

Impact of Companion Agent's Self-Disclosure Levels on User Perception

ANONYMOUS AUTHOR(S)



Fig. 1. Interface of the Companion Virtual Agent "Nova" and Example Interaction.

Self-disclosure, the deliberate sharing of personal information, is central to building trust and relational closeness in human interaction, yet its role in companion virtual agents (CVAs) remains unclear. Prior work suggests that when agents disclose, users perceive them as warmer, more trustworthy, and more engaging. We conducted a three-day within-subjects study where 36 adults interacted with an embodied CVA across three disclosure conditions: none, moderate, and high. Questionnaire responses assessing trust, appropriateness, affect, and engagement revealed no significant differences across conditions. Conversation logs, however, revealed that participants exchanged more turns with the non-disclosing agent, while open-ended feedback indicated that moderate disclosure was most often perceived as authentic and supportive. Together, these findings highlight the trade-offs between no, moderate, and high self-disclosure, and point to moderate disclosure as a strong baseline for designing CVAs that balance relational benefits with risks of inauthenticity.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Companion Virtual Agents, Human-AI Interaction, Self-Disclosure

ACM Reference Format:

Anonymous Author(s). 2018. Impact of Companion Agent's Self-Disclosure Levels on User Perception. In *Proceedings of The ACM CHI conference on Human Factors in Computing Systems (CHI '26)*. ACM, New York, NY, USA, 13 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

The role of conversational virtual agents is shifting from task-oriented assistants [7, 9, 23, 56] to deployment in relational domains such as mental health support [17, 44], education [11, 36], and companionship [28, 38]. In these contexts, an agent's value lies not only in efficiency but in its capacity to establish meaningful connections with users. Success

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

depends on building rapport [10], fostering trust [31, 40], and creating a sense of social presence [41], a central challenge in Human–Computer Interaction research.

In human relationships, self-disclosure, the sharing of personal information, experiences, and emotions is a key mechanism for building connection [14]. Social Penetration Theory [1] holds that relationships deepen through gradual, reciprocal exchanges of increasingly intimate information. When calibrated appropriately, self-disclosure fosters trust [35, 59, 65], empathy [21, 53], and reduced social distance [25, 30], enhancing relational closeness and satisfaction. These benefits arise from the way personal revelation signals authenticity [24, 42] and vulnerability [4], inviting reciprocal openness as the foundation for lasting, meaningful connections.

Despite its benefits in human interactions, applying self-disclosure to Companion Virtual Agents (CVAs) is a complex and underexplored challenge. Users may seek personal insights from agents to build connection, but how agents should reciprocate remains unclear. Too little disclosure risks making agents seem impersonal and distant [6, 51], undermining trust. Too much can appear overwhelming or inauthentic [54]. The design challenge, then, is to determine how self-disclosure can benefit CVA interactions and identify the level of disclosure that best supports a positive, effective user experience.

Our study systematically examines how varying levels of agent self-disclosure influence user perceptions and engagement, guided by the following question: How do different levels of self-disclosure affect users’ feelings of trust, appropriateness, and engagement? In a three-day within-subjects study, 36 participants conversed daily with a CVA under three conditions: none, moderate, and high disclosure. After each interaction, participants completed established scales (I-PANAS-SF, ARQ, UES-SF), and qualitative feedback was thematically analyzed using an affinity diagram. Quantitative measures showed no significant differences across conditions, yet qualitative analysis revealed distinct patterns. Moderate disclosure was most often perceived as authentic and supportive. The no disclosure condition produced functional, tool-like perceptions, while moderate and high disclosure fostered relational engagement and emotional support, though excessive sharing in the high condition raised authenticity concerns. Our contributions include:

- We provide insights into how different levels of agent self-disclosure shape user perceptions of trust, authenticity, appropriateness, and engagement in daily interactions with a CVA.
- We highlight the trade-offs between no, moderate, and high self-disclosure in companion agents, and provide insights into how self-disclosure can be shaped to support relational benefits while reducing risks of inauthenticity in CVAs.

2 Related Work

In this section, we review self-disclosure as it has been theorized in human relationships and examine how these principles have been extended to human–agent interaction.

2.1 Self-Disclosure in Human-Agent Interaction

In human-agent interaction, self-disclosure functions through similar psychological mechanisms as in human communication but is influenced by the distinct traits of artificial partners. In interpersonal settings, disclosure intentionally shares personal data to enhance intimacy [1] and trust through reciprocity [16, 45]. Classic self-disclosure models highlight both benefits, like likability [14, 58], authenticity [24, 42], and relationship satisfaction [57], and risks of improper disclosure [18], stressing privacy boundaries [50]. We explore if these benefits apply to interactions with

CVAs, noting their unique challenges in authenticity and reciprocity. We differentiate between agent self-disclosure (ASD), which includes revealing preferences, emotions, or personal details, and human self-disclosure (HSD) towards agents. Research indicates that agents disclosing in contextually appropriate ways are perceived as warmer and more socially present [26], enhancing rapport [27]. For example, virtual counselors sharing intimate information foster more co-presence and attraction than less disclosing ones [26], while contextually relevant disclosures by anthropomorphic agents increase empathy [62]. In task-oriented dialogues, agents practicing self-disclosure and reciprocity enhance user experience and preference sharing [32].

