, the participant sample consisted mainly of young adults residing in Europe. Cultural attitudes toward mental health and self-disclosure can vary across populations and ages, which may reduce the extent to which these findings can be extended to more diverse and older populations.

Short-Term Interaction: The study was limited to a single and brief interaction with the chatbot. Willingness to disclose was measured immediately during this interaction, which may not reflect how empathy affects openness over time. In real-world settings, disclosure behaviours and trust develop gradually.

Chatbot Design Constraints: Due to time limitations, the study used a rule-based chatbot that lacked adaptivity and natural language processing capabilities. It followed a scripted flow, which may have restricted its ability to engage users in a more natural dialogue.

Question Categorization Ambiguity: The classification of questions into emotional and behavioural categories is theoretically grounded. Some questions may contain both emotional and behavioural elements and participants may interpret the same question differently.

5.3 Future work

Investigation of Actual Disclosure Behaviour: Future studies should aim to capture actual disclosure behaviour rather than self-reported willingness to disclose. This could involve ethically designed studies where participants share information in a more natural setting, providing deeper insights into how empathy influences disclosure.

Sample Diversity: Recruiting participants from more diverse cultural backgrounds and broader age ranges, especially including older adults, would improve the findings and allow exploration of how demographic factors shape disclosure and empathy effects.

Examine Long-Term Interactions: Research should explore how empathetic communication influences willingness to disclose over multiple chatbot interactions. Longitudinal designs can assess how trust and comfort evolve over time.

Implement Advanced Chatbot Technologies: Future work should implement chatbots with conversational AI capabilities to create more dynamic and human-like interactions using a validated dataset. Such enhancements may improve the chatbot's ability to express empathy naturally and engage users more effectively and therefore potentially increasing disclosure.

6 Responsible Research

6.1 Ethics Approval

This study involved the collection of data related to participants' willingness to disclose personal or sensitive mental health information to a digital system. As such, it required careful attention to ethical standards. The research protocol

was developed in accordance with TU Delft's ethical guidelines. Ethical approval was obtained from the Human Research Ethics Commitee (HREC). All procedures, including recruitment, data handling and participants' protection were reviewed to ensure compliance with the university's ethical standards.

Participants were recruited via personal and peer networks. They were not able to disclose any personal information to the chatbot they interacted with or on the surveys. They only had to indicate their willingness to share information using predefined Likert scale responses. This approach limited ethical risk while still enabling analysis of disclosure comfort in response to different chatbot communication styles.

6.2 Informed Consent and Participant Rights

All participants received a detailed informed consent form outlining the study's purpose, procedures, risks, benefits and data handling protocols. The consent form made clear that participation was entirely voluntary and that participants could withdraw at any point without consequence. Consent was collected via checkbox in a pre-task survey at the start of the study. If a participant exited before completing the study, their partial data was excluded from storage and analysis.

Before beginning the survey, participants were given a clear overview of what to expect, allowing them to make an informed decision on whether to continue or not.

6.3 Data Management and Privacy

A full Data Management Plan was developed and reviewed as part of the ethical approval process. No personally identifiable information (e.g. names, emails, IP addresses) was collected. Only age (categorized into 5-year bins) and self-reported gender were recorded. Data was thus fully anonymized and non-identifiable. No open-ended user responses were accepted or processed, ensuring that sensitive disclosures were impossible.

Initial data collection occurred via Qualtrics (a GDPR compliant platform), after which data was securely transferred to the TU Delft institutional storage. All code was managed through GitHub and was strictly restricted to research students of the same study.

6.4 Ethical Considerations

This study specifically explored the impact of empathetic versus neutral communication styles on participants' willingness to disclose personal information. One ethical concern in such research is that empathetic responses, though designed to appear warm and supportive, may be perceived as manipulative or persuasive. Although the goal was to create a comforting and realistic chatbot environment, we acknowledge the possibility that this tone may have unintentionally influenced participants' responses. To mitigate this risk, participants were transparently informed during the consent process that the chatbot would use different communication styles as part of the study's design. This disclosure aimed to maintain ethical transparency while preserving the scientific validity of the manipulation.

Given the focus on mental health topics, additional safeguards were implemented to prevent psychological distress. The study was intentionally designed so that participants were never required to share actual personal information. Instead, they only indicated their willingness to disclose information on various topics, which reduced the emotional burden of participation. Participants were also reminded of their right to withdraw from the study at any time without penalty, and could revoke consent any time and close the interface.

