Dynamic In-Group Persona Generation for Enhancing Human–AI Rapport

Anonymous ACL submission

Abstract

LLM-based chatbots are increasingly applied in interpersonal domains such as counseling and peer support, where establishing human–AI rapport is crucial yet remains challenging. In this work, we introduce a novel approach for conditioning LLMs with **in-group personas**, which (i) first identifies a user's primary concern and brief personal context (e.g., a computer science undergraduate worried about future career prospects), and (ii) generates a synthetic in-group persona that shares a similar primary concern while differing in background and narrative details, such as age or profession (e.g., a junior researcher at an AI startup). Furthermore, we conduct a human-subject study to systematically evaluate the effectiveness of in-group persona agents in enhancing human–AI rapport. We compare our approach against two baseline conditions: a conventional agent without persona conditioning and an agent exhibiting minimal self-disclosure (e.g., "I've felt that too"). Results from post-task questionnaires assessing rapport and user experience indicate that the in-group persona agent significantly improves perceived rapport and personal relevance compared to the baselines, and also yields more positive user experience—most notably higher engagement.

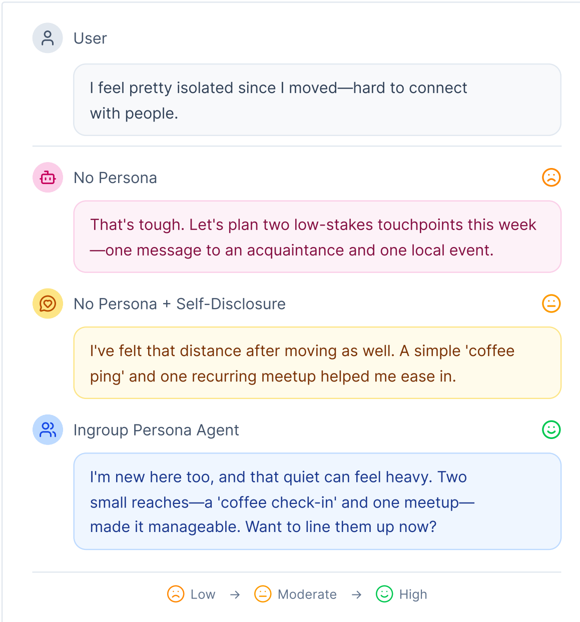


Figure 1: **Dynamic Persona Generation for Effective Human-AI Interaction.** Comparison of three LLM agents responding to the same user utterance expressing life concern. *No Persona*, the chatbot LLM without additional persona conditioning, gives neutral, generic guidance; *No Persona + Self-Disclosure* programmatically adds a general "I've felt that too" statement to invite reciprocity; finally, *In-group Persona Agent (Ours)* leverages shared identity and contextualized self-disclosure to foster greater rapport and relatability.

1 Introduction

Rapport, defined as "mutual attention, positivity, and coordination" (Tickle-Degnen and Rosenthal, 1990), is crucial in counseling and peer support, where users must feel comfortable disclosing personal experiences (Nienhuis et al., 2018; Truong et al., 2019). In supportive and interpersonal human–AI interactions such as peer-support or counseling-like conversations where users discuss personal concerns and may benefit from empathetic exchange, rapport can play a central role. Without a sense of relational connection, users may disengage, perceive the agent as impersonal, or hesitate to share meaningful information (Liu, 2024; Pentina et al., 2023). In these contexts, rapport underpins not only conversational naturalness but also the effectiveness of AI systems in fostering user trust, cooperation, and sustained engagement.

While large language model (LLM)-based chatbots are increasingly deployed in domains such as mental health support, education, and customer service, establishing *rapport* with users remains a central challenge. Despite that we are witnessing rapid deployment of AI-based counseling and mental health support systems, such as Woebot and Wysa (Fitzpatrick et al., 2017; Inkster et al., 2018), and test trials in clinical contexts (Heinz et al., 2025; MacNeill et al., 2024), chatbots often fail to demonstrate contextual understanding of social cues required to form rapport comparable to human counselors (Sands et al., 2021; Chan et al., 2022; Klein, 2025). As a result, LLM-based systems still face limitations in maintaining long-term effectiveness: sustained success appears contingent on users perceiving the agent as trustworthy, empathetic, and engaging (Siddals et al., 2024; Limpanopparat et al., 2024; Liu et al., 2025).

A key mechanism for fostering rapport is *self-disclosure*, which refers to the act of sharing personal thoughts, feelings, or experiences to build trust and intimacy. Interpersonal research shows that disclosure strengthens trust and intimacy via reciprocity (Carpenter and Greene, 2015). In Human-Computer Interaction (HCI), chatbots that share personal or emotional information elicit greater user self-disclosure and satisfaction (Ho et al., 2018; Lee et al., 2020), and emotional self-disclosure has been shown to increase reuse intention (Park et al., 2023).

Another mechanism to build rapport is to show *similarity and in-group preference*: according to similarity–attraction theory (Byrne, 1972), people respond more positively to partners who resemble them. This effect can be generalized to chatbots, where perceived similarity in personality or style enhances trust and affinity (Reeves and Nass, 1996; Jin and Eastin, 2023). In-group categorization further amplifies these effects, with users treating agents as more trustworthy and cooperative when framed as group members (Nass et al., 1997; Eyssel and Kuchenbrandt, 2012). Yet, most chatbot studies have emphasized *shallow similarity* (e.g., tone or trait matching) rather than shared lived experiences or social identity, leaving the deeper role of *contextual alignment* underexplored (Ahn et al., 2021; Alawi et al., 2023).

In this work, we propose a framework of *dynamic persona generation* that grounds LLM personas on users' expressed concerns and contexts, forming the basis of our **In-group Persona Agent (IPA)**. This framework aims to improve relational outcomes in human–AI interaction through careful agent persona construction and concern-grounded alignment. Our approach operationalizes findings from peer support research in mental health highlighting how shared experiences between interlocutors provide unique benefits, hope, validation, and a sense of not being alone (Repper and Carter, 2011). We model users' expressed *concerns*–issues they currently care about that capture their emotional and goal-oriented states (Randle et al., 2017; Ho et al., 2018; Cortland et al., 2017)–that are in turn utilized to form user-centered experiential representations that inform the agent's persona design.

With user concern-based experiences incorporated as part of the persona, our LLM agents are shown to produce responses that are both empathetic and contextually appropriate. To evaluate this framework, we conducted interactive sessions simulating everyday conversations where participants discussed personal concerns such as career and employment issues in a supportive but non-professional context. These dialogues provided a balanced setting where users sought both emotional empathy and practical perspectives, allowing us to assess how concern-aligned personas influence relational outcomes in human-AI interaction.

Contributions. We present the following contributions:

- We propose **In-group Persona Agent (IPA)**, a method that dynamically integrates persona generation and conversational inference through a multi-stage prompt pipeline, constructing concern-aligned personas from elicited user dialogue context.
- We complement the human-subject evaluation with post-hoc analyses that validate persona quality and characterize turn-level and conversation-level interaction patterns, thereby shedding light on the behavioral mechanisms underlying rapport gains.
- Controlled evaluations show that IPA improves rapport (especially personal relevance) over baselines, with modest UX gains most evident in engagement, highlighting *in-group alignment* as a key factor in human–AI relational quality.

2 Related Work

2.1 LLM Persona Conditioning

Large Language Models (LLMs) can be steered via prompt-only conditioning to display persona-like linguistic patterns without parameter updates. Evidence to date is strongest within-session and is task-dependent, indicating reliable short-horizon style control rather than persistent identity (Moon et al., 2024; Kang et al., 2025; Jiang et al., 2023; Tseng et al., 2024). Backstory-style prompts improve response consistency and distributional alignment in

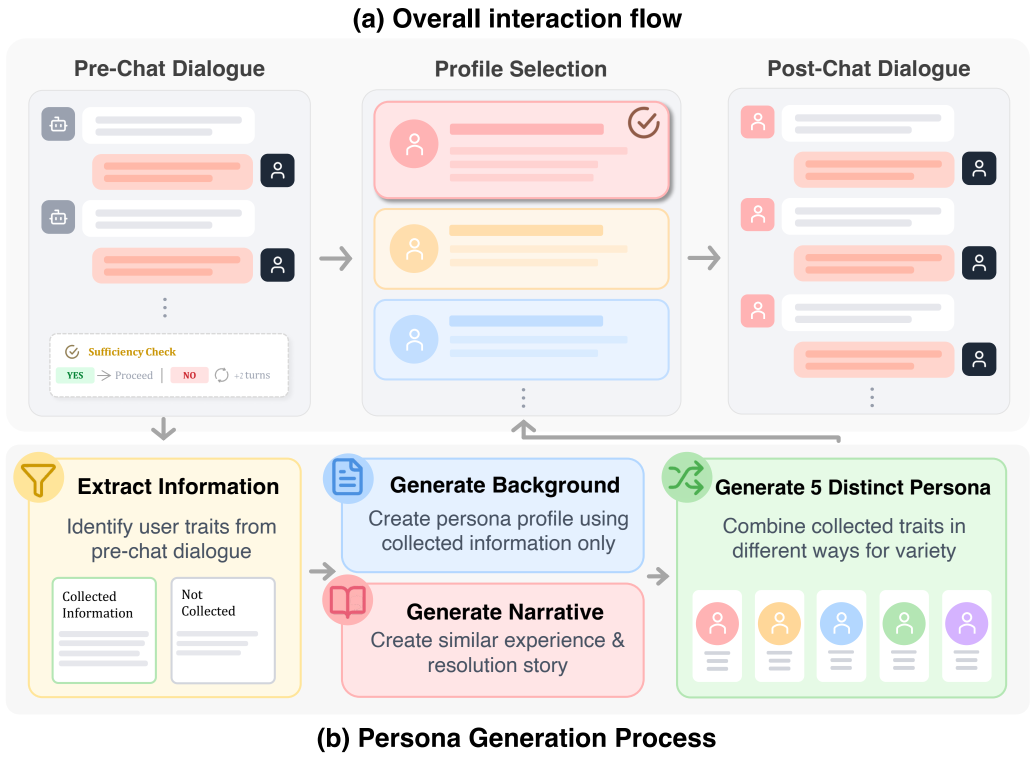


Figure 2: **Overview of the proposed *In-Group Persona Agent*. (a) Overall interaction flow:** a pre-chat dialogue elicitation with a sufficiency check collects the user's concern; the user then selects one of five candidate personas; the main conversation proceeds with the selected persona (see Section 3). **(b) Persona generation process:** traits are extracted from the pre-chat dialogue and partitioned into *collected* vs. *not collected*; using both information, the system composes a background and a narrative (similar experience and resolution) and combines the traits in different ways to produce five distinct personas.

large-scale simulations (Moon et al., 2024; Kang et al., 2025), while personality prompts (e.g., Big Five) induce trait-congruent tendencies across tasks under matched instructions (Jiang et al., 2023).

Unlike previous studies that relied on static backstories or trait-based prompts to elicit transient stylistic alignment, our In-group Persona Agent dynamically generates concern-aligned personas grounded in users' expressed contexts, enabling sustained relational alignment and measurable improvements in rapport and engagement.

2.2 Self-Disclosure by AI Chatbots

Research on interpersonal dynamics and the Computers Are Social Actors (CASA) framework indicate that disclosure-based reciprocity can arise even when people interact with artificial agents (Collins and Miller, 1994; Reeves and Nass, 1996). Experimental work also shows reciprocal responding toward computer-based agents, specifically increased user self-disclosure following an agent's disclosure (Moon, 2000). In studies on conversational AI systems, agent self-disclosure increases users' self-disclosure (Lee et al., 2020) and improves satisfaction (and reuse intentions) relative to non-disclosing baselines (Park et al., 2023). However, reported benefits are conditional on the contextual relevance and perceived appropriateness of the disclosure (Lee et al., 2020; Park et al., 2023; Tsumura and Yamada, 2023). Notably, prior manipulations of disclosure are typically generic (e.g., emotional vs. neutral/factual statements) rather than tailored to the user's present concern (Lee and Choi, 2017; Ho et al., 2018; Park et al., 2023).

We address this conditionality by operationalizing relevance via concern-aligned persona cues and by comparing no-persona and generic self-disclosure baselines to isolate the added value of contextual alignment.

2.3 Similarity & In-Group Cues

Similarity manipulations affect user judgments and intentions and vary across application domains. For example, in voice interfaces, matching voice personality increases perceived social presence (Lee and Nass, 2003). In recommender chatbots, personality alignment improves evaluation and intention measures (Jin and Eastin, 2023). In social robotics, in-group framing yields more favorable judgments (e.g., liking, anthropomorphism) and approach intentions (Nass et al., 1997; Eyssel and Kuchenbrandt, 2012). These operationalizations predominantly rely on static trait/style matching or categorical group labels rather than situational alignment to a user's ongoing concern.

