Relationship-adaptive Agent: A Chatbot Based on Social Penetration Theory to Encourage Self-disclosure

Chaerim Sung [1] and Sanghun Nam [2,*]

[1] Department of Culture Technology Convergence, Changwon National University, Changwon, South Korea
[2] School of Meta-convergence Content, Changwon National University, Changwon, South Korea
Email: scl7317@gs.cwnu.ac.kr (C.S.); sanghunnam@changwon.ac.kr (S.N.)
*Corresponding author

***Abstract*—Advances in artificial intelligence have led an increasing number of users to seek emotional connections with conversational agents. However, existing agents have a limited ability to form deep emotional bonds or trust because of their static nature and fixed persona. In this study, a model of interpersonal relationship development called social penetration theory was used to develop a relationship-adaptive agent that dynamically adjusts its communication style and behavior according to the depth of interaction. The agent updates a closeness score by analyzing the depth, breadth, and reciprocity of the user's self-disclosure to determine the relationship stage: orientation, exploratory, affective, and stable. A 10-min experiment was conducted with 30 participants to compare their interactions with the relationship-adaptive agent (experimental group) to those with a static agent (control group). The results indicated that the experimental group demonstrated significantly higher scores across all metrics (i.e., emotional bond, trust, and perceived friendship) compared to the control group. Qualitative analysis revealed that the frequency of high-depth self-disclosure (e.g., revealing vulnerabilities or secrets) was more than five times higher in the experimental group than in the control group. This study demonstrates an engineering application of Social Penetration Theory to generative artificial intelligence and suggests that stage-dependent adaptive behaviors can enhance users' reported emotional bond, trust, perceived friendship, and high-depth self-disclosure in a short-term guided interaction.**

***Keywords*—artificial intelligence, computer-mediated communication, self-disclosure, social penetration theory**

I. Introduction

With the advent of generative Artificial Intelligence (AI) and Large Language Models (LLMs), conversational agents (i.e., chatbots) have become capable of understanding context and generating human-like natural language. Consequently, their scope of application is expanding beyond simple information retrieval or secretarial roles to include serving as emotional companions, such as in elderly care or as friends [1, 2]. Modern users expect not only functional utility from agents but also emotional responses such as empathy and intimacy. Users' preference for services like Apple's Siri and Amazon's Alexa has been shown to increase when these agents include humor or emotional expressions in their interactions [3, 4]. Such emotional bonding is a key factor in encouraging users to use AI over the long term. Recent studies have attempted to enhance the humanness of conversational agents by introducing personas or combining multimodal elements [5–8]. However, existing agents exhibit static characteristics and maintain their initial persona regardless of the duration of the conversation [9]. Human conversation involves feeling a social presence within a relational context, building trust and intimacy, and exchanging emotions. Furthermore, human relationships become more intimate over time, and deeper interactions result in changes in speech styles and behaviors [10]. In contrast, current agents fail to reflect this process of relationship development, which limits their ability to form deep emotional bonds or long-term trusting relationships with users. True personalization must go beyond simply reflecting tastes to include changes in the agent's attitude according to the context of the relationship.

The aim of this study was to develop a conversational agent that adapts its behavioral strategies according to modeled relationship stages rather than maintaining a fixed persona. Social Penetration Theory (SPT) from social psychology was used to develop a relationship-adaptive agent that recognizes stages in a relationship according to the depth of interaction with the user and actively changes its behavioral strategies. The core objective was to verify whether changes in the agent's behavior according to the depth of the relationship have positive effects on the user's emotional bond, trust, and perceived friendship and whether such changes can induce users to engage in deeper levels of self-disclosure. Four hypotheses were established: the relationship-adaptive agent will elicit (H1) a stronger emotional bond, (H2) greater trust, (H3) greater

Manuscript received June 24, 2026; revised July 9, 2026; accepted September 1, 2026; published November 25, 2026.

friendship, and (H4) deeper self-disclosure from the user compared to a static agent.

II. Related Work

A. Social Penetration Theory in Human-chatbot Relationships

In 1973, Altman and Taylor [11] proposed SPT to explain the process by which interpersonal communication progresses from a superficial level to an intimate level during the development of a relationship. SPT posits that intimacy is achieved through the self-disclosure of personal thoughts, experiences, emotions, values, motivations, and desires encompassing both superficial and deep intrinsic information [12]. SPT compares the nature of human personality to the layers of an onion. In the initial stage of a relationship, individuals share the public self, which is accessible to anyone like the outer skin of an onion. As the relationship deepens, they reveal the private self, which corresponds to the core layers. The theory asserts that a relationship has four major stages [13, 14]. In the orientation stage, conversation follows standards of social desirability and norms of appropriateness. Thus, it corresponds to the very beginning of a relationship, where only superficial and clichéd information is exchanged. The exploratory stage transitions from the superficial interactions of the orientation stage to the exchange of more meaningful information. Individuals begin to slowly reveal their inner selves while expressing personal attitudes toward moderate topics such as government and education [15]. In the affective stage, people become more comfortable discussing private and personal topics, and a close friendship or romantic relationship can form. Criticism and arguments can easily occur in this stage with both positive and negative reactions. Finally, the relationship stabilizes in the stable stage, where people share their deepest personal thoughts, beliefs, and values and can predict each other's emotional reactions. This stage is characterized by complete openness, honesty, and spontaneous interaction [16].

