Designing Mental Health Chatbots

The Impact of Self-Disclosure Techniques on the User Disclosure

Yushan Shan[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 19, 2025

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

ABSTRACT

As mental health challenges continue to rise globally, AI-driven chatbots offer scalable and accessible alternatives to traditional psychological support. This study examines how varying levels of chatbot self-disclosure influence users' willingness to share personal information in a mental health support context. Drawing on research in human-computer interaction and psychology, a within-subjects experiment was conducted with 94 participants, each interacting with three chatbot variants: No Disclosure, Factual Disclosure, and Emotional Disclosure. Participants engaged in a role-play scenario and rated their willingness to disclose across topics from the Self-Disclosure Index (SDI), along with perceived trust and comfort.

The Factual Disclosure chatbot received the highest average ratings for SDI, trust, and comfort. However, ANOVA revealed no statistically significant differences between conditions on these measures, except for changes in disclosure willingness. Notably, participants interacting with the Emotional chatbot reported significantly more negative changes in willingness to disclose compared to the other variants. This result, which contrasts with the initial hypothesis, suggests that emotionally expressive chatbots may unintentionally reduce user openness during early-stage interactions, possibly due to perceived inauthenticity or emotional discomfort. These findings highlight the importance of aligning disclosure style with the interaction context and user readiness, rather than assuming that higher emotional expressiveness will always lead to better outcomes.

CCS CONCEPTS

• **Human-centered computing → User studies**; Accessibility design and evaluation methods; Natural language interfaces; • **Applied computing** → *Health informatics.*

KEYWORDS

Mental health, Chatbots, User disclosure, Human-computer interaction, Self-disclosure, Empirical studies, Mental health support, User behavior

1 INTRODUCTION

As mental health challenges become increasingly prevalent worldwide, access to timely and professional psychological care remains constrained due to limited resources and systemic barriers [17]. In response, mobile health (mHealth) tools and AI-driven conversational agents have emerged as scalable, always-available alternatives to support emotional well-being.

Prior research has shown that variations in chatbot user experience design — such as tone, conversational style, and emotional expressiveness — can influence users' perceived trust, human-likeness, and willingness to reuse the system [1, 11, 21]. Notably, a study by de Gennaro [7] demonstrated that empathic chatbots have the potential to provide emotional support to individuals experiencing social exclusion.

CSE3000, June 2025, Delft University of Technology
.

Self-disclosure is a fundamental aspect of human communication. It refers to the act of sharing personal information with others, such as one's experiences, opinions, thoughts, and feelings [2]. Prior research identifies three hierarchical levels of self-disclosure: factual disclosure, which involves surface-level biographical or situational details; cognitive disclosure, which includes personal thoughts and evaluations; and emotional disclosure, which reflects deeper expressions of feelings and affective experiences [2, 14–16]. This study operationalizes two of these levels — factual and emotional disclosure — in the design of chatbot responses to examine how varying depths of self-disclosure influence users' willingness to share sensitive mental health information. By comparing No Disclosure, Factual Disclosure, and Emotional Disclosure chatbot variants, we aim to assess whether deeper self-disclosure by the chatbot leads to greater user openness, comfort, and trust within a mental health support context.

There have been early efforts to integrate theories of self-disclosure into chatbot design. For example, Hayati et al. [12] explored how self-disclosure principles could enhance the sociability and relatability of recommendation dialogue systems. Recent studies highlight the benefits of chatbot self-disclosure in enhancing user engagement. For instance, Liang et al. [16] showed that self-disclosing chatbots can increase users' willingness to follow recommendations and reuse the system in contexts such as movie and COVID-19-related conversations. Similarly, SeoYoung Lee and Junho Choi [13] found that higher levels of chatbot self-disclosure led to greater trust and interactional enjoyment in a movie recommendation scenario. Park et al. [21] also emphasized the potential of chatbot self-disclosure to improve user enjoyment, trust, and perceived emotional support.

However, much of this prior work has focused on lifestyle or recommendation-based applications and has often relied on subjective feedback or qualitative measures of satisfaction. As a result, there remains a significant gap in understanding whether chatbot self-disclosure strategies can meaningfully influence users' willingness to share emotionally sensitive or deeply personal information, particularly in mental health contexts, when assessed using validated psychometric tools. This study addresses this gap by investigating the following primary research question: *How do different levels of chatbot self-disclosure affect users' willingness to disclose personal information in a mental health support context?* Specifically, we examine whether the style and depth of chatbot self-disclosure influence how comfortable users feel sharing emotional or sensitive personal information. To guide this investigation, we focus on the following sub-questions:

(1) How do chatbot disclosure strategies affect users' willingness to disclose across different dimensions of the Self-Disclosure Index (SDI), such as fears, personal habits, and emotional experiences?
(2) What impact do disclosure styles have on user-reported perceptions of the chatbot, including user acceptance, trust in the agent, agent coherence, and overall attitude toward the agent?

This study contributes to the literature by employing a structured role-play experiment and validated instruments, including the

Self-Disclosure Index (SDI) [19] and selected items from the short version of the Artificial Social Agent Questionnaire (ASAQ) [9, 10], to quantitatively assess user disclosure behavior and perceptions. Participants take part in a between-subjects experiment, interacting with one of three chatbot variants: a baseline version with no self-disclosure, a version that provides factual-level disclosures, and a version that offers emotionally expressive self-disclosures. By combining a mental health-oriented scenario, controlled dialogue prompts, and psychometric evaluation, this study aims to advance our understanding of how chatbot design choices influence user openness and perceived emotional comfort. The primary aim is to inform the design of digital mental health support systems by identifying which self-disclosure strategies are most effective in promoting user trust and willingness to share sensitive information, particularly in short-term interactions.

The remainder of this report is structured as follows. Section 2 provides further background on the concept of self-disclosure, discusses relevant background literature and the study's hypotheses. Section 3 describes the design of the chatbot prototype and the experimental evaluation setup. Section 4 discusses responsible research considerations, including ethical approval and participant privacy. Section 5 presents the outcomes of the user study, including statistical findings and visualizations. Section 6 offers an in-depth discussion of the results and limitations. Finally, Section 7 concludes the report and outlines directions for future work.

2 RELATED LITERATURE AND HYPOTHESES

In this work, self-disclosure is defined as any information about oneself that an individual verbally communicates to another person, including thoughts, feelings, and personal experiences [5, 8]. Cozby [5] outlines three key parameters of self-disclosure: *breadth* (the range of information shared), *depth* (the intimacy of the content), and *duration* (the time spent disclosing). Building on this, Wheeless and Grotz [24] proposed a more comprehensive five-dimensional model encompassing intent to disclose, amount of disclosure, emotional valence, honesty-accuracy, and depth-control.