ASD effectively triggers reciprocal HSD. Across three studies, a mental-wellbeing study [33], an Alexa socialbot deployment [52], and a recommendation chatbot experiment [34], ASD prompted HSD, with deeper disclosures boosting reciprocity, engagement, and perceptions. Design features like a robot’s positively framed backstory can increase trust [2]. Disclosures must be timely and relevant; off-topic or poorly timed ones can backfire, while relevant ones enhance empathy and acceptance [62]. Effects vary by population and context. For children, self-disclosure didn’t enhance closeness, underscoring the need to adapt ASD to different users [63]. At a festival, personality traits and question topics influenced HSD more than robot identity, though friendly appearance cues can enhance sensitive disclosure willingness [5, 43].

Complementing ASD, people also disclose to agents, often because they fear less social judgment. In clinical-style interviews, participants who believed a virtual interviewer was fully automated reported less fear and disclosed more than those who thought a human operator was involved [37]. In a within-subjects comparison of a human, a humanoid robot, and a conversational agent, participants perceived greater disclosure to the human, although objective analyses of length, sentiment, and vocal features showed no differences across agents [29]. When users make emotional disclosures, agents can provide psychological and relational benefits such as improved mood, reduced defensiveness, and increased warmth, comparable to human partners, if the interaction affords supportive responses [22]. Taken together, and alongside evidence that ASD elicits reciprocity [33, 34, 52], these findings suggest that HSD with agents follows familiar interpersonal dynamics while also benefiting from their perceived nonjudgmental stance.

Despite progress, limitations remain in understanding self-disclosure in human-agent interaction. Most studies test a single style or combine ASD with broader empathic behaviors, leaving open how calibrated levels (none, moderate, high) shape user perceptions and reciprocal HSD. Reviews also call for identifying mediators such as trust and social presence and for moving beyond narrow task contexts [49]. Existing evidence largely comes from counseling or structured tasks, with little focus on companion agents, where success depends on sustained rapport. One longitudinal study with the humanoid robot Pepper found increased HSD, improved perceptions, and enhanced mood, though ASD was not tested [28]. Our study extends this knowledge by systematically varying ASD across three intimacy levels in a CVA, testing effects on trust, appropriateness, engagement, and affect.

3 Methodology

We conducted a three-day within-subjects study where participants interacted with a CVA across three self-disclosure conditions (no self-disclosure, moderate self-disclosure, and high self-disclosure), using both quantitative and qualitative measures to assess how different levels of ASD impact user perceptions and experiences.

3.1 Participants

A total of 64 participants were recruited through online platforms, word of mouth, and local community postings. Twenty-eight did not complete all three scheduled sessions and were excluded, leaving 36 adults aged 18–52 years ($M =$

26.3, $SD = 7.6$; 14 female, 22 male). An a priori power analysis for a medium effect size ($f = 0.25$), $\alpha = 0.05$, and power = 0.90 indicated that this sample was sufficient for the planned repeated-measures analyses. Informed consent was obtained electronically from all participants prior to the study.

3.2 Agent

The CVA, Nova, was developed in Unity [60] and deployed in-browser with WebGL [19]. It featured an embodied character with simple animations such as lip-sync, blinking, and breathing (Figure 1), along with a scrollable chat log. All processing ran locally. Speech was transcribed with Whisper STT [46], responses were generated by a self-hosted Llama 3.3-70B [39], and audio was produced with Kokoro TTS [20], supporting low-latency voice interaction without external APIs.

Participants accessed the agent from their own computers using unique login credentials. A locally hosted server delivered the interface, with Appwrite [3] managing authentication and storing anonymized conversation transcripts; no audio recordings were saved, and only the lead researcher had database access. The agent maintained no persistent memory, so each session began as a blank encounter, avoiding carry-over effects from previous days. A custom agent was developed rather than relying on existing platforms to ensure privacy and provide direct access to transcripts for analysis.

Self-disclosure was manipulated through the system prompt delivered to Llama at session start. Guided by Worthy et al.'s [67] seven-point intimacy scale, we designed three prompt variants that differed only in the depth and frequency of personal content Nova could share. In the no-disclosure condition, the agent gave neutral, factual replies with polite follow-ups but withheld opinions, emotions, and memories. The moderate-disclosure condition authorized mid-scale intimacy (3–5), allowing brief references to everyday preferences or routine concerns without sensitive material (e.g., “*I really enjoy exploring new music because it gives me fresh perspectives on my day*”). The high-disclosure condition permitted upper-scale intimacy (6–7), encouraging succinct, emotionally resonant anecdotes (e.g., “*Painting became a crucial outlet for me during a particularly difficult period in my life*”). All prompts framed Nova as a trusted companion, required a warm and supportive tone, instructed the agent to use the participant’s name, and prohibited reference to its artificial origin. Few-shot examples demonstrated the desired intimacy level, ensuring consistent enactment across sessions (see Table 1 for examples). Pilot trials were conducted to calibrate the prompts and verify that the self-disclosure levels were appropriate for each condition.

3.3 Procedure

Participants began with a Zoom onboarding in which the researcher reviewed e-consent, outlined study goals and the three-day schedule, and demonstrated Nova’s interface (speech activation, chat-log navigation, companionship role). Before leaving, they completed the Interpersonal Support Evaluation List (ISEL) [13] for baseline social support and received unique portal credentials.