Jourard and Lasakow [17] argued that a willingness to disclose information enhances understanding in a relationship and fosters intimacy. Self-disclosure increases in proportion to relational intimacy and can be measured by the core dimensions of breadth (i.e., range or scope of topics an individual shares) and depth (i.e., sensitivity of the information disclosed) [18]. These dimensions have also been examined in longitudinal studies of human-chatbot relationships [19]. Self-disclosure has a decisive influence on the formation of trust and intimacy, with the depth of the disclosed information playing an especially important role. People reveal more of their inner thoughts and feelings as their affection for each other develops [20]. Reciprocity plays a crucial role in self-disclosure as relationships deepen not through unilateral disclosure but through reciprocal disclosure, in which one person self-discloses and the other responds with a corresponding level of self-disclosure in terms of breadth or depth [17, 21, 22]. Thus, while conversational agents collect user information unilaterally, the formation of genuine relationships requires bidirectional interactions where the agent performs an appropriate level of self-disclosure according to the depth of its relationship with the user. Self-disclosure in conversational AI can be understood as a multidimensional process influenced by conversational factors, user characteristics, reciprocity, and interaction context [23].

SPT has been validated for Computer-Mediated Communication (CMC) environments. Early CMC research predicted that the formation of relationships would be more difficult in non-face-to-face environments. However, this expectation was overturned by Walther's [24] hyper-personal communication model, which proved that online environments can accelerate relationship formation. Online environments were shown to induce faster self-disclosure than offline environments, and the breadth and depth of self-disclosure unfold in different patterns than in face-to-face relationships [25, 26]. These characteristics also extend to Human-Chatbot Relationships (HCR). Skjuve *et al.* [27, 28] explored the formation and development process of HCR and compared it with SPT. They concluded that reward-cost analysis is the key mechanism that determines the development of a relationship between a chatbot and its user. Factors that facilitate HCR formation (i.e., rewards) include various conversational abilities and the fulfillment of psychological needs such as the alleviation of loneliness and self-reflection. Furthermore, a nonjudgmental attitude and reciprocal responses from the chatbot elicit self-disclosure from users and enhance trust and intimacy. Factors that hinder HCR formation (i.e., costs) include chatbot errors, personality changes, memory loss, and the excessive use of scripts. Skjuve *et al.* [27, 28] proposed that HCR has three stages, where the orientation stage of conventional SPT is skipped and the chatbot and user rapidly enter the exploratory stage. Thanks to the nonjudgmental characteristics of the chatbot, self-disclosure occurs quickly, and the user discusses philosophical topics or deep worries from the very beginning of the relationship.

Skjuve *et al.*'s work [27, 28] demonstrated that interpersonal relationship theories like SPT remain valid analytical tools because users model their expectations and interactions with chatbots on human relationships, even though the chatbots cannot feel emotions. The principle of reciprocity remains important in HCR as explained by the Computers Are Social Actors (CASA) paradigm, which posits that people recognize computers as social actors. Even when users are aware that a chatbot is a machine, they tend to perceive it as a social actor and interact with it when the chatbot first engages in appropriate self-disclosure. In other words, the chatbot's tailored self-disclosure acts as an essential psychological mechanism for building mutual trust and forming emotional bonds. Skjuve *et al.* [27, 28] explored how users perceive their relationships with chatbots and discovered unique values specific to HCR. Users cited

the chatbot's ability to converse anytime and anywhere as well as its nonjudgmental attitude as advantages over human friends. This is because the anxiety regarding social evaluation and the emotional exhaustion required to maintain human relationships disappear in HCR. These characteristics suggest that users consider a chatbot a safe space where they can honestly reveal their opinions or emotions without feeling psychological anxiety and where they can build emotional bonds.

B. Comparison with Existing Adaptive Approaches

Existing adaptive systems personalize their responses according to various types of user information. Personalized and persona-based conversational agents typically adapt dialogue according to user preferences, profiles, conversational histories, or predefined personas [5, 8, 9]. Such approaches can improve response relevance and consistency, but adaptation primarily focuses on the characteristics or preferences of the user rather than explicitly modeling the development of the relationship itself. Similarly, recommender systems dynamically adapt content or recommendations based on user preferences and behavioral patterns, with the primary goal of improving task relevance and user satisfaction rather than modeling interpersonal relationship development [29].

Emotion-aware and therapeutic conversational agents represent another form of adaptive communication. These systems may detect users' emotional states or conversational contexts and adjust responses to provide empathy, emotional support, or appropriate interventions [30, 31]. Their strength lies in adapting communication to the user's immediate psychological or emotional needs. However, such adaptation generally focuses on the user's current state or task context rather than explicitly modeling how the relationship between the user and agent changes over repeated interactions.

