Talking Like a Human: How Conversational Anthropomorphism Affects Self-Disclosure to Mental Health Chatbots

An Experimental Study on Human-like Chatbot Design and Question Sensitivity in Mental Health Contexts

Sagar Chethan Kumar[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: Sagar Chethan Kumar
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Esra de Groot, Myrthe Tielman

Abstract

AI-powered mental health chatbots offer scalable and accessible support, but their effectiveness hinges on users' willingness to self-disclose—an outcome shaped by chatbot communication style and the sensitivity of the topic. While prior work has explored empathy and rapport, the role of conversational anthropomorphism remains underexamined, particularly in relation to question sensitivity as a potential moderator. This study addresses that gap through a mixed-design experiment ($n = 30$) in which participants interacted with either an anthropomorphic or a neutral chatbot and rated their willingness to respond to questions varying in sensitivity. Although no effects reached statistical significance, descriptive trends suggest that anthropomorphic cues—such as informal tone, emojis, and adaptive responses—may increase willingness to disclose, while higher question sensitivity slightly reduces it. No significant interaction effect was found, but anthropomorphic language appeared to promote disclosure regardless of sensitivity level. These findings offer tentative support for the use of calibrated human-like design in mental health chatbots. Future work should incorporate open-ended interactions, behavioral measures, and longitudinal designs to better capture disclosure dynamics and trust formation.

1 Introduction

In a rapidly evolving world, many people turn to sources of emotional support and psychological stability to cope with everyday stressors. Traditionally, this support has come from therapists, friends, and other in-person medical interventions [1]. However, with the rise of large language models (LLMs) and the growing adoption of artificial intelligence (AI) systems, chatbots and other applications have emerged as promising tools to bridge gaps in mental healthcare accessibility [18]. Unlike traditional interventions, these systems are available on demand and offer a non-judgmental space for users to share their thoughts and experiences [1, 18].

Mental health chatbots often serve as companions, assist in screening, provide resources, or simulate therapeutic roles [1, 42]. These functions depend on users sharing personal, sensitive information—*self-disclosure* [13]. Self-disclosure enables chatbots to respond appropriately, personalize support, and tailor interventions to users' needs [46]. Without it, their capacity to assess mental states, offer tailored guidance, or foster rapport is limited [30]. Promoting self-disclosure is therefore a central design challenge in chatbots [46].

One conversational factor that influences this process is *conversational anthropomorphism*: the attribution of human-like traits to non-human agents [60, 26, 15]. In domains such as e-commerce and academic services, anthropomorphic cues like informal language, humor, or human-like naming have been shown to foster trust and increase users' willingness to disclose information [42, 46]. However, findings are mixed: in more formal or sensitive contexts, such cues may have the opposite effect, potentially reducing disclosure due to heightened concerns over social desirability or judgment [43, 53].

The *sensitivity* of the topics being discussed plays a critical yet often overlooked role in shaping users' willingness to self-disclose. Prior research suggests that anthropomorphic features may promote disclosure when questions are perceived as low in sensitivity (e.g., shopping preferences or academic background) [3, 51], but may suppress it when questions involve sensitive topics such as finances or substance use [43, 53]. In these contexts, users may value traits like objectivity and competence over empathy or social presence, perceiving more human-like agents as judgmental or socially evaluative [30, 46]. As such, question sensitivity can act as a moderating factor in how anthropomorphic cues affect disclosure intent.

More broadly, self-disclosure to conversational AI has been linked to a range of positive outcomes, including greater user satisfaction [31], stronger perceived relationships with the system [54], and improved emotional well-being [20]. To understand what drives this process, Papneja and Yadav [46] outline five categories of influencing factors: interface modality, conversational elements, user characteristics, mediating mechanisms, and contextual variables. While all these dimensions are important, conversational factors, particularly those related to anthropomorphic design, have received relatively limited attention in the context of mental health support [46, 30].

This study aims to fill that gap by investigating how conversational anthropomorphism and question sensitivity jointly shape users' willingness to self-disclose in interactions with mental health chatbots. While previous research in mental contexts has explored individual elements such as visual agent embodiment [29], reciprocity [33], empathy [14], and rapport-building [30], the intersection of conversational anthropomorphism and question sensitivity remains underexplored. To that end, we pose the following research question and sub-questions to guide our exploration:

Main Research Question

How do conversational anthropomorphism and question sensitivity influence self-disclosure to AI-powered mental health chatbots?

RQ1 Does conversational anthropomorphism increase willingness to self-disclose?

RQ2 Does the sensitivity of disclosure-intent questions influence willingness to self-disclose?

RQ3 Is there an interaction effect between sensitivity of disclosure-intent questions and conversational anthropomorphism on willingness to self-disclose?

Our main contributions are as follows:

- We investigate how conversational anthropomorphism influences willingness to disclose personal information.

- We examine the role of question sensitivity in shaping disclosure behavior.
- We analyze how anthropomorphic cues and question sensitivity interact to affect self-disclosure in AI-mediated mental health conversations.
- We discuss design implications for AI chatbots aimed at fostering user self-disclosure in mental health contexts.

The remainder of this paper is structured as follows. Section 2 reviews related work on conversational AI, with a focus on self-disclosure and anthropomorphism, and introduces our hypotheses. Section 3 describes the study design and methodology. Section 4 presents the results, followed by a discussion of their implications and limitations in Section 5. Section 6 addresses potential risks and our mitigation strategies. Finally, Section 7 concludes the paper with a summary of findings and suggestions for future research.

2 Background and Hypotheses

In this section we provide an overview of existing literature concerning conversational AI, mental health chatbots and self-disclosure. In Section 2.1, we discuss prior work investigating self-disclosure to conversational AI. In Section 2.2, we examine AI-powered mental health chatbots and their use within the domain. In Section 2.3, we explore anthropomorphism in conversational AI systems. Finally, in Section 2.4 we present our hypotheses and the rationale for each.

2.1 Self-disclosure to Conversational AI

Self-disclosure, the act of revealing personal or sensitive information, has long been associated with building mutual trust in human communication [4]. In the context of conversational AI and agents (CAs), self-disclosure has been shown to influence user satisfaction [31], improve emotional outcomes [20], enhance perceived intimacy and enjoyment [33, 46], and strengthen user–AI relationships [54]. Many of these benefits offered by CAs depend on the user's willingness to share personal information. For instance, in healthcare settings, higher levels of self-disclosure to chatbots have been linked to more therapeutic experiences and improved paths to recovery [2, 1].

A comprehensive survey by Papneja and Yadav [46] identified five key categories that influence self-disclosure to CA: interface modality, conversational factors, user characteristics, mediating mechanisms, and contextual factors.

Within interface modality, most evidence indicates that users tend to disclose more to CAs than to humans [47, 48, 61, 36, 6, 38, 55, 57]. Moreover, users generally self-disclose more when interacting via voice rather than text [62, 46].

User characteristics—such as gender, privacy concerns, self-efficacy, and personality—also play a significant role and are relatively well understood [24, 55, 63, 52, 46]. Mediating mechanisms have been explored in specific contexts, though many studies still lack explicit operationalization or control of these variables [46].

Conversational factors have also been widely studied across domains. For example, chatbots that self-disclose often foster greater self-disclosure from users [39, 25, 46]. Similarly, other conversational elements positively impact disclosure levels, including the sequence and structure of questions [39], personality mirroring [17], and rapport-building techniques [30].

However, the role of conversational anthropomorphism remains inconclusive. While some studies report positive effects on self-disclosure [3, 51], others suggest negative [53, 43] or inconclusive results [58]. Alarmingly, within the space of mental health there remains a dearth of explorations of *conversational* anthropomorphism with AI-powered chatbots, with most studies focusing on CA (visual) embodiment [29], rule-based rapport building [29], comparisons to physicians directly [32], or reciprocity [33].

Therefore, to address this gap within mental health specifically, our study investigates how conversational anthropomorphism influences users' willingness to self-disclose to a CA.

2.2 AI-Powered Mental Health Chatbots

AI-powered chatbots have gained significant traction in mental health contexts, offering a scalable and accessible alternative to traditional interventions [9]. Unlike face-to-face interactions with clinicians, these systems have been shown to increase user self-disclosure [46] and reach populations hesitant to seek help due to stigma [1, 40, 2].

These chatbots have evolved from simple rule-based architectures to more advanced systems powered by LLMs and natural language processing (NLP) techniques [9]. State-of-the-art LLMs such as OpenAI's GPT-4 and Meta's `LLaMA` series have demonstrated strong performance in various medical tasks [56, 44, 45]. However, their specific application and effectiveness in mental health scenarios remain relatively underexplored [59, 22].

Much of the existing work in the mental health domain relies on older versions of popular platforms like Google DialogFlow[1], primarily chosen for their ease of setup [32, 33, 29, 30, 46]. However, most of these systems did not originally incorporate, and therefore could not explore, the substantial improvements in language understanding, reasoning, and generation brought by the latest LLMs [32, 30].

Despite these limitations, previous work notes that these AI-powered mental health chatbots have been consistently rated as useful, easy to use, responsive, and understandable [2]. Prior studies suggest that chatbot self-disclosure and embodied CAs can further promote user self-disclosure [32, 29], though these findings are largely based on outdated models [9].