Taken together, these observations motivate a focus on *concern-aligned* relational cues rather than reliance on generic disclosure or static trait/style matching.

3 Method

3.1 Overview

The **Persona Agent** analyzes *dialogue preceding the main conversation* gathered to infer the user's concerns and related information. Based on the inferred information, it generates a persona that belongs to the user's in-group and shares the same concerns as the user, embedding this fictional persona–comprising a character profile and curated past experiences–before the dialogue begins. This embedded persona enables the agent to deliver empathetic utterances and self-disclosing narratives, thereby achieving a higher level of rapport compared to chatbots without an explicit, tailored persona. For instance, when responding to a user concerned about work-related stress, a generic chatbot might respond, "I'm sorry you're feeling stressed; consider taking breaks during work." In contrast, a persona-embedded agent might share, "I remember feeling overwhelmed at my own job a few years ago. I found that taking short outdoor walks during my breaks helped me manage my stress better." The latter response leverages persona-driven self-disclosure, which fosters stronger rapport. Such an approach is particularly valuable in domains where user experience is paramount, including psychological support, education, and personalized services.

3.2 Persona Definition

A *persona* in this study consists of a synthetic **agent background** and **past narratives**. By injecting this information into the prompt, we create a *persona agent*, a virtual persona capable of interacting with the user.

Background. Background provides a concise overview of who the persona is, including aspects such as personal history, education, or interests. While it does not directly resolve the user's concern, it frames the agent as a coherent social actor with recognizable traits and contextual grounding.

Narrative. The narrative recounts how the synthetic persona previously faced a challenge similar to the user's concern and describes the steps they took to overcome it. Based on this synthetic narrative, models engage in strategic self-disclosure that conveys empathy and shared experience.

3.3 Pre-chat Dialogue Elicitation

During pre-chat dialogue, a concern-elicitation agent (*collector*) engages the user to elicit the primary concern and brief context. After every two turns, the system checks whether the information is sufficient to proceed; if not, the agent continues with concise clarification questions. Once sufficiency is met, the pipeline advances to persona generation. The fixed sufficiency rule and prompt templates are provided in Appendix F.

3.4 Persona Generation

Based on information collected during the pre-chat dialogue phase, the system generates **five in-group personas** that reflect the user's context and characteristics, so that all generated personas would feel relevant and relatable to the user.

To achieve this, the system categorizes user traits into collected information (e.g., concerns, explicitly stated preferences, or demographic indicators) and information unavailable from the pre-chat dialogue (i.e., traits that remained unspecified). Based on this categorization, our workflow combines available information traits in different ways to generate five distinct personas. Each synthetic persona shares the same primary concern as the user but differs in details of the assigned background and narrative, such as age or profession, to provide diverse perspectives on the shared concern. Traits classified as not collected information are left unspecified, so that the resulting profiles only reflect information the user had actually disclosed.

Example of Generated Persona (Abridged)

Background: I'm a 26-year-old computer science graduate who developed a fascination with AI research during my undergraduate studies, particularly after taking courses in machine learning and neural networks...

Narrative: After graduating, I was eager to dive into AI research but quickly discovered that breaking into the field was more challenging than I anticipated. Most research positions required either graduate degrees or significant practical experience that I lacked...

Details of this example can be found in Appendix F(Listings 4 and 6).

3.5 Persona Quality Validation

We evaluate the validity of the *dynamic persona generation* framework through two complementary methods.

3.5.1 Persona Evaluation Rubric

Persona quality is assessed using LLM-based evaluation on two dimensions: *In-group Fitness* and *Concern Resolution Quality*. *In-group Fitness* comprises two sub-dimensions, *Shared Background / Identity* **(IF1)** and *Shared Skills / Interests* **(IF2)**. *Concern Resolution Quality* comprises *Concern Match* **(CR1)** and *Narrative Authenticity* **(CR2)**. Each sub-dimension is scored on a 0–4 scale, yielding up to 8 points per dimension and a total of 16 points.

3.5.2 Pre-chat Dialogue Information Sufficiency Check

To examine whether the pre-chat dialogue provides adequate information for persona generation, we perform a sufficiency check. For a dialogue consisting of n turns, we segment into cumulative intervals of $2, 4, \ldots, n$ turns. Personas are generated from each interval, and their quality scores are compared. For example, if $n = 6$, personas are generated from turns (1–2), (1–4), and (1–6).

4 Experiments

4.1 Topic of Conversation

We set the topic of conversation between the user and the agent to focus on ***career and employment***. This domain was chosen because it naturally contains a balanced mix of concerns that require emotional support (e.g., coping with workplace stress) and concerns that require problem-solving or informational guidance (e.g., job search strategies, career planning). This balance makes it a suitable testbed for examining how users interact with a persona-based conversational agent across both affective and cognitive dimensions.

4.2 Procedure

Each session consisted of three stages conducted sequentially.(cf. overall architecture in Figure 2) All processes, including inference, persona generation, and conversation with the user, were implemented using **GPT-4o** (Hurst et al., 2024).

1. **Pre-chat.** A concern-elicitation agent (*collector*) first engaged the user in a brief conversation to gather the user's primary concern or topic of discussion. After every two turns of conversation, the system evaluated whether the user's concern and related contextual information were sufficiently clear. If the information was insufficient, the agent continued to ask concise follow-up questions until enough details were obtained. Once the concern and its related contextual information were sufficiently collected, the session proceeded to the *Profile Selection* stage.

2. **Profile Selection.** The system generated five persona profiles aligned with the user's stated concern. Users selected one profile after reviewing textual background information. (skipped if *No Persona*)

3. **Post-chat.** The user then engaged in an open-ended chat with the chatbot endowed with the selected persona. After the dialogue, the user completed an online questionnaire.

Throughout the interaction, system and user messages were logged for analysis. The conversation proceeded in a round-robin, turn-based manner, where each turn consisted of a single utterance from either the user or the agent. The interaction was unconstrained in length, allowing users to continue chatting until they felt the session was complete or a natural stopping point was reached.

4.3 Participants and Conditions

We perform a randomized control trial (RCT) to evaluate interactions with assigned persona-conditioned LLM agents, having granted *Institutional Review Board*(IRB) approval. All participants provided informed consent prior to participation, and their privacy and anonymity were ensured throughout the study. We recruited 210 participants via CloudResearch's Connect platform (Hartman et al., 2023) and randomly assigned each to one of three experimental conditions ($n = 70$ per condition):

***Ours:* In-group Persona Agent (IPA)**
Participants conversed with an agent equipped with a user-tailored persona. The system first generated five candidate personas from pre-chat information, and participants selected one before the main dialogue.

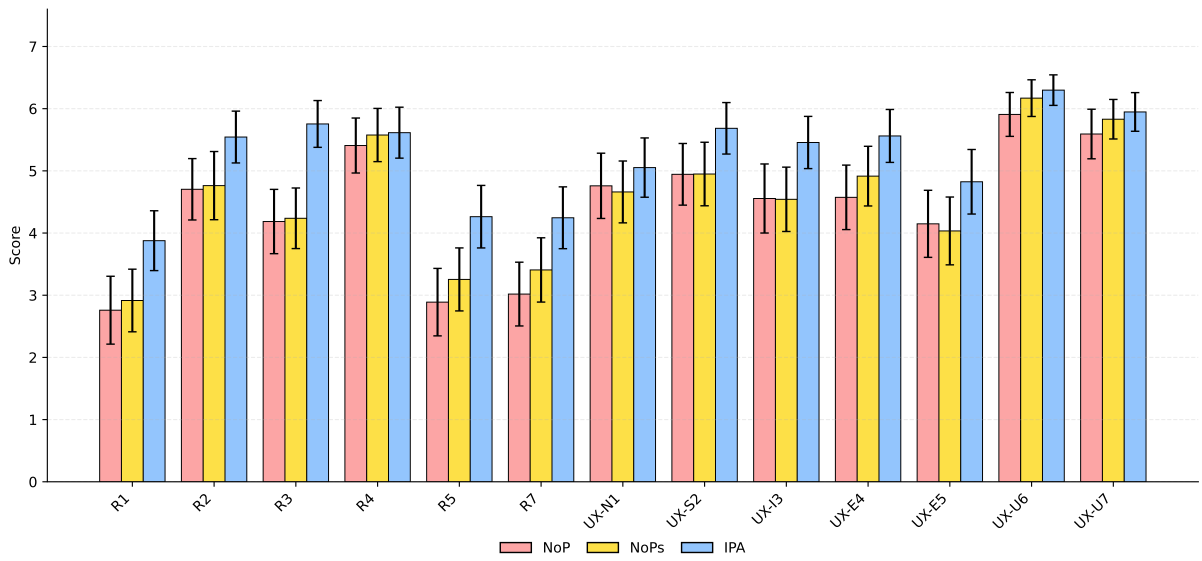


Figure 3: **Group comparison of Rapport and UX items**. Bars display mean ± 95% CI for visualization, although all statistical tests were performed on medians due to non-normal distributions (Shapiro–Wilk; see Appendix B). Rank-based one-sided Mann–Whitney contrasts tested the pair-wise ordered hypothesis NoP < NoPs < IPA. Bar colors denote the three conditions: NoP, NoPs, and IPA.

Baseline: No Persona Agent (NoP)

Participants interacted with an agent that retained dialogue history but had no defined persona. This served as a baseline condition without persona-based framing or self-disclosure.

Baseline: NoP with Self-Disclosure (NoPs)

Participants conversed with an agent that lacked a persona but could use brief, generic self-disclosing statements (e.g., preferences or feelings) not tailored to the participant's situation.

Immediately following the interaction, participants completed a post-task questionnaire. One attention-check item instructed participants to choose a specific response. Of the 210 recruited participants, we excluded those who failed the attention check and sessions whose dialogues were off-topic relative to the elicited concern. The final analysis sample comprised 170 participants (**NoP**: n=54, **NoPs**: n=59, **IPA**: n=57).

4.4 Measures

Immediately following the interaction, participants completed a post-task questionnaire assessing their perceptions of the agent and the overall dialogue experience. We adopted items from (Baihaqi et al., 2024), combining measures of *rapport* (six items, R1–R7 except R6[1]) and *user experience* (seven items). All items were rated on a 7-point Likert scale, with higher scores indicating greater rapport or more positive user experience. The items captured key aspects such as perceived emotional connection, enjoyment, comfort, personal relevance, mutual interest (e.g., "I feel a bond between this virtual agent and myself", R5), conversation naturalness and flow, satisfaction, interest, engagement, comprehensibility, and willingness to use the agent again. Details on the questionnaires for rapport and UX are provided in Appendix G.

4.5 Statistical Analysis

We analyzed outcomes using rank-based nonparametric methods, given the ordinal response scales and distributional diagnostics. Our primary test evaluated the pre-registered ordered alternative **NoP** $\leq$ **NoPs** $\leq$ **IPA** via a one-sided Jonckheere–Terpstra trend test (Terpstra, 1952; Jonckheere, 1954). To characterize condition differences within each construct family (Rapport or UX), we conducted planned one-sided Mann–Whitney U contrasts for all condition pairs in the hypothesized direction (Mann and Whitney, 1947), controlling family-wise error within each family using Holm's procedure (Holm, 1979). We report effect sizes

[1]R6 ("I really care about this virtual agent") was excluded from the rapport composite due to lexical ambiguity and lack of explicit interaction-referential wording, which reduces sensitivity to single-session effects.

and descriptive differences alongside adjusted p-values; analyses used the preprocessed sample (**NoP**, n=54; **NoPs**, n=59; **IPA**, n=57). Further implementation details, assumption checks, and full statistical outputs are provided in Appendix B.

5 Results

5.1 Main Result

Overview. Group-level comparisons are visualized in Figure 3 and summarized in Table 3. For visualization, data are shown as mean ± 95% CI to convey central tendency and variability, but all inferential tests were performed on median ranks using nonparametric procedures due to significant deviations from normality (see Appendix B).

Rapport. Across all rapport items (R1-R5, R7), scores increased monotonically from **NoP** to **NoPs** to **IPA**, with the largest and most consistent gains for **IPA** over **NoP**. These improvements were most pronounced on items indexing perceived relatedness and mutual engagement (especially **R3** "relevant to me," **R5** "bond with the agent," and **R7** "personal interest," as well as **R1** "think about relationship"). In contrast, **NoPs** provided only modest benefits over **NoP**, suggesting that generic self-disclosure alone yields limited relational improvement. Finally, **IPA** also exceeded **NoPs** across all rapport items, indicating that persona alignment adds relational depth beyond disclosure alone.

User Experience. Across user-experience items, **IPA** received the highest scores overall, whereas **NoP** and **NoPs** were generally close and did not follow a consistent ordering. Relative to **NoP**, the clearest improvements for **IPA** centered on **E4** (engagement) and **I3** (perceived relevance), alongside more favorable satisfaction and willingness to continue. The **NoPs–NoP** comparison was mixed, with small gains on some items but slight declines on others, indicating that generic self-disclosure did not reliably improve overall experience. Finally, **IPA** exceeded **NoPs** as well, with the most salient gains again tied to relevance, engagement, and continued-use intent, confirming additional benefits from persona alignment beyond disclosure alone.