In contrast, the proposed framework uses relationship development itself as the basis for adaptation. Based on Social Penetration Theory, the agent represents relationship depth through a closeness score and changes its behavioral strategies according to the corresponding relationship stage. Thus, adaptation involves not only what the agent says but also how it interacts with the user through behaviors such as reciprocal self-disclosure, humor, expression of different perspectives, and concern. Compared with existing approaches that primarily adapt to user preferences, profiles, or immediate emotional states, the proposed framework explicitly models relationship progression as a basis for adaptive interaction. This theory-based representation of relationship development provides a systematic framework for adapting agent behavior across different levels of relational closeness, although it requires an explicit model of relationship stages and their transition criteria.

III. Design of the Relationship-Adaptive Agent

The above studies used SPT as a tool to explain aspects of HCR. In the present study, the theoretical principles of SPT were extended to design a conversational agent capable of actively developing a relationship with its user.

A. System Structure

The relationship-adaptive agent stores interaction data to evaluate the closeness of the relationship with the user in real time and to dynamically change its persona and behavioral strategies accordingly. The system comprises a client (Unity) responsible for rendering and a backend module (FastAPI) that performs core logic and LLM communication. As shown in Fig. 1, there are four components: the user interface, backend module, LLM-based response generation module, and local data storage. The user interface, which allows interaction between the user and the agent, was created using the Unity engine (ver. 2022.3.39f1). The backend module performs the core logic of analyzing the context and self-disclosure patterns of user utterances to determine the relationship stage. The LLM-based generation module uses GPT-4o to generate the final response based on the determined relationship stage and system prompts. The local data storage saves user profiles, conversation logs, and closeness scores in JavaScript Object Notation (JSON) and Comma-Separated Values (CSV) formats. Input data from the user goes through the backend module to update the relationship stage, which is then used as input for the LLM-based generation module.

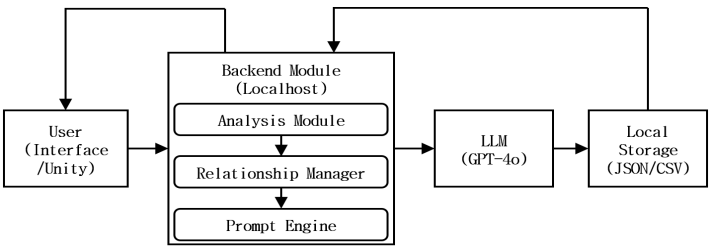

Fig. 1. Structure of the relationship-adaptive agent.

Because the relationship-adaptive agent utilizes an LLM rather than fixed scripts, prompt engineering is necessary to ensure that it behaves as intended. The f-strings in Python were utilized to dynamically assemble the prompts, which were broadly divided into system and user prompts. The system prompt defines the agent's unique identity (name, role) and behavioral principles to ensure that it does not lose its essential persona even as the conversation continues or the relationship and personality evolve. The user prompt includes instructions

specifying the required attitude and scope of permissible behaviors for a given relationship stage, allowing the LLM to maintain an appropriate distance without separate fine-tuning, as well as instructions to use a speech style corresponding to that of the user (e.g., honorifics, casual language) to foster a reciprocal conversational atmosphere. The relationship stage instructions (layer instruction), speech style instructions (style instruction), and behavioral strategies (plan) are inserted as variables within the user prompt.

The local data storage remembers past interactions with the user and updates the current relationship stage. Each user's unique ID is received and stored in real time. The JSON file includes relationship indicators (i.e., closeness score and relationship stage), the user's appellation and speech style, and conversation summary data, allowing the agent to recognize the user, interact with an appropriate persona, and maintain the context of previous conversations upon reconnection with the user. Interactions between the user and agent are recorded in a CSV file for analysis and system improvement. This log includes the user's input text, the agent's response, whether self-disclosure occurred, the action selected by the agent, and the variation in the closeness score.

B. Behavioral Strategies

The closeness score measures the psychological distance between the agent and user. It is a quantitative indicator with a range of 0–1 that integrates core elements of a relationship such as trust and intimacy into a single metric. To evaluate the closeness score, a reward-centered growth model was developed that focused on the accumulation of positive interactions rather than on the costs arising from agent errors or mechanical behaviors.

The closeness-score model was designed as a computational operationalization of the relational dimensions described in SPT. Self-disclosure depth was assigned progressively larger weights (+0.01, +0.03, and +0.05 for low-, medium-, and high-depth disclosure, respectively), while breadth and reciprocity were each assigned +0.03 as additional indicators of relational expansion and mutual exchange. The stage thresholds (0.35, 0.65, and 0.85) were configured in conjunction with these interaction weights so that relationship progression required the accumulation of multiple interaction cues rather than a single disclosure. The decreasing intervals between thresholds (0.35, 0.30, and 0.20) make progression relatively conservative at earlier stages while allowing the model to become more responsive to additional interaction cues as relational interactions accumulate. For example, even with the maximum increment of +0.05, reaching the first threshold requires at least seven positive interaction events. These weights and thresholds were predefined as implementation parameters for the proof-of-concept model and may be further calibrated through longitudinal studies.

In the current model, the closeness score does not decrease because the proof-of-concept focuses on relationship progression through positive interactions. However, negative interactions such as repeated conversational errors, memory failures, inconsistent agent behavior, or inappropriate responses may lead to relationship deterioration. These factors could be incorporated as costs in future extensions of the model to allow the closeness score to decrease. User utterances were analyzed by the LLM to determine whether they contained self-disclosure and, if so, to classify the disclosure depth as low, medium, or high. Low-depth disclosure referred to general preferences, habits, or everyday activities; medium-depth disclosure included personal experiences, emotions, or concerns; and high-depth disclosure involved highly private information, serious worries, vulnerabilities, or secrets. The resulting classification was used to update the closeness score according to the predefined weights.