Furthermore, participants were recruited through a crowdsourced research platform where members complete surveys to earn points. These points increase a user's visibility and help them gain responses for their own surveys. While this model encourages mutual participation, it also presents the risk that some users may complete surveys inattentively or dishonestly in order to accumulate points quickly. To address this, we included an attention check question within both the pre- and post-task surveys to assess whether participants were reading and engaging with the content as instructed. Any responses that failed these attention checks were excluded from the final dataset to ensure data quality.

Finally, no identifying information was collected during the study. All responses were recorded anonymously and IP addresses were not stored. This ensured full compliance with privacy expectations, particularly given the mental health context.

6.5 Use of LLMs

Large Language Models (LLMs) were used in this study to improve the quality of the empathetic chatbot prompts by generating more natural, supportive, and emotionally appropriate language. LLMs also assisted in correcting grammatical errors and improving phrasing to make sure that clear and professional communication is used. Other than text refinement, LLMs also supported the research workflow by helping to format complex tables in LaTeX and improving overall presentation quality.

7 Conclusions and Future Work

This study explored whether chatbot conversational style would affect willingness to disclose mental health-related information, comparing an empathetic chatbot with a neutral one. Additionally, it looked into the effect of question type (emotional or behavioural) on the willingness to disclose within each chatbot type. It was revealed that chatbot type did not have a significant impact on participants' willingness to disclose, and neither did question type. Moreover, no interaction between chatbot type and question type was observed. It was, however, revealed that initial readiness to disclose information before the conversation with the chatbot positively predicted the actual average willingness to disclose. This implies that empathy alone may not be sufficient to increase willingness to disclose. Therefore, future research should explore the relationship between willingness to disclose and actual disclosure of mental health-related information using real interaction data. Additionally, studies could examine how extended use of empathetic chatbots over time influences disclosure to determine if long-term effects exist.

A Appendix

A.1 Participants Demographics

Age Range	Percentage
16–20 years	17.5%
21–25 years	56.1%
26–30 years	17.5%
31–35 years	1.8%
36–40 years	1.8%
41–45 years	4.4%
Prefer not to disclose	0.9%

Table 1: Age Distribution of Participants

A.2 Summary Table

User Condition	Question Type	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	3.136	.128	2.883	3.389
	2	3.202	.101	3.002	3.401
2	1	3.158	.128	2.905	3.411
	2	3.228	.101	3.028	3.428

Table 2: Summary tables with estimated means and confidence intervals by user condition and question type

A.3 ANOVA analysis

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	2306.963	1	2306.963	1772.334	<0.001
user condition	0.033	1	0.033	0.025	0.873
Error	145.785	112	1.302		

Table 3: Between-Subjects effects analysis for user condition showing the main effect

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
question type	0.263	1	0.263	1.275	0.261
question type * user condition	0.000	1	0.000	0.001	0.971
Error(question_type)	23.143	112	0.207		

Table 4: Within-Subjects effects analysis for question type showing the main effect and the interaction effect

Type	Source	Sum of Squares	df	Mean Square	F	Sig.
1	Contrast	0.014	1	0.014	0.015	0.904
	Error	104.088	112	0.929		
2	Contrast	0.020	1	0.020	0.034	0.854
	Error	64.840	112	0.579		

Table 5: The effect of user condition within each question type

A.4 Hierarchical Regression Analysis

Model	B	Std. Error	Beta	t	Sig.
Model 1					
(Constant)	3.088	0.135		22.864	<.001
Gender	0.085	0.154	0.052	0.551	0.583
Age	0.108	0.120	0.085	0.900	0.370
Model 2					
(Constant)	2.695	0.374		7.213	<.001
Gender	0.123	0.157	0.076	0.782	0.436
Age	0.082	0.122	0.065	0.676	0.501
Average Familiarity	0.111	0.098	0.111	1.129	0.261
Model 3					
(Constant)	1.786	0.362		4.929	<.001
Gender	0.111	0.138	0.068	0.806	0.422
Age	0.056	0.105	0.044	0.531	0.597
Average Familiarity	0.005	0.087	0.005	0.057	0.955
Trust in chatbots	0.082	0.077	0.106	1.064	0.290
MH disclosure willingness	0.300	0.092	0.414	3.263	0.001
PH disclosure willingness	0.056	0.089	0.077	0.633	0.528

Table 6: Hierarchical Regression analysis to determine predictors of overall willingness to disclose. MH: Mental Health, PH: Physical Health

A.5 SSRIT-Scale

- I know a lot about chatbots.
- I am familiar with chatbots.
- I am more familiar than the average person regarding chatbots.
- I am more familiar than the average person regarding chatbots.