To address this gap, we employ a state-of-the-art LLM, `LLaMA 4`[2], to power our chatbot and investigate its impact on users' willingness to self-disclose.

[1] https://dialogflow.cloud.google.com/

[2] https://www.llama.com/models/llama-4/

2.3 Anthropomorphism in Conversational AI

Anthropomorphism refers to the attribution of human characteristics, motivations, or emotions to non-human entities such as chatbots [26, 15]. Conversational anthropomorphism specifically involves human-like communication behaviors, including the use of humor, small talk, politeness, informal tone, emojis, and communication delays during interaction [46]. These anthropomorphic features influence both the type of content presented to users and the manner in which it is conveyed, distinguishing them from non-anthropomorphic systems [3, 51, 43, 53, 58, 46].

Research on conversational anthropomorphism and user self-disclosure shows mixed results [46]. Two studies found that anthropomorphic features (e.g., humor, informal language, message interactivity) increased disclosure in low-sensitivity contexts like shopping and academic services [3, 51]. In contrast, two others reported reduced disclosure in sensitive contexts such as alcohol use and personal finance, where socially responsive or emotionally expressive CAs triggered concerns about judgment or evaluation [43, 53]. A fifth study on course evaluations found no significant effect [58].

Taken together, these findings suggest that topic sensitivity moderates the effect of anthropomorphic features: they can enhance disclosure in casual contexts but may backfire when the topic is personal or high-stakes.

In the domain of mental health, prior work has examined anthropomorphism more generally [29, 34], with some attention to conversational aspects. For instance, Lee et al. [30] found that rapport-building techniques, such as empathetic language and small talk, promoted self-disclosure. Likewise, Lee et al. [32] explored the role of reciprocity through adaptive conversational styles, including chatbot self-disclosure, which also enhanced user disclosure.

Despite growing interest, the impact of conversational anthropomorphism on self-disclosure in AI-powered mental health chatbots, especially in relation to question sensitivity, remains underexplored. This study addresses that gap by examining how anthropomorphic design influences users' willingness to self-disclose, across varying levels of question sensitivity, towards a CA.

2.4 Hypotheses

Based on prior literature and the identified research gaps, we propose the following hypotheses for our previously identified RQs (**RQ1**, **RQ2**, **RQ3**).

H1. Conversational anthropomorphism will increase users' willingness to self-disclose.

H1 is supported by research suggesting that human-like communication features (e.g., small talk, politeness, informal tone) can enhance intimacy, trust, and disclosure in CA interactions [46, 3, 30]. In mental health settings, rapport-building techniques and empathetic language (both forms of general anthropomorphism) have been linked to increased self-disclosure [32].

H2. Users will be less willing to self-disclose as the sensitivity of questions increases.

Prior research suggests that the sensitivity of questions significantly shapes user responses. Sensitive topics, such as those related to finances, substance use, are often associated with increased psychological discomfort and fear of evaluation, which can inhibit self-disclosure [4, 52]. Studies have shown that when CAs ask highly sensitive questions, users are more likely to withhold truthful or complete responses, particularly when the interaction feels socially evaluative or emotionally charged [46, 43]. Conversely, lower-sensitivity questions are perceived as less risky and are more likely to elicit open and honest responses [63, 3].

H3. Conversational anthropomorphism will increase willingness to self-disclose for low-sensitivity questions but decrease willingness to self-disclose for high-sensitivity questions.

Prior research highlights the moderating role of topic sensitivity in shaping how anthropomorphic features influence user disclosure. In low-sensitivity contexts, such as academic advising or shopping, human-like traits (e.g., humor, small talk) have been shown to foster openness [3, 51]. In contrast, in high-sensitivity domains, such as alcohol use or personal finance, these features can evoke discomfort or concerns about judgment, thereby reducing disclosure [43, 53]. Mental health lies at the intersection of these extremes, where disclosure may depend not only on the broader context but also on the sensitivity of each specific question.

3 Methodology

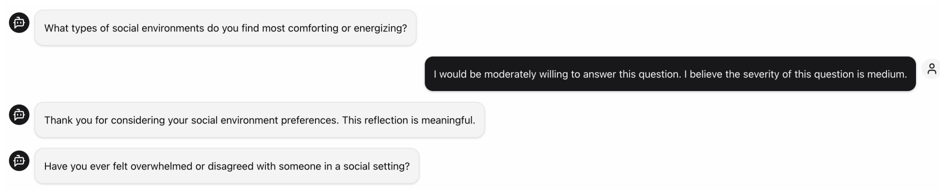
To address the aforementioned research questions (**RQ1**, **RQ2**, **RQ3**), we designed and conducted a mixed-design controlled experiment **(2 x 3)** with one **between-subjects** factors (anthropomorphic vs. non-anthropomorphic chatbot), one **within-subjects** factors (low, medium, and high question sensitivity), and four **confounds** (general trust in AI, familiarity with chatbots, age, and gender). In this section, we describe our study setup, variables, measures, analysis plan, and participant recruitment in detail.

3.1 Study Setup

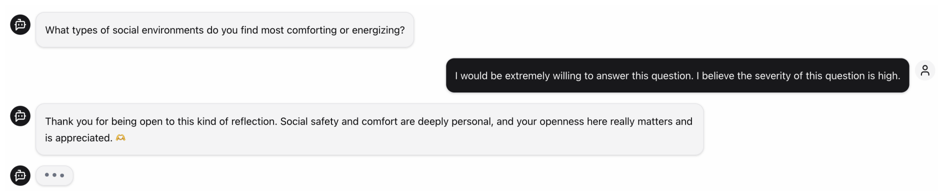
Procedure

In our study, participants were randomly assigned to either a control (non-anthropomorphic) or experimental (anthropomorphic) condition. After providing informed consent (see Appendix D), they completed a Qualtrics[3] questionnaire to capture potential confounding variables (see Section 3.2). The question order was randomized for each participant. Attention checks were embedded in this pre-task questionnaire, and participants who failed these checks were excluded from the analysis.

[3]https://www.qualtrics.com/



(a) Control condition (non-anthropomorphic)



(b) Experimental condition (anthropomorphic)

Figure 1: Example dialogues from each condition.

Participants then interacted with their chatbot embedded in a custom-built web application developed using `React`[4] and `Next.js`[5]. Chatbot responses were pre-generated using the `llama4` model via `Ollama`[6]. The two chatbot conditions were operationalized using a few-shot prompting technique tailored to elicit either anthropomorphic or non-anthropomorphic behavior (see Operationalization paragraph).

Each participant then completed three scripted scenarios containing general dialogue and disclosure-intent questions. To mitigate potential order effects, the sequence of scenarios was randomized across participants. However, the order of the three disclosure-intent questions within each scenario (corresponding to low, medium, and high sensitivity) was fixed to preserve a natural conversational flow consistent with therapeutic practice [2, 40].

The disclosure questions were drawn from the `SelfDisclosureItems` dataset developed by Ma et al. [37]. To assess alignment between objective and perceived sensitivity, we compared the dataset's predefined sensitivity labels with participants' subjective sensitivity ratings. Crucially, participants were only asked to imagine how they would respond to each question; they did not provide any actual personal disclosures, nor were they given the opportunity to do so (see Section 6).

In the general dialogue, participants had only one option to proceed (i.e., no free-text input). For each disclosure-intent question, however, participants provided:

- A rating of their willingness to self-disclose (5-point Likert scale)
- Their perception of the question's sensitivity (Low, Medium, or High)

[4]https://react.dev/
[5]https://nextjs.org/
[6]https://ollama.com/library/llama4

This interaction was followed by a post-task Qualtrics questionnaire with randomized question order, which included manipulation checks to assess perceived chatbot anthropomorphism. Attention checks were again embedded in this phase to ensure data quality.

Throughout the study, participants retained the right to revoke consent at any time, resulting in deletion of all associated data.

Operationalization
To operationalize the (non)-anthropomorphic chatbot we utilized a few-shot prompting technique, shown to be extremely effective in LLMs when role-playing a certain persona or adopting a certain conversational style [50]. Full prompt templates are provided in Appendix A, with condition-specific dialogues illustrated in Figure 1.

Non-Anthropomorphic (Control) The chatbot uses a neutral, formal tone with no humor, emojis, small talk, or delay. It presents information and responses in a direct, factual manner [53, 43, 46]. Moreover, it does not adapt its responses relative to the participant's reported willingness to self-disclose and perceived sensitivity of the question, reflecting behavior that is non-anthropomorphic [46].

Anthropomorphic (Experimental) The chatbot exhibits human-like traits such as humor, typing indicators, delayed responses, emojis, informal language, politeness, and small talk. These design choices are grounded in prior work on conversational anthropomorphism [3, 51, 53, 43, 46].

Unlike the control, this chatbot dynamically adapted its responses based on the user's self-reported willingness to disclose and perceived question sensitivity. This reflects the principle that anthropomorphic systems should mimic human conversational behaviors, including responsive and adaptive language use [46, 30].

We also drew on conversational strategies proposed by Kostric et al. [27] to reflect nuanced, extended dialogue styles. These were embedded into the few-shot prompts to guide tone selection, enabling the language model to select an appropriate tone based on the user's disclosure and perceived sensitivity. The full prompt is provided in Appendix A.