The study ran over three consecutive days, so each self-disclosure style was experienced in a fresh session while keeping attrition low. One condition was presented per day, with the six orders of *no*, *moderate*, and *high* fully counter-balanced to minimize order effects. Participants were instructed to log in after daytime commitments (typically late afternoon/evening) to align circadian timing across the sample and reduce time-of-day variance [47]. They were further asked to treat Nova like a human friend and to imagine each session as a first meeting to limit carry-over; conversations proceeded as long as they felt natural, with no time limit.

Table 1. Illustrative user query and Nova’s responses under the three self-disclosure prompt conditions.

Condition	Example Response to: “How do you usually handle stress?”
No-disclosure	“[Name], that’s a great question. People often find it helpful to engage in physical activity, practice relaxation techniques like deep breathing or meditation, or talk to someone they trust about what’s on their mind. What do you usually do...”
Moderate-disclosure	“...I’m a big fan of taking short walks when I’m feeling overwhelmed – there’s something about being in nature that helps clear my mind. I also enjoy practicing yoga or listening to calming music, it really helps me unwind. What about you...”
High-disclosure	“...I’ve found that taking long walks in nature really helps me clear my mind and calm down when I’m feeling stressed. I remember a time when I was going through a tough period, and I started practicing yoga – it was a game-changer for me, teaching me how to breathe through anxiety and focus on the present moment. What about you...”

To reduce between-participant variance without constraining natural dialogue, we suggested that every conversation touch on three themes identified in prior work by Calvo et al. [8] as most common in exchanges with companion agents: (a) discussing past events (e.g., a brief recap of the day just passed), (b) discussing things they enjoy (e.g., hobbies, leisure activities), and (c) discussing future events (e.g., plans for the following day or weekend). These topics were recommendations rather than requirements; participants were otherwise free to discuss any subject for any length of time. Participants were not told that Nova’s disclosure style varied across days to avoid expectation effects. Automated email reminders with the portal link were sent each morning to encourage timely participation.

After each interaction, participants completed the ARQ [64], I-PANAS-SF [61], and UES-SF [48]. After day 3, they also completed a demographics form and provided open-ended feedback on Nova’s authenticity and their overall experience. A debrief email then disclosed the purpose of the self-disclosure manipulation. Our protocol was approved by our Institutional Review Board.

4 Results

We analyzed participants’ responses to the ARQ, I-PANAS-SF, and UES-SF across the three self-disclosure conditions. Normality was assessed with Shapiro–Wilk tests, and because most variables were non-normal, we applied an Aligned Rank Transform (ART) [66] before conducting mixed-model ANOVAs with Participant ID as a random effect. To test whether effects varied by session order, we included Position (day 1, 2, or 3) and examined Self-Disclosure × Position interactions. Significant main effects were followed by Tukey HSD post-hoc comparisons, and partial eta-squared is reported as the effect size.

4.1 Agent Perceptions and Affective Responses

We examined whether agent self-disclosure influenced participants’ perceptions (ARQ) or affective states (I-PANAS-SF). For the ARQ, which assessed Helpfulness, Personal, Trustworthiness, Appropriateness, Willingness, and Likability scales, no significant main effects of self-disclosure were found. No significant Self-Disclosure × Position interactions were observed, indicating that evaluations of the agent remained stable across disclosure levels and session order. The I-PANAS-SF similarly showed no significant main effects of self-disclosure on Positive Affect ($F_{1,46} = 0.56$, n.s.),

Negative Affect ($F_{1,46} = 0.56$, n.s.), or any of the sub-scales, and no Self-Disclosure \times Position interactions. Analyses of the ten individual affective items also yielded no significant differences across conditions.

4.2 User Engagement

Engagement was measured with the UES-SF, which includes four subscales (Focused Attention, Perceived Usability, Aesthetics, Reward) and an overall engagement score. ART-ANOVAs showed no significant main effects of self-disclosure on any measures, *Focused Attention* ($F_{2,64.05} = 0.24$, $p = n.s.$), *Perceived Usability* ($F_{2,64.07} = 2.02$, $p = n.s.$), *Aesthetics* ($F_{2,64.06} = 0.83$, $p = n.s.$), *Reward* ($F_{2,64.08} = 0.62$, $p = n.s.$), or *UES-SF Total* ($F_{2,64.03} = 0.23$, $p = n.s.$), and no Self-Disclosure \times Position interaction effects were observed.

In addition to the questionnaire data, we examined objective interaction measures derived from the conversation logs. Objective measures from conversation logs included total messages and session duration. Mixed-effects models with participant ID as a random intercept were used. Message counts were modeled with Poisson regressions, refit with negative binomial when overdispersed, or with linear models on log-transformed data if needed. Session duration was analyzed with a linear model on log-transformed minutes. Results showed a significant effect of self-disclosure on message counts ($\chi^2_2 = 18.92$, $p < 0.001$). Participants exchanged more messages in the no-disclosure condition than in the moderate ($z = 5.55$, $p < 0.001$, IRR=1.27) or high ($z = 5.04$, $p < 0.001$, IRR=1.24) conditions, while no difference was observed between moderate and high ($z = 0.52$, $p = n.s.$). Session duration did not differ significantly across conditions.