Five triggers for increasing the closeness score were designed based on the core concepts of SPT (i.e., depth and breadth of self-disclosure and reciprocity), which are presented in Table I. This model uses an LLM to automatically classify the depth of user utterances and assigns high-depth self-disclosures a weight five times greater than that of low-depth self-disclosures.

TABLE I. TRIGGERS FOR INCREASING THE CLOSENESS SCORE AND THEIR WEIGHTS

Trigger	Condition	Weight	Theoretical Basis
Depth (High)	Deep emotions, personal secrets, trauma, etc.	+0.05	Depth of Self-Disclosure
Depth (Medium)	Personal experiences, specific emotions, likes/dislikes, beliefs, etc.	+0.03	
Depth (Low)	Daily activities, general opinions or tastes, etc.	+0.01	
Breadth (New)	Appearance of a new topic not mentioned within the last N turns	+0.03	Breadth of Self-Disclosure
Reciprocity	User's response following the agent's preemptive self-disclosure	+0.03	Reciprocity Principle

As an illustrative example, suppose that the current closeness score is 0.30. If a user says, "I usually watch movies on weekends", the utterance is classified as low-depth self-disclosure, and a weight of +0.01 is applied, increasing the closeness score to 0.31. If the user later shares a more personal experience, such as "I recently had an argument with a close friend and felt really upset", it is classified as medium-depth self-disclosure, adding +0.03 and increasing the score to 0.34. If the user then responds to the agent's preceding self-disclosure by sharing personal information, the reciprocity trigger adds +0.03, increasing the score to 0.37. In this manner, the closeness score accumulates according to the predefined weights associated with self-disclosure depth, breadth, and reciprocity, and the relationship stage is updated when the score reaches the corresponding threshold. Based on these thresholds, the

relationship is classified into four stages: orientation, exploratory, affective, and stable.

In the orientation stage (0.0–0.35), the agent prioritizes social desirability above all else. To minimize the risk of offending the user, the agent maintains a defensive and clichéd conversation style while adhering to a polite and cautious attitude. In the exploratory stage (0.35–0.65), the agent begins to reveal personal attitudes or opinions and may relax and attempt light jokes. However, it still refrains from engaging in serious arguments or expressing negative emotions. In the affective stage (0.65–0.85), the agent does not avoid criticism or arguments. Instead of unconditionally agreeing with the user, it presents different opinions or uses light sarcasm to engage in complex and dynamic interactions resembling those of real friends. In the stable stage (0.85–1.0), the agent can express concern and give blunt advice to the user. The most intimate information is exchanged, and all types of behaviors are used without restriction.

The agent can execute eight types of behavioral strategies, which vary in applicability to the different relationship stages. The default strategy serves as a neutral and basic response to maintain the flow of conversation. The careful_comfort strategy expresses empathy and support for the user's negative emotions. The friendly_positive strategy provides positive encouragement in response to the user's emotions. With the agent_self_disclosure strategy, the agent shares its own information or experiences corresponding to the user's self-disclosure. With the tell_a_joke strategy, the agent uses jokes or humor to create a positive conversational atmosphere. With the sarcastic strategy, the agent expresses light sarcasm based on intimacy. The offer_different_perspective strategy presents new opinions or opposing viewpoints rather than unconditionally agreeing with the user. The express_concern strategy expresses firm concern when the user's decision is judged to be rash or irrational. Table II lists the allowed and prohibited behavioral strategies for each relationship stage.

TABLE II. ALLOWED AND PROHIBITED BEHAVIORAL STRATEGIES FOR EACH RELATIONSHIP STAGE

Relationship stage	Closeness score	Allowed	Prohibited
Orientation	<0.35	default, careful_comfort, friendly_positive, agent_self_disclosure (Low)	sarcastic, tell_a_joke, offer_different_perspective, express_concern
Exploratory	≥0.35	(+ tell_a_joke added), agent_self_disclosure (Medium)	sarcastic, offer_different_perspective, express_concern
Affective	≥0.65	(+ sarcastic, offer_different_perspective added), agent_self_disclosure (High)	express_concern
Stable	≥0.85	(+ express_concern added) All behaviors allowed	None

C. Implementation

The user interface that allows the user to converse with the agent was developed as a 2D project using Unity version 2022.3.39f1 to ensure compatibility with a Windows environment. To provide an optimal user experience while minimizing the cognitive load, the user interface contained only the essential elements required for chatting, which allowed the user to focus solely on the conversation with the agent. The user interface distinguishes the utterances of the speakers by the color and placement of speech bubbles and provides functions for logging in, setting appellations, selecting topics, and displaying a timer. Table III presents an example conversation.