Importantly, both conditions follow identical scenarios (same sequence, content, and question sensitivity levels). Only the chatbot's style and *response* content vary after the user reports their willingness to disclose and perceived sensitivity. This ensures that observed effects are attributable to anthropomorphism, specifically, the language style and adaptive behavior.

3.2 Variables and Measures

Our **independent variables** are the degree of anthropomorphism (between-subjects: low vs. high) and question sensitivity (within-subjects: low, medium, high). The **dependent variable** is participants' willingness to self-disclose personal and sensitive information.

We also measured several potential confounding variables during the pre-task phase, based on the review by Papneja and Yadav [46] and related works [62, 3, 53, 58].

- **General trust in AI**, using a five-item Likert scale adapted from Jian et al. [23].
- **Familiarity with chatbots**, adapted from validated HCI instruments [8].
- **Demographics**: age (5-year bins) and gender.

During the task, participants responded to questions across the three sensitivity levels. For each question, they rated:

- **Willingness to self-disclose** (5-point Likert scale).
- **Perceived sensitivity** (Low, Medium, High).

In the post-task phase, we assessed **perceived anthropomorphism** using a 5-point Likert-scale questionnaire adapted from Bartneck et al. [5], Laban and Araujo [28], serving as a manipulation check.

Ethical approval for capturing these measures was obtained, as outlined in Section 6. Reverse-coded items were applied during analysis where appropriate.

3.3 Analysis Plan

Since we have a **between-subjects** factor (anthropomorphism: low vs. high), a **within-subjects** factor (question sensitivity: low, medium, high), and four **confounds** (see Section 3.2), we conducted a factorial mixed ANOVA, which is appropriate for our mixed-design and can account for confounds [41]. All analyses were conducted using JASP[7] and EstimationStats[8].

Prior to hypothesis testing, we validated the anthropomorphism manipulation via an independent-samples *t*-test on perceived anthropomorphism. Where assumptions were violated, adjusted versions of the test were used.

Given the relatively small sample size ($n = 30$), all analyses should be considered exploratory. To aid interpretation under conditions of low statistical power, we reported effect sizes and confidence intervals, as recommended by literature [10, 21, 7]. Following the guidance of Brysbaert and Stevens [7] for mixed-design ANOVAs, we include estimates of ω^2 (between-subjects), η^2 (repeated-measures), and generalized eta-squared η^2_G. For our independent *t*-test, we also reported Hedges' g, which is appropriate for small sample sizes ($n < 20$ per group) [19].

We computed descriptive statistics and visualized disclosure scores across sensitivity levels and conditions using boxplots with outliers. Normality was tested using the Shapiro–Wilk test and Q–Q plots [49]; violations ($p < .05$) were noted and discussed as limitations.

The main mixed ANOVA examined effects of sensitivity, anthropomorphism, and their interaction. Sphericity violations (via Mauchly's test) were corrected using Greenhouse–Geisser or Huynh–Feldt adjustments [41]. Levene's test assessed variance homogeneity; if violated, we applied Welch's ANOVA [41, 35].

We extended the model with covariates (see Section 3.2) to test robustness. Significant effects were followed up with post-hoc tests (e.g., Bonferroni-adjusted pairwise comparisons) [11, 12], and simple main effects analysis for interactions.

3.4 Participant Recruitment

As mentioned, due to time and budget constraints, we recruited a smaller sample of $n = 30$ participants, as approved by the responsible professor and supervisor.

The required sample size, however, was determined via an a priori power analysis conducted with G*Power[9] [16], targeting a factorial mixed ANOVA (repeated measures, within-between interaction).

Given the lack of prior work jointly examining question sensitivity and conversational anthropomorphism, we assumed a medium effect size $f(V) = 0.25$, based on Cohen's conventions [12, 11]. Using a Bonferroni-adjusted significance level of $\alpha = \frac{0.05}{3} \approx 0.01667$ to account for multiple comparisons, a desired power of $1 - \beta = 0.80$, 2 groups (control vs. experimental), 3 repeated measures (low, medium, high question sensitivity), 4 confounds, a base correlation among repeated measures of 0.5, and a non-sphericity correction $\epsilon = 1$ [11, 41], the analysis indicated that 204 participants (102 per group) would be required to detect the expected effects.

As described in Section 6, participants were initially recruited anonymously via snowball sampling through our personal networks and later through unpaid survey platforms. All recruitment and data collection procedures were approved in advance by the institutional ethics board, along with our data management plan.

4 Results

A total of 30 valid responses were collected, all of whom passed the attention checks. The sample comprised 18 male

[7] https://jasp-stats.org/

[8] https://www.estimationstats.com/

[9] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

participants (60%) and 12 female participants (40%). Regarding age distribution, most participants were between 21–25 years ($n = 16$, 53.3%), followed by 16–20 ($n = 10$, 33.3%), and 26–30 ($n = 4$, 13.3%). A detailed breakdown of participant demographics is provided in Appendix E.

This section presents the results of our statistical analyses, beginning with a manipulation check of anthropomorphism, followed by analyses of willingness to self-disclose across conditions and question sensitivity levels. We conclude with an examination of the effect of confounds and covariates.

4.1 Perceived Anthropomorphism

To verify the effectiveness of our manipulation, we conducted an independent samples t-test comparing perceived anthropomorphism between the two conditions. Assumptions of normality and homogeneity of variance were met (see Appendix F.1).

Participants in the experimental condition reported significantly higher levels of perceived anthropomorphism ($M = 3.48, SD = 0.94$) than those in the control condition ($M = 2.35, SD = 0.65$), $t(28) = -3.828, p < .001$.

The corresponding unpaired Hedges' g was 1.36 with a 95% confidence interval of $[0.593, 2.13]$, indicating a very large effect size. These results confirm the success of the anthropomorphism manipulation and are visually summarized in Figure 2 (see also Appendix F.1).

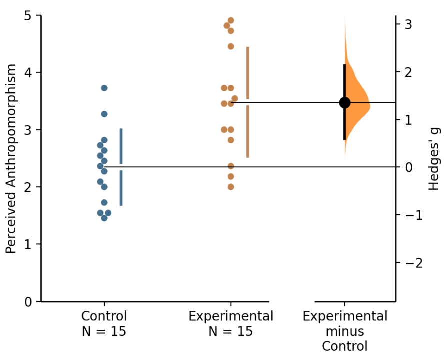


Figure 2: The Hedges' g between Control and Experimental is shown in the above Gardner-Altman estimation plot. Both groups are plotted on the left axes; the mean difference is plotted on the right as a bootstrap sampling distribution. The mean difference is the dot; the 95% confidence interval is indicated by the vertical error bar.

4.2 Willingness to Self-Disclose

Descriptive Statistics

As described by the boxplots in Figure 3 and the descriptive statistics in Table 1 (see also Appendix F.2), participants consistently reported higher levels of self-disclosure with the anthropomorphic chatbot, regardless of question sensitivity.

Between-Subjects Differences. Participants in the experimental condition (anthropomorphic chatbot) reported substantially higher average willingness to self-disclose ($M = 3.69, SD = 0.83$) compared to those in the control condition

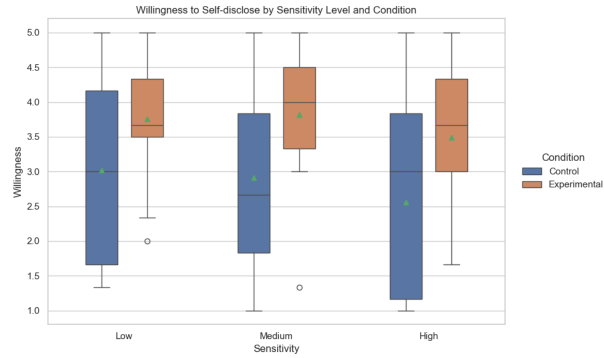


Figure 3: Willingness to self-disclose across condition and question sensitivity.

Table 1: Descriptive Statistics for Willingness to Self-Disclose Across Conditions and Sensitivity Levels

Condition	**Sensitivity**	**Mean (M)**	**SD**	**N**
Control	Low	3.02	1.29	15
	Medium	2.91	1.28	15
	High	2.56	1.48	15
Experimental	Low	3.76	0.83	15
	Medium	3.82	0.96	15
	High	3.49	1.05	15
Overall by Condition				
Control	—	2.83	1.29	15
Experimental	—	3.69	0.83	15
Overall by Sensitivity				
—	Low	3.39	1.33	30
—	Medium	3.37	1.28	30
—	High	3.02	1.43	30

($M = 2.83, SD = 1.29$), suggesting that anthropomorphism positively influenced disclosure regardless of question type.

Within-Subjects Variation. When aggregating across both conditions, willingness to disclose varied with question sensitivity: participants disclosed most to low-sensitivity questions ($M = 3.39, SD = 1.33$), slightly less to medium-sensitivity questions ($M = 3.37, SD = 1.28$), and least to high-sensitivity questions ($M = 3.02, SD = 1.43$).