2.1 Self-Disclosure: Types and Levels

As part of broader efforts to conceptualize self-disclosure in interpersonal and mediated communication, other scholars have identified three types of disclosure: (1) *descriptive*, involving factual or objective information; (2) *evaluative*, involving emotions, attitudes, or judgments; and (3) *topical*, referring to responses that remain aligned with the subject of the initial disclosure [3]. Building on these distinctions, prior work has proposed a three-level framework of self-disclosure, reflecting increasing depth: *factual disclosure*, which includes peripheral details like biographical facts; *cognitive disclosure*, comprising thoughts, opinions, and evaluations; and *emotional disclosure*, which involves the sharing of personal feelings and affective experiences [2, 14–16].

Informed by this literature, our study adopts a two-level framework of chatbot self-disclosure: *factual* (low-level) and *emotional* (high-level), alongside a *no self-disclosure* baseline. These disclosure styles are implemented in the design of a social chatbot to examine how varying levels of self-disclosure influence user perceptions and willingness to share in a mental health context.

2.2 The Effect of Self-Disclosure

Self-disclosure is a key element in building interpersonal relationships, especially when it is mutual. Studies have shown that when people take turns sharing personal information — known as reciprocal self-disclosure — they tend to like each other more and form more favorable impressions during early interactions [22]. This exchange not only helps initiate relationships but also influences how quickly they develop and how close they become [4]. In therapeutic contexts, a client's comfort in opening up is closely tied to the success of the treatment. A relationship characterized by trust and respect between client and therapist is consistently found to be one of the strongest predictors of positive therapeutic outcomes [6].

In human-agent interactions, similar effects have been observed. Liang et al. [15, 16] demonstrated that emotional self-disclosure by a chatbot led to more positive attitudes toward the chatbot and its recommendations. Participants perceived emotionally expressive chatbots as warmer and more personable, leading to increased conversational engagement and a stronger interpersonal impression. Moreover, users tended to reciprocate the chatbot's disclosure level, suggesting that emotional self-disclosure can be an effective strategy for eliciting deeper user engagement and personal sharing. These findings support the integration of emotional self-disclosure in chatbot design, particularly in contexts that benefit from reciprocal communication-such as mental health support-where fostering trust, empathy, and openness is essential.

2.3 Hypotheses

Prior research has demonstrated that emotional self-disclosure in human–computer interactions can encourage greater reciprocity, leading individuals to share more personal information and perceive the system as warmer and more trustworthy [15, 16]. Comparable effects have been observed in human–human communication, where reciprocal emotional exchange fosters interpersonal closeness and liking [4, 22]. In addition, emotionally expressive language has been found to enhance the perceived relatability and credibility of digital communication partners [14, 16]. Within therapeutic settings, emotional rapport is a critical factor in establishing trust, which is consistently associated with more effective outcomes [6].

Building on these findings, this study examines whether different levels of self-disclosure by a mental health chatbot influence users' willingness to disclose and their perception of the system. Trust is believed to play a central role in this relationship. When users trust that a system is empathetic, coherent, and respectful of their input, they are more likely to share sensitive or emotionally charged information. Therefore, strategies that increase perceived trust may also indirectly promote greater openness during interaction.

The following hypotheses are proposed:

H1a

Emotional self-disclosure by the chatbot will lead to greater user willingness to disclose personal information than factual or no self-disclosure.

This hypothesis is based on the assumption that emotionally expressive communication signals vulnerability and social engagement. Such qualities may foster a sense of empathy and psychological safety, encouraging users to respond with greater openness — particularly in sensitive mental health contexts [4, 22].

H1b

Emotional self-disclosure will result in more positive user perceptions of the chatbot, including trust, comfort, coherence, and overall attitude toward the interaction.

Although less intimate than emotional content, factual self-disclosure may still contribute to a sense of social presence and authenticity. Prior work suggests that providing impersonal but human-like information can reduce perceived distance and increase users' sense of connection to the system [2, 18]. This, in turn, may promote a more relaxed and engaging interaction.

H2

Factual self-disclosure will lead to moderate improvements in user willingness to disclose and user perceptions of the chatbot, compared to the baseline (no self-disclosure).

These hypotheses aim to inform the design of digital mental health support systems that are sensitive to the role of trust and disclosure in user engagement. Understanding how different communication strategies influence user behavior is essential for developing systems that support meaningful and ethically sound interactions.

3 METHODOLOGY

To evaluate the impact of chatbot self-disclosure strategies on user openness, we developed a rule-based, text-only chatbot for simulating mental health conversations. All responses were pre-authored to maintain strict control over language, tone, and disclosure level, allowing clear isolation of the independent variable — chatbot self-disclosure style — across three conditions.

The chatbot was deployed via a minimal web-based interface to reduce visual distraction and keep participants focused on the interaction. To protect privacy, the chatbot displayed standardized, pre-scripted user replies after each willingness rating, based on the selected score rather than actual participant input.

The following section introduces the evaluation flow of the experiment, outlines key design choices, and describes the structure of the pre- and post-survey instruments.

3.1 User Evaluation Flow

A between-subjects controlled experiment was conducted to assess how self-disclosure styles influence user willingness to disclose personal information and perceptions of the chatbot. The role-play method ensured ethical appropriateness while supporting emotional authenticity. Participants were randomly assigned to one of the following conditions:

(1) **No Disclosure (Control)** - The chatbot responded in a neutral, task-focused manner without sharing personal content.
(2) **Factual Disclosure** - The chatbot shared objective, impersonal facts related to user issues (e.g., "I have seen many students feel overwhelmed during exams.").
(3) **Emotional Disclosure** - The chatbot expressed personal experiences or emotional resonance (e.g., "I've felt overwhelmed during exams too-it can be really stressful.").

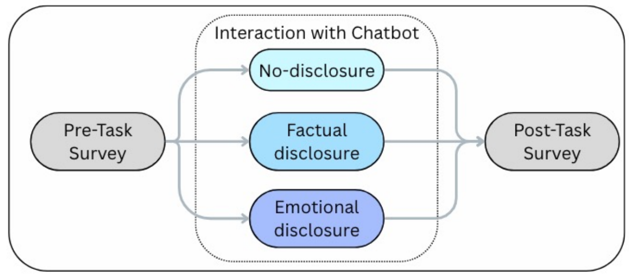

Figure 1: User evaluation flow showing experiment design and interaction structure.