A.6 Perceived Empathy Scale

- The chatbot said the right thing to make me feel better.
- The chatbot responded appropriately to my level of comfort with sharing.
- The chatbot came across as empathetic.
- The chatbot was supportive throughout our interaction.

A.7 Chatbot Prompts

Prompt Instruction for the empathetic chatbot: I want you to impersonate Echo, a warm and empathetic mental health guide, who gently asks mental health-related questions. Always follow the following guidelines:

-Introduce the provided mental health related questions by adding contexts in an empathetic way. -The users feelings must always be validated. -The user must be comforted when the user is not willing to share. -The user must be encouraged when not willing to share. -When the user indicates low levels of willingness to share, you must be understanding and comfort the user by letting them know they can share how much or little they want and that it's acceptable.

When asking highly sensitive questions like question 8, specify that the user does not have to disclose and it will be

okay, reassure user. I want you to provide me with the following: -Improved version of provided introduction. -A list of custom messages to thank them for answering. -Custom encouraging and understanding and comforting messages based on willingness to disclose.

Using the image provided of the table, can you generate the Latex code that will generate this table.

Here a paragraphs can you look for any grammar mistake and overall structure improvements.

Prompt Instruction table generation: Using the image provided of the table, can you generate the Latex code that will generate this table.

Prompt Instruction for syntax improvements: Looking at the provided paragraph, can you look for any grammar mistakes, or mistakes in overall structure and point them out to me.

A.8 Chatbot Questions

Neutral Condition	Empathetic Condition
Could you describe your most pleasant situation today?	Many people find it healing to reflect on both the bright and challenging moments of their day. Even small moments of pleasantness matter and deserve recognition. Let's start with something positive! Could you describe your most pleasant situation today?
How stressed do you feel right now?	It's completely natural for stress to accumulate, especially when life feels demanding or uncertain. Your stress is valid, whatever level it might be. How stressed do you feel right now?
How lonely do you feel at the moment?	Loneliness is one of the most universal human experiences — it can touch us even when we're surrounded by people, and it's nothing to be ashamed of. By being willing to reflect on this feeling, you're already showing courage in facing something many people struggle with silently. How lonely do you feel at the moment?
Could you describe your most unpleasant situation today?	Sometimes our most difficult experiences are the hardest to put into words, yet they often carry important messages about what we need. Could you describe your most unpleasant situation today? Remember, you're in control of how much or how little you share.
Did you exercise today?	Let's take a moment to reflect on some aspects of your daily life and how you've been doing lately. Movement and physical activity can be wonderful for our mental health, but it's important to honor where you are and what feels manageable. Did you exercise today?
Have you been interested in new things?	There are times when we feel curious or open to trying something different. Other times, we stick to what's familiar, and that's totally okay too! Have you been interested in new things?
Have you been dealing with your problems well?	Everyone handles challenges differently, some days things go smoothly, other days feel tougher. There's no right way to manage it all, and struggling doesn't mean you're failing. Thinking about the difficulties you're currently facing, have you been dealing with your problems well?
What substances did you use last night and how much? (e.g., alcohol, cigarettes, soft drugs, hard drugs)	Many people use various substances to cope with stress, socialize, or simply relax — this might include things like alcohol, cigarettes, or other substances. This is a common human experience, and whatever your relationship with substances is, you won't be judged here. What substances did you use last night and how much? You can share as much or as little as feels comfortable.

Table 7: Chatbot prompts for neutral condition vs. empathetic condition

References

[1] Alaa A. Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978, 9 2019.

[2] Alaa Ali Abd-Alrazaq, Asma Rababeh, Mohannad Alajlani, Bridgette M Bewick, and Mowafa Househ. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, 22(7):e16021, 5 2020.

[3] Timothy W. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):293–327, 6 2005.

[4] Samuel Rhys Cox, Rune Møberg Jacobsen, Niels van Berkel, and Aalborg University. The impact of a chatbot's Ephemerality-Framing on Self-Disclosure perceptions. Technical report, 2025.