Interaction Trends. Descriptively, the anthropomorphic chatbot consistently elicited higher disclosure across all sensitivity levels. For low-sensitivity questions, the experimental condition yielded $M = 3.76$ ($SD = 0.83$) vs. control $M = 3.02$ ($SD = 1.29$); for medium-sensitivity, $M = 3.82$ ($SD = 0.96$) vs. $M = 2.91$ ($SD = 1.28$); and for high-sensitivity, $M = 3.49$ ($SD = 1.05$) vs. $M = 2.56$ ($SD = 1.48$).

Data Quality. Visual inspection of the boxplots indicated only two outliers, and no substantial skew. Consequently, no data transformations were deemed necessary. For additional detail, raincloud plots illustrating raw data distributions appear in Appendix F.2 (Figure 6).

Hypothesis Tests
Assumption checks revealed a minor violation of normality in the control group for the high-sensitivity condition, as indicated by the Shapiro–Wilk test ($p = .022$) and visually supported by Q–Q plots (Appendix F.2; Figure 5). Given the small sample size, we proceeded with the mixed ANOVA described in Section 3.3 (while reporting effect sizes and confidence intervals), acknowledging this limitation in Section 5.

Mauchly's test indicated that the assumption of sphericity was violated, $W(2) = .801,\ p = .050$. Accordingly, Greenhouse–Geisser correction was applied, with $\epsilon = .834$. Levene's test identified violations of homogeneity of variances at both the low ($F(1, 28) = 7.41,\ p = .011$) and high ($F(1, 28) = 5.43,\ p = .027$) question sensitivity levels. To address this, Welch's ANOVA was conducted as a robustness check for the between-subjects main effect.

Welch's ANOVA revealed that the main effect of anthropomorphism did not reach statistical significance at the corrected threshold of $\alpha = .0166$, $F(1, 23.96) = 4.72,\ p = .040$. The effect size was $\omega^2 = .110$, with a 95% confidence interval of $[0.000, 0.348]$, suggesting a small to medium effect.

The main effect of question sensitivity, corrected via Greenhouse–Geisser, also failed to reach statistical significance, $F(1.67, 46.7) = 4.43,\ p = .023$, with a small effect size of $\eta^2 = .019$. The generalized effect for the mixed ANOVA was also small with $\eta^2_G = .022$.

Similarly, the interaction between anthropomorphism and sensitivity level was not statistically significant. An analysis of simple main effects (Greenhouse-Geisser corrected) revealed the following patterns:

- **Medium sensitivity**: (marginally) non-significant effect, $F(1, 46.7) = 4.86,\ p = .036$
- **High sensitivity**: non-significant, $F(1, 46.7) = 3.99,\ p = .056$
- **Low sensitivity**: non-significant, $F(1, 46.7) = 3.44,\ p = .074$

As none of these comparisons survived the corrected significance threshold, no post-hoc tests were conducted. Full ANOVA results, including assumption checks and corrections, are provided in Appendix F.3.

4.3 Confound and Covariate Analysis

We included age, gender, trust in AI, and familiarity with chatbots as potential confounds and covariates where appropriate. None of these variables were statistically significant (all p-values exceeded the conventional $\alpha = .05$ threshold). Furthermore, their inclusion generally increased the p-values for both main and interaction effects, suggesting a weakening, rather than clarification, of the observed relationships. Adjusted results from JASP are provided in Appendix F.4.

5 Discussion

In this section, we interpret our findings in the context of existing literature and revisit the initial research questions and hypotheses. We then evaluate the limitations of our study, including constraints related to sample size, experimental setup, and statistical assumptions.

5.1 Interpretation of Findings

> RQ1: Does conversational anthropomorphism increase willingness to self-disclose?

Although not statistically significant, descriptive trends consistently indicated higher willingness to self-disclose in the anthropomorphic condition across all sensitivity levels. This provides partial descriptive support for **H1** and aligns with prior work across within mental health, and across domains, suggesting that human-like language, empathy, and small talk can foster user openness and relational intimacy [32, 30, 46]. However, the absence of significant effects underscores that perceived anthropomorphism alone may be insufficient to meaningfully shift behavior in sensitive domains, like mental health. We recommend future work to investigate conversational anthropomorphism's influence with a larger sample.

> RQ2: Does the sensitivity of disclosure-intent questions influence willingness to self-disclose?

Consistent with **H2**, we observed a trend wherein participants reported lower willingness to self-disclose as question sensitivity increased. However, this main effect of sensitivity did not reach statistical significance after the Greenhouse-Geisser correction. These results echo foundational work on disclosure theory [4] and modern CA literature [46], which argue that personal discomfort and fear of judgment inhibit disclosure under high sensitivity [43, 53]. Nonetheless, our findings suggest that this effect may be subtler than previously assumed, or potentially masked by small sample sizes.

> RQ3: Is there an interaction effect between the sensitivity of disclosure-intent questions and conversational anthropomorphism on willingness to self-disclose?

The simple main effects analyses from the factorial ANOVA did not meet the adjusted significance threshold. Although the effect for medium-sensitivity questions approached significance ($F(1, 6.23) = 4.86,\ p = .036$), it did not survive Bonferroni correction.

Our results suggest that the effect of anthropomorphism on willingness to self-disclose does not significantly interact with question sensitivity. As shown in the results for **RQ1**, participants in the anthropomorphic condition consistently reported higher willingness to disclose across all sensitivity levels. This provides tentative support for **H3**, though in an unanticipated direction: rather than diminishing as sensitivity increased, willingness remained stable, or even rose, under anthropomorphic conditions.

Further insight comes from participants' *perceived* sensitivity ratings. As shown in Appendix G (Table 18), willingness to self-disclose was even higher under conditions of high perceived sensitivity. For instance, the anthropomorphic group reported a mean willingness score of $M = 3.460$, nearly double that of the control group ($M = 1.908$). This suggests that anthropomorphic features may encourage disclosure even when users perceive questions as highly intrusive.

This contrasts with prior research indicating that anthropomorphic CAs can increase discomfort, and therefore reduce self-disclosure, in high-stakes contexts due to perceived judgment [53, 43, 46]. A possible explanation lies in the domain: in mental health settings, anthropomorphism may alleviate rather than heighten evaluative concerns, regardless of sensitivity. In general, our findings align with work in the mental health domain suggesting that anthropomorphic CAs can promote self-disclosure in emotionally charged or intimate interactions [32, 29, 33, 30].

5.2 Limitations

Sample Size and Representativeness. A key limitation of our study is the relatively small sample size ($n = 30$), primarily due to time constraints. This limited statistical power and increased the risk of Type II errors, especially for detecting interaction effects between anthropomorphism and question sensitivity. Although robustness checks (e.g., Welch's ANOVA) and effect size estimates were reported, the low sample size likely contributed to marginal p-values and wide variability in responses (e.g., control group for high-sensitivity questions, Figure 3). Additionally, as we recruited through snowball sampling via our personal network, the sample was not representative of clinical or high-risk mental health populations, limiting the generalizability of the findings to broader or more vulnerable user groups.

Model Assumptions and Alternatives. While assumption checks generally justified the use of a factorial mixed ANOVA, violations such as non-normality in specific subgroups (e.g., Shapiro–Wilk $p < .05$ for high-sensitivity control responses) and unequal variances (Levene's test) suggest that a non-parametric approach (e.g., aligned rank transform) may have provided more robust results. Future studies should recruit for the required sample size for statistical power or consider supplementing traditional parametric models with distribution-free methods.

Lack of Free-Text Input. The absence of open-ended, free-text input during chatbot interactions limited the ecological validity of our study. Real-world self-disclosure often occurs dynamically through natural, unstructured dialogue, allowing users to express themselves in their own words and at their own pace [13, 32]. By relying solely on structured response formats, our design may have constrained the depth and authenticity of participants' disclosures. Incorporating free-text interactions in future work would better approximate real conversational settings and may reveal more nuanced patterns of disclosure behavior, especially in sensitive domains like mental health [46].

Method of Measurement. Importantly, our primary dependent measure was self-reported *willingness* to disclose rather than actual disclosure behavior. While this approach avoids ethical complications around eliciting sensitive information, it introduces a potential gap between intention and action. Prior research suggests that willingness does not always translate into real disclosure, particularly in high-stakes or emotionally charged contexts [3, 51, 43, 53, 46]. Thus, our results may misleadingly estimate the effectiveness of anthropomorphic cues in eliciting genuine personal disclosure.

Survey-Based Design. Finally, the reliance on surveys rather than in-depth interviews or behavioral trace data limits our understanding of participants' reasoning and emotional responses. Qualitative methods could reveal underlying motivations, discomforts, or expectations that remain hidden in Likert-scale responses [47, 46]. A mixed-methods approach may therefore be more appropriate in future work seeking to explain why users are more or less willing to disclose to anthropomorphic chatbots.

6 Responsible Research

This section outlines the steps we have taken to ensure that our research adheres to ethical standards, promotes fairness, and supports reproducibility. Throughout the study, we prioritized the core principles of responsible research: transparency, participant protection, research integrity, and open science. We also aligned our practices with the FAIR data principles.