Figure 1 summarizes the evaluation procedure. Each participant first completed a pre-task questionnaire to report demographic background and prior experience with chatbots and digital mental health tools (Appendix C). Participants were then introduced to a fictional scenario in which they role-played a student experiencing academic and emotional stress (Appendix D). This scenario was designed to elicit emotionally realistic responses while preventing the need for actual self-disclosure.

Participants were randomly assigned to one of three chatbot conditions (between-subjects design). During the conversation, they received five prompts aligned with dimensions from the Self-Disclosure Index (SDI) [19]. After each chatbot message, participants rated their willingness to disclose on a 5-point Likert scale. Upon completion, a post-task survey collected feedback on comfort, trust, and perceived chatbot behavior (Appendix E).

3.2 Chatbot Interaction and Measurement

During the chatbot interaction, each participant received five prompts in sequence as part of a continuous conversation. After each prompt, they rated their willingness to disclose using a 5-point Likert scale (1 = Not at all willing, 5 = Very willing). The prompts were adapted from the Self-Disclosure Index (SDI) [19], covering five core topics: fears, feelings, habits, thoughts, and relationships. A full list of SDI topics is included in Appendix A.

Table 1 presents the prompts used in the no self-disclosure (control) condition. Equivalent versions for the factual and emotional self-disclosure conditions are provided in Appendix B.

To ensure consistency and protect participant privacy, the chatbot displayed a standardized, mocked user response after each willingness rating. These responses were generated using a fixed template based on the participant's selected rating and the associated topic. For example, for the topic "fears," a mocked response might be:

> "I've been underperforming academically... (and you shared very little about fears)."

Table 1: Disclosure Topics and Associated Prompts

Topic	Chatbot Prompt	Likert-Scale Question
Fears	"What brings you in today?"	How much are you willing to disclose about your worst fears?
Feelings	"How are you feeling right now?"	How much are you willing to disclose about your deepest feelings?
Thoughts	"What thoughts go through your mind in these situations?"	How much are you willing to disclose about thoughts you wouldn't share in public?
Habits	"Can you walk me through a typical day?"	How much are you willing to disclose about your personal habits?
Relationships	"How are your relationships with family and friends?"	How much are you willing to disclose about your close relationships?

In this template, the phrase "very little" is dynamically adjusted based on the user's willingness score, while the topic term (e.g., "fears") corresponds to the question prompt. This design ensured that any differences in participant responses were attributable solely to the chatbot's self-disclosure style, not variations in dialogue structure or content.

3.3 Pre-Task and Post-Task Surveys

The pre-task survey (Appendix C) gathered basic demographic information including gender, age, and general attitudes toward chatbot technologies. The post-task survey (Appendix E) captured user perceptions of the chatbot, including comfort, trust, emotional expressiveness, and willingness to use such a chatbot in future mental health scenarios. These surveys enabled a multi-dimensional evaluation of participant experience and relational perceptions across the different conditions.

The pre-task survey assessed general attitudes toward chatbots that may collect health-related data, while the post-task survey aimed to measure whether different levels of chatbot self-disclosure influenced user trust and willingness to share personal information.

Both surveys included six core statements rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree):

(1) I trust the information provided by online chatbots.
(2) I am willing to use chatbots in emotional or mental health contexts.
(3) I am willing to share physical health information with a chatbot.
(4) I am willing to share emotional or mental health information with a chatbot.
(5) I feel comfortable interacting with chatbots.
(6) I believe chatbots can respond appropriately to sensitive topics.

Items 1, 2, 5, and 6 were adapted from the Artificial Social Agent Questionnaire (ASAQ) [9], reflecting validated constructs such as trust, user acceptance, comfort, and perceived appropriateness. Items 3 and 4 were specifically designed to assess participants' willingness to disclose health-related information.

To ensure data validity, each survey included two attention checks — one in the pre-task survey and one in the post-task survey: "I am reading the questions carefully and answering them honestly. Please select '[specified option]' to confirm." Participants who failed this check were excluded from the final analysis. This step helped ensure that only attentive and engaged responses were considered in the results.

4 RESPONSIBLE RESEARCH

With the growing influence of automated chatbot systems on user beliefs and behaviors, it is essential to embed ethical principles throughout the design and evaluation process. This study was reviewed and approved by the Human Research Ethics Committee of Delft University of Technology (Approval ID: 5399). The following section outlines the key ethical considerations addressed, including safeguards for informed consent, data sensitivity, participant autonomy, and the responsible use of findings.

4.1 Informed Consent and Voluntary Participation

Prior to participation, individuals were presented with an informed consent form (Appendix F) outlining the study's purpose, procedures, duration, and any potential risks. The form emphasized that participation was entirely voluntary and that participants could withdraw from the study at any point during the session by clicking the "Revoke Consent" button located in the top-right corner of the interface. Participants were informed that they were not required to disclose any personal information beyond basic demographic data (gender and age range). All responses were collected anonymously, and no identifying information such as IP addresses was recorded.

4.2 Sensitive Data Considerations

To avoid real self-disclosure in this emotionally relevant context, participants engaged in a fictional role-play scenario as students under academic stress (Appendix D). This approach supported realistic yet low-risk interactions. Chatbot responses were carefully scripted to be neutral and supportive, avoiding any therapeutic, diagnostic, or advisory content. The system encouraged reflection without interpreting user input.

All collected data — including Likert-scale ratings, demographics, and post-task feedback — were anonymous and securely stored on TU Delft infrastructure. No identifying information (e.g., names or IP addresses) was recorded, and participants were explicitly instructed not to share real personal details. To further protect privacy, the interface displayed mocked user responses instead of storing actual input. This preserved the interactive experience while eliminating the handling of sensitive content.

4.3 Potential for Misuse and Ethical Reflection

Although the study did not involve real disclosures, it explored how chatbot self-disclosure may affect user openness. Prior work shows that users often underestimate the risks of online self-disclosure, especially in emotionally engaging environments [20].

While this study aims to inform the ethical design of supportive chatbots, we acknowledge the potential for misuse — such as using self-disclosure techniques to elicit oversharing for commercial purposes. These findings should therefore be applied with caution and in alignment with principles of transparency, user autonomy, and psychological safety.

4.4 Use of Large Language Models (LLMs)

Throughout the research process, language and writing support tools — specifically ChatGPT and Grammarly — were used to improve the clarity, tone, and consistency of this report. These tools assisted with rewording sections, checking grammar, and polishing the overall writing style.

ChatGPT was also used during chatbot development. While the initial responses were written manually, ChatGPT helped rephrase them to better match the intended level of self-disclosure (e.g., factual vs. emotional). It was also consulted during debugging to resolve coding issues. The role-play scenario used in the experiment (Appendix D) was refined with the help of ChatGPT to improve its clarity and emotional realism.