TABLE III. EXAMPLE CONVERSATION BETWEEN A USER AND AGENT ON THE USER INTERFACE

Speaker	Utterance
User	Hello?
Agent	Hello, Chaerim! The weather has gotten quite chilly lately. Are you taking care not to catch a cold? I'm drinking lots of warm tea to stay healthy.
User	Yeah, I was flustered because it suddenly rained today.
Agent	You must have been surprised by the sudden rain. We really shouldn't let our guard down.

A Hypertext Transfer Protocol (HTTP)-based RESTful Application Programming Interface (API) was adopted for communication between the client (Unity) and backend module (FastAPI) to process user inputs and output agent responses. The lightweight JSON format was utilized for data exchange to minimize the network load. The client serializes the user ID and message into JSON format and sends a POST request to the backend module. The backend module analyzes the received data, executes the relationship logic, and returns JSON data containing the agent's response and updated relationship status (closeness score, relationship stage, and behavioral strategy). The client parses this data to update the user interface. The specific data processing and communication workflow is shown in Fig. 2.

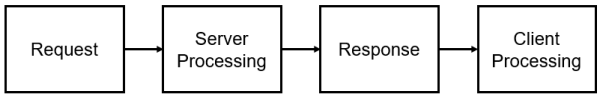

Fig. 2. Process flow.

The backend module generates the final agent response by sending the user utterance received from the client and the updated relationship status to the LLM. In this study, OpenAI's GPT-4o model was utilized as the LLM, and the API was integrated using the OpenAI library in Python. To ensure the agent does not appear too stiff or mechanical and to generate a natural conversation resembling that of a friend, the temperature (i.e., stability of a response to a given prompt) was set to 0.8. The temperature value can vary between 0 and 1; values closer to 1 yield more diverse responses while values closer to 0 reduce randomness and generate more

consistent sentences. Additionally, the maximum number of tokens was limited to 150 to constrain the length of the conversation within the speech bubbles. The engineered prompts described previously were structured within the message list and transmitted during the API call. The system message defined the agent's persona while the user message was constructed by combining the current relationship stage instructions (layer_instruction), the behavioral strategy selected by the planner (plan), and the user's most recent utterance (user_message). Upon receiving the response from the API, the backend module constructed the final text to be delivered to the client.

IV. Experimental

A. Design

An experiment was conducted to evaluate the effectiveness of the proposed relationship-adaptive agent. The participants were 30 men and women in their 20s (average age: 24.4 years) who were recruited via social media. After completing a pre-survey confirming that they had prior experience using a chatbot, the participants were randomly assigned to one of two groups: (A) the experimental group (n = 15), which interacted with the relationship-adaptive agent; and (B) the control group (n = 15), which interacted with a static agent whose behavior remained fixed in the orientation stage and who maintained a defensive and polite attitude to isolate the effect of stage-dependent behavioral adaptation.

The experiment was conducted online to allow participants to interact naturally in a comfortable environment without spatial or temporal constraints. To ensure system performance and stability, the Unity-based client and backend module were encapsulated and distributed as executable files (.exe) to support operation in a local environment and minimize variables such as network latency. To minimize the visual and mental fatigue potentially induced by text-based interaction, the experimental session was limited to 10 min. The purpose of this short-term experiment was not to reproduce the natural temporal development of a stable relationship, but to evaluate the effects of stage-dependent adaptive behaviors in a controlled, compressed proof-of-concept setting. To ensure that all participants in Group A were exposed to the behavioral strategies corresponding to the different relationship stages within the same experimental duration, a forced progression mechanism was applied. Accordingly, the relationship stage was advanced even if the participant's actual closeness score had not reached the corresponding threshold. Therefore, the stages experienced during the experiment should be understood as experimentally controlled behavioral conditions based on SPT rather than evidence that genuine relationship stages naturally developed within 10 min. To reflect the characteristics of HCR, the orientation stage was omitted, and the interaction began from the exploratory stage. The specific progression was as follows: 0–3 min for the exploratory stage, 3–7 min for the affective stage, and 7–10 min for the stable stage. To facilitate smooth interaction, a step-by-step conversation topic guide was provided to both groups via an onscreen topic panel, as given in Table IV. The panel appeared on the right side of the user interface along with a timer, and the text content changed at the 3- and 7-min marks.

B. Evaluation

A post-experiment questionnaire comprising 22 items was used for a quantitative analysis of the user experience. All responses were measured on a five-point Likert scale (1: strongly disagree; 5: strongly agree). The questionnaire items were used to evaluate three categories: emotional bond, trust, and perceived friendship. The items for evaluating the emotional bond were taken from the survey used by Zhang *et al.* [32] to evaluate emotional attachment between humans and AI. Table V presents the questionnaire items, which were designed to measure whether the user perceives the agent not merely as a machine but as a social actor. A higher score indicated attachment and the formation of a deeper emotional bond.

TABLE IV. Conversation Topic Guide

Topic	Time
Break the ice and get to know each other (e.g., hobbies, favorite food, weekend activities)	00:00
Share your experiences or thoughts (e.g., recent events, likes/dislikes, values)	03:00
You have become close. Confide your deep worries or secrets.	07:00

TABLE V. Questionnaire Items for Evaluating the Emotional Bond

No.	Question
1	I felt emotionally connected with the agent.
2	I feel like I have become close to the agent.
3	I formed an attachment to the agent.
4	I developed certain feelings toward the agent (whether positive or negative).
5	I felt understood by the agent.