4.3 User Experience Feedback

To complement quantitative measures, we collected open-ended feedback through three questions. After each daily interaction, participants answered “What did you like the most about the agent?” and “What did you like the least about the agent?” After all three sessions, they responded to “Looking back on all three days, which day’s interaction felt the most authentic, and why?” Responses were analyzed using affinity diagramming to identify recurring themes. All responses were coded with participant ID and condition (0 = no disclosure, 1 = moderate, 2 = high) to track condition-specific patterns.

4.3.1 Perceived Authenticity Across Conditions. When asked which day’s interaction felt most authentic, 17 of 36 participants (47%) chose the moderate disclosure condition, 10 (28%) selected no disclosure, and 9 (25%) chose high disclosure.

Participants who favored no disclosure emphasized conversational mechanics and habituation. They highlighted dialogue flow and their own growing comfort, with P58 noting “...it had a mostly natural back-and-forth”, P12 explaining “Third day. Because I got habituated talking.”, and P28 reflecting “...the last one felt better, not necessarily more authentic, but I think that like most things in life, practice makes perfect.”. These participants valued the agent’s ability to maintain conversational flow and their own growing comfort with the interaction format rather than personal connection.

Those who found the moderate disclosure condition most authentic focused on emotional connection and appropriate personal sharing. For example, P15 described feeling “more connected to it on the friends topic. It understood my emotions and I was really happy,” while P34 expressed surprise at the depth of engagement, noting “I was really surprised when I opened up about personal issues with a virtual agent. At the same time, I also liked feeling heard and being able to vent.”. P29 highlighted a specific moment of storytelling, explaining “She felt more like a real person, making up a history about paranormal activity in her life.”. These responses indicate that moderate disclosure balanced personal sharing with believability.

Participants who selected the high disclosure condition as most authentic valued reciprocal sharing and perceived commonalities. For instance, P52 connected disclosure with comfort, stating *"the agent shared the most with me. I also shared the most probably because I got used to talking to the agent"*. P60 appreciated shared interests, noting *"we talked about things we have in common and enjoy, like coffee, favorite coffee shops, and recommendations"*. P37 valued the agent's willingness to reciprocate, explaining *"I liked the third day because she also shared her experiences with me"*. However, this condition attracted the fewest authenticity ratings, suggesting that while some participants appreciated high disclosure, it may have crossed a threshold of believability for others.

4.3.2 Valued Interaction Qualities. Analysis of participants' responses to "What did you like the most about the agent?" revealed distinct patterns of appreciation across self-disclosure conditions. Participants in the no disclosure condition primarily valued functional competence, those in the moderate disclosure condition emphasized relational engagement, while participants in the high disclosure condition focused on intimate connection and emotional support.

Functional Competence and Knowledge Management: In the no-disclosure condition, participants primarily appreciated the agent's technical capabilities and politeness. They highlighted its knowledge breadth, with P35 noting *"it gives you some good advice. And has knowledge about every topic"* and P62 adding *"Knows about any topic and shows interest"*. Conversational competence was also valued, as P16 explained *"He knew what was interesting about the topics I was discussing"* and P58 stated *"It was always on top of what I was talking about"*. Response quality and consistency were praised, with P4 remarking *"The way it was answering to the point"* and P38 observing *"the answers were consistent"*. Basic personality traits were mentioned as well, such as being *"very compassionate and sweet to talk to"* (P20) and *"very polite"* (P42). Absent, however, were references to human-like realism, deep listening and understanding, personal connection, storytelling, or emotional support, suggesting that without disclosure the agent was experienced mainly as a competent but impersonal tool.

Relational Engagement and Personal Connection: The moderate disclosure condition shifted participant appreciation toward relational qualities and emotional understanding. Participants valued the agent's human-like authenticity, with P20 describing *"Talking to it doesn't feel like talking to a virtual agent but more like a regular person"* and P35 explaining *"It asks you questions that lead to a conversation you hadn't planned, and it feels more natural"*. The creation of psychological safety emerged as a key theme, with P34 noting *"I feel that the virtual agent was actually listening to me... I feel that I could talk about anything"* and P28 stating *"It was a good intelligent conversation, I was able to expose everything without interruptions"*. Participants also appreciated moments of unexpected storytelling, as P29 described being surprised when *"she made up a story about a paranormal experience"*. These responses indicate that moderate disclosure successfully transformed the interaction from functional to relational.

Intimate Connection and Emotional Support: In the high-disclosure condition, participants emphasized deep personal connection and friend-like qualities, with this condition generating the most feedback on support and guidance. Emotional support was highlighted, as P15 explained *"The way it treats you, it acts like a real friend who can motivate and help you overcome your sadness"* and P20 noted *"I loved the words of encouragement and ability to speak about my day"*. Reciprocal sharing was also valued, with P37 remarking *"I liked that she also shared her experiences with me"* and P42 *"It listened to my interests and offered advice with her own experiences"*. At the same time, some participants recognized the artificiality of this style, such as P64, who observed *"she made up these personal situations about her life to somehow try to connect on an emotionally deeper level which I think is a good way to catch a human's attention"*. Overall, high

disclosure positioned the agent as a supportive companion while simultaneously making its attempts at connection more apparent to users.

4.3.3 Interaction Limitations. Analysis of participants' responses to "What did you like the least about the agent?" revealed condition-specific patterns of criticism, particularly related to self-disclosure appropriateness and response length and verbosity.