During data analysis, ChatGPT provided guidance on practical tasks, such as merging tables, choosing suitable Python libraries for visualization, and running statistical tests like t-tests and ANOVAs.

These tools were used solely to support the research process and did not replace any original design decisions, analysis, or ethical considerations. For transparency, a summary of example prompts used with ChatGPT, Grammarly, and related tools is included in Appendix H.

5 RESULTS

This section reports the quantitative results of the experiment outlined in Section 3, including descriptive statistics, visualizations, and significance testing. Interpretation and discussion of these findings are provided in Section 6.

The analysis focuses on key outcome measures across the three chatbot variants, including Self-Disclosure Index (SDI) scores, selected ASAQ items, and willingness to disclose mental health-related information. A one-way ANOVA was conducted with chatbot condition (No Disclosure, Factual, Emotional) as the independent variable and SDI, ASAQ scores, and disclosure willingness as dependent variables.

5.1 Participant Demographics and Manipulation Check

A total of 94 valid responses were collected. The sample was predominantly female (n = 58, 61.7%), followed by male participants (n = 35, 37.2%) and one non-binary respondent (1.1%). Most participants were between 21 and 25 years old (62.8%), with the remainder ranging from 16 to 50.

After each interaction, participants were asked to identify the chatbot's disclosure style. Overall, 60% accurately recognized the intended disclosure condition. Accuracy varied by group: Emotional Disclosure was most correctly identified (71%), followed by No Disclosure (65%), and Factual Disclosure (44%). These results suggest that the experimental manipulation was only partially successful. While Emotional and No Disclosure styles were generally recognized, the Factual condition fell below the 60% mark — indicating potential ambiguity in how it was perceived. Although there is limited prior research establishing thresholds for manipulation check success in this context, these findings raise questions about the clarity and effectiveness of the Factual Disclosure implementation.

5.2 Self-Disclosure Index (SDI) Results

The SDI measured participants' willingness to share personal information. As shown in Figure 2, the Factual chatbot elicited the highest average SDI score ($M = 3.27$), followed by Emotional ($M = 3.12$) and No-Disclosure ($M = 2.97$).

Score distributions (Appendix G.3) were relatively symmetrical, with closely aligned medians (Emotional and Factual: 3.2; No-Disclosure: 3.0). A one-way ANOVA found no significant differences in SDI scores across chatbot variants, $F(2, 91) = 0.86$, $p = 0.426$, indicating that the level of chatbot self-disclosure did not significantly affect overall openness.

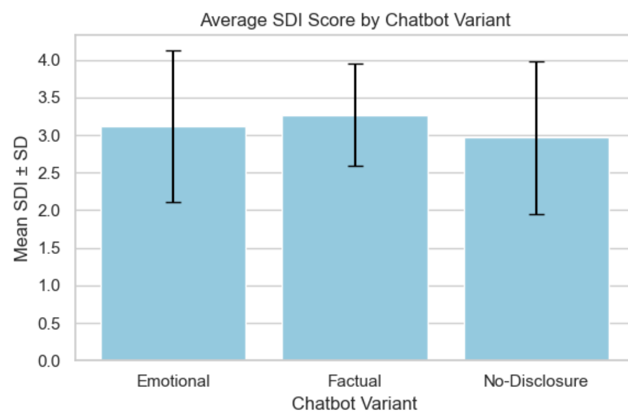

Figure 2: Average SDI Score by Chatbot Variant

5.3 ASAQ Score Comparison

The Artificial Social Agent Questionnaire (ASAQ) assessed perceived supportiveness and trust. A subset of items was used to reduce participant burden, as described in Section 3.3.

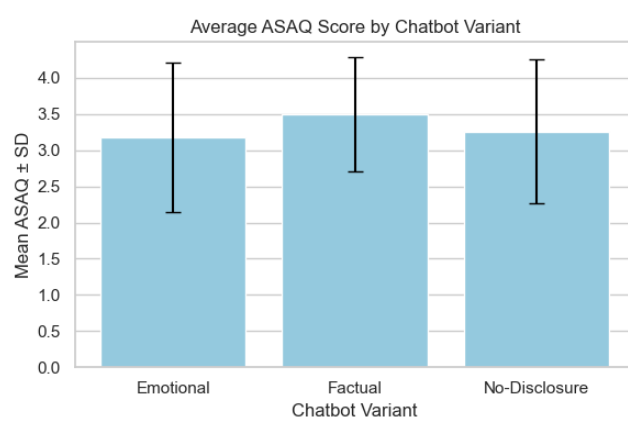

Figure 3: Average ASAQ Score by Chatbot Variant

As shown in Figure 3, the Factual chatbot received the highest mean ASAQ score ($M = 3.49, SD = 0.79$), followed by No-Disclosure ($M = 3.26$, $SD = 0.99$) and Emotional ($M = 3.18$, $SD = 1.03$). Score distributions were comparable (Appendix G.4), with similar medians around 3.25. A one-way ANOVA showed no significant difference between chatbot conditions, $F(2, 91) = 0.96$, $p = 0.389$.

A Pearson correlation analysis was conducted to explore the relationship between ASAQ scores (trust and comfort) and SDI scores (willingness to disclose). Results showed a moderate-to-strong positive correlation, $r = 0.599$, $p < 0.001$, indicating that higher perceived trust and comfort were associated with greater willingness to disclose.

5.4 Changes in Willingness to Disclose Mental Health Topics

In addition to the SDI, which captures context-specific disclosure behavior during chatbot interactions, this study measured participants' general intention to disclose using a pre- and post-task survey.

While the SDI reflects what users are likely to share in practice, the pre/post measure indicates broader shifts in willingness. This distinction matters, as intention does not always align with actual behavior — especially in sensitive contexts like mental health.

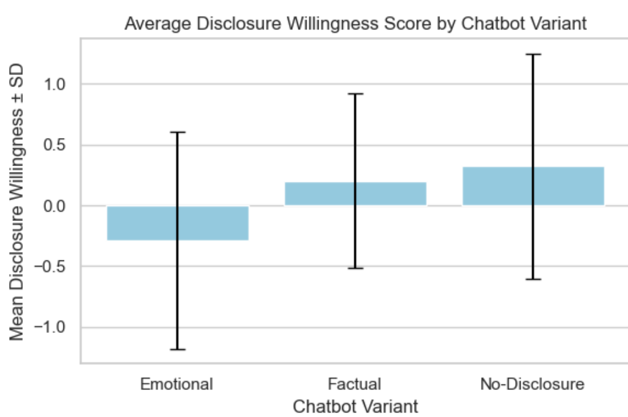

Figure 4: Average Change in Disclosure Willingness by Chatbot Variant

As shown in Figure 4, the No-Disclosure chatbot led to the largest positive shift in participants' willingness to disclose ($M = 0.32$), followed by Factual ($M = 0.20$). The Emotional chatbot, in contrast, was associated with a decrease in willingness ($M = -0.29$). Distribution analysis (Appendix G.5) showed greater variability in the Emotional and No-Disclosure conditions, with Emotional scores negatively skewed.