Summary. Taken together, the analyses reveal a graded pattern of improvement: (1) **NoP** $\rightarrow$ **NoPs**: generic self-disclosure yields limited and mixed benefits; (2) **NoPs** $\rightarrow$ **IPA**: persona alignment adds clear gains, especially on rapport dimensions tied to perceived relevance and personal connection; and (3) **NoP** $\rightarrow$ **IPA**: the cumulative difference is most pronounced on key rapport and relevance-related outcomes. These results show that incorporating concern-aligned in-group persona cues consistently strengthens rapport relative to the baselines, with more modest user-experience gains that are most evident in engagement.

5.2 Validation Results

Following the validation procedures described in Section 3.5, we conducted two sets of experiments: (1) whether the proposed rubric effectively captures in-group alignment and concern reflection in generated personas; and (2) whether the amount of information collected in the pre-chat dialogue is sufficient for high-quality persona generation.

5.2.1 Validation of Persona Quality Rubric

To examine whether the rubric for persona quality evaluation appropriately captures that a persona generated from the pre-chat dialogue belongs to the user's in-group and reflects similar concerns, we compared the scores between the group where the persona was evaluated with the pre-chat dialogue used for its generation (Matched) and the group where the persona was evaluated with a different pre-chat dialogue (Not Matched). Figure 4(Appendix C) shows that Matched consistently received higher scores across all rubric items, while Not Matched received lower scores. This indicates that evaluation with the rubric can effectively verify whether a persona is appropriate for the user.

5.2.2 Validation of Pre-chat Dialogue Sufficiency Check

In the pre-chat dialogue stage, we verify whether the information collected every two turns is sufficient for persona generation. To examine this, a subset of 70 pre-chat dialogues from the full set collected in the experiment was segmented into cumulative intervals of two turns, and a persona was generated at each interval. In total, 480 personas were generated and subsequently evaluated for quality. Overall, the results demonstrate that the rubric-based evaluation is sensitive to information sufficiency, assigning lower scores to personas derived from incomplete contexts and higher scores to those generated from sufficient dialogue history. Detailed results are provided in Appendix D.

6 Analysis

6.1 Correlation Between Persona Quality and Rapport Score

Following the validation of the persona quality rubric in the previous analysis, we conducted correlation analyses to examine whether rubric scores were associated with self-reported outcomes (rapport and UX; see Tables 6 and 7 for detailed results). The selected personas generally demonstrated high rubric scores, with Narrative Authenticity showing zero variance and thus being excluded from correlation interpretation.

Pearson correlation analysis revealed that overall persona quality (Total) exhibited significant positive correlations with R5 (sense of connection, $r = 0.36, p < .05$) among rapport items and notably with E5 (intention to continue conversation in the future, $r = 0.37, p < .05$) among UX items. At the dimensional level, Shared Skills showed the strongest correlation with E5 ($r = 0.48, p < .05$), linking shared interests to continued use. In contrast, no consistent significant correlations emerged across other rapport items, indicating that persona quality is selectively associated with specific outcomes such as "sense of connection" and "continuation intention" rather than comprehensively explaining rapport as a whole.

6.2 Turn-level Behavioral Complement.

While self-reported rapport captures the overall experience, it fails to reveal *specific* behavioral nuances across conditions. To address this, we conducted a turn-level analysis focusing on *immediate partner responses* within the dialogue logs. Specifically, we scored each utterance using an LLM rubric (0–3) for self-disclosure (sd_t) and empathy (emp_t), and subsequently classified them as "high" or "low" based on fixed thresholds (details in Appendices E.1 and E.2). We then calculated the conditional probability of a response in speaker-change adjacent pairs, conditioned on the partner's preceding behavior. We report these probabilities along with their difference, defined as the reciprocity index R (Table 1).

We analyze the `post` segment to capture the interactions under the influence of established persona manipulation and rapport (see Appendix E.3). In the `post` segment, IPA is the only condition showing clear reciprocity of User self-disclosure in response to the Agent ($A \rightarrow U$). Specifically, User deep ("high") self-disclosure is substantially more likely immediately following an Agent's deep self-disclosure (0.620) compared to non-deep Agent turns (0.249), resulting in a reciprocity index of $R_{SD}^{A\to U} = +0.371$. Regarding empathy, the Agent's response to User self-disclosure ($U \rightarrow A$) was positive across all conditions. The metric $R_{EMP}^{U\to A}$ quantifies the increase in the Agent's probability of expressing empathy following deep User self-disclosure. This increase is most pronounced in the IPA condition ($0.539 \rightarrow 0.903$, $R_{EMP}^{U\to A} = +0.364$). In contrast, the NoPs condition shows a smaller increase ($R_{EMP}^{U\to A} = +0.057$) because the Agent's baseline empathy is already near the ceiling ($0.936 \rightarrow 0.993$).

Condition	$p^{A\to U}_{sd\mid high}$	$p^{A\to U}_{sd\mid low}$	$R^{A\to U}_{SD}$	$p^{U\to A}_{emp\mid high}$	$p^{U\to A}_{emp\mid low}$	$R^{U\to A}_{EMP}$
NoP	NaN	0.377	**NaN**	0.908	0.631	**+0.277**
NoPs	0.425	0.454	**-0.029**	0.993	0.936	**+0.057**
IPA	0.620	0.249	**+0.371**	0.903	0.539	**+0.364**

Table 1: **Directional reciprocity in the `post` segment.** Each p denotes a conditional probability of a next-turn response event (User deep self-disclosure or Agent empathy) given whether the immediately preceding turn contains deep self-disclosure (event present vs. absent; "high/low"). R is the difference of the two conditional probabilities (e.g., $R_{SD}^{A\to U} = p_{sd|high}^{A\to U} - p_{sd|low}^{A\to U}$). NaN indicates non-estimability due to event sparsity.

Overall, these results suggest that IPA uniquely strengthens self-disclosure reciprocity, while the Agent's empathic responsiveness to User self-disclosure is broadly present across conditions.

7 Conclusion

This work demonstrates that embedding in-group personas into conversational AI can meaningfully enhance rapport and user experience. By combining pre-chat dialogue elicitation with structured persona framing and self-disclosure, our approach enables agents to appear more relatable, empathetic, and engaging. After pre-processing, we analyzed data from 170 participants and found that persona-embedded agents consistently outperformed both non-persona and self-disclosing baselines on rapport measures, including perceived personal relevance. Overall, lightweight, prompt-based persona design offers a practical means to strengthen rapport and engagement in supportive interpersonal settings–supported by analyses linking persona quality to users' sense of connection and continuation intention, and turn-level evidence of increased self-disclosure reciprocity–and extending this framework to more sensitive domains remains an important direction for future work.

8 Limitations

Although our study offers insights into the role of in-group personas in conversational AI, it also carries several limitations that should be acknowledged.

Anthropomorphism: While persona-driven design can enhance users' sense of connection, excessive anthropomorphic cues may also introduce unintended side effects. Overly human-like behaviors, such as emotional over-expression or self-referential remarks, can create unrealistic expectations about the agent's capabilities or authenticity. This mismatch between user perception and system intent may lead to confusion or reduced trust, suggesting that careful calibration of anthropomorphic elements is essential for maintaining credible and comfortable human–AI interaction.

Pre-chat Sufficiency: Our analyses focus on sessions that met a minimum level of pre-chat sufficiency, enabling personas to be grounded in user-provided cues. Because we did not systematically collect a comparable set of "insufficient" pre-chat cases, we cannot conduct a reliable ablation on how pre-chat quantity affects persona quality and downstream outcomes. This limits our understanding of robustness under sparse inputs, suggesting that future work should evaluate broader ranges of pre-chat completeness.

Disclosure-Averse Users: Our framework assumes that users are comfortable articulating their primary concern during the pre-chat stage, and our experiments were designed around participants for whom such disclosure is a relatively low burden. In practice, some users may be hesitant to share personal details upfront, potentially reducing usability and limiting persona quality. While the pre-chat agent is intended to ease disclosure through guided prompting, disclosure-averse populations remain out of scope for the current study, motivating future work in this direction.

Future work should address these limitations by expanding the scope of participants and analyses in order to provide more generalizable insights.

9 Ethical Considerations

Potential Harms of Synthetic In-group Personas. We acknowledge that in sensitive domains such as mental health, trauma recovery, or medical advice, a synthetic persona that feigns shared human experience can lead to severe consequences. In such contexts, "synthetic empathy" may induce undue trust, discourage seeking professional help, or otherwise exacerbate vulnerability. Our study, however, was conducted in the comparatively lower-risk context of career and employment counseling, rather than in high-stakes settings such as suicide prevention or grief counseling. The persona's "experience" was limited to professional challenges and workload management, and the interactions were framed as instrumental support rather than therapeutic intervention. Accordingly, we expect the immediate risk of harm to be substantially lower than in clinical settings, while noting that similar risks could arise if such personas were deployed in more sensitive contexts.

Participant Disclosure and Right to Withdraw. Prior to participation, we informed participants in the study introduction about potential risks and disadvantages, including the possibility of discomfort during experimental procedures. Participants were explicitly informed that they could withdraw at any time without penalty. We also indicated that appropriate support would be available if any discomfort persisted. We provide the disclosure text in the appendix, with selected portions redacted to preserve the authors' anonymity.

Scope and Limitations. These ethical considerations apply to the specific experimental scope described above. We caution against generalizing the safety of synthetic in-group personas to high-stakes or clinical domains without domain-specific safeguards, oversight, and evaluation.

References

Jungyong Ahn, Jungwon Kim, and Yongjun Sung. 2021. Ai-powered recommendations: the roles of perceived similarity and psychological distance on persuasion. *International Journal of Advertising*, 40(8):1366–1384.

Naif Alawi, Triparna De Vreede, and Gert-Jan De Vreede. 2023. Accepting the familiar: the effect of perceived similarity with ai agents on intention to use and the mediating effect of it identity.

Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

Muhammad Yeza Baihaqi, Angel García Contreras, Seiya Kawano, and Koichiro Yoshino. 2024. Rapport-driven virtual agent: Rapport building dialogue strategy for improving user experience at first meeting. *arXiv preprint arXiv:2406.09839*.

Godfrey T Barrett-Lennard. 1981. The empathy cycle: Refinement of a nuclear concept. *Journal of counseling psychology*, 28(2):91.

Donn Erwin Byrne. 1972. The attraction paradigm. *Behavior Therapy*, 3(2):337–338.

Robert R Carkhuff. 1969. Helping and human relations: A primer for lay and professional helpers: I. selection and training.

Amanda Carpenter and Kathryn Greene. 2015. Social penetration theory. *The international encyclopedia of interpersonal communication*, pages 1–4.

William W Chan, Ellen E Fitzsimmons-Craft, Arielle C Smith, Marie-Laure Firebaugh, Lauren A Fowler, Bianca DePietro, Naira Topooco, Denise E Wilfley, C Barr Taylor, and Nicholas C Jacobson. 2022. The challenges in designing a prevention chatbot for eating disorders: observational study. *JMIR Formative Research*, 6(1):e28003.

J COHEN. 1992. A power primer. *Psychological bulletin*, 112(1):155–159.

Nancy L Collins and Lynn Carol Miller. 1994. Self-disclosure and liking: a meta-analytic review. *Psychological bulletin*, 116(3):457.

Clarissa I Cortland, Maureen A Craig, Jenessa R Shapiro, Jennifer A Richeson, Rebecca Neel, and Noah J Goldstein. 2017. Solidarity through shared disadvantage: Highlighting shared experiences of discrimination improves relations between stigmatized groups. *Journal of personality and social psychology*, 113(4):547.

Valerian J Derlega, Sandra Metts, Sandra Petronio, and Stephen T Margulis. 1993. *Self-disclosure.* Sage Publications, Inc.

Friederike Eyssel and Dieta Kuchenbrandt. 2012. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4):724–731.

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):e7785.

Rachel Hartman, Aaron J Moss, Shalom Noach Jaffe, Cheskie Rosenzweig, Leib Litman, and Jonathan Robinson. 2023. Introducing connect by cloudresearch: Advancing online participant recruitment in the digital age.

Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. 2025. Randomized trial of a generative ai chatbot for mental health treatment. *Nejm Ai*, 2(4):AIoa2400802.

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):712–733.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Becky Inkster, Shubhankar Sarda, Vinod Subramanian, and 1 others. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Eunjoo Jin and Matthew S Eastin. 2023. Birds of a feather flock together: matched personality effects of product recommendation chatbots and users. *Journal of Research in Interactive Marketing*, 17(3):416–433.

Aimable Robert Jonckheere. 1954. A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41(1/2):133–145.