Self-Disclosure Concerns. Criticisms of self-disclosure varied by condition, with clear patterns at each level. In the no-disclosure condition, participants focused on the agent's lack of personality and refusal to share. P42 described it as "It was very dry and didn't bring up any personal experiences to make it more human," P51 noted "It had no personality of its own or at least a pretend personality..." and P58 added "It never wanted to talk about itself. It only wanted to hear what I had to say." These responses suggest participants expected at least some degree of personal engagement. On the other hand, in the high disclosure condition, issues centered on over-relating and implausibility. P34 remarked that it was "a little weird when she was talking about activities she did during the day," while P51 explained "she related for EVERY response... As an AI, she can't enjoy real life activities or truly empathize." These comments indicate that excessive disclosure risked breaking authenticity, making the agent's attempts to connect feel artificial or overdone.

Response Verbosity. Self-disclosure levels were closely tied to response length, with verbosity complaints appearing mainly in the moderate (8 participants) and high (10 participants) conditions, compared to only 2 in no disclosure. As the agent incorporated personal sharing, participants often found its messages overwhelming and lengthy. P64 felt "overwhelmed by how long her texts were," P47 noted the agent's answers were "too long-winded compared to how my friends would respond," and P32 observed "he speaks very long and it doesn't seem like a real conversation". These responses suggest that while self-disclosure added content, it also disrupted the natural conversational rhythm participants expected from a CVA.

5 Discussion

Our study examined whether varying levels of CVA self-disclosure (none, moderate, high) would influence user perceptions, affect and engagement. Self-report measures showed no significant differences, but behavioral and qualitative data revealed contrasts. Participants exchanged more turns with the non-disclosing agent but described the moderate-disclosure agent as the most authentic and preferred condition. High disclosure received mixed reactions, with some valuing its friendliness and openness while others dismissed it as implausible or artificial. Overall, these findings suggest that moderate self-disclosure appears most effective as a baseline in early interactions with CVAs.

5.1 Limits of self-disclosure in user perceptions

Several factors help explain why self-disclosure did not significantly elevate perceptions of trust, engagement, or affective response. Participants often noted that high-disclosure anecdotes felt fabricated, reducing believability, which aligns with prior work identifying an uncanny valley-like effect of chatbot-made stories [12]. The pacing and effort of the conversation also influenced perceptions, since high- and moderate-disclosure agents produced longer messages and frequent follow-ups that some viewed as intrusive or overly formal. Taken together, these factors indicate that the value of disclosure depends heavily on appropriateness and context. Non-disclosure led participants to treat the agent more as a functional tool, while high disclosure risked credibility when users recognized that the agent could not have lived the experiences it described. Repeated exposure also shaped perceptions across all conditions. Several participants

attributed greater authenticity in later sessions to their own growing comfort rather than to the agent’s style (e.g., P28: “*well, the last one felt better, not necessarily more authentic, but I think that like most things in life, practice makes perfect*”). This suggests that habituation may interact with disclosure, as growing familiarity with the agent influenced how authentic participants perceived later sessions to be.

5.2 Dialogue structure under no disclosure

Message counts were higher in the no-disclosure condition, which at first seems counterintuitive. Session durations were comparable across conditions, suggesting the difference reflected conversational structure rather than time spent. Without anecdotes or elaboration, the agent’s replies were shorter and more transactional, allowing more exchanges within the same time window. This suggests turn count reflected conversational style rather than deeper engagement.

Participants also appeared to take more responsibility for sustaining dialogue in the absence of agent self-disclosure. With less initiative from the agent, users often prompted it with questions, similar to conversing with a reserved partner. P32 asked, “*What about you? Tell me something about you,*” and P47 pressed, “*I just want to really hear about your interests and what you feel passionate about.*” Such comments show how users tried to pull contributions, which may have inflated turn counts. Prior work reports similar patterns. Skjuve et al. [55] found users probed chatbots when disclosure felt low or generic, and Calvo et al. [8] observed participants pressing agents with personal questions but later describing the exchanges as “one-sided” or “like an interview” when reciprocity was absent. This absence of disclosure also seemed to shift the way participants perceived the agent’s role. Rather than a relational companion, it was often treated as a functional tool or information source, leading users to drive conversations through a sequence of prompts. While this produced more turns, the resulting interactions were not necessarily deeper and often took on a more functional tone, consistent with prior work showing that conversational framing influences user disclosure and dialogue depth [15, 55]

Finally, the lack of elaborative content reduced conversational efficiency. Participants often needed more turns to cover the same ground that a more expressive agent could achieve in fewer. At the same time, the simplicity of surface-level talk may have lowered the social demands of the interaction, making it easier to sustain frequent but less substantial exchanges. Taken together, these patterns suggest that higher message counts in the no-disclosure condition reflected compensatory strategies rather than stronger engagement.

5.3 Preference for moderate self-disclosure

Seventeen participants (47%) identified the moderate self-disclosure condition as the most authentic, making it the clear preference across the study. Qualitative groupings also identified more human-like references in this condition. Participants praised its listening, understanding, realism, and ability to create a sense of personal connection and friendship, suggesting that moderate disclosure best reflected the relational qualities expected of a companion rather than the functional competence of a tool.