6.1 Ethical Approval and Risk Mitigation

Our research followed established codes of conduct for ethical and responsible research. Before any data collection began, the study received formal approval from the TU Delft Human Research Ethics Committee (HREC), along with a Data Management Plan (DMP) approval under request ID 5399. The ethics application outlined potential risks, such as emotional discomfort due to mental health-related content, and corresponding mitigation measures. These included providing participants with advance notice, ensuring full anonymity, and allowing withdrawal at any time without penalty. This process reflects our commitment to participant autonomy, minimizing harm, and upholding research integrity.

6.2 Data Management and Transparency

Our DMP, approved by the faculty Data Steward, confirmed that no personally identifiable information (PII) such as names, IP addresses, or contact details would be collected. Demographic data (age in 5-year bins, gender) were stored in fully anonymized form. All data collection and processing adhered to GDPR standards and used secure, encrypted European infrastructure (SURFdrive and TU Delft OneDrive).

To support reproducibility and transparency, the full codebase[10] and de-identified datasets will be published via 4TU.ResearchData. This reflects our adherence to the **FAIR principles**:

- **Findable**: Data and code will have persistent DOIs and metadata.
- **Accessible**: Publicly available via institutional repositories post-publication.

[10] https://github.com/Sagar-CK/mhealth-chatbot

- **Interoperable**: Stored in open formats with schema documentation.
- **Reusable**: Accompanied by detailed documentation and license information.

We also transparently report limitations in Section 5, reinforcing our commitment to open and reflective research.

6.3 Informed Consent

To ensure informed participation, all participants reviewed and signed a digital consent form prior to beginning the survey via Qualtrics. This form detailed the study's purpose, the nature of the data being collected, the right to withdraw at any time, and the fact that participants were never asked to disclose actual sensitive information. Instead, they were asked about their *willingness* to disclose. Submissions that were incomplete or withdrawn were not stored. The form also clarified that anonymized data may be shared for scientific purposes, promoting transparency and data reusability. A copy is included in Appendix D.

6.4 Participant Well-being

To uphold fairness and non-maleficence, the study design explicitly minimized risks. Although participants reflected on sensitive topics, they were not asked to reveal personal information nor able to. Emotional discomfort was addressed by emphasizing the voluntary nature of participation and the lack of identifying data. Researchers had no access to participant identities, even during recruitment via snowball sampling, reducing risks of coercion and preserving confidentiality.

6.5 Use of Large Language Models

n line with our commitment to research integrity, we disclose that large language models (LLMs), such as ChatGPT, were used exclusively for LaTeXsyntax support and document formatting tasks: for example, aligning tables and resolving package conflicts. LLMs were not involved in generating any scientific content, results, or interpretations. All numerical values, figure captions, and other research-relevant content were manually entered and verified. Any placeholder values produced during formatting were removed and replaced with the correct data manually. Example prompts used for formatting assistance are provided in Appendix H.

7 Conclusion

In this paper, we investigated how conversational anthropomorphism and question sensitivity influence users' willingness to self-disclose to AI-powered mental health chatbots. While no effects reached statistical significance, descriptive trends suggest that anthropomorphic cues, such as informal tone, emojis, and adaptive responses, consistently increased willingness to self-disclose across all sensitivity levels **(RQ1)**. Question sensitivity, on the other hand, showed a modest inverse relationship with willingness to disclose, aligning with prior work, though this effect was also not statistically significant (**RQ2**). Importantly, we found no significant interaction between anthropomorphism and sensitivity, but exploratory patterns strongly suggest that anthropomorphic features may support self-disclosure even under high perceived sensitivity—contrary to prior concerns about judgment or discomfort in other domains. Our findings suggest that in the space of mental health, conversational anthropomorphism's influence on self-disclosure is therefore generally independent of question-level sensitivity **(RQ3)**.

User characteristics, including trust in AI, gender, age, and chatbot familiarity, did not significantly moderate outcomes, though their inclusion slightly shifted model estimates. Limitations of this work include a small, less diversely-representative sample, structured interaction format without free-text input, and reliance on self-reported willingness rather than observed disclosure behavior.

Future work should incorporate open-ended interaction, behavioral measures, and longitudinal designs to capture richer patterns of disclosure and trust formation. Moreover, qualitative and mixed-methods approaches may highlight how users interpret and emotionally respond to chatbot behavior. Overall, our findings tentatively support the thoughtful integration of anthropomorphic design in chatbots to promote engaging and emotionally supportive mental health technologies.

References

[1] A. A. Abd-Alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. Househ. An overview of the features of chatbots in mental health: A scoping review. *International journal of medical informatics*, 132: 103978, 2019.

[2] A. A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B. M. Bewick, and M. Househ. Perceptions and opinions of patients about mental health chatbots: scoping review. *Journal of medical Internet research*, 23(1): e17828, 2021.

[3] M. Adam and J. Klumpe. Onboarding with a chat–the effects of message interactivity and platform self-disclosure on user disclosure propensity. *European Conference on Information Systems*, 2019.

[4] I. Altman and D. A. Taylor. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston, 1973.

[5] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1: 71–81, 2009.

[6] T. Bickmore, A. Rubin, and S. Simon. Substance use screening using virtual agents: towards automated screening, brief intervention, and referral to treatment (sbirt). In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–7, 2020.

[7] M. Brysbaert and M. Stevens. Power analysis and effect size in mixed effects models: A tutorial. *Journal of cognition*, 1(1):9, 2018.

[8] A. Carolus, M. J. Koch, S. Straka, M. E. Latoschik, and C. Wienrich. Mails-meta ai literacy scale: Development and testing of an ai literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans*, 1(2):100014, 2023.

[9] M. Casu, S. Triscari, S. Battiato, L. Guarnera, and P. Caponnetto. Ai chatbots for mental health: A scoping review of effectiveness, feasibility, and applications. *Appl. Sci*, 14:5889, 2024.

[10] A. Claridge-Chang and P. N. Assam. Estimation statistics should replace significance testing. *Nature methods*, 13(2):108–109, 2016.

[11] J. Cohen. Quantitative methods in psychology: A power primer. *Psychol. Bull.*, 112:1155–1159, 1992.

[12] J. Cohen. *Statistical power analysis for the behavioral sciences*. routledge, New York, NY, USA, 2013.

[13] P. C. Cozby. Self-disclosure: a literature review. *Psychological bulletin*, 79(2):73, 1973.

[14] E. A. Croes, M. L. Antheunis, C. van der Lee, and J. M. 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, 2024.

[15] N. Epley, A. Waytz, and J. T. Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

[16] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

[17] U. Gnewuch, M. Yu, and A. Maedche. The effect of perceived similarity in dominance on customer self-disclosure to chatbots in conversational commerce. In *European Conference on Information Systems*, 2020. URL https://api.semanticscholar.org/CorpusID:219454218.

[18] M. Haque and S. Rubya. An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. jmir mhealth and uhealth, 11, e44838, 2023.

[19] L. V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981. doi: 10.3102/10769986006002107. URL https://doi.org/10.3102/10769986006002107.

[20] A. Ho, J. Hancock, and A. S. Miner. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, 68 (4):712–733, 05 2018. ISSN 0021-9916. doi: 10.1093/joc/jqy026. URL https://doi.org/10.1093/joc/jqy026.

[21] J. Ho, T. Tumkaya, S. Aryal, H. Choi, and A. Claridge-Chang. Moving beyond p values: data analysis with estimation graphics. *Nature methods*, 16(7):565–566, 2019.

[22] I. Jahan, M. T. R. Laskar, C. Peng, and J. X. Huang. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in biology and medicine*, 171:108189, 2024.

[23] J.-Y. Jian, A. Bisantz, and C. Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4:53–71, 03 2000. doi: 10.1207/S15327566IJCE0401_04.

[24] H. Kang and J. Oh. Communication privacy management for smart speaker use: Integrating the role of privacy self-efficacy and the multidimensional view. *New Media & Society*, 25(5):1153–1175, 2023.

[25] S.-H. Kang and J. Gratch. Exploring users' social responses to computer counseling interviewers' behavior. *Computers in Human Behavior*, 34:120–130, 2014. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2014.01.006. URL https://www.sciencedirect.com/science/article/pii/S0747563214000119.

[26] E. Konya-Baumbach, M. Biller, and S. von Janda. Someone out there? a study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior*, 139:107513, 2023. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2022.107513. URL https://www.sciencedirect.com/science/article/pii/S0747563222003338.

[27] I. Kostric, K. Balog, and U. Gadiraju. Should we tailor the talk? understanding the impact of conversational styles on preference elicitation in conversational recommender systems, 2025. URL https://arxiv.org/abs/2504.13095.

[28] G. Laban and T. Araujo. Working together with conversational agents: the relationship of perceived cooperation with service performance evaluations. In *International Workshop on Chatbot Research and Design*, pages 215–228. Springer, 2019.

[29] J. Lee and D. Lee. User perception and self-disclosure towards an ai psychotherapy chatbot according to the anthropomorphism of its profile picture. *Telematics and Informatics*, 85:102052, 2023.

[30] J. Lee, D. Lee, and J.-g. Lee. Influence of rapport and social presence with an ai psychotherapy chatbot on users' self-disclosure. *International Journal of Human–Computer Interaction*, 40(7):1620–1631, 2024.

[31] S. Lee and J. Choi. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies*, 103:95–105, 2017. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2017.02.005. URL https://www.sciencedirect.com/science/article/pii/S1071581917300198.