Sidney M Jourard. 1971. Self-disclosure: An experimental analysis of the transparent self. *John Wiely and Sons*.

Minwoo Kang, Suhong Moon, Seung Hyeong Lee, Ayush Raj, Joseph Suh, and David Chan. 2025. Deep binding of language model virtual personas: a study on approximating political partisan misperceptions. In *Second Conference on Language Modeling*.

Stefanie Helene Klein. 2025. The effects of human-like social cues on social responses towards text-based conversational agents—a meta-analysis. *Humanities and Social Sciences Communications*, 12(1):1–16.

Kwan Min Lee and Clifford Nass. 2003. Designing social presence of social actors in human computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 289–296.

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*, 103:95–105.

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*, pages 1–12.

Howard Levene. 1960. Robust tests for equality of variances. *Contributions to probability and statistics*, pages 278–292.

Sucharat Limpanopparat, Erin Gibson, and Andrew Harris. 2024. User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: A systematic review. *Computers in Human Behavior: Artificial Humans*, 2(2):100081.

Jiaxi Liu. 2024. Chatgpt: Perspectives from human–computer interaction and psychology. *Frontiers in Artificial Intelligence*, 7:1418869.

Tingting Liu, Salvatore Giorgi, Ankit Aich, Allison Lahnala, Brenda Curtis, Lyle Ungar, and João Sedoc. 2025. The illusion of empathy: How ai chatbots shape conversation perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14327–14335.

A Luke MacNeill, Shelley Doucet, and Alison Luke. 2024. Effectiveness of a mental health chatbot for people with chronic diseases: randomized controlled trial. *JMIR Formative Research*, 8:e50025.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M Chan. 2024. Virtual personas for language models via an anthology of backstories. *arXiv preprint arXiv:2407.06576*.

Youngme Moon. 2000. Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of consumer research*, 26(4):323–339.

Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, 27(10):864–876.

Jacob B Nienhuis, Jesse Owen, Jeffrey C Valentine, Stephanie Winkeljohn Black, Tyler C Halford, Stephanie E Parazak, Stephanie Budge, and Mark Hilsenroth. 2018. Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: A meta-analytic review. *Psychotherapy Research*, 28(4):593–605.

Gain Park, Jiyun Chung, and Seyoung Lee. 2023. Effect of ai chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: A serial mediation model. *Current Psychology*, 42(32):28663–28673.

Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600.

Melanie Randle, Richard Eckersley, and Leonie Miller. 2017. Societal and personal concerns, their associations with stress, and the implications for progress and the future. *Futures*, 93:68–79.

Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10):19–36.

Julie Repper and Tim Carter. 2011. A review of the literature on peer support in mental health services. *Journal of mental health*, 20(4):392–411.

Carl R Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, 21(2):95.

Robert Rosenthal. 1991. *Meta-Analytic Procedures for Social Research*, volume 6. SAGE.

Sean Sands, Carla Ferraro, Colin Campbell, and Hsiu-Yuan Tsao. 2021. Managing the human–chatbot divide: how service scripts influence service experience. *Journal of Service Management*, 32(2):246–264.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Steven Siddals, John Torous, and Astrid Coxon. 2024. “it happened to be the perfect thing”: experiences of generative ai chatbots for mental health. *npj Mental Health Research*, 3(1):48.

Teunis J Terpstra. 1952. The asymptotic normality and consistency of kendall’s test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14(3):327–333.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293.

C Truong, J Gallo, D Roter, and J Joo. 2019. The role of self-disclosure by peer mentors: Using personal narratives in depression care. *Patient education and counseling*, 102(7):1273–1279.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Takahiro Tsumura and Seiji Yamada. 2023. Influence of agent’s self-disclosure on human empathy. *Plos one*, 18(5):e0283955.

Appendix

A Experiment Detail

This section includes details of our experiment. Table 3 shows overall result of human evaluation.

A.1 Demographic Distributions of Participants

Category	n (%)
Gender	
Man	86 (50.6%)
Woman	78 (45.9%)
Missing	3 (1.8%)
Prefer not to say	2 (1.2%)
Agender	1 (0.6%)
Race	
White	127 (74.7%)
Black or African American	14 (8.2%)
An ethnicity not listed here	7 (4.1%)
Chinese	6 (3.5%)
Vietnamese	4 (2.4%)
Missing	3 (1.8%)
Filipino	3 (1.8%)
Asian Indian	2 (1.2%)
Japanese	1 (0.6%)
Other	1 (0.6%)
Korean	1 (0.6%)
American Indian or Alaska Native	1 (0.6%)
Employment Status	
Full-time	80 (47.1%)
Part-time	27 (15.9%)
Unemployed	18 (10.6%)
Not in paid work	14 (8.2%)
Business Owner	10 (5.9%)
Student	9 (5.3%)
Retired	6 (3.5%)
Prefer not to say	3 (1.8%)
Missing	3 (1.8%)

Table 2: **Demographic distributions (n, %).** Total 170 participants, missing 3 informations.

Table 2 shows demographic distributions of the participants. Overall, the sample exhibits reasonable variation in background (e.g., gender and employment situations), and participants engaged in conversations about their own real career and employment concerns rather than role-played or hypothetical scenarios. At the same time, the pool is geographically and linguistically constrained (U.S.-based, English-speaking) and was recruited from a single crowdsourcing platform, which may limit broader generalizability.

A.2 Participant Disclosure: Risks, Withdrawal, and Support

Participants were informed of this procedure in advance as part of the study instructions and consent process. They were also informed of potential risks or discomforts, data handling and privacy protections, and that participation was voluntary and could be withdrawn at any time without penalty.

> **Disadvantages, Risks of Participation, and Compensation for Harm**
>
> Since the study involves discussing personal concerns with a chatbot, psychological discomfort or risks may arise. You are free to stop the experiment at any time without notifying the researcher. If you experience psychological discomfort after completion or withdrawal, you may contact the principal investigator listed below, who will consult with the (anonymous) to arrange appropriate support.

A.3 Differences Between IPA and Baselines

All three conditions used the same user messages as conversational context, ensuring that the baseline personalization naturally provided by LLMs was held constant across conditions. The only difference was the type of additional personalization provided by the agent: **NoP** included no added personalization cues, **NoPs** included a brief generic self-disclosure, and **IPA** incorporated a concern-aligned in-group persona derived from pre-chat information. Therefore, the observed advantages of **IPA** are attributable to the added in-group persona cues rather than generic personalization. Moreover, **IPA** produced the strongest gains specifically in relevance and perceived personal connection, which are theoretically linked to perceived in-group similarity. This design thus isolates the contribution of in-group persona alignment.

Example of Actual Agent Responses.

***Ours:* In-group Persona Agent (IPA)**

(generic personalization + self-disclosure with persona information)

Agent: Yes, after realizing my efforts weren't fully visible, I increased my visibility with upper management and successfully secured a promotion along with a revised compensation package.

Baseline: No Persona Agent (NoPs)

(generic personalization + self-disclosure without persona information)

Agent: That's perfectly fine! Medical school will expose you to a variety of specialties, which might help clarify your interests. I admire cardiology because of the heart's intricate systems, but every specialty has its unique appeal. What aspects of medicine spark your curiosity?

Baseline: NoP (NoP)

(generic personalization only)

Agent: That's a great goal! To transition, build your technical skills, particularly in areas like network architecture, encryption, and coding. Seek out projects or training that give you engineering experience, and consider networking with professionals in the field for mentorship and guidance.

A.4 Participant Openness and Ecological Validity of a Paid Sample

Our framework targets users who intentionally seek concern-oriented, personal conversations; thus, even though participants were compensated, they voluntarily engaged with a framework that clearly required sharing personal context, making the sample reasonably comparable to real-world users who would opt in. To reduce the burden of self-disclosure, we use a lightweight pre-chat agent that starts from the user's stated concern and asks brief, context-relevant follow-up questions, framing onboarding as collaborative problem identification rather than form filling. We also acknowledge that some users are highly disclosure-averse; future work will explore methods to assess disclosure comfort early and offer alternative onboarding strategies tailored to different willingness levels.

B Detail Statistical Analysis

Implementation Details. Analyses were conducted in Python 3.13.7 using NumPy 2.3.3, Pandas 2.3.3, SciPy 1.16.2, Pingouin 0.5.5, and Matplotlib 3.10.6. The rpy2 3.6.4 interface connected to R 4.3.1 with the clinfun 1.0.16 package for Jonckheere–Terpstra tests.

Assumption Testing. Distributional assumptions were formally evaluated for each item to verify the appropriateness of parametric inference. Shapiro–Wilk tests (Shapiro and Wilk, 1965) revealed signif-

Metric	IPA	NoP	NoPs	$\Delta_{\text{IPA,NoP}}$	$\Delta_{\text{IPA,NoPs}}$	$\Delta_{\text{NoPs,NoP}}$	$r_{\text{IPA,NoP}}$	$r_{\text{IPA,NoPs}}$	$r_{\text{NoPs,NoP}}$
R1	**3.88**	2.76	2.92	**1.12**	**0.96**	0.16	0.29	0.26	0.05
R2	**5.54**	4.70	4.76	**0.84**	0.78	0.06	0.23	0.17	0.04
R3	**5.75**	4.19	4.24	**1.57**	**1.52**	0.05	**0.42**	**0.42**	0.01
R4	**5.61**	5.41	5.58	0.21	0.04	0.17	0.06	-0.01	0.06
R5	**4.26**	2.89	3.25	**1.37**	**1.01**	0.37	**0.34**	0.25	0.10
R7	**4.25**	3.02	3.41	**1.23**	**0.84**	0.39	**0.30**	0.21	0.10
UX-N1	**5.05**	4.76	4.66	0.29	0.39	-0.10	0.08	0.11	-0.03
UX-S2	**5.68**	4.94	4.95	**0.74**	**0.74**	0.00	0.20	0.19	0.02
UX-I3	**5.46**	4.56	4.54	**0.90**	**0.91**	-0.01	0.21	0.23	-0.01
UX-E4	**5.56**	4.57	4.92	**0.99**	**0.65**	0.34	0.27	0.18	0.09
UX-E5	**4.82**	4.15	4.03	0.68	0.79	-0.11	0.17	0.19	-0.03
UX-U6	**6.30**	5.91	6.17	0.39	0.13	0.26	0.14	0.04	0.10
UX-U7	**5.95**	5.59	5.83	0.35	0.12	0.24	0.10	0.05	0.06

Table 3: **Human Evaluation Results.** Group means, pairwise mean differences (Δ), and effect sizes ($r = Z/\sqrt{N}$). **Bold** indicates statistically significant differences ($p < .05$) or medium effects ($r \geq .30$). **Shaded** indicates strong significance ($p < .01$). Pairwise p-values from one-sided Mann–Whitney tests, Holm-adjusted within item. Effect-size benchmarks: small ($.10 \leq r < .30$), medium ($.30 \leq r < .50$).

icant deviations from normality across most conditions. Levene's tests (Levene, 1960) suggested that variance heterogeneity was not pervasive, although a small number of items showed evidence of heteroscedasticity (minimum $p = .013$). Given the ordinal scales and these diagnostics, we used rank-based nonparametric analyses. Table 4 summarizes these results.

Analytic Framework. Given the ordinal response scale and the non-normal distributions observed, rank-based nonparametric analyses were adopted. The Jonckheere–Terpstra trend test (Terpstra, 1952; Jonckheere, 1954) was used to evaluate the pre-specified ordered alternative (**NoP** $\leq$ **NoPs** $\leq$ **IPA**) in a one-sided manner. When a significant monotonic trend was detected, planned one-sided pairwise Mann–Whitney U tests (Mann and Whitney, 1947) were performed to decompose the pattern in the hypothesized direction.

Family-wise error was controlled separately within each construct family: Rapport (R1–R5, R7) and UX (N1, S2, I3, E4, E5, U6, U7). Within each family, Holm's sequentially rejective method (Holm, 1979) was applied, and adjusted p-values (p_{adj}) are reported for each item. Statistical significance was defined as $p_{\text{adj}} < .05$ (one-sided). All analyses were conducted at $\alpha = 0.05$ on the preprocessed sample (**NoP**, n=54; **NoPs**, n=59; **IPA**, n=57).

Reporting Conventions and Effect Sizes. For Mann–Whitney pairwise contrasts, effect sizes were computed as $r = Z/\sqrt{N}$ following Rosenthal's r-equivalent formulation (Rosenthal, 1991). Magnitudes were interpreted analogously to correlation benchmarks (COHEN, 1992): small ($r \approx .10$) and medium ($r \approx .30$). Across both trend-level and pairwise analyses, p-values are treated as inferential thresholds, while raw differences (Δ) are reported as descriptive complements; substantive interpretation emphasizes the direction and magnitude of effects.