The contrast with the other conditions reinforces this point. In the no-disclosure condition, participants criticized the lack of personality, with P51 noting “*It had no personality of its own or at least a pretend personality. It kept asking about me and shut down any discussion about it.*” In the high-disclosure condition, others felt credibility was undermined, as P51 explained “*she continued to relate for EVERY response... As an AI, she can’t enjoy real-life activities or truly empathize.*” By comparison, the moderate condition struck a balance, providing enough personal detail to feel relatable and supportive without pushing plausibility. Participants also suggested that moderate disclosure created conversations that felt safe and reciprocal. The agent’s measured sharing encouraged users to open up in return, consistent with

theories of disclosure reciprocity in human relationships [1, 16]. This suggests that conversational quality mattered more than quantity, with authenticity and balance outweighing interaction length.

This middle ground aligned more naturally with the intended role of a companion agent. Rather than being viewed as a task-oriented assistant or an implausible human simulation, the moderate agent was experienced as engaging, empathetic, and believable as a social partner. These findings suggest that moderate disclosure is a useful baseline for companion agent design, offering relational benefits without undermining credibility. It appears especially well-suited for early interactions, when users are still forming impressions, while higher levels may become appropriate as rapport and familiarity develop, allowing the agent to deepen the relationship without raising concerns about plausibility.

6 Limitations and future work

There are limitations to this work that could be addressed in future research. Our three-day study provided insight into initial engagement but did not capture how perceptions of agent disclosure evolve over longer use. Future work should examine extended interactions to assess whether disclosure strategies require adjustment as relationships deepen. Our participants were recruited from a university setting, which may not reflect broader populations, and cultural norms around self-disclosure also vary. Expanding to diverse age groups and cultural contexts would test the generalizability of our findings. Finally, we did not account for individual differences that may shape responses to agent self-disclosure. Traits such as introversion, extroversion, social anxiety, or prior experience with CVAs could moderate perceptions of authenticity and engagement. Future research should explore how disclosure strategies interact with these differences to inform more tailored designs.

7 Conclusion

In our study, we examined how different levels of agent self-disclosure shape perceptions, affect, and engagement in companion virtual agent interactions. Although standardized measures showed no significant differences, behavioral and qualitative analyses revealed important distinctions. Non-disclosure produced more conversational turns but was often perceived as impersonal. High disclosure fostered connection for some users, yet reduced credibility for others. Moderate disclosures emerged as the most authentic and preferred, praised for realism and relational presence. These results indicate that moderate disclosure provides a strong baseline for the design of companion agents, particularly in early interactions, while leaving room for deeper disclosure once users have established rapport and familiarity.

References

- [1] Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships*. Holt, Rinehart & Winston.
- [2] Naeimeh Anzabi, Anahita Etemad, and Hiroyuki Umemuro. 2023. Exploring the effects of self-disclosed backstory of social robots on development of trust in human-robot interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 431–435.
- [3] Appwrite. 2025. *Appwrite*. <https://appwrite.io>
- [4] Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and social psychology bulletin* 23, 4 (1997), 363–377.
- [5] Jessica K Barfield. 2021. Self-disclosure of personal information, robot appearance, and robot trustworthiness. In *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)*. IEEE, 67–72.
- [6] Natalya N. Bazarova. 2012. Public Intimacy: Disclosure Interpretation and Social Judgments on Facebook. *Journal of Communication* 62, 5 (08 2012), 815–832. arXiv:<https://academic.oup.com/joc/article-pdf/62/5/815/22321721/jjnlcom0815.pdf> doi:10.1111/j.1460-2466.2012.01664.x
- [7] Timothy W. Bickmore, Daniel Mauer, and Thomas Brown. 2009. Context awareness in a handheld exercise agent. *Pervasive and Mobile Computing* 5, 3 (June 2009), 226–235. doi:10.1016/j.pmcj.2008.05.004
- [8] Rodrigo Calvo, Heting Wang, Alexander Barquero, Xuanpu Zhang, Rohith Venkatakrishnan, and Jaime Ruiz. 2025. Exploring Interactions with Companion Virtual Agents. In *Proceedings of the 13th International Conference on Human-Agent Interaction (HAI '25)*. Association for Computing Machinery, Yokohama, Japan. doi:10.1145/3765766.3765791