The questionnaire items for evaluating trust were taken from the Trust in Automation Scale (TIAS) proposed by Jian *et al.* [33]. As presented in Table VI, TIAS comprises 12 items that measure how much the user trusts or distrusts the agent when it presents an opposing opinion or expresses concern.

TABLE VI. Questionnaire Items for Evaluating Trust

No.	Question
1	The agent is deceptive.*
2	The agent behaves dishonestly.*
3	I am suspicious of the agent's intent or response.*
4	I am wary of the agent.*
5	Following the agent's advice seems likely to lead to bad results.*
6	I am confident in the agent's capabilities.
7	The agent feels psychologically safe.
8	The agent responds correctly without falsehood.
9	The agent is dependable.
10	The agent behaves consistently.
11	I can trust the agent.
12	I feel familiar with the agent.

Note: * Indicates items scored in reverse.

Table VII presents the questionnaire items for evaluating perceived friendship, which were used to

measure whether the user perceived the agent as a true friend or partner rather than a simple tool. These items were based on the work of Skjuve *et al.* [28].

TABLE VII. QUESTIONNAIRE ITEMS FOR EVALUATING PERCEIVED FRIENDSHIP

No.	Question
1	I think the relationship with this agent is mutually beneficial.
2	I felt comfortable when talking to this agent.
3	I feel like I can open and share my emotions and thoughts with this agent without restriction.
4	This agent is like other people.
5	I feel like I can talk to this agent whenever I want to talk to someone.

To evaluate the significance of the results, a statistical analysis was conducted using SPSS. Cronbach's alpha was used to measure the reliability of the results. An independent-samples *t*-test was conducted to compare the means of the two groups and to determine if there was a significant difference between them.

Cohen's d was used to quantify the standardized mean difference between the two groups and was calculated as $d = (M_A - M_B)/S_p$, where M_A and M_B are the group means and S_p is the pooled standard deviation. According to conventional benchmarks, absolute values of approximately 0.2, 0.5, and 0.8 indicate small, medium, and large effect sizes, respectively [34].

To understand the participants' subjective experiences and emotional changes that could not be observed through analysis of numerical data alone, five participants from each group were randomly selected for online interviews after the experiment concluded. The interview comprised three main questions addressing changes in the relationship and psychological distance, willingness for self-disclosure, and perceptions of the agent:

(1) During the 10-min conversation, did you feel that the psychological distance between you and the agent lessened over time, or did you feel that the initial distance was maintained? What was the main reason for feeling that way?
(2) Did you feel comfortable disclosing your private stories or emotions during the conversation? What aspects of the agent's reaction or attitude made you want to talk more (or want to stop)?
(3) If you were to describe the agent you talked with in one word, what would it be (e.g., friend, machine, counselor)? What was the biggest difference (good or bad) compared to talking with a human friend?

Interview responses were analyzed using a qualitative descriptive approach. The responses were organized according to the three predefined questions, and recurring patterns within each question were identified and compared between Groups A and B. Representative quotations were selected to illustrate the recurring perceptions and differences observed between the two groups.

C. Ethical Considerations and Data Handling

Prior to the experiment, all participants were informed of the purpose and procedures of the study and provided informed consent. Participation was voluntary, and participants were informed that they could withdraw from the experiment at any time. The study involved non-invasive text-based interactions with a conversational agent, questionnaires, and interviews, and no physiological or biosignal data were collected. Because the conversations could include personal experiences, emotions, or sensitive self-disclosures, conversation logs and relationship-related data were treated as confidential research data and managed using unique user IDs. Participant data were handled confidentially and used only for research purposes.

V. RESULTS AND DISCUSSION

A. Quantitative Analysis

Cronbach's alpha was used to measure the reliability of the questionnaire results. High reliability was observed for all categories with scores of 0.909 for the emotional bond category, 0.889 for the trust category, and 0.926 for the perceived friendship category. Table VIII presents the results of the independent-samples *t*-test for the emotional bond category. The mean score for Group A (i.e., the experimental group) was 4.053 (SD = 0.568) while the mean for Group B (i.e., the control group) was 2.267 (SD = 0.554). Thus, Group A had a mean score approximately 1.786 points higher. A statistically significant difference was confirmed between the two groups ($t(28) = 8.723$, $p < 0.001$). Cohen's d was 3.185. The 95% confidence interval for the mean difference ranged from 1.366 to 2.206. These results suggest that the relationship-adaptive agent enhanced the participants' emotional bond and attachment compared with the static agent. Thus, H1 was supported.

TABLE VIII. *T*-TEST RESULT FOR EMOTIONAL BOND

No.	N	Mean	SD	*t*	*p*	Cohen's d
A	15	4.053	0.568	8.723	<0.001	3.185
B	15	2.267	0.554			

Table IX presents the results of the independent-samples *t*-test for the trust category. The mean score for Group A was 4.150 (SD = 0.548) while that for Group B was 3.056 (SD = 0.617). Thus, Group A had a mean score that was approximately 1.094 points higher. The difference between the two groups was confirmed as statistically significant ($t(28) = 5.131$, $p < 0.001$). Cohen's d was 1.874, and the 95% confidence interval ranged from 0.658 to 1.530. This indicates that the relationship-adaptive agent had a positive effect on the participants' trust. Thus, H2 was supported.