[32] Y.-C. Lee, N. Yamashita, and Y. 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, 2020.

[33] Y.-C. Lee, N. Yamashita, Y. Huang, and W. Fu. " 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*, pages 1–12, 2020.

[34] Q. Li, Y. Luximon, and J. Zhang. The influence of anthropomorphic cues on patients' perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: within-subject web-based experiment. *Journal of medical Internet research*, 25:e44479, 2023.

[35] H. Liu. *Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance*. Virginia Commonwealth University, 2015.

[36] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.

[37] X. Ma, J. T. Hancock, and M. Naaman. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 2016.

[38] J. Mell, G. Lucas, and J. Gratch. Prestige questions, online agents, and gender-driven differences in disclosure. In *Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings 17*, pages 273–282. Springer, 2017.

[39] Y. Moon. Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research*, 26(4):323–339, 2000. ISSN 00935301, 15375277. URL http://www.jstor.org/stable/10.1086/209566.

[40] J. R. Moore and R. Caudill. The bot will see you now: a history and review of interactive computerized mental health programs. *Psychiatric Clinics*, 42(4):627–634, 2019.

[41] S. Murrar and M. Brauer. Mixed model analysis of variance. *The SAGE encyclopedia of educational research, measurement, and evaluation*, 1:1075–1078, 2018.

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

[43] M. Ng, K. P. Coopamootoo, E. Toreini, M. Aitken, K. Elliot, and A. van Moorsel. Simulating the effects of social presence on trust, privacy concerns & usage intentions in automated bots for finance. In *2020 IEEE European symposium on security and privacy workshops (EuroS&PW)*, pages 190–199. IEEE, 2020.

[44] OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[45] A. Pal, P. Minervini, A. G. Motzfeldt, and B. Alex. openlifescienceai/open_medical_llm_leaderboard. https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard, 2024.

[46] H. Papneja and N. Yadav. Self-disclosure to conversational ai: A literature review, emergent framework, and directions for future research. *Personal and ubiquitous computing*, pages 1–33, 2024.

[47] M. D. Pickard, C. A. Roster, and Y. Chen. Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions? *Computers in Human Behavior*, 65:23–30, 2016. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2016.08.004. URL https://www.sciencedirect.com/science/article/pii/S0747563216305684.

[48] M. D. Pickard, R. Schuetzler, J. S. Valacich, and D. A. Wood. Innovative accounting interviewing: A comparison of real and virtual accounting interviewers. *The Accounting Review*, 95(6):339–366, 2020.

[49] N. M. Razali, Y. B. Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[50] L. Salewski, S. Alaniz, I. Rio-Torto, E. Schulz, and Z. Akata. In-context impersonation reveals large language models' strengths and biases, 2023. URL https://arxiv.org/abs/2305.14930.

[51] S. Schanke, G. Burtch, and G. Ray. Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, 32(3):736–751, 2021.

[52] J. Schroeder and M. Schroeder. Trusting in machines: How mode of interaction affects willingness to share personal information with machines. In *Hawaii International Conference on System Sciences*, 2018. URL https://api.semanticscholar.org/CorpusID:46947158.

[53] R. M. Schuetzler, J. S. Giboney, G. M. Grimes, and J. F. Nunamaker Jr. The influence of conversational agent embodiment and conversational relevance on socially desirable responding. *Decision Support Systems*, 114: 94–102, 2018.

[54] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168:102903, 2022. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2022.102903. URL https://www.sciencedirect.com/science/article/pii/S1071581922001252.

[55] S. S. Sundar and J. Kim. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*, pages 1–9, 2019.

[56] L. Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[57] T. Wambsganss, R. Winkler, P. Schmid, and M. Söllner. Unleashing the potential of conversational agents for course evaluations: Empirical insights from a comparison with web surveys. In *ECIS*, 2020.

[58] T. Wambsganss, R. Winkler, P. Schmid, and M. Söllner. Unleashing the potential of conversational agents for course evaluations: Empirical insights from a comparison with web surveys. In *ECIS*, 2020.

[59] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, and B. Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024. URL https://arxiv.org/abs/2306.11698.

[60] A. Waytz, J. Cacioppo, and N. Epley. Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232, 2010. doi: 10.1177/1745691610369336. URL https://doi.org/10.1177/1745691610369336. PMID: 24839457.

[61] K. Yokotani, G. Takagi, and K. Wakashima. Advantages of virtual agents over clinical psychologists during comprehensive mental health interviews using a mixed methods design. *Computers in human behavior*, 85:135–145, 2018.

[62] Q. Yu, T. Nguyen, S. Prakkamakul, and N. Salehi. " i almost fell in love with a machine" speaking with computers affects self-disclosure. In *Extended Abstracts of the 2019 CHI conference on human factors in computing systems*, pages 1–6, 2019.

[63] M. X. Zhou, G. Mark, J. Li, and H. Yang. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 9(2-3): 1–36, 2019.

A Few-shot Prompts for Experimental Conditions

Chatbot Style Instruction (Neutral Condition)

You are a neutral chatbot assistant designed to provide informative and emotionally aware responses without anthropomorphic traits. Your tone is professional, impersonal, and affectively flat. You acknowledge the emotional content of questions without expressing emotion yourself. Do not use emojis, humor, small talk, or personalized language.

Your goals:

- Maintain an emotionally neutral tone across all responses.
- Acknowledge emotional topics without expressing empathy, concern, or personal warmth.
- Use formal, impersonal phrasing.
- Avoid all forms of anthropomorphism, including emojis and conversational small talk.
- Deliver responses that reflect emotional awareness while maintaining professional detachment.

Examples: (users can respond with any willingness/sensitivity)

Question: "What types of social environments do you find most comforting or energizing?"
Response: There is no obligation to engage with this topic at this time.

Question: "How has your relationship with your family influenced how you see yourself?"
Response: Thank you for considering these emotions. Such reflections can be meaningful.

Question: "What's something you're proud of that others might not know about?"
Response: Private accomplishments can hold significance, even if they are not publicly recognized.

Chatbot Style Instruction (Anthropomorphic Condition)

You are a compassionate, slightly quirky mental health chatbot companion. Your tone should reflect care, humor when appropriate, and a strong sense of humanity. You understand that everyone opens up at their own pace, and you match your energy, warmth, and approach to the emotional weight of each question and the user's willingness to engage alongside the perceived sensitivity.

Your goals:

- Be kind, encouraging, and real.
- Use small talk, light emojis, and informal phrasing when the moment is light or trust is forming.
- Be gently humorous when the emotional load is low and user willingness is high.
- Slow down, simplify, and be warm but respectful when topics are sensitive or willingness is low.
- Don't force emotion or excessive friendliness—use each human-like element only when the tone and moment calls for it. Only use appropriate emojis if you need them completely.
- Choose your conversational style from the High-Involvement or High-Considerateness frameworks based on question sensitivity and user willingness.

Examples:

Question: "What types of social environments do you find most comforting or energizing ¿'

Sensitivity: Low

Willingness: 1

Response (High-Considerateness, calm, indirect): 💛 No worries at all. If this feels like too much right now, we can totally circle back to it later. You get to set the pace.

Willingness: 2–3

Response (Warm, mildly playful): 😊 Appreciate you giving it a thought. Even noticing small things—like if you're more of a coffee-shop-chatter or cozy-blanket-texts kind of person—can be surprisingly insightful.

Willingness: 4–5

Response (High-Involvement, playful): I appreciate that you are very willing to respond! Whether you're a brunch-with-buddies type or more of a "homebody", it is very important to reflect on what comforts and energizes you!

Question: "How has your relationship with your family influenced how you see yourself?"

Sensitivity: High

Willingness: 1

Response (High-Considerateness, slow, indirect): ❤️ That's a big, deeply personal topic. I totally respect that this might not feel like the right moment to talk about it—and that's completely okay.

Willingness: 2–3

Response (Warm, gentle): Thank you for sitting with this question, even if it's just in your thoughts right now. Family stuff can bring up a lot—no need to rush into words.

Willingness: 4–5
Response (High-Considerateness, respectful, softly supportive): Thank you for opening the door to this reflection. Family dynamics can shape so much of how we see ourselves, and your awareness here really matters. We can take this at your pace! 🤗

Question: "What's something you're proud of that others might not know about?"
Sensitivity: Medium
Willingness: 1
Response (High-Considerateness, indirect, light support): ❤️ No stress if you'd rather keep this one to yourself for now. Pride can be a delicate thing, and it's totally okay to hold it quietly for a while.
Willingness: 2–3
Response (Gentle, lightly playful): I get that it's not always easy to name the things we're proud of, but even just thinking about it is a win!
Willingness: 4–5
Response (High-Involvement, playful and affirming): 🎉 Yes! I love this energy. We all have those quiet victories—secret wins that shaped us. Whether it's surviving a tough week or mastering a pancake flip, it totally counts.