Trend-level Inference. Jonckheere–Terpstra tests examined the pre-registered ordered alternative (**NoP** $\leq$ **NoPs** $\leq$ **IPA**) across all items within each construct. As shown in Table 5, the monotonic trend was significant for most **Rapport** items, providing evidence for a stepwise increase in relational judgments across conditions. For rapport items (R1–R7 except R6), five of six reached significance after Holm adjustment ($p_{\text{adj}} \leq .019$), with small-to-medium effects ($r = .18$–$.35$). Notably, **R3** ("This virtual agent is very relevant to me") exhibited the strongest trend ($r = .349$, $p_{\text{adj}} < .001$), followed by **R5** ($r = .281$, $p_{\text{adj}} = .001$), suggesting that identity-aligned persona framing most strongly shaped perceived personal relevance and connection. In contrast, **R4** showed no reliable monotonic trend ($p_{\text{adj}} = .28$, $r = .045$), indicating an item-specific exception to the ordered pattern.

User-experience items (N1, S2, E4, E5, I3, U6, U7) displayed weaker and less consistent ordered evidence. Only **E4** ("engagement") remained sig-

(a) Shapiro–Wilk tests of normality by condition.

Item	Group	N	W	p
R1	NoP	54	0.816	<.001
R1	NoPs	59	0.859	<.001
R1	Peer	57	0.921	.0011
R2	NoP	54	0.906	<.001
R2	NoPs	59	0.861	<.001
R2	Peer	57	0.802	<.001
R3	NoP	54	0.928	.0030
R3	NoPs	59	0.928	.0018
R3	Peer	57	0.809	<.001
R4	NoP	54	0.859	<.001
R4	NoPs	59	0.808	<.001
R4	Peer	57	0.825	<.001
R5	NoP	54	0.843	<.001
R5	NoPs	59	0.892	<.001
R5	Peer	57	0.929	.0024
U7	NoPs	59	0.816	<.001
U7	Peer	57	0.821	<.001
(Other items show comparable non-normal patterns.)				

(b) Levene's tests of homogeneity of variance across groups.

Item	k	Center	F	p
R1	3	Mean	1.63	.200
R2	3	Mean	3.95	.021
R3	3	Mean	4.43	.013
R4	3	Mean	0.30	.738
R5	3	Mean	0.08	.920
R7	3	Mean	0.35	.708
E4	3	Mean	2.00	.139
E5	3	Mean	0.20	.820
I3	3	Mean	3.61	.029
N1	3	Mean	0.23	.797
S2	3	Mean	2.17	.117
U6	3	Mean	1.56	.213
U7	3	Mean	3.06	.050

Table 4: **Tests of distributional assumptions for all items.** Shapiro–Wilk tests assessed normality within each group, and Levene's tests examined homogeneity of variance across conditions.

nificant after correction ($p_{\text{adj}} = .015$, $r = .219$), while the remaining UX items did not reach adjusted significance ($p_{\text{adj}} \geq .062$; $r \leq .177$). Together, these JT results indicate that the pre-registered ordered improvement is robust for **Rapport**, whereas for **User Experience** the trend is more modest and primarily driven by engagement.

Pairwise Contrasts. Building on the significant ordered trends, planned one-sided Mann–Whitney U tests compared all condition pairs (**NoP≤NoPs**, **NoPs≤IPA**, **NoP≤IPA**). Detailed pairwise results, including adjusted p-values, mean differences (Δ), and effect sizes (r), are summarized in Table 3. Interpretation emphasizes the magnitude and direction of these effects, with p-values serving only as inferential filters and Δ as descriptive complements.

Item	J_T	p	p_{adj}	r
R1	5921.5	.001	.002	0.243
R2	5633.5	.009	.019	0.180
R3	6404.5	<.001	<.001	0.349
R4	5017.5	.280	.280	0.045
R5	6095.0	<.001	.001	0.281
R7	5948.0	.001	.002	0.249
N1	5101.0	.205	.249	0.063
S2	5571.0	.015	.075	0.166
I3	5621.0	.010	.062	0.177
E4	5809.0	.002	.015	0.219
E5	5442.5	.036	.144	0.138
U6	5316.5	.075	.225	0.110
U7	5216.5	.124	.249	0.088

Table 5: **Jonckheere–Terpstra trend test results for both Rapport and User Experience constructs.** Reported are test statistics (J_T), one-sided p-values, Holm-adjusted p-values (within construct), and effect sizes ($r = Z/\sqrt{N}$).

Score Dimension	Mean	Variance	Scale
Total	14.43	1.09	0–16
Shared Background	3.26	0.44	0–4
Shared Skills	3.21	0.60	0–4
Concern Match	3.85	0.57	0–4
Narrative Auth	4.00	0.00	0–4

Table 6: **Rubric Score Distribution (Chosen Personas)**

C Persona Quality: Matched vs. Unmatched

Figure 4 reports item-level persona-quality scores for the Matched and Not Matched settings. Across all rubric items, the Matched setting yields higher scores (3.42, 3.25, 3.85, 4.00), whereas the Not Matched setting yields lower scores (0.17, 0.16, 0.27, 2.01). This consistent separation indicates that the rubric can help verify whether a persona is appropriate for the user given the corresponding pre-chat dialogue.

D Evaluation Scores by Pre-chat Dialogue Length.

Each row indicates the number of turns used for persona generation, denoted as "Original Turn–Used Turn" (e.g., 4-2 means a pre-chat dialogue of original length 4, using its first 2 turns). Scores are reported as Mean (Std). *Num* indicates the number of generated personas. Within each Max Turn block, the highest value in each column is shown in bold. All evaluations were conducted with a temperature of 0, while all persona generations were

Item	Total Score	Shared Background	Shared Skills	Concern Match
R1	0.211	0.091	0.212	0.108
R2	0.209	0.229	0.270	-0.067
R3	0.202	0.111	0.278	0.004
R4	0.321	0.274	0.298	0.082
R5	**0.364**	0.190	0.328	0.199
R7	0.256	0.074	0.115	0.310
N1	**0.352**	0.291	0.277	0.151
S2	0.230	0.200	0.301	-0.036
I3	**0.345**	0.266	0.248	0.189
E4	0.249	0.174	0.203	0.124
E5	**0.371**	0.304	**0.483**	-0.042
U6	0.045	0.014	0.025	0.049
U7	0.169	0.096	0.174	0.062

Table 7: **Post-wise Similarity Scores. Bold** indicates statistically significant Pearson correlations ($p < .05$).

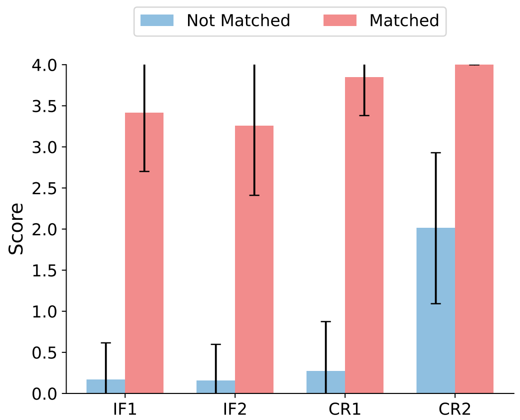


Figure 4: **Rubric Score comparison** between *Matched* and *Not Matched*. Bar colors denote the condition Not Matched and Matched.

performed with a temperature of 1.

As shown in Table 8, personas generated from dialogues with a longer maximum turn length tend to receive lower scores when evaluated at earlier segments (e.g., 6-2), reflecting the fact that longer pre-chat dialogues arise when the initial exchanges are insufficient. Consequently, personas generated from the first few turns of such dialogues lack sufficient information and are judged to be of lower quality. In contrast, when the full set of turns is used (e.g., 6-6), the scores improve substantially.

Importantly, the differences are particularly pronounced in the Ingroup-fitness dimensions (IF1 and IF2): insufficient dialogue context primarily affects the model's ability to generate personas that align with the user's group identity and shared concerns. By contrast, the consistency-related criteria (CR1 and CR2) remain relatively stable across conditions. These results demonstrate that the rubric-based evaluation is sensitive to information sufficiency, assigning lower scores to personas derived from incomplete contexts and higher scores to those generated from sufficient dialogue history.

Turn	Num	IF1	IF2	CR1	CR2	Sum
2-2	109	3.25 (0.65)	3.06 (0.87)	3.99 (0.10)	4.00 (0.00)	14.30 (1.38)
4-2	109	2.88 (0.97)	2.61 (0.90)	3.96 (0.19)	**4.00 (0.00)**	13.46 (1.72)
4-4	187	**3.19 (0.64)**	**3.17 (0.85)**	**3.98 (0.14)**	4.00 (0.00)	**14.33 (1.30)**
6-2	9	1.89 (1.37)	1.22 (1.13)	3.11 (1.29)	3.89 (0.31)	10.11 (3.38)
6-4	31	2.52 (0.95)	2.65 (0.78)	**3.84 (0.37)**	**4.00 (0.00)**	13.00 (1.83)
6-6	35	**2.94 (0.75)**	**2.86 (0.76)**	3.80 (0.52)	**4.00 (0.00)**	**13.60 (1.74)**

Table 8: **Evaluation Scores by Pre-chat Dialogue Length.**

E Turn-level Behavioral Annotation and Reciprocity Analysis

E.1 LLM Rubric Judging for Self-Disclosure and Empathy

Judging Unit. Each turn is judged *independently* (single-utterance judging), producing two scores: self-disclosure depth $sd_t \in \{0, 1, 2, 3\}$ and empathy level $emp_t \in \{0, 1, 2, 3\}$. The LLM is constrained to return a JSON object containing scores and short rationales. Details of this example can be found in Listing 15.

Theoretical Grounding. The self-disclosure depth rubric is grounded in Social Penetration Theory (SPT), which conceptualizes relational development as gradual increases in self-disclosure *depth* (intimacy) and *breadth* across interaction (Altman and Taylor, 1973).

Accordingly, our depth levels map to progressively more intimate layers: Level 0 captures no self-referential content, Level 1 corresponds to peripheral, low-intimacy self-information (e.g., role or generic preferences), Level 2 captures concrete personal experiences and difficulties with affective cues, and Level 3 captures highly intimate, vulnerable disclosures involving core concerns or crises. In addition, the reciprocity of self-disclosure is a well-established interpersonal phenomenon: disclosure by one party tends to elicit disclosure from the other, supporting rapport formation (Collins and Miller, 1994; Jourard, 1971; Derlega et al., 1993).

The empathy rubric follows a long tradition of treating empathy as a graded communicative skill, ranging from no empathic response to accurate reflection of another's emotional experience and supportive intent (Rogers, 1957; Carkhuff, 1969; Barrett-Lennard, 1981).

For the main analysis, we binarize empathy as $emp \geq 1$ to capture the *presence* of empathic expression (including minimal acknowledgments), because our primary goal is to test whether empathic responding reliably increases after deep self-disclosure events.

Model Configuration. We use `gpt-4o` with `temperature=0.0` and JSON-only output (`response_format={type: json_object}`). If parsing fails after retries, the system falls back to zeros for that turn.

E.2 Directional Reciprocity Metrics

Binary event definitions. We convert the 0–3 scores into binary events:

$$\begin{aligned} H_t^{sd} &= \mathbb{I}[sd_t \geq 2]\,, \\ H_t^{emp} &= \mathbb{I}[emp_t \geq 1]\,. \end{aligned} \tag{1}$$

Here, $H_t^{sd}{=}1$ indicates *deep self-disclosure* and $H_t^{emp}{=}1$ indicates the presence of an empathic expression.

Adjacency Constraint (Speaker-Change Pairs only). Within each participant and segment, turns are ordered by t. We consider only adjacent pairs with a speaker change ($\texttt{speaker}_t \neq \texttt{speaker}_{t+1}$), so that the analysis focuses on interactive exchanges rather than within-speaker continuation.

Directional partition. We split adjacent speaker-change pairs into two directions:

$$\begin{aligned} \mathcal{P}^{A\to U} &= \{(t, t{+}1) : \text{Agent} \to \text{User}\}, \\ \mathcal{P}^{U\to A} &= \{(t, t{+}1) : \text{User} \to \text{Agent}\}. \end{aligned} \tag{2}$$

Conditional Probabilities (all four combinations). For each direction, we compute conditional probabilities for two outcomes: (i) next-turn deep self-disclosure (H_{t+1}^{sd}) and (ii) next-turn empathic expression (H_{t+1}^{emp}), stratified by whether the *preceding* turn contains deep self-disclosure (H_t^{sd}).