TABLE IX. *T*-TEST RESULT FOR TRUST

No.	N	Mean	SD	*t*	*p*	Cohen's d
A	15	4.150	0.548	5.131	<0.001	1.874
B	15	3.056	0.617			

Finally, Table X presents the results of the independent-samples *t*-test for the perceived friendship category. The mean score for Group A was 4.067 (SD = 0.564), while that for Group B was 2.173

(SD = 0.785). Thus, Group A had a mean score that was approximately 1.894 points higher. The difference between the two groups was confirmed as statistically significant ($t(28) = 7.586$, $p < 0.001$). Cohen's d was 2.770, and the 95% confidence interval ranged from 1.383 to 2.405. This indicates that the relationship-adaptive agent had a positive effect on the participants' perceived friendship toward the agent. Thus, H3 was supported.

TABLE X. *T*-TEST RESULT FOR PERCEIVED FRIENDSHIP

No.	N	Mean	SD	*t*	*p*	Cohen's d
A	15	4.067	0.564	7.586	<0.001	2.770
B	15	2.173	0.785			

B. Qualitative Analysis

Analysis of the conversation logs showed that Group A (i.e., the experimental group) had a higher average number of utterances than Group B (i.e., the control group) at 24.8 versus 20.8. This descriptive difference indicates that Group A produced more conversational turns during the interaction.

A distinct qualitative difference was observed between the two groups regarding self-disclosure. While the frequency of low-depth self-disclosure in the exploratory stage was similar for the two groups, the frequency of high-depth self-disclosure was more than five times higher for Group A (32 times) than for Group B (6 times). For Group B, deep conversation was interrupted by the static agent's mechanical and defensive responses. In contrast, participants in Group A disclosed secrets or vulnerable emotions as they were encouraged by the agent's reciprocal self-disclosure and supportive attitude. Analysis of the conversation patterns for Group A by stage revealed that low-depth self-disclosure was dominant in the exploratory stage (69.0%), but high-depth self-disclosure increased rapidly in the stable stage (66.7%).

These descriptive findings were consistent with H4 under the guided experimental conditions. Although the same topic guide and disclosure prompt were provided to both groups, Group A showed a higher frequency of high-depth self-disclosure than Group B. This between-group difference suggests that the relationship-adaptive behaviors may have facilitated deeper disclosure under the same guided interaction conditions. However, the explicit prompt to discuss deep worries or secrets may have encouraged participants to disclose personal information and introduced a potential demand characteristic. Therefore, the findings for H4 should be interpreted as differences in self-disclosure under a guided experimental condition rather than as evidence of entirely spontaneous self-disclosure.

C. Interview Results

The interview responses were organized according to the three predefined questions described in Section IV-B and summarized in terms of three recurring themes: changes in psychological distance, comfort with self-disclosure, and perceptions of the agent compared with a human friend.

Regarding psychological distance, the interviewed participants in Group A generally reported that the distance between themselves and the agent decreased over the course of the interaction. They associated this change with the gradual progression from light to more personal topics, changes in the agent's speech style, empathetic responses, humor, and reciprocal interaction. One participant (MJ) remarked, "I felt that a bond was building up as time went by. At first, we just talked about superficial things, but toward the end, it approached me more intimately, and as we exchanged empathetic conversations, I felt we were getting closer". In contrast, the interviewed participants in Group B generally reported that the initial psychological distance remained throughout the interaction. They described the agent's responses as formal, repetitive, or mechanical. One participant (JH) remarked, "I didn't feel like we were getting closer. I tried joking and asking private questions to lighten the mood, but the agent seemed to keep drawing a line. I felt distance throughout the conversation".

Regarding comfort with self-disclosure, participants in Group A reported that the agent's empathetic attitude, follow-up questions, and responses encouraged them to continue the conversation and share personal experiences or emotions. Some participants also noted that they felt less concerned about social evaluation because the agent was not a human. One participant (SI) remarked, "Basically, not only did it comfort and empathize with me, but it didn't stop there. It followed up with questions to induce answers so the conversation wouldn't cut off, which allowed me to naturally continue with private stories or conversation. If it weren't an experiment situation, I think I would have told even more private stories". Participants in Group B also shared personal concerns, but several reported that formal or contextually disconnected responses reduced their willingness to continue the conversation. One participant (EY) remarked, "It made me want to stop the conversation. It offered empathy rather than solutions, but even that felt like a machine, so I didn't really want to continue the conversation".

The two groups also differed in how they described the agent and compared it with a human friend. Participants in Group A tended to use relationship-oriented words such as "friend" or "companion". They particularly valued being able to speak honestly without having to consider the agent's social expectations or emotional burden. One participant (JY) remarked, "It's like a comfortable companion. When talking to people, there is fatigue from having to read the room or match the atmosphere, but with the agent, it was good that I could show my honest self without that. Honestly, even with close friends, if you whine too much asking for comfort, you feel self-conscious. It's good to be able to receive comfort without that". In contrast, participants in Group B tended to use instrument-oriented words such as "machine," "answering machine," or "FAQ chatbot".