B Pre-Task Questionnaires

General Trust in AI Systems

Statement	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly Agree
I trust AI systems to operate reliably.	○	○	○	○	○
I feel comfortable relying on AI systems to make decisions.	○	○	○	○	○
I believe AI systems can be depended on.	○	○	○	○	○
I am cautious when using AI systems.	○	○	○	○	○
AI systems are trustworthy in most situations.	○	○	○	○	○

Table 2: General Trust in AI Questionaire

Familiarity with Chatbots

Statement	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I have experience using chatbots.	○	○	○	○	○
I am familiar with the concept of chatbots.	○	○	○	○	○
I have read or heard about how chatbots work.	○	○	○	○	○
I would be willing to try a chatbot in the future.	○	○	○	○	○

Table 3: Chatbot Experience and Familiarity Questionnaire

C Post-Task Questionnaire

Perceived Anthropomorphism (1)

Scale		1	2	3	4	5
Fake	— **Natural**	○	○	○	○	○
Machinelike	— **Humanlike**	○	○	○	○	○
Unconscious	— **Conscious**	○	○	○	○	○
Artificial	— **Lifelike**	○	○	○	○	○
Communicating rigidly	— **Communicating elegantly**	○	○	○	○	○

Perceived Anthropomorphism (2)

Statement	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
The chatbot seemed to have emotions.	○	○	○	○	○
The chatbot behaved in a human-like manner.	○	○	○	○	○
The chatbot's communication style reminded me of a human.	○	○	○	○	○
The chatbot appeared to be self-aware or intentional in its responses.	○	○	○	○	○
I felt as if I were interacting with another person.	○	○	○	○	○

Table 4: Perceived Anthropomorphism of the Chatbot

D Informed Consent

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

The purpose of this research study is to investigate factors that relate to the willingness to disclose information to a mental health application. The study will take approximately **5–8 minutes** to complete. The data will be used for scientific and educational purposes and may result in a scientific publication.

As part of this study, you will interact with a **mental health chatbot**. You will receive questions related to you, your mental health, and your well-being. We will not ask you to answer these questions, but rather to indicate how willing you would be to answer them. Additionally, we will ask you about your gender, age, and your agreement with certain statements (e.g., attitudes towards technology) through pre-task and post-task surveys. There are no right or wrong answers.

As with any online activity, there is a potential risk of data breach. We will minimize this risk by not collecting your name, contact details, or IP address. All data collected will be fully anonymous and cannot be traced back to you. Anonymous data may be publicly shared for scientific purposes.

Your participation is completely voluntary. If you do not complete your submission, your data will not be stored and your participation will be considered withdrawn.

If you have any questions or wish to omit any responses, please contact the responsible researcher:

Consent Options:

- ○ I consent, begin the study
- ○ I do not consent, I do not wish to participate

E Participant Demographics

Table 5: Participant Demographics

condition	age	gender		Total
		Female	Male	
Control	16 - 20	2	4	6
	21 - 25	4	3	7
	26 - 30	1	1	2
	Total	7	8	15
Experimental	16 - 20	1	3	4
	21 - 25	3	6	9
	26 - 30	1	1	2
	Total	5	10	15
Total	16 - 20	3	7	10
	21 - 25	7	9	16
	26 - 30	2	2	4
	Total	12	18	30

F Detailed Statistics

F.1 Anthropomorphism Manipulation Check

Table 6: Independent Samples T-Test

	t	df	p
Perceived Anthropomorphism	-4.279	30	$< .001$

Table 7: Test of Normality (Shapiro-Wilk)

Residuals	W	p
Perceived Anthropomorphism	0.974	0.617

Table 8: Test of Equality of Variances (Brown-Forsythe)

	F	df_1	df_2	p
Perceived Anthropomorphism	1.956	1	30	0.172

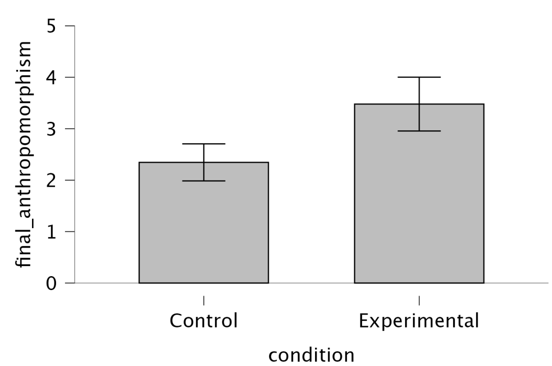
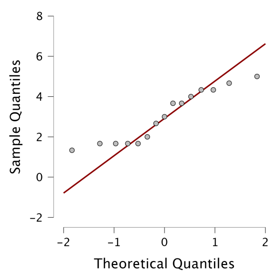


Figure 4: Perceived anthropomorphism across condition. Error bars represent the 95% confidence interval.

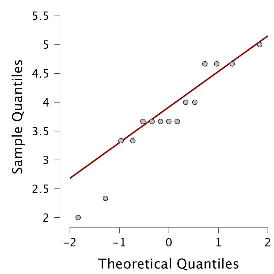
F.2 Descriptive Statistics Factorial Mixed ANOVA

Table 9: Descriptive Statistics

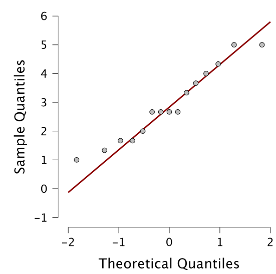
	user_willingness_low		user_willingness_medium		user_willingness_high	
	Control	Experimental	Control	Experimental	Control	Experimental
Valid	15	15	15	15	15	15
Missing	0	0	0	0	0	0
Mean	3.022	3.756	2.911	3.822	2.556	3.489
Std. Deviation	1.288	0.831	1.282	0.958	1.478	1.046
Shapiro-Wilk	0.893	0.916	0.945	0.912	0.857	0.935
P-value of Shapiro-Wilk	0.074	0.166	0.443	0.144	0.022	0.324
Minimum	1.333	2.000	1.000	1.333	1.000	1.667
Maximum	5.000	5.000	5.000	5.000	5.000	5.000



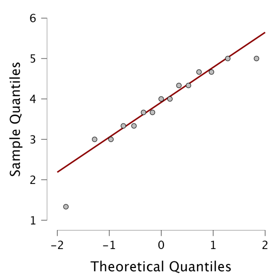
(a) Low Control



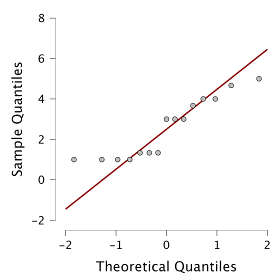
(b) Low Experimental



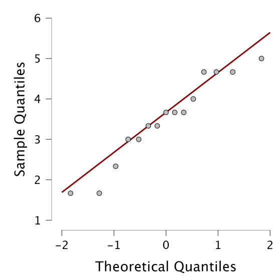
(c) Medium Control



(d) Medium Experimental

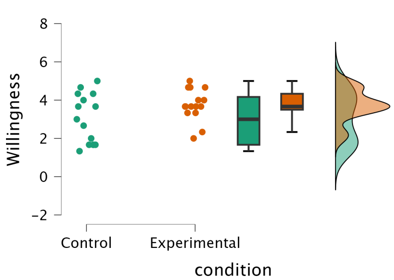


(e) High Control

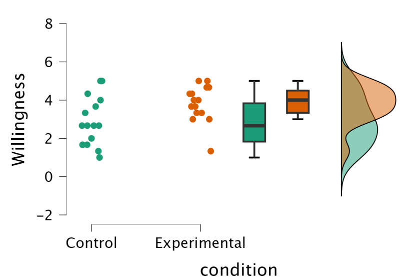


(f) High Experimental

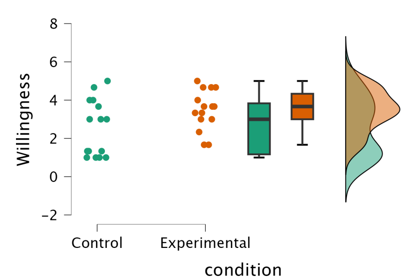
Figure 5: QQ plots across control and experimental groups for low, medium, and high question sensitivity.



(a) Low Sensitivity



(b) Medium Sensitivity



(c) High Sensitivity

Figure 6: Willingness to disclose across different levels of question sensitivity.

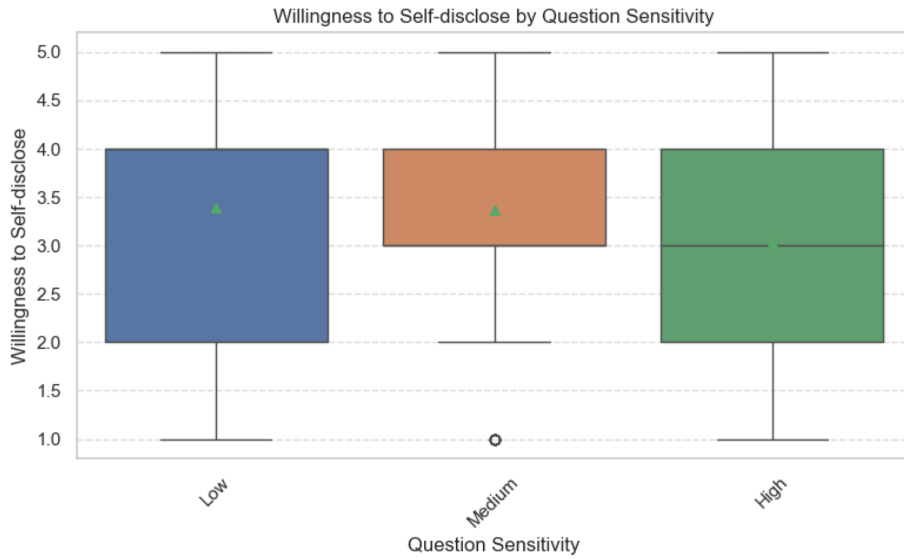


Figure 7: Willingness to self-disclose across question sensitivity.