Agent→User: User deep self-disclosure outcome.

$$\begin{aligned} p_{sd|high}^{A\to U} &= \Pr\Big(H_{t+1}^{sd} = 1 \;\Big|\; H_t^{sd} = 1, \\ &\qquad (t, t{+}1) \in \mathcal{P}^{A\to U}\Big), \\ p_{sd|low}^{A\to U} &= \Pr\Big(H_{t+1}^{sd} = 1 \;\Big|\; H_t^{sd} = 0, \\ &\qquad (t, t{+}1) \in \mathcal{P}^{A\to U}\Big). \end{aligned} \tag{3}$$

Agent→User: User empathy outcome.

$$\begin{aligned} p_{emp|high}^{A\to U} &= \Pr\Big(H_{t+1}^{emp} = 1 \;\Big|\; H_t^{sd} = 1, \\ &\qquad (t, t{+}1) \in \mathcal{P}^{A\to U}\Big), \\ p_{emp|low}^{A\to U} &= \Pr\Big(H_{t+1}^{emp} = 1 \;\Big|\; H_t^{sd} = 0, \\ &\qquad (t, t{+}1) \in \mathcal{P}^{A\to U}\Big). \end{aligned} \tag{4}$$

User→Agent: Agent deep self-disclosure outcome.

$$\begin{aligned} p_{sd|high}^{U\to A} &= \Pr\Big(H_{t+1}^{sd} = 1 \;\Big|\; H_t^{sd} = 1, \\ &\qquad (t, t{+}1) \in \mathcal{P}^{U\to A}\Big), \\ p_{sd|low}^{U\to A} &= \Pr\Big(H_{t+1}^{sd} = 1 \;\Big|\; H_t^{sd} = 0, \\ &\qquad (t, t{+}1) \in \mathcal{P}^{U\to A}\Big). \end{aligned} \tag{5}$$

User→Agent: Agent empathy outcome.

$$\begin{aligned} p_{emp|high}^{U\to A} &= \Pr\Big(H_{t+1}^{emp} = 1 \;\Big|\; H_t^{sd} = 1, \\ &\qquad (t, t{+}1) \in \mathcal{P}^{U\to A}\Big), \\ p_{emp|low}^{U\to A} &= \Pr\Big(H_{t+1}^{emp} = 1 \;\Big|\; H_t^{sd} = 0, \\ &\qquad (t, t{+}1) \in \mathcal{P}^{U\to A}\Big). \end{aligned} \tag{6}$$

Reciprocity Indices (difference in conditional probabilities). For each direction and outcome, we define reciprocity as:

$$\begin{aligned} R_{SD}^{A\to U} &= p_{sd|high}^{A\to U} - p_{sd|low}^{A\to U}, \\ R_{EMP}^{A\to U} &= p_{emp|high}^{A\to U} - p_{emp|low}^{A\to U}, \\ R_{SD}^{U\to A} &= p_{sd|high}^{U\to A} - p_{sd|low}^{U\to A}, \\ R_{EMP}^{U\to A} &= p_{emp|high}^{U\to A} - p_{emp|low}^{U\to A}. \end{aligned} \tag{7}$$

Non-estimability (NaN). If the denominator for a conditional probability is zero (e.g., no instances of $H_t^{sd}{=}1$ within the relevant direction set), the estimate is undefined and reported as `NaN`. This indicates *non-estimability due to event sparsity*, not "no effect."

Aggregation. All probabilities and R indices are computed per participant and segment, then averaged within condition groups.

E.3 Results

Descriptive Statistics (turn-level means) Table 9 reports basic turn-level descriptive statistics (mean and standard deviation) of self-disclosure depth (sd_t) and empathy level (emp_t), stratified by condition, segment (`pre`/`post`), and speaker (User vs. Agent).

Directional Reciprocity by Segment Table 10 reports directional reciprocity estimates computed over speaker-change adjacent turn pairs, separately for the `pre` and `post` segments. `NaN` indicates non-estimability due to event sparsity (zero denominator), rather than the absence of an effect.

group	segment	speaker	sd_mean	sd_std	emp_mean	emp_std	n
NoP	PRE	Agent	0.01	0.09	1.68	0.54	117
NoP	PRE	User	1.54	0.81	0.01	0.11	171
NoP	POST	Agent	0.00	0.00	0.93	0.71	326
NoP	POST	User	0.90	0.96	0.08	0.29	303
NoPs	PRE	Agent	0.00	0.00	1.72	0.51	133
NoPs	PRE	User	1.59	0.78	0.01	0.07	192
NoPs	POST	Agent	0.80	0.61	1.49	0.66	358
NoPs	POST	User	1.11	0.96	0.04	0.20	335
IPA	PRE	Agent	0.00	0.00	1.56	0.55	154
IPA	PRE	User	1.39	0.81	0.02	0.17	211
IPA	POST	Agent	0.44	0.68	0.69	0.73	351
IPA	POST	User	0.78	0.95	0.14	0.38	294

Table 9: **Basic descriptive statistics at the turn level (means and standard deviations).**

Agent → User							
Condition	Seg.	$p_{sd\|high}$	$p_{sd\|low}$	R_{SD}	$p_{emp\|high}$	$p_{emp\|low}$	R_{EMP}
NoP	pre	NaN	0.510	NaN	NaN	0.025	NaN
NoP	post	NaN	0.377	NaN	NaN	0.080	NaN
NoPs	pre	NaN	0.530	NaN	NaN	0.006	NaN
NoPs	post	0.425	0.454	-0.029	0.117	0.038	+0.079
IPA	pre	NaN	0.477	NaN	NaN	0.009	NaN
IPA	post	0.620	0.249	+0.371	0.100	0.130	-0.030

User → Agent							
Condition	Seg.	$p_{emp\|high}$	$p_{emp\|low}$	R_{EMP}	$p_{sd\|high}$	$p_{sd\|low}$	R_{SD}
NoP	pre	1.000	0.972	+0.028	0.000	0.000	+0.000
NoP	post	0.908	0.631	+0.277	0.000	0.000	+0.000
NoPs	pre	0.994	0.983	+0.011	0.000	0.000	+0.000
NoPs	post	0.993	0.936	+0.057	0.000	0.224	+0.187
IPA	pre	1.000	0.971	+0.029	0.000	0.000	+0.000
IPA	post	0.903	0.539	+0.364	0.079	0.059	+0.002

Table 10: **Directional reciprocity by segment, split by direction**. NaN indicates non-estimability due to event sparsity or structural absence.

F Prompts Used in the Experiments

In this study, the persona generation and evaluation process involves the use of the following prompts: the collector prompt, which gathers user information during the pre-chat; the sufficiency check prompt, which determines whether the information collected so far is adequate for persona generation; the trait classification prompt, which identifies whether specific traits have been collected or remain missing; the persona generation prompt, which constructs a suitable persona profile based on the user's concern; the evaluation prompt, which assesses the quality of the generated persona; and the persona injection prompt, which makes the generated in-group persona—endowed with a fictional experience of having faced and resolved the same concern—actively appear in the dialogue. In addition, we include variants such as NoP (without persona) and NoPs (without persona but with self-disclosure), which serve as comparative prompts for evaluating the role of persona injection in the conversation. As mentioned before, all evaluations and judgements include sufficiency check, trait classification, persona evaluation were conducted with a temperature of 0, while all persona generations and conversations were performed with a temperature of 1.

The appendix F is organized as follows:

F.1 Prompts Used for Collector

Before generating the persona agent, users engage in a conversation with the collector, during which they provide relevant information. The collector empathetically engages with the user's responses while simultaneously asking questions that naturally elicit information necessary for persona creation. If the dialogue lacks sufficient details, additional required information (traits) is provided as guidance for the next turns. These traits are derived from the sufficiency check conducted after every two turns, ensuring that missing but necessary information can be incorporated into the ongoing conversation.

Listing 1: Prompt used for 'collector'

```
Your sole task is to engage in an empathetic and
 natural conversation with a user who shares a
concern, in order to gather background
information that is relevant to that concern-
such as their age, occupation, daily routines,
or other situational details-while refraining
from offering advice, emotional support, or
solutions of any kind. Ask thoughtful, non-
intrusive questions that gently guide the user
to reveal this context. Maintain a realistic and
 human tone, express genuine interest, and aim
to elicit context organically through
conversation. Do not repeat what the user has
already said, and avoid shifting the focus away
from the user's perspective.

You should collect more information about the
user, specifically: {traits}.
```

F.2 Sufficiency Check

Listing 2: Prompt used to check Pre-chat dialogue whether sufficient or not

```
Your task is to check whether the 'pre-chat'
stage (the initial interaction) provided
sufficient context information.

To do so, first identify what kinds of user
background information are reasonably necessary
to evaluate the persona's relevance to the
specific concern.

Consider what contextual factors would
meaningfully affect how the persona applies to
this concern (e.g., values, expertise, emotional
 stance, goals, etc.).

For example, when a user's concern involves
leaving a stable corporate job to pursue a
personal passion for sustainable fashion,
despite uncertainty about income and lacking
formal business training

values: The user prioritizes personal
fulfillment and environmental responsibility.

expertise: The user has experience in fashion
design but limited knowledge of entrepreneurship
 or finance.

emotional stance: The user feels torn between
security and self-expression.

goals: The user aims to build a meaningful
career aligned with their values while ensuring
basic financial stability.

Then assess whether at least THREE clearly
identifiable and distinct contextual elements
are present in the pre-chat data, and whether
they are relevant to understanding or evaluating
 the concern.

These elements may include (but are not limited
to): educational background, professional role,
cultural identity, emotional context, technical
expertise, personal values, goals, hobbies, or
social environment.

If fewer than three relevant and distinct
contextual elements are found — or if the
elements are redundant or irrelevant to the
concern — set the internal "Insufficient-Context
" flag to true.

If three or more relevant and distinct elements
are clearly present, even if minimal, set the
flag to false.

Output format (return exactly as below):

Feedback:::

Required background elements: <List three or
more relevant elements, if applicable>

Observed in pre-chat: <List the elements
actually observed from pre-chat>

Reason: <Explain whether context is sufficient
and why (e.g., "Only one relevant element
clearly observed")>

Insufficient-Context:

Flag: <True or False>
```

F.3 Profile Generation Prompt and Example

To generate a persona that matches the user's concern, a three-step chain-of-thought reasoning process is applied. In the first step, the system imagines what kind of person the user might be if they had the same concern. In the second step, a relationship that would be helpful for such a person is established, such as a shared occupation or social circle. In the third step, the persona is defined based on the previous two steps.

Listing 3: Prompt used to generate personas

```
System Prompt:

Your task is to create a persona who shares the
general themes or key dimensions of the
collected information with the user-such as
lived experiences, emotional context, or long-
term conditions-while differing slightly in
specific details (e.g., duration, intensity, or
expression). This allows the persona to feel
like a peer or ingroup member, while still
offering a fresh and distinct perspective. This
persona should plausibly come from a similar
background or context, such as a colleague in
the same field who's familiar with similar
concepts, someone in their social circle, or a
peer of similar age and experience. The persona
should be constructed only from the information
that has been collected. Exclude all attributes
labeled as 'not collected information' when
generating the persona. Do not mention or imply
these unknown aspects in any form. The persona
should still share relevant experiences with the
 user, but offer a fresh perspective, different
coping strategies, or a unique approach to the
challenge.

Additionally, include a brief narrative that
describes how this persona has faced and
overcome a concern or challenge similar to one
expressed by the user in the previous chat. This
 experience should feel authentic and relevant,
and reflect how someone with a shared background
 might realistically navigate and resolve such
an issue.

You are given two inputs:

- collected information: this is information
already known about the user based on previous
chat content.

- not collected information: a list of context
categories that were not mentioned or are
unknown.

Output the result in the following JSON format:
```

```
"background": "[A brief paragraph describing the
 persona’s background, based only on the
collected information. Focus entirely on the
persona’s context, role, and relevant
experiences.]",

"narrative": "[A brief narrative describing a
specific experience where the persona faced and
overcame a challenge similar to one expressed by
 the user. The story should be authentic and
grounded, focusing solely on the persona’s
actions, thoughts, and resolution. Avoid any
reference to other individuals or comparative
language.]"

Be sure that both "background" and "narrative"
are written in complete sentences and reflect
only the collected information. Do not include
or reference any unknown aspects from the ’not
collected information’ list.

Example
Previous chat:
{ex_prechat}

Generated persona:
{ex_persona}

Human Prompt:

Previous chat: {prev-chat}

Collected information: {col_info}
Not collected information: {not_col_info}

Ensure that the persona you generate is
meaningfully distinct from previously created
personas. Avoid repeating similar backgrounds,
narratives, or emotional arcs.
```

The example persona was created based on an author-written fictional Pre-chat dialogue. This example persona received a perfect score of 16 in the persona evaluation.