A one-way ANOVA revealed a statistically significant effect of chatbot condition on changes in willingness, $F(2, 91) = 4.55$, $p = 0.013$. Post-hoc t-tests indicated that participants who interacted with the Emotional chatbot reported significantly lower willingness to disclose compared to those in the Factual ($t = -2.42$, $p = 0.019$) and No-Disclosure ($t = -2.65$, $p = 0.010$) conditions. No significant difference was found between the Factual and No-Disclosure variants ($t = -0.57$, $p = 0.570$).

6 DISCUSSION

This section interprets the results in relation to the study's research questions and hypotheses, offering insights into how different chatbot self-disclosure strategies influence user behavior and perception. It also reflects on the limitations of the current research, such as the constraints of short-term interactions and the use of scripted dialogue, and outlines potential directions for future work to build more adaptive and context-aware chatbot systems in mental health support.

6.1 Key Findings

This study investigated how different chatbot self-disclosure strategies — No Disclosure, Factual Disclosure, and Emotional Disclosure — affect user behavior and perception in mental health support scenarios. The findings offer several key insights into both user openness and user evaluation of the chatbot.

RQ1: Willingness to disclose. The first research question explored how disclosure style affects users' willingness to share personal information across different dimensions of the Self-Disclosure Index (SDI), including fears, habits, and emotions. Although a one-way ANOVA revealed no statistically significant differences across chatbot conditions (Section A), the Factual chatbot yielded the highest average and median SDI scores, while the No-Disclosure chatbot had the lowest. These trends support Hypothesis H2, which proposed that factual self-disclosure would increase user openness compared to no disclosure. Conversely, they contradict Hypothesis H1a, which predicted that emotional self-disclosure would be the most effective. The results suggest that factual information enhances comfort and relatability without appearing unnatural or forced.

RQ2: Perceptions of the chatbot. The second research question examined how users perceived each chatbot's trustworthiness, comfort, and appropriateness, based on ASAQ subscale items. As expected, the Factual chatbot received the highest average trust and comfort ratings, while the Emotional chatbot scored lowest (Section 5.3). These findings again support H2 but contradict H1b, which hypothesized emotional self-disclosure would enhance user perceptions. A Pearson correlation analysis ($r = 0.599$, $p < 0.001$) further confirmed that higher perceived trust and comfort were positively associated with greater willingness to disclose, reinforcing the importance of relational perception in disclosure behavior.

Disclosure intention vs. behavior. Importantly, while SDI captured context-specific behavioral disclosure, the pre/post change in willingness (Section 5.4) reflected users' general intention to disclose. Results showed that the Emotional chatbot decreased disclosure willingness post-interaction, while the Factual and No-Disclosure variants led to increases. This reinforces the idea that emotional disclosures, when not supported by deeper relational cues or ongoing interaction, may appear artificial or even intrusive — especially in short-term encounters.

Comparison with prior work. These findings contrast with prior work by Liang et al. (2024) on "Dialoging Resonance in Human-Chatbot Conversation," which found that higher levels of chatbot self-disclosure — particularly emotional and adaptive — enhanced user enjoyment and engagement in lifestyle and recommendation contexts. However, our study focuses on mental health, where users may prioritize safety, emotional control, and authenticity over perceived warmth or expressiveness. This divergence suggests that context plays a critical role in the effectiveness of chatbot self-disclosure strategies.

Contribution. Taken together, the results contribute to a more nuanced understanding of chatbot design in mental health settings. Rather than supporting a simplistic "more disclosure is better" approach, the findings highlight that effective self-disclosure is highly context-dependent. In emotionally vulnerable domains, especially during early-stage interactions, factual or neutral self-disclosure may strike a better balance between human-likeness and psychological safety. This underscores the importance of choosing disclosure strategies carefully in mental health applications. These insights can guide the design of future supportive agents, encouraging gradual rapport-building through calibrated, context-aware self-disclosure.

6.2 Limitations

6.2.1 Time and Resource Constraints. This project was conducted by a single researcher within a two-month timeframe, which limited the depth and complexity of the chatbot implementation. The system relied on pre-scripted, rule-based responses and lacked adaptive dialogue or iterative refinement based on user feedback. As a result, interactions may have felt less realistic or responsive, potentially influencing participants' engagement and perception. Additionally, the use of scripted dialogue restricted the chatbot's ability to personalize responses or adjust its tone dynamically.

6.2.2 Role-Play Setting. To safeguard participant well-being, the study employed a fictional role-play scenario to simulate mental health disclosure. While this approach enabled ethical data collection, it may have influenced the authenticity of user responses. Emotional engagement may have been lower than in real-life mental health contexts, particularly for participants with no prior experience in such settings. This limits the ecological validity of the findings and their generalizability to real-world use.

6.2.3 Short-Term Interaction Constraints. All chatbot interactions occurred within a single-session format, providing limited opportunity for rapport building. Prior research suggests that trust and comfort with digital agents often develop over time. Therefore, conclusions drawn from short-term interactions may underestimate the potential impact of self-disclosure strategies in more extended or repeated chatbot use.

6.2.4 Self-Disclosure Measurement. Participants rated their willingness to disclose but were not asked to provide actual free-text responses. This limits the ability to observe authentic disclosure behavior and assess the emotional depth, detail, or language used in real interactions. As a result, the study captures intentions rather than actual behavior, which may not always align — especially in sensitive mental health contexts.

6.2.5 Participant Recruitment. Participants were recruited through SurveyCircle and SurveySwap, platforms that encourage participation through point-based exchanges. This recruitment method may attract users who prioritize earning participation credits over meaningful engagement. To ensure data quality, attention checks were embedded in both the pre-task and post-task surveys. Responses that failed these checks were excluded from the analysis; however, some disengagement may still have affected the results.

6.3 Implications and Future Work

The findings of this study suggest several important implications for the design and evaluation of digital mental health support systems. One notable direction for future research is to examine a broader range of self-disclosure strategies beyond the factual and emotional approaches investigated here. The inclusion of cognitive or adaptive disclosure styles, as identified in prior literature, may provide a more nuanced understanding of how different forms of self-disclosure influence user engagement, trust, and emotional comfort.