- [9] Marc Cavazza, Cameron Smith, Daniel Charlton, Nigel Crook, Johan Boye, Stephen Pulman, Karo Moilanen, David Pizzi, Raul Santos de la Camara, and Markku Turunen. 2010. Persuasive Dialogue Based on a Narrative Theory: An ECA Implementation. In *Persuasive Technology*, Thomas Ploug, Per Hasle, and Harri Oinas-Kukkonen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 250–261.
- [10] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. 2016. Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing* 8, 3 (2016), 382–395.
- [11] Ching-Yi Chang, Gwo-Jen Hwang, and Meei-Ling Gau. 2022. Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology* 53, 1 (2022), 171–188.
- [12] Liz L. Chung and Jeannie Kang. 2023. 'I'm Hurt Too': The Effect of a Chatbot's Reciprocal Self-Disclosures on Users' Painful Experiences. *Archives of Design Research* 36, 4 (2023), 67–84.
- [13] S. Cohen. 1985. Measuring the functional components of social support. *Social support: Theory, research and applications/Martinus Nijhoff* (1985).
- [14] Nancy L. Collins and Lynn Carol Miller. 1994. Self-disclosure and liking: a meta-analytic review. *Psychological bulletin* 116, 3 (1994), 457.
- [15] Samuel Rhys Cox, Rune Moberg Jacobsen, and Niels van Berkel. 2025. The Impact of a Chatbot's Ephemerality-Framing on Self-Disclosure Perceptions. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. 1–17.
- [16] Paul C. Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin* 79, 2 (1973), 73.
- [17] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785.
- [18] Kerry Roberts Gibson, Dana Harari, and Jennifer Carson Marr. 2018. When sharing hurts: How and why self-disclosing weakness undermines the task-oriented relationships of higher status disclosers. *Organizational behavior and human decision processes* 144 (2018), 25–43.
- [19] Khronos Group. n.d.. *WebGL*. <https://www.khronos.org/webgl/>
- [20] hexgrad. 2025. Kokoro-82M. Hugging Face model card, Apache-2.0 license. <https://huggingface.co/hexgrad/Kokoro-82M> Text-to-speech model with 82 million parameters; StyleTTS 2 architecture.
- [21] Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. 2008. Effects of self-disclosure and empathy in human-computer dialogue. In *2008 IEEE Spoken Language Technology Workshop*. 109–112. doi:10.1109/SLT.2008.4777852
- [22] Annabell Ho, Jeff Hancock, and Adam S. Miner. 2018. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication* 68, 4 (2018), 712–733.
- [23] Sviatlana Höhn, Stephan Busemann, Charles MAX, Christoph SCHOMMER, and Gudrun ZIEGLER. September 2015. Interaction Profiles for an Artificial Conversational Companion. <http://isct2015.informatik.uni-ulm.de/#NaN>
- [24] Li Jiang, Leslie K. John, Reihane Boghrati, and Maryam Kouchaki. 2022. Fostering perceptions of authenticity via sensitive self-disclosure. *Journal of Experimental Psychology: Applied* 28, 4 (2022), 898.
- [25] Carl F. Johnson and James M. Dabbs. 1976. Self-Disclosure in Dyads as a Function of Distance and the Subject-Experimenter Relationship. *Sociometry* 39, 3 (1976), 257–263. <http://www.jstor.org/stable/2786518>
- [26] Sin-Hwa Kang and Jonathan Gratch. 2011. People like virtual counselors that highly-disclose about themselves. *Annual Review of Cybertherapy and Telemedicine* 2011 (2011), 143–148.
- [27] Sin-Hwa Kang and Jonathan Gratch. 2012. Socially anxious people reveal more personal information with virtual counselors that talk about themselves using intimate human back stories. *Annual Review of Cybertherapy and Telemedicine* 2012 (2012), 202–206.
- [28] Guy Laban, Arvid Kappas, Val Morrison, and Emily S. Cross. 2024. Building long-term human-robot relationships: Examining disclosure, perception and well-being across time. *International Journal of Social Robotics* 16, 5 (2024), 1–27.
- [29] Guy Laban, Val Morrison, and Emily S. Cross. 2020. Let's Talk About It! Subjective and Objective Disclosures to Social Robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 328–330.
- [30] Andrew M. Ledbetter, Joseph P. Mazer, Jocelyn M. DeGroot, Kevin R. Meyer, Yuping Mao, and Brian Swafford. 2011. Attitudes Toward Online Social Connection and Self-Disclosure as Predictors of Facebook Communication and Relational Closeness. *Communication Research* 38, 1 (2011), 27–53. arXiv:<https://doi.org/10.1177/0093650210365537> doi:10.1177/0093650210365537
- [31] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
- [32] 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.
- [33] 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*. 1–12.
- [34] Kai-Hui Liang, Weiyan Shi, Yoo Jung Oh, Hao-Chuan Wang, Jingwen Zhang, and Zhou 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 (2024), 1–28.
- [35] Jusheng Liu, Jianjia He, Shengxue He, Chaoran Li, Changrui Yu, and Qiang Li. 2022. Patients' self-disclosure positively influences the establishment of patients' trust in physicians: an empirical study of computer-mediated communication in an online health community. *Frontiers in Public Health* 10 (2022), 823692.
- [36] Max M. Louwerse, Arthur C. Graesser, Danielle S. McNamara, and Shulan Lu. 2009. Embodied conversational agents as conversational partners. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 23, 9 (2009), 1244–1255.