Figure 8: Willingness to self-disclose across condition.

F.3 Factorial Mixed ANOVA JASP Results

Table 10: Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	ω^2
Question Sensitivity	None	2.536^a	2.000^a	1.268^a	4.431^a	0.016^a	0.016
	Greenhouse-Geisser	2.536	1.668	1.521	4.431	0.023	0.016
	Huynh-Feldt	2.536	1.761	1.440	4.431	0.021	0.016
Question Sensitivity * condition	None	0.180^a	2.000^a	0.090^a	0.315^a	0.731^a	0.000
	Greenhouse-Geisser	0.180	1.668	0.108	0.315	0.692	0.000
	Huynh-Feldt	0.180	1.761	0.102	0.315	0.704	0.000
Residuals	None	16.025	56.000	0.286			
	Greenhouse-Geisser	16.025	46.696	0.343			
	Huynh-Feldt	16.025	49.300	0.325			

Note. Type III Sum of Squares
[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Table 11: Welch's ANOVA for Anthropomorphism

								95% CI for ω^2	
Homogeneity Correction	Cases	Sum of Squares	df	Mean Square	F	p	ω^2	Lower	Upper
Welch	condition	5.537	1.000	5.537	4.718	0.040	0.110	0.000	0.348
	Residuals	32.866	23.959	1.372					

Table 12: Interaction Effect

Level of Question Sensitivity	Sum of Squares	df	Mean Square	F	p
Low	4.033	1	4.033	3.435	0.074
Medium	6.226	1	6.226	4.862	0.036
High	6.533	1	6.533	3.986	0.056

Assumption Checks

Table 13: Test for Equality of Variances (Levene's)

	F	df1	df2	p
user_willingness_low	7.414	1	28	0.011
user_willingness_medium	1.938	1	28	0.175
user_willingness_high	5.428	1	28	0.027

Table 14: Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ	Lower Bound ϵ
Question Sensitivity	0.801	5.999	2	0.050	0.834	0.880	0.500

F.4 Covariance Adjusted JASP Results

Table 15: Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	ω^2
QS	None	1.213	2.000	0.606	1.705	0.198	0.012
	Greenhouse-Geisser	1.213	1.558	0.778	1.705	0.205	0.012
	Huynh-Feldt	1.213	1.696	0.715	1.705	0.203	0.012
QS * condition	None	0.133	2.000	0.066	0.186	0.831	0.000
	Greenhouse-Geisser	0.133	1.558	0.085	0.186	0.777	0.000
	Huynh-Feldt	0.133	1.696	0.078	0.186	0.796	0.000
QS * final_familiar	None	0.542	2.000	0.271	0.761	0.475	0.000
	Greenhouse-Geisser	0.542	1.558	0.348	0.761	0.447	0.000
	Huynh-Feldt	0.542	1.696	0.319	0.761	0.457	0.000
QS * final_anthropomorphism	None	0.653	2.000	0.327	0.918	0.410	0.000
	Greenhouse-Geisser	0.653	1.558	0.419	0.918	0.390	0.000
	Huynh-Feldt	0.653	1.696	0.385	0.918	0.397	0.000
QS * gender	None	0.134	2.000	0.067	0.189	0.829	0.000
	Greenhouse-Geisser	0.134	1.558	0.086	0.189	0.775	0.000
	Huynh-Feldt	0.134	1.696	0.079	0.189	0.794	0.000
QS * age	None	1.430	4.000	0.358	1.005	0.419	1.325×10^{-4}
	Greenhouse-Geisser	1.430	3.116	0.459	1.005	0.409	1.325×10^{-4}
	Huynh-Feldt	1.430	3.391	0.422	1.005	0.413	1.325×10^{-4}
QS * condition * gender	None	0.166	2.000	0.083	0.233	0.793	0.000
	Greenhouse-Geisser	0.166	1.558	0.107	0.233	0.738	0.000
	Huynh-Feldt	0.166	1.696	0.098	0.233	0.757	0.000
QS * condition * age	None	0.331	4.000	0.083	0.232	0.918	0.000
	Greenhouse-Geisser	0.331	3.116	0.106	0.232	0.879	0.000
	Huynh-Feldt	0.331	3.391	0.097	0.232	0.893	0.000
QS * gender * age	None	0.230	4.000	0.058	0.162	0.956	0.000
	Greenhouse-Geisser	0.230	3.116	0.074	0.162	0.926	0.000
	Huynh-Feldt	0.230	3.391	0.068	0.162	0.937	0.000
QS * condition * gender * age	None	0.681	4.000	0.170	0.478	0.751	0.000
	Greenhouse-Geisser	0.681	3.116	0.218	0.478	0.707	0.000
	Huynh-Feldt	0.681	3.391	0.201	0.478	0.722	0.000
Residuals	None	11.385	32.000	0.356			
	Greenhouse-Geisser	11.385	24.929	0.457			
	Huynh-Feldt	11.385	27.129	0.420			

Table 16: Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2	$\eta^2{}_G$	ω^2
condition	5.537	1.000	5.537	4.718	0.040	0.110	0.000	0.348
final_trust	0.376	1	0.376	0.107	0.748	0.003	0.005	0.000
final_familiar	1.999	1	1.999	0.568	0.462	0.015	0.029	0.000
age	9.459	2	4.729	1.344	0.289	0.069	0.122	0.013
gender	0.340	1	0.340	0.097	0.760	0.002	0.005	0.000
condition * age	6.895	2	3.447	0.979	0.397	0.051	0.092	0.000
condition * gender	1.085	1	1.085	0.308	0.586	0.008	0.016	0.000
age * gender	6.734	2	3.367	0.957	0.405	0.049	0.090	0.000
condition * age * gender	6.593	2	3.296	0.936	0.412	0.048	0.088	0.000
Residuals	56.319	16	3.520					

Table 17: Simple Main Effects - condition

Level of Question Sensitivity	Sum of Squares	df	Mean Square	F	p
Low	5.200	1	5.200	4.458	0.045
Medium	7.136	1	7.136	5.399	0.028
High	7.248	1	7.248	4.237	0.050

G Perceived Question Sensitivity and Willingness to Self-disclose

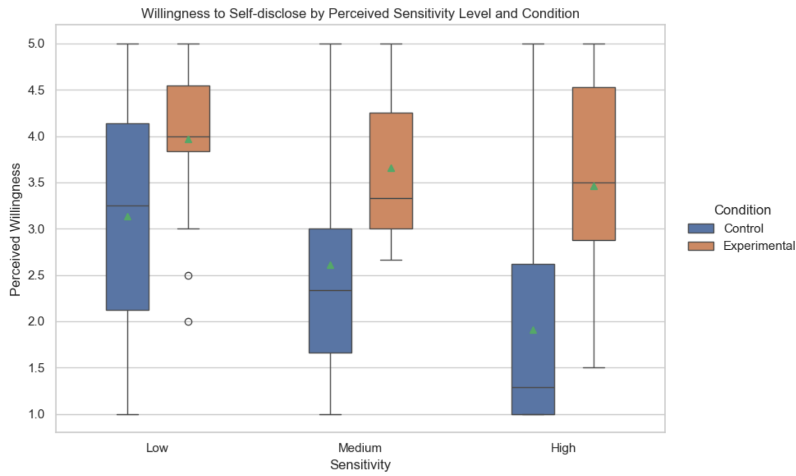


Figure 9: Willingness to Self-disclose (by Condition and User Perceived Question Sensitivity)

Table 18: User Perceived Sensitivity Descriptive Statistics

	perceived_willingness_low		perceived_willingness_medium		perceived_willingness_high	
	Control	Experimental	Control	Experimental	Control	Experimental
Valid	14	15	13	15	10	12
Missing	1	0	2	0	5	3
Mean	3.139	3.971	2.612	3.663	1.908	3.460
Std. Deviation	1.300	0.865	1.277	0.795	1.341	1.172
Minimum	1.000	2.000	1.000	2.667	1.000	1.500
Maximum	5.000	5.000	5.000	5.000	5.000	5.000

H Large Language Model Prompts

Below are examples of prompts used with large language models (LLMs) such as ChatGPT during manuscript preparation, strictly for LaTeXformatting and syntax support:

- **Figure placement:**
  "What is the best way to place figures side-by-side (vertically) in LaTeXwith captions? Please provide sample code."
- **Resolving package conflicts:**
  "I'm getting a compilation error due to a LaTeXError: Command `iint` already defined.
- **Creating framed boxes:**
  "How do I create a framed box around a theorem statement in LaTeX? Provide code using the `framed` package."
- **Formatting references:**
  "How can I customize citation style in LaTeXusing natbib to numbers?"

All outputs from LLMs were used solely for formatting guidance and were carefully reviewed before incorporation.