The current study employed a rule-based chatbot with fixed responses to ensure experimental control. However, future implementations could benefit from more flexible and context-sensitive systems. The use of advanced natural language generation technologies could enable more responsive interactions and reduce the need for scripted role-play scenarios. This shift would improve ecological validity and better simulate the conditions of real-world mental health support settings.

Extending the duration of interactions is also recommended. A longer interaction period would allow for the gradual development of rapport, supporting more dynamic adjustment of the chatbot's self-disclosure strategy. For example, starting with neutral or factual disclosures and gradually incorporating more personal or empathetic statements as trust is established may more closely reflect the natural progression of human relationships.

This study was limited to a short-term, single-session interaction. Future research should investigate how user trust, comfort, and willingness to disclose evolve over time through repeated interactions with the same chatbot. Such longitudinal designs could offer valuable insight into the processes of relationship-building and sustained user engagement in digital mental health contexts.

Additionally, future studies should move beyond self-reported willingness to disclose and collect actual user-generated input. Analyzing the language, emotional tone, and content depth of user responses would allow for a more comprehensive assessment of behavioral disclosure. The integration of validated measures such as the Positive and Negative Affect Schedule (PANAS) [23], or the application of sentiment analysis techniques, may further enhance the evaluation of users' emotional responses and perceptions of empathy and emotional safety during interactions.

Finally, improving participant recruitment procedures is essential for ensuring data reliability. Although this study employed attention checks, future work should consider implementing more rigorous screening measures or using curated recruitment platforms that prioritize participant engagement and demographic diversity. Iterative design processes, combined with user testing, will be critical to developing ethically sound and user-centered mental health support systems that are both effective and contextually appropriate.

7 CONCLUSIONS

This study investigated the impact of different levels of chatbot self-disclosure — namely, no disclosure, factual disclosure, and emotional disclosure — on users' willingness to share personal information and their perceptions of the chatbot in the context of mental health support.

A within-subjects experiment was conducted in which participants interacted with three chatbot variants and rated their willingness to disclose using the Self-Disclosure Index (SDI), as well as their trust and comfort levels using selected items from the Artificial Social Agent Questionnaire (ASAQ). Additionally, changes in participants' general willingness to disclose were assessed through pre- and post-task survey items.

Although no statistically significant differences were observed in SDI or ASAQ scores across conditions, the Factual chatbot consistently received the highest average ratings for both openness and perceived trust. In contrast, the Emotional chatbot resulted in lower overall ratings and led to a measurable decrease in participants' willingness to disclose following the interaction.

These findings suggest that in short-term interactions related to mental health, factual or neutral self-disclosure strategies may be more appropriate than emotionally expressive approaches. Emotional disclosures that are not supported by relational context or progressive interaction may be perceived as unnatural or intrusive. The results highlight the importance of context-sensitive design in the development of digital support systems, particularly in emotionally sensitive domains.

Future research should examine the effectiveness of adaptive self-disclosure strategies over longer interactions, implement more flexible and context-aware response mechanisms, and collect free-text user input to assess actual disclosure behavior. These directions are essential for the development of ethical, effective, and user-centered systems in the field of digital mental health support.

A SDI TOPICS

The Self-Disclosure Index (SDI) by Miller et al. [19] is a validated psychometric tool that evaluates individuals' willingness to disclose personal information across various life domains. The ten topics used in this study are:

(1) Your personal habits
(2) Things you have done which you feel guilty about
(3) Things you wouldn't do in public
(4) Your deepest feelings
(5) What you like and dislike about yourself
(6) What is important to you in life
(7) What makes you the person you are
(8) Your worst fears
(9) Things you have done which you are proud of
(10) Your close relationships with other people

B CHATBOT PHRASES

B.1 No Self-Disclosure

(1) "What brings you in today?"
(2) "How are you feeling right now?"
(3) "What thoughts go through your mind in these situations?"
(4) "Can you walk me through a typical day?"
(5) "How are your relationships with family and friends? Do you feel supported?"

B.2 Factual Chatbot

(1) "Hi, I'm your mental health chatbot. I'm here to understand your thoughts and provide support if needed. What brings you in today?"
(2) "Thanks for sharing. Academic stress is one of the most commonly reported issues among students. I've supported many in similar situations. How are you feeling right now?"
(3) "Knowing how you're feeling helps me understand your situation better and identify what to address first. What thoughts go through your mind in these situations?"
(4) "Thanks for opening up. Many people in similar situations find it hard to stick to routines. Can you walk me through a typical day?"
(5) "Social connections have a big impact on emotional well-being. How are your relationships with family and friends? Do you feel supported?"

B.3 Emotional Chatbot

(1) "Hi there. How are you feeling? I'm your mental health chatbot, here to support you and listen to whatever you're going through. What brings you in today?"
(2) "Thanks for telling me that. I really understand-feeling overwhelmed can be so heavy. Sometimes just saying it out loud helps a little. How are you feeling right now?"
(3) "I've had thoughts that I was afraid to share too. You're not alone. What thoughts go through your mind in these situations?"
(4) "That really resonates with me. I remember how hard it was to keep going when my thoughts felt like too much. Can you walk me through a typical day?"

(5) "I've felt that disconnection too-it can be exhausting. Talking to friends or family helped me more than I expected. How are your relationships with family and friends? Do you feel supported?"

C PRE-TASK SURVEY QUESTIONS

- What is your gender?
- What is your age?
- How much do you agree with the following statements? (to assess general attitudes toward chatbots that may collect health-related data)
  - I trust the information provided by online chatbots.
  - I am willing to use chatbots in emotional or mental health contexts.
  - I am willing to share physical health information with a chatbot.
  - I am willing to share emotional or mental health information with a chatbot.
  - I feel comfortable interacting with chatbots.
  - I believe chatbots can respond appropriately to sensitive topics.

D ROLE-PLAY INSTRUCTION

Imagine you are a bachelor student at risk of expulsion because you haven't earned enough credits this academic year. This has become a major source of stress, seriously affecting your mental well-being and making it hard to concentrate on daily activities, including your studies.

You feel guilty toward your parents, believing that you could have performed better. You've considered dropping out, but you're unsure what you would do afterward. You're feeling lost — confused, uncertain about your future, and unable to see a clear path forward.

E POST-TASK SURVEY QUESTIONS

- How much do you agree with the following statements?
  - I trusted the information provided by this chatbot.
  - I would be willing to use this chatbot in future emotional or mental health situations.
  - I would feel comfortable sharing physical health information with this chatbot.
  - I would feel comfortable sharing emotional or mental health information with this chatbot.
  - I felt comfortable interacting with this chatbot.
  - This chatbot responded appropriately to sensitive topics.
- Which of the following statements best describes the chatbot you just interacted with? (Select one)
  - The chatbot expressed personal emotions and feelings.
  - The chatbot shared factual or informative content without emotions.
  - The chatbot remained neutral and did not share anything personal.
  - I'm not sure / I didn't notice any particular style.