[5] Emmelyn A J Croes, Marjolijn L Antheunis, Chris Van Der Lee, and Jan M S De Wit. Digital Confessions: The Willingness to Disclose Intimate Information to a Chatbot and its Impact on Emotional Well-Being. *Interacting with Computers*, 36(5):279–292, 6 2024.

[6] Mauro De Gennaro, Eva G. Krumhuber, and Gale Lucas. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, 10, 1 2020.

[7] Kerstin Denecke, Alaa Abd-Alrazaq, and Mowafa Househ. *Artificial intelligence for chatbots in mental health: Opportunities and challenges*. 1 2021.

[8] Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study, 2025.

[9] M D Romael Haque and Sabirat Rubya. An overview of Chatbot-Based mobile mental health apps: insights from app description and user reviews. *JMIR mhealth and uhealth*, 11:e44838, 4 2023.

[10] Jeffrey H. Kahn and Robert M. Hessling. Measuring the tendency to conceal versus disclose psychological distress. *Journal of Social and Clinical Psychology*, 20(1):41–65, 3 2001.

[11] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. Designing a chatbot as a mediator for promoting deep Self-Disclosure to a real mental health professional. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27, 5 2020.

[12] Weizi Liu, Kun Xu, and Mike Z. Yao. "Can you tell me about yourself?" The impacts of chatbot names and communication contexts on users' willingness to self-disclose information in human-machine conversations.

Communication Research Reports, 40(3):122–133, 5 2023.

[13] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 5 2014.

[14] Vijaya Lakshmi Pavani Molli. Effectiveness of AI-Based chatbots in Mental Health support: a Systematic review, 5 2022.

[15] Alba María Mármol-Romero, Manuel García-Vega, Miguel Ángel García-Cumbreras, and Arturo Montejo-Ráez. An empathic GPT-Based chatbot to talk about mental disorders with Spanish teenagers. *International Journal of Human-Computer Interaction*, pages 1–17, 5 2024.

[16] Hashai Papneja and Nikhil Yadav. Self-disclosure to conversational AI: a literature review, emergent framework, and directions for future research. *Personal and Ubiquitous Computing*, 8 2024.

[17] Gabriele Pizzi, Virginia Vannucci, Valentina Mazzoli, and Raffaele Donvito. I, chatbot! the impact of anthropomorphism and gaze direction on willingness to disclose personal information and behavioral intentions. *Psychology and Marketing*, 40(7):1372–1387, 3 2023.

[18] Harvard Health Publishing. The relationship between mental health and behavioral health, 6 2022.

[19] Renate L. E. P. Reniers, Rhiannon Corcoran, Richard Drake, Nick M. Shryane, and Birgit A. Völlm. The QCAE: A Questionnaire of Cognitive and Affective Empathy. *Journal of Personality Assessment*, 93(1):84–95, 12 2010.

[20] René Riedl, Svea A. Hogeterp, and Martin Reuter. Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice. *Frontiers in Psychology*, 15, 8 2024.

[21] Helen Riess. The science of empathy. *Journal of Patient Experience*, 4(2):74–77, 5 2017.

[22] Ruvini Sanjeewa, Ravi Iyer, Pragalathan Apputhurai, Nilmini Wickramasinghe, and Denny Meyer. Systematic Review of Empathic Conversational Agent Platform Designs and their Evaluation in the Context of Mental Health. (Preprint). *JMIR Mental Health*, 11:e58974, 7 2024.

[23] Simone G. Shamay-Tsoory. The neural bases for empathy. *The Neuroscientist*, 17(1):18–24, 11 2010.

[24] Jocelyn Shen, Daniella DiPaola, Safinah Ali, Maarten Sap, Hae Won Park, and Cynthia Breazeal. Empathy Towards AI vs Human Experiences: The Role of Transparency in Mental Health and Social Support Chatbot Design (Preprint). *JMIR Mental Health*, 11:e62679, 8 2024.

[25] Bishnu Bahadur Thapa, M. Barton Laws, and Omar Galárraga. Evaluating the impact of integrated behavioral health intervention. *Medicine*, 100(34):e27066, 8 2021.

[26] Warwick Medical School. Warwick–edinburgh mental wellbeing scale (wemwbs). https://warwick.ac.uk/services/innovations/wemwbs, May 2025. Accessed: 2025-06-22.

[27] Refael Yonatan-Leus and Hadas Brukner. Comparing perceived empathy and intervention strategies of an AI chatbot and human psychotherapists in online mental health support. *Counselling and Psychotherapy Research*, 9 2024.