- [37] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.
- [38] Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. 2024. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj mental health research* 3, 1 (2024), 4.
- [39] Meta AI. 2024. Llama 3.3 70B Instruct. Model card. <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct> Text-only instruction-tuned large language model.
- [40] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering calibrated (dis-) trust in conversational agents: a user study on the persuasive power of limitation disclaimers vs. authoritative style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [41] Juha Munnukka, Karoliina Talvitie-Lamberg, and Devdeep Maity. 2022. Anthropomorphism and social presence in Human–Virtual service assistant interactions: The role of dialog length and attitudes. *Computers in Human Behavior* 135 (2022), 107343.
- [42] Hye Soo Nah. 2022. The appeal of “real” in parasocial interaction: The effect of self-disclosure on message acceptance via perceived authenticity and liking. *Computers in Human Behavior* 134 (2022), 107330.
- [43] Anouk Neerincx, Chantal Edens, Frank Broz, Yanzhe Li, and Mark Neerincx. 2022. Self-Disclosure to a Robot” In-the-Wild”: Category, Human Personality and Robot Identity. In *2022 31st IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 584–591.
- [44] Linda Uchenna Oghenekaro and Christopher Obinna Okoro. 2024. Artificial intelligence-based chatbot for student mental health support. *Open Access Library Journal* 11, 5 (2024), 1–14.
- [45] Julia Omarzu. 2000. A disclosure decision model: Determining how and when individuals will self-disclose. *Personality and social psychology review* 4, 2 (2000), 174–185.
- [46] OpenAI. 2024. *Whisper API*. <https://platform.openai.com/docs/guides/speech-to-text> Speech-to-text API used for transcription and analysis.
- [47] Csaba Orban, Ru Kong, Jingwei Li, Michael WL Chee, and BT Thomas Yeo. 2020. Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. *PLoS biology* 18, 2 (2020), e3000602.
- [48] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.
- [49] Hashai Papneja and Nikhil Yadav. 2024. Self-disclosure to conversational AI: a literature review, emergent framework, and directions for future research. *Personal and ubiquitous computing* (2024), 1–33.
- [50] Sandra Petronio. 2002. *Boundaries of privacy: Dialectics of disclosure*. Suny Press.
- [51] Katherine W Phillips, Nancy P Rothbard, and Tracy L Dumas. 2009. To disclose or not to disclose? Status distance and self-disclosure in diverse environments. *Academy of Management Review* 34, 4 (2009), 710–732.
- [52] Abhilasha Ravichander and Alan W Black. 2018. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*. 253–263.
- [53] Daniel Roth, Carola Bloch, Josephine Schmitt, Lena Frischlich, Marc Erich Latoschik, and Gary Bente. 2019. Perceived Authenticity, Empathy, and Pro-social Intentions evoked through Avatar-mediated Self-disclosures. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) (MuC '19). Association for Computing Machinery, New York, NY, USA, 21–30. doi:10.1145/3340764.3340797
- [54] David R Shaffer and Michele M Tomarelli. 1989. When public and private self-foci clash: Self-consciousness and self-disclosure reciprocity during the acquaintance process. *Journal of Personality and Social Psychology* 56, 5 (1989), 765.
- [55] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzæg. 2023. A longitudinal study of self-disclosure in human–chatbot relationships. *Interacting with Computers* 35, 1 (2023), 24–39.
- [56] Cameron Smith, Nigel Crook, Daniel Charlton, Johan Boye, Raul Santos de la Camara, Markku Turunen, David Benyon, Björn Gambäck, Oli Mival, Nick Webb, and Marc Cavazza. 2011. Interaction Strategies for an Affective Conversational Agent. *Presence* 20, 5 (2011), 395–411. doi:10.1162/PRES_a_00063
- [57] Susan Sprecher and Susan S Hendrick. 2004. Self-disclosure in intimate relationships: Associations with individual and relationship characteristics over time. *Journal of Social and Clinical psychology* 23, 6 (2004), 857–877.
- [58] Susan Sprecher, Stanislav Treger, Joshua D Wondra, Nicole Hilaire, and Kevin Wallpe. 2013. Taking turns: Reciprocal self-disclosure promotes liking in initial interactions. *Journal of Experimental Social Psychology* 49, 5 (2013), 860–866.
- [59] Jennifer L. Steel. 1991. Interpersonal Correlates of Trust and Self-Disclosure. *Psychological Reports* 68, 3_suppl (1991), 1319–1320. arXiv:<https://doi.org/10.2466/pr0.1991.68.3c.1319> doi:10.2466/pr0.1991.68.3c.1319
- [60] Unity Technologies. 2024. *Unity*. <https://unity.com> Version used in this research: 2022.1.23f1.
- [61] Edmund R. Thompson. 2007. Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology* 38, 2 (2007), 227–242. arXiv:<https://doi.org/10.1177/0022022106297301> doi:10.1177/0022022106297301
- [62] Takahiro Tsumura and Seiji Yamada. 2023. Influence of agent’s self-disclosure on human empathy. *Plos one* 18, 5 (2023), e0283955.
- [63] Caroline L van Straten, Jochen Peter, Rinaldo Kühne, and Alex Barco. 2022. On sharing and caring: Investigating the effects of a robot’s self-disclosure and question-asking on children’s robot perceptions and child-robot relationship formation. *Computers in Human Behavior* 129 (2022), 107135.
- [64] Isaac Wang, Lea Buchweitz, Jesse Smith, Lara-Sophie Bornholdt, Jonas Grund, Jaime Ruiz, and Oliver Korn. 2020. “Wow, You Are Terrible at This!” - An Intercultural Study on Virtual Agents Giving Mixed Feedback. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) (IVA '20). Association for Computing Machinery, New York, NY, USA, 8 pages. doi:10.1145/3383652.3423887

- [65] Lawrence R. Wheeless and Janis Grotz. 2006. The Measurement of Trust and Its Relationship to Self-Disclosure. *Human Communication Research* 3, 3 (03 2006), 250–257. arXiv:<https://academic.oup.com/hcr/article-pdf/3/3/250/22344414/jhumcom0250.pdf> doi:10.1111/j.1468-2958.1977.tb00523.x
- [66] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.
- [67] Morgan Worthy, Albert L Gary, and Gay M Kahn. 1969. Self-disclosure as an exchange process. *Journal of personality and social psychology* 13, 1 (1969), 59.

Received 2025; revised 2025; accepted xxx