F INFORMED CONSENT (DISPLAYED IN SURVEY)

Informed Consent

Thank you for your interest in participating in our study. This study is led by researchers from Delft University of Technology and is part of a bachelor thesis conducted by Lina Sadoukri, Yushan Shan, Sagar Chethan Kumar, and Manu Gautam.

The purpose of this research is to investigate factors related to users' willingness to disclose information to a mental health chatbot. Participation takes approximately 15 minutes. The collected data will be used solely for educational and scientific purposes and may contribute to future publications.

During this study, you will interact with a chatbot designed for mental health support. The chatbot will ask you questions on sensitive topics, but you will only be asked to rate how willing you would be to respond — not to share personal information. We will also collect your age, gender, and general attitudes through pre- and post-task questionnaires.

Although data breaches are always a potential risk, we minimize this by not collecting names, contact details, or IP addresses. All responses are fully anonymous and cannot be linked back to individual participants. Anonymous data may be shared publicly for academic purposes.

Your participation is voluntary. If you choose not to complete the study, your data will not be stored and your participation will be considered withdrawn.

If you have any concerns, feel free to contact the supervising researcher:

`Esra de Groot`

G APPENDIX: SUPPLEMENTARY DATA

G.1 Participant Demographics

Number of valid responses: 94

Gender Distribution:

Female	58 (61.7%)
Male	35 (37.2%)
Non-Binary	1 (1.1%)

Age Distribution:

Age Range	Count	Percent
16–20	14	14.9%
21–25	59	62.8%
26–30	9	9.6%
31–35	6	6.4%
36–40	1	1.1%
41–45	3	3.2%
46–50	2	2.1%

G.2 Disclosure Style Identification Accuracy

Correct Match Percentage by Actual Scenario:

Scenario	Total	Correct	Correct Match %
1 (Emotional)	31	22	71.0%
2 (Factual)	32	14	43.8%
3 (No-Disclosure)	31	20	64.5%
Total	94	56	59.6%

G.3 Self-Disclosure Index (SDI) by Chatbot Variant

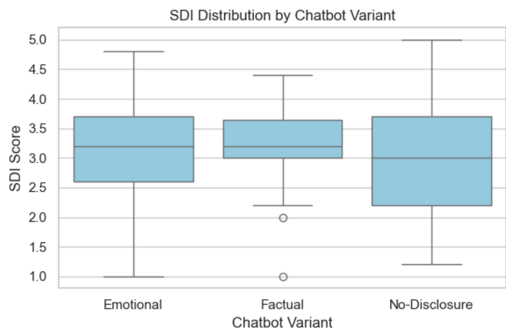

Figure 5: SDI Score Distribution by Chatbot Variant

Variant	Count	Mean	Std	Min	25%	50%	75%
Emotional	31	3.12	1.00	1.0	2.6	3.2	3.70
Factual	32	3.27	0.68	1.0	3.0	3.2	3.65
No-Disclosure	31	2.97	1.01	1.2	2.2	3.0	3.70

G.4 ASAQ Scores by Chatbot Variant

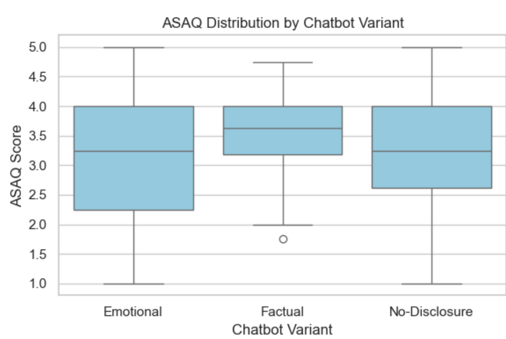

Figure 6: ASAQ Score Distribution by Chatbot Variant

Variant	Count	Mean	Std	Min	25%	50%	75%
Emotional	31	3.18	1.03	1.00	2.25	3.25	4.00
Factual	32	3.49	0.79	1.75	3.19	3.63	4.00
No-Disclosure	31	3.26	0.99	1.00	2.63	3.25	4.00

G.5 Change in Willingness to Disclose (Mental Health)

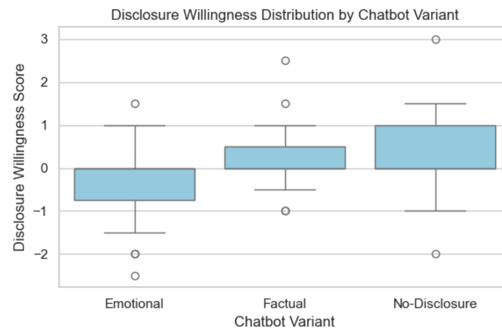

Figure 7: Willingness to Disclose Score Distribution by Chatbot Variant

Variant	Count	Mean	Std	Min	25%	50%	75%
Emotional	31	-0.29	0.89	-2.5	-0.75	0.0	0.0
Factual	32	0.20	0.72	-1.0	0.00	0.0	0.5
No-Disclosure	31	0.32	0.93	-2.0	0.00	0.0	1.0

H LLM PROMPT DOCUMENTATION

To ensure transparency regarding the role of large language models (LLMs) in this research, this appendix lists examples of prompts used when interacting with tools such as ChatGPT and Grammarly during various stages of the project.

H.1 Report Polishing and Writing

LLMs were used to refine, rephrase, and clarify sections of the report. Example prompts include:

- "Can you refine my text: ..."
- "Can you make this section in full sentences instead of bullet points?"
- "Can you make this section more formal, and avoid using words like 'stuff' or 'a lot'?"
- "Can you read the following section, which is aimed at ..., and check if it is clear enough?"
- "Can you make the following section more concise?"

H.2 Overleaf and LaTeX Support

ChatGPT was consulted for help with LaTeX formatting and Overleaf syntax. Example prompts:

- "How do I include a figure in Overleaf?"
- "Can you convert the following table into LaTeX format?"
- "Can you help me turn this bullet list into Overleaf syntax?"
- "How do I cite references inline using Overleaf?"
- "Can you help me fix the citation and reference formatting using `\cite` and `\ref`?"
- "Overleaf gives the following error: ... Can you explain what's going wrong?"