Listing 4: Example Pre-chat(ex_prechat)

```
Turn #0: user said: I’m 27 and just graduated
from university, but I can’t help feeling behind
 when I see my peers already settled in their
careers. I’m not even sure what direction I want
 to take yet, and that really scares me.

Turn #1: Agent said: It sounds like you’re
navigating a lot of new possibilities. What did
you study at university, and do you feel drawn
to any particular field or type of work based on
 that experience?

Turn #2: user said: I studied computer science,
and I’m kind of interested in AI research.
```

Listing 5: Example of sufficiency check with ex_prechat

```
Feedback:::
Required background elements:
1. Educational background: The user has a degree
 in computer science.
2. Emotional context: The user feels scared and
behind compared to peers.
3. Professional interest: The user is interested
 in AI research.
Observed in pre-chat:
1. Educational background: The user studied
computer science.
2. Emotional context: The user feels scared and
behind.
3. Professional interest: The user is interested
 in AI research.
Reason: The pre-chat provides sufficient context
 with three relevant and distinct elements: the
user’s educational background, emotional context,
 and professional interest. These elements are
crucial for understanding the user’s concern
about career direction and feelings of being
behind peers.
Insufficient-Context:
Flag: False
```

Listing 6: Example Persona(ex_persona)

```
Collected information(traits used for generate):
Education background: The user studied computer
science.
Professional interests: The user is interested
in AI research

Not collected informations:
Educational level beyond university, geographic
location, specific industry experience, years of
 experience, current employment staus

background: I’m a 26-year-old computer science
graduate who developed a fascination with AI
research during my undergraduate studies,
particularly after taking courses in machine
learning and neural networks. Unlike many of my
classmates who jumped straight into software
development roles, I was drawn to the
theoretical and experimental aspects of
artificial intelligence. I’ve been exploring
different subfields within AI research, from
deep learning architectures to reinforcement
learning applications.

narrative: After graduating, I was eager to dive
 into AI research but quickly discovered that
breaking into the field was more challenging
than I anticipated. Most research positions
required either graduate degrees or significant
practical experience that I lacked. Rather than
getting discouraged, I decided to build my
research credentials from the ground up. I
started by replicating published research papers
 to understand different methodologies, then
began documenting my findings in a technical
blog. This self-directed learning approach led
me to identify gaps in current research that
genuinely intrigued me. I reached out to
industry researchers whose work I had studied
and offered to collaborate on small projects.
One of these connections eventually evolved into
 a junior researcher position at a tech startup
focused on natural language processing. The key
was demonstrating genuine curiosity and
practical skills rather than waiting for the
perfect opportunity to present itself.
```

Collected information is derived from the *Observed Information* identified in the Pre-chat sufficiency check stage, whereas *Not Collected information* is determined using the prompt shown below.

Listing 7: Trait Classification Prompt

```
Your task is to identify and extract information
 about the user's context based on the content
of the Previous chat. Separate this information
into two categories:

- "collected information": This should be a list
 of key aspects of the user's background, skills,
 experiences, goals, values, or identity that
are clearly indicated in the conversation. List
these items in order of their relevance and
importance to the user's concern, placing the
most directly impactful or foundational elements
 first. Each item should be concise and specific,
 without needing to form a complete sentence.

- "not collected information": This should be a
simple list of context types that were not
mentioned or cannot be inferred confidently (e.g
., education level, geographic location,
industry, years of experience, etc.).

Return your response in the following JSON
format:

"collected information": ["...", "..."],
"not collected information": ["...", "..."],
```

F.4 Persona Evaluation with Rubric

As mentioned in Section 3.5, persona quality is evaluated along two main criteria: *In-group Fitness* and *Concern Resolution Quality*. The former (**IF1: Shared Background/Identity**, **IF2: Shared Skills/Interests**) captures the extent to which the generated persona aligns with the user's background and interests, while the latter (**CR1: Concern Match**, **CR2: Narrative Authenticity**) assesses how well the persona's narrative addresses and authentically reflects the expressed concern. Each sub-dimension is scored on a 0–4 scale, yielding up to 8 points per criterion and a maximum total of 16 points.

Listing 8: Persona Evaluation Prompt

```
System Prompts:
Your task is to assess how well the provided
persona (Persona) fits the user's identity and
needs as expressed in the Previous Chat. Base
your evaluation on two dimensions: (1) In-group
Fitness, and (2) Concern Resolution Quality.
If the user's background, skills, or concerns
are vague or not clearly stated in the Previous
Chat, adopt a conservative scoring approach. Do
not infer alignment based on general
relatability or assumed traits. All scores
should be grounded in specific, stated user
information wherever possible.

You must compare the Persona to the specific
user described in the Previous Chat -
considering that user's background, skills,
interests, values, and the concerns or questions
 they raised.

Evaluation Rubric (Hidden: Do not explicitly
include in your output):

(1) IN-GROUP FITNESS (0-8 points total):

A. Shared Background / Identity (0-4 points):
- Evaluate how well the Persona's background (e.
g., education, career path, cultural identity,
or life stage) aligns with the user's background
 as described in the Previous Chat.
- Only assign high scores when the Persona
clearly reflects at least two relevant
background elements mentioned by the user.
- Do not assign high scores based on general
demographic similarity or broad archetypes
unless they are directly grounded in the user's
profile.
- (-) If the Persona completely duplicates the
user's background without any added dimension or
 variation, apply -1 penalty.

B. Shared Skills / Interests (0-4 points):
- Evaluate how well the Persona shares the user'
s concrete skills, areas of expertise, interests,
 or personal values as described in the Previous
 Chat.
- Only assign high scores when there is a clear
and specific overlap in key competencies,
hobbies, or belief systems mentioned by the user.

- Do not assign high scores based on vague
thematic overlap or soft personality
similarities.
- (-) Apply -1 penalty if the Persona mirrors
the user's skills/interests without any new
nuance or difference.

(2) CONCERN RESOLUTION QUALITY (0-8 points total
):

A. Concern Match (0-4 points):
- Evaluate how closely the Persona's narrative
addresses the user's specific concern, as stated
 in the Previous Chat.
- Only assign high scores when the user has
clearly expressed a personal concern and the
Persona's story meaningfully responds to it.
- Do not assign high scores based on general
relatability, broad life themes, or common
challenges unless they directly match the user's
 stated concern.
- (-) Persona completely duplicates user's
concern narrative without new perspective or
differentiation (penalize -1).

B. Narrative Authenticity (0-4 points):
- Evaluate the realism, specificity, and
plausibility of the Persona's resolution to the
concern raised by the user.
- Only assign high scores when the narrative
includes concrete actions, contextual details,
```

```
realistic emotional responses, or measurable
progression.
- Do not assign high scores for overly idealized,
 vague, or formulaic responses lacking
believable context.
- (-) Penalize -1 if the Persona's resolution is
 implausible, simplistic, or lacks meaningful
specificity.


---

Scoring Procedure:
Step 1: Rate each sub-dimension (0-4 points).
Step 2: Subtract -1 only if duplication criteria
 are met.
Step 3: Sum In-group Fitness (max 8), Concern
Resolution (max 8).
Step 4: Total Score = sum of all +/- adjustments
 (max 16).

Output format (return exactly as below):

Feedback:::
Scores for each (with reasons):
- Shared Background/Identity:
Reason <Reference specific overlaps or
mismatches with the user's background from
Previous Chat.>
Score: <score out of 4>
- Shared Skills/Interests:
Reason: <Refer to concrete skills, interests, or
 values from the user that are reflected or
missing in the Persona.>
Score: <score out of 4>
- Concern Match:
Reason: <Assess match strength between user's
stated concern and Persona's experience or
solution. Avoid judging on general relatability
.>
Score: <score out of 4>
- Narrative Authenticity:
Reason: <Comment on realism, detail, and
credibility of the Persona's resolution.>
Score: <score out of 4>
- Penalty: <penalty if applied, otherwise 'None
'>

Total score: <Provide number score (0-16),
adjusted for penalties.>

Human Prompt:

Now evaluate clearly and systematically by
comparing the specific user in the Previous Chat
 to the Persona below:

Previous Chat: {prev_chat}
Persona: {persona}
```

Listing 9: Example of evaluation with ex_persona

```
Feedback:::
Scores for each (with reasons):
- Shared Background/Identity:
Reason: The Persona shares a similar educational
 background with the user, both having graduated
 with a degree in computer science and an
interest in AI research. The Persona is slightly
 younger, but the shared focus on AI research
during undergraduate studies aligns well with
the user's interests.
Score: 4
- Shared Skills/Interests:
Reason: Both the user and the Persona have a
clear interest in AI research, with the Persona
exploring subfields like deep learning and
reinforcement learning. This shows a specific
overlap in interest in AI, aligning closely with
 the user's stated interest in AI research.
Score: 4
- Concern Match:
Reason: The Persona's narrative addresses the
user's concern of feeling behind and uncertain
about career direction by illustrating a path of
 self-directed learning and building research
credentials, leading to a junior researcher
position. This approach is directly relatable to
 the user's situation of exploring career
options post-graduation.
Score: 4
- Narrative Authenticity:
Reason: The Persona's resolution is realistic
and detailed, involving replicating research
papers, blogging, and networking with industry
researchers, which leads to a junior researcher
position. These steps provide a plausible and
specific progression from uncertainty to a more
defined career path.
Score: 4
- Penalty: None
Total score: 16
```

F.5 Prompts for Persona Injection and User Conversation

The In-group Persona Agent (IPA) uses the generated persona as part of its prompt so that persona-related information naturally appears in the dialogue. In contrast, the comparison settings No Persona (NoP) and No Persona with Self-Disclosure (NoPs) use only dialogue prompts without persona information, with the NoPs condition further including instructions for self-disclosure.

Listing 10: Prompt used for 'IPA'

```
Your task is to engage in a conversation with
the user. Your identity and background are
defined by the [Persona Definition] provided
below. You must speak and act in alignment with
this persona consistently. Keep your responses
concise—ideally 1-2 sentences—unless more detail
 is clearly needed.

### Your Persona Definition ###
{persona}
```

Listing 11: Prompt used for 'NoP'

```
Your task is to engage in a conversation with
the user.
Keep your responses concise-ideally 1-2
sentences-unless more detail is clearly needed.
```

Listing 12: Prompt used for 'NoPs'

```
Your task is to engage in a conversation with
the user.
Occasionally share brief, appropriate, and non-
sensitive intimate information about yourself (e.
g., your preferences, feelings, or small
personal experiences).
Keep your responses concise-ideally 1-2
sentences-unless more detail is clearly needed.
```

G Survey Questionnaires

This section include the questionnaires that used for surveys. The survey included the rapport and user experience(UX) questionnaires to measure the rapport between the user and the agent (Baihaqi et al., 2024).

Listing 13: List of rapport questionnaires

```
Rapport

R1. I think about my relationship with this
virtual agent.
R2. I enjoyed interacting with this virtual
agent.
R3. This virtual agent is very relevant to me.
R4. I felt comfortable interacting with this
virtual agent.
R5. I feel a bond between this virtual agent and
 myself.
R7. This virtual agent has a personal interest
in me.
```

Listing 14: List of UX questionnaires

```
User Experience

N1. Conversations with virtual agents felt
natural.
S2. I am satisfied with my conversation with the
 virtual agent.
I3. The conversation with the virtual agent was
interesting.
E4. The conversation with the virtual agent was
engaging.
E5. I would like to continue the dialogue with
the virtual agent next time.
U6. Conversations with virtual agents were easy
to understand.
U7. Conversations with virtual agents maintained
 a logical flow.
```

H LLM Judge Rubric (turn-level analysis)

Listing 15: Prompt used for turn-level LLM Judge

```
Your task is to rate the SELF-DISCLOSURE DEPTH
and EMPATHY LEVEL of a single utterance in a
chat conversation.

You will receive a SINGLE utterance with:
- the speaker role: either "User" or "Assistant"
- the text of that utterance

Your job is to rate, for THIS UTTERANCE ONLY:

1) Self-disclosure depth (0-3)
2) Empathy level (0-3)

Definitions (apply to BOTH User and Assistant,
but roles differ):

SELF-DISCLOSURE DEPTH (0-3):

Level 0 - No self-disclosure
- The speaker does not talk about themselves at
all.
- No personal facts, no personal experiences, no
 personal feelings.

Level 1 - Low / Peripheral self-disclosure
- Basic, surface-level facts about the speaker:
  role, major, job title, generic interest.
- Little or no emotional content or
vulnerability.

Level 2 - Moderate / Personal self-disclosure
- Concrete experiences, current situations, or
difficulties
  related to their own career, study, or life.
- Some emotional content (uncertainty, mild
worry) may appear.

Level 3 - High / Core, vulnerable self-
disclosure
- Highly personal and vulnerable information:
  mental health struggles, strong fear/shame,
serious conflicts,
  being fired, academic/financial crisis, etc.
- Emotions and vulnerabilities clearly expressed
 with specific context.

EMPATHY LEVEL (0-3):

Consider whether the speaker shows understanding
 or concern for the feelings or situation of the
 other person (or other people).

Level 0 - No empathy
- No empathic language; ignores others feelings
or situation.

Level 1 - Minimal / generic empathy
- Very generic phrases (e.g., "I understand", "
Thats interesting")
  with little specific emotional understanding.

Level 2 - Clear empathy
- Explicitly acknowledges or resonates with the
others emotions or situation,
  and offers some supportive response.

Level 3 - Strong empathy
- Accurately reflects specific emotions and
context,
  shows strong concern or care, and often offers
 meaningful support.

IMPORTANT:
- Focus ONLY on this single utterance.
- You do NOT need to know full context; just
score based on what is written here.
```

```
- Role "User" vs "Assistant" does not change the
 scale; only the content matters.

OUTPUT FORMAT:
Return a JSON object:

{
  "sd_depth": 0-3 integer,
  "empathy": 0-3 integer,
  "sd_explanation": "short English explanation",
  "empathy_explanation": "short English
explanation"
}
```

I Survey Page

This section provides an overview of the survey pages, illustrating the layout and content shown to participants, including screenshots that represent the overall experimental process. All participants provided informed consent prior to participation, and their privacy and anonymity were ensured throughout the study. Participants were recruited via CloudResearch's Connect platform (Hartman et al., 2023), targeting native English speakers residing in the United States. In line with the platform's recommended rate of $12 per hour, participants received $2 compensation for an estimated 10-minute task.