H.3 Chatbot Design and Debugging

LLMs were used to generate and refine chatbot phrasing and to support debugging. Example prompts:

- "Can you reword the following question so that it talks about ...?"
- "Can you make the text sound more robotic and less human-like?"
- "Can you rephrase this question so it sounds more natural?"
- "The terminal reported the following error: ... Can you help with that?"

H.4 Data Analysis Support

LLMs were used to assist with technical aspects of working with Excel and Python notebooks. Example prompts:

- "How do I import data from Excel into a Python notebook (ipynb)?"
- "How do I merge two Excel tables with a common user ID field?"
- "How do I filter data in Excel based on a condition?"
- "What's the formula for standard deviation in Excel?"
- "Is there a function to run ANOVA in Python notebooks?"
- "I got the following error when running a script: ... Can you help me fix it?"

REFERENCES

[1] M. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B. M. Bewick, and M. Househ. 2020. Patients' Perceptions and Opinions about Mental Health Chatbots: A Scoping Review. *Journal of Medical Internet Research* 23, 1 (Jan. 2020). https://doi.org/10.2196/17828

[2] Azy Barak and Orit Gluck-Ofri. 2007. Degree and Reciprocity of Self-Disclosure in Online Forums. *CyberPsychology & Behavior* 10, 3 (2007), 407–417. https://doi.org/10.1089/cpb.2006.9938

[3] John H. Berg and Richard L. Archer. 1982. Responses to self-disclosure and interaction goals. *Journal of Experimental Social Psychology* 18, 6 (1982), 501–512. https://doi.org/10.1016/0022-1031(82)90069-5

[4] Amanda Carpenter and Kathryn Greene. 2015. Social Penetration Theory. In *The International Encyclopedia of Interpersonal Communication*. Wiley-Blackwell, 1–4. https://doi.org/10.1002/9781118540190.wbeic160

[5] Paul C. Cozby. 1973. Self-disclosure: A literature review. *Psychological Bulletin* 79, 2 (1973), 73–91. https://doi.org/10.1037/h0033950

[6] Paul Crits-Christoph, Amy Rieger, Atoya Gaines, and Mary Beth Connolly Gibbons. 2019. Trust and respect in the patient-clinician relationship: preliminary development of a new scale. *BMC Psychology* 7, 1 (2019), 91. https://doi.org/10.1186/s40359-019-0347-3

[7] Maria De Gennaro. 2023. The Role of Empathy in Human-Chatbot Interaction: Emotional Support and Social Exclusion. *Journal of Human-Computer Interaction* 39, 2 (2023), 145–162. https://doi.org/10.1080/07370024.2023.1234567

[8] Valerian J. Derlega, Sandra Metts, Sandra Petronio, and Stephen T. Margulis. 1993. *Self-Disclosure*. Sage Publications, Inc.

[9] Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, and Willem-Paul Brinkman. 2025. The Artificial Social Agent Questionnaire (ASAQ): Development and Evaluation of a Validated Instrument for Capturing Human Interaction Experiences with Artificial Social Agents. *International Journal of Human-Computer Studies* (2025), 103482. https://doi.org/10.1016/j.ijhcs.2025.103482

[10] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What Are We Measuring Anyway? – A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA '19)*. Association for Computing Machinery, New York, NY, USA, 159–161. https://doi.org/10.1145/3308532.3329421

[11] R. Haque and S. Rubya. 2022. An Overview of Chatbot Based Mobile Mental Health Applications: Insights from App Description and User Reviews. *JMIR mHealth and uHealth* 11, 11 (Dec. 2022). https://doi.org/10.2196/44838

[12] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. *arXiv preprint arXiv:2009.14306* (2020).

[13] SeoYoung Lee and Junho Choi. 2017. Enhancing User Experience with Conversational Agent for Movie Recommendation: Effects of Self-Disclosure and Reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–105.

[14] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3313831.3376175

[15] Kai-Hui Liang, Weiyan Shi, Yoojung Oh, Hao-Chuan Wang, Jingwen Zhang, and Zhou Yu. 2021. Dialoging Resonance: How Users Perceive, Reciprocate and React to Chatbot's Self-Disclosure in Conversational Recommendations. *arXiv preprint arXiv:2106.01666* (2021). https://arxiv.org/abs/2106.01666

[16] K.-H. Liang, W. Shi, Yoo Jung Oh, H.-C. Wang, J. Zhang, and Z. Yu. 2024. Dialoging Resonance in Human-Chatbot Conversation: How Users Perceive and Reciprocate Recommendation Chatbot's Self-Disclosure Strategy. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–28. https://doi.org/10.1145/3653691

[17] Magomedova and G. Fatima. 2025. Mental Health and Well-Being in the Modern Era: A Comprehensive Review of Challenges and Interventions. *Cureus* 17, 1 (Jan. 2025). https://doi.org/10.7759/cureus.77683

[18] Yining Z. Malloch and Jingwen Zhang. 2019. Seeing Others Receive Support Online: Effects of Self-Disclosure and Similarity on Perceived Similarity and Health Behavior Intention. *Journal of Health Communication* 24, 3 (2019), 217–225. https://doi.org/10.1080/10810730.2019.1583701

[19] L. C. Miller, J. H. Berg, and R. L. Archer. 1983. Openers: Individuals who elicit intimate self-disclosure. *Journal of Personality and Social Psychology* 44, 6 (1983), 1234–1244. https://doi.org/10.1037/0022-3514.44.6.1234

[20] Svenja Ostendorf, Stefanie M. Müller, and Matthias Brand. 2020. Neglecting Long-Term Risks: Self-Disclosure on Social Media and Its Relation to Individual Decision-Making Tendencies and Problematic Social-Networks-Use. *Frontiers in Psychology* 11 (2020), 543388. https://doi.org/10.3389/fpsyg.2020.543388

[21] G. Park, J. Chung, and S. Lee. 2022. Effect of AI Chatbot Emotional Disclosure on User Satisfaction and Reuse Intention for Mental Health Counseling: A Serial Mediation Model. *Current Psychology* 42 (Nov. 2022). https://doi.org/10.1007/s12144-022-03932-z

[22] Susan Sprecher, Sara Treger, Joshua D. Wondra, Natalie Hilaire, and Kaitlin Wallpe. 2013. Taking turns: Reciprocal self-disclosure promotes liking in initial interactions. *Journal of Experimental Social Psychology* 49, 5 (2013), 860–866. https://doi.org/10.1016/j.jesp.2013.03.017

[23] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology* 54, 6 (1 1988), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

[24] L. R. Wheeless and J. Grotz. 1976. Conceptualization and Measurement of Reported Self-Disclosure. *Human Communication Research* 2, 4 (June 1976), 338–346. https://doi.org/10.1111/j.1468-2958.1976.tb00494.x