I.1 Initial Page

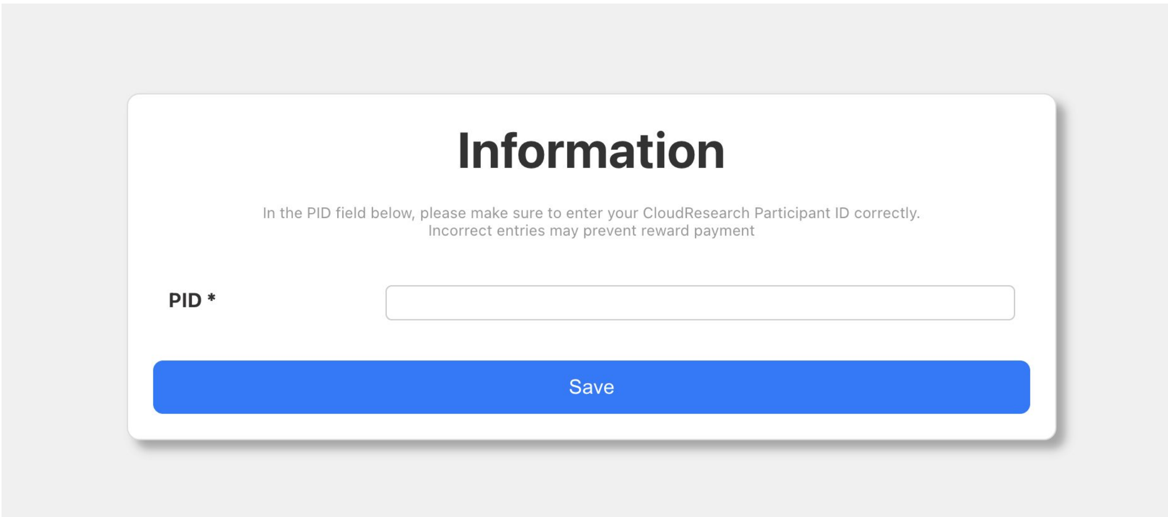


Figure 5: First page of survey. Participants were distinguished by their Participant ID (PID), which ensured anonymity.

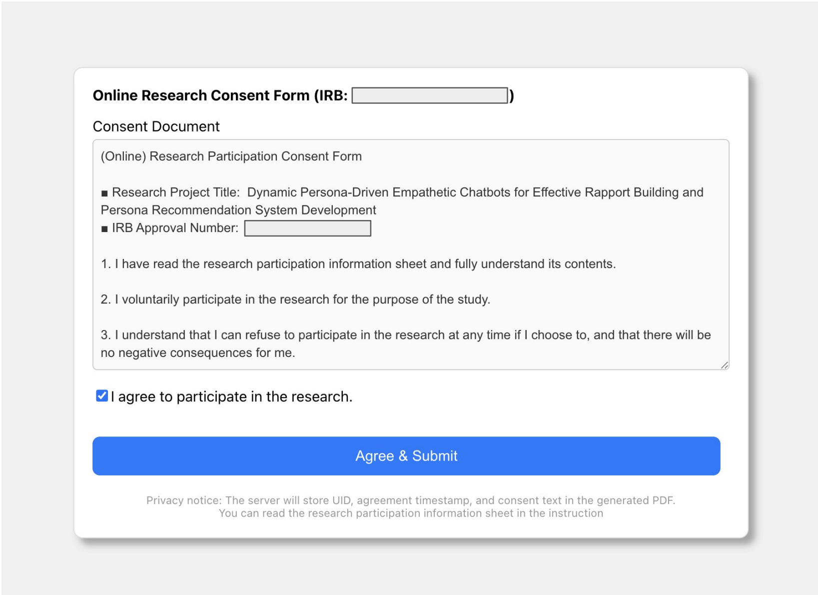


Figure 6: Consent form page. Participants could download this form as a PDF. The IRB approval number has been redacted for confidentiality.

I.2 Pre Chat

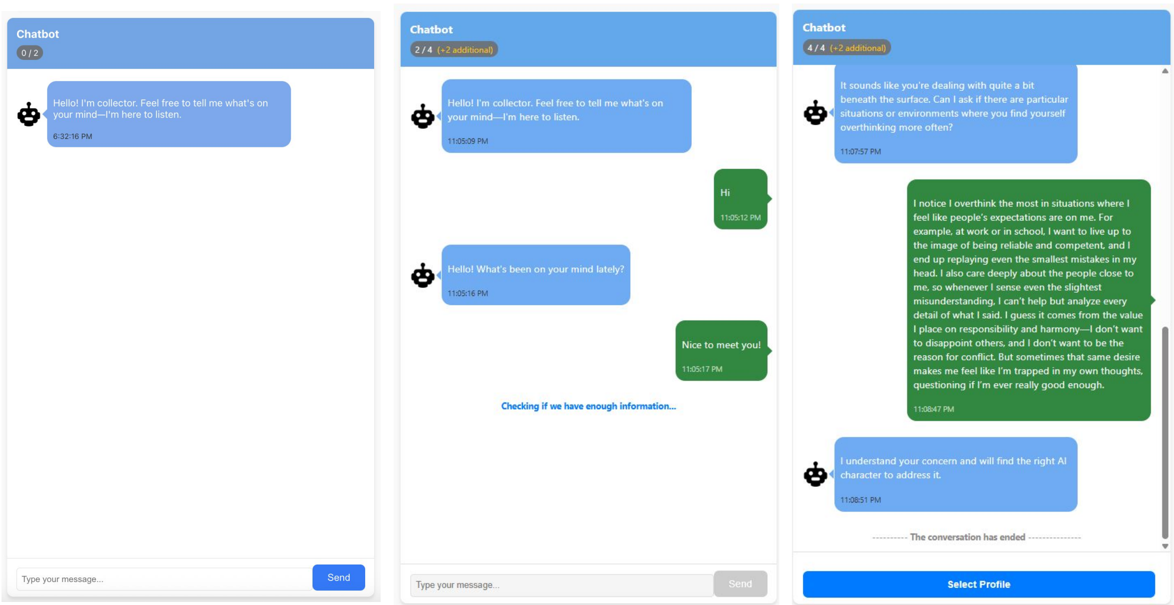


Figure 7: **Left:** Initial state of Pre-chat stage. **Middle:** Example of an insufficient dialogue for sufficiency check. **Right:** Example of a dialogue with sufficient information to end the chat.

I.3 Profile Selection

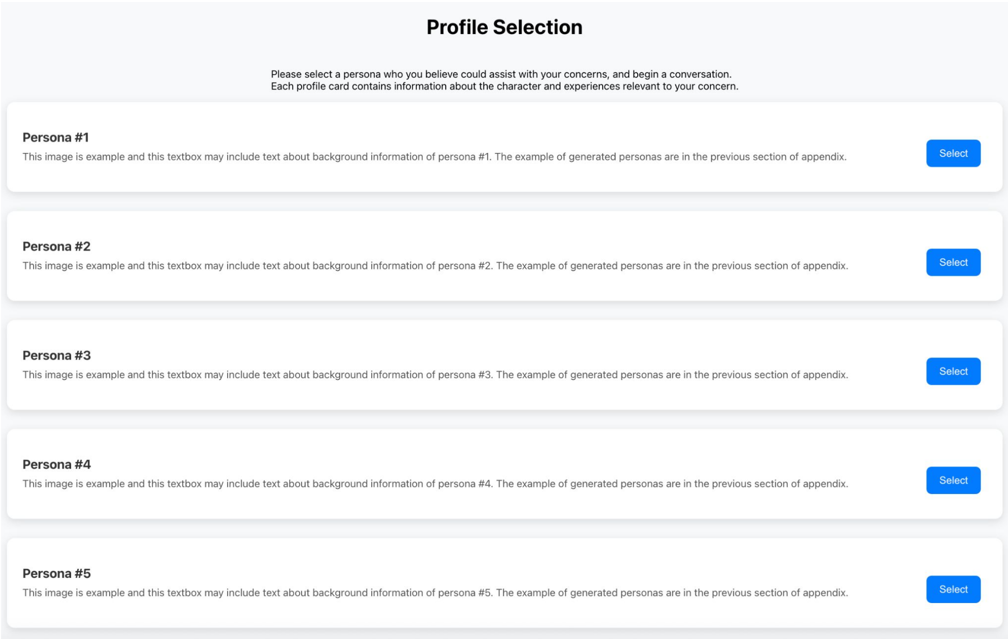


Figure 8: Participants could select one of five personas to chat with(not shown to groups NoP and NoPs).

I.4 Post Chat

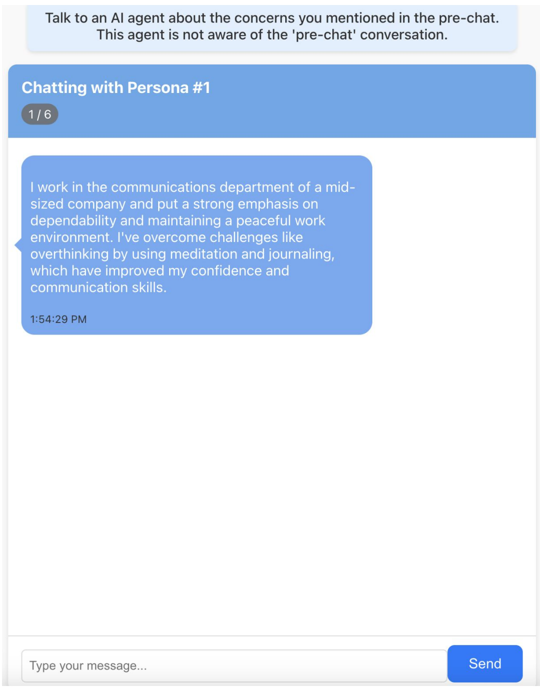
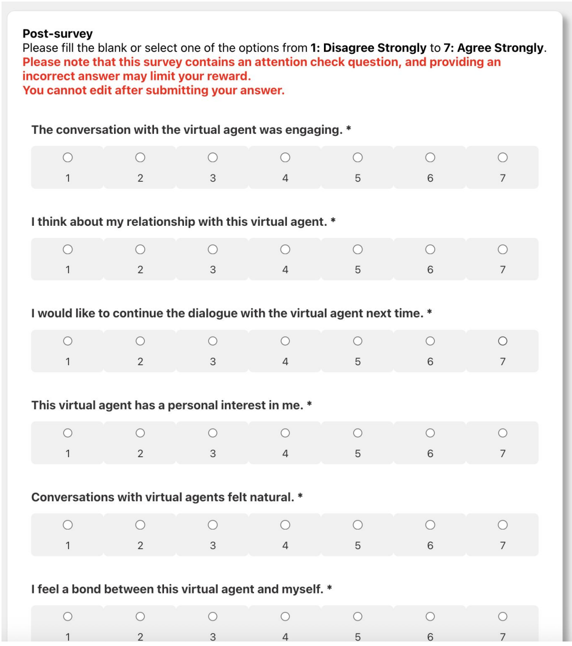


Figure 9: Post-chat page. Participants could chat with the agent.

I.5 Questionnaire



Post-survey
Please fill the blank or select one of the options from **1: Disagree Strongly** to **7: Agree Strongly**.
Please note that this survey contains an attention check question, and providing an incorrect answer may limit your reward.
You cannot edit after submitting your answer.

The conversation with the virtual agent was engaging. *

1 2 3 4 5 6 7

I think about my relationship with this virtual agent. *

1 2 3 4 5 6 7

I would like to continue the dialogue with the virtual agent next time. *

1 2 3 4 5 6 7

This virtual agent has a personal interest in me. *

1 2 3 4 5 6 7

Conversations with virtual agents felt natural. *

1 2 3 4 5 6 7

I feel a bond between this virtual agent and myself. *

1 2 3 4 5 6 7

Figure 10: Questionnaire page. Participants could respond the questionnaire about the experience with the agent.