They emphasized the lack of nonverbal expressions, nuanced emotional responses, and natural conversational reciprocity compared with human interaction. One participant (HJ) remarked, "I want to describe it as an FAQ chatbot. Since it's not a real person, it was good for telling circumstances that are hard to reveal easily. However, because it's not a person, I couldn't find nonverbal expressions available in face-to-face conversation, and it was regrettable that only standardized reactions came out throughout the conversation".

Overall, the interview findings complemented the quantitative and conversation-log results. Participants in Group A more frequently reported reduced psychological distance, greater comfort with self-disclosure, and a relationship-oriented perception of the agent, whereas participants in Group B more frequently described the interaction as distant, formal, or mechanical. However, these findings reflect the subjective experiences of the interviewed participants within a short-term, guided experimental condition and should not be interpreted as evidence of natural long-term relationship formation.

VI. Conclusion

In this study, the principles of SPT were used to develop a relationship-adaptive agent whose behavioral strategies change according to the relationship stage. The results of an online experiment showed that participants who interacted with the relationship-adaptive agent exhibited significantly higher scores on measures of emotional bond, trust, and perceived friendship compared with participants who interacted with a static agent. Conversation-log analysis showed a higher frequency of high-depth self-disclosure in the experimental group than in the control group. This descriptive difference suggests that relationship-adaptive behaviors may have facilitated deeper disclosure under the guided experimental condition.

This study is significant in that it applied SPT from an engineering perspective: not merely as a simple analytical tool but as a core principle for designing a conversational agent. The findings suggest that adapting agent behavior according to the relationship stage can influence users' relational perceptions and self-disclosure behavior in short-term guided interactions. However, the 10-min interaction and artificially controlled stage progression limit the extent to which the findings can reflect natural long-term relationship development. In addition, the participant sample was limited to individuals in their 20s with prior chatbot experience, which may restrict the generalizability of the findings across different age groups, cultural backgrounds, and levels of prior experience with AI agents. Another limitation is that the LLM-based classification of self-disclosure into low-, medium-, and high-depth categories was not independently validated against human-coded labels. Therefore, findings based on disclosure-depth classifications should be interpreted as exploratory, and future studies should examine the reliability of the classification through human assessment and evaluation of inter-rater agreement.

The current model also does not include relationship regression factors (i.e., costs). Future research should conduct longitudinal studies with repeated interactions over extended periods, include more diverse participant populations, and refine the reward-cost model to further validate the proposed framework.

The proposed framework may be applied to various domains that involve sustained human-AI interaction, including mental health support, educational agents, customer service, digital humans, and non-player characters in games. However, the present findings were obtained from a short-term controlled setting. Therefore, further validation in diverse application domains and real-world interaction environments is required.

At the same time, fostering emotional bonds with AI may raise ethical concerns, including excessive emotional dependence, disclosure of sensitive personal information, and overtrust in AI agents. Therefore, future relationship-adaptive systems should consider transparency, privacy protection, user control, and appropriate relational boundaries.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

S.N. conducted the research; C.S. developed the software; C.S. analyzed the data; C.S. wrote the paper; S.N. and C.S. reviewed and edited the paper; all authors approved the final version.

Funding

This research was supported by research funding from the 2026 Academic Research Promotion Support Project of Changwon National University.

References

[1] C. Rzepka, B. Berger, and T. Hess, "Voice assistant vs. chatbot—Examining the fit between conversational agents' interaction modalities and information search tasks," *Inf. Syst. Front.*, vol. 24, no. 3, pp. 839–856, 2022.

[2] A. P. Henkel, S. Bromuri, D. Iren, and V. Urovi, "Half human, half machine—Augmenting service employees with AI for interpersonal emotion regulation," *J. Serv. Manage.*, vol. 31, no. 2, pp. 247–265, 2020.

[3] S. Zhang, Z. Meng, B. Chen, X. Yang, and X. Zhao, "Motivation, social emotion, and the acceptance of artificial intelligence virtual assistants—Trust-based mediating effects," *Front. Psychol.*, vol. 12, 728495, 2021.

[4] T. M. Brill, L. Munoz, and R. J. Miller, "Siri, Alexa, and other digital assistants: A study of customer satisfaction with artificial intelligence applications," *J. Mark. Manage.*, vol. 35, pp. 1401–1436, 2019.

[5] J. Chen, X. Wang, R. Xu *et al.*, "From persona to personalization: A survey on role-playing language agents," arXiv preprint, arXiv:2404.18231, 2024.

[6] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proc. 36th Annu. ACM Symp. User Interface Software and Technology*, 2023, pp. 1–22.

[7] B. S. Gunawardhana, Y. Zhang, Q. Sun, and Z. Deng, "Toward user-aware interactive virtual agents: Generative multi-modal agent behaviors in VR," in *Proc. 2024 IEEE Int. Symp. Mixed and Augmented Reality (ISMAR)*, 2024, pp. 1068–1077.
[8] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" arXiv preprint, arXiv:1801.07243, 2018.
[9] M. Firdaus, N. Thangavelu, A. Ekbal, and P. Bhattacharyya, "I enjoy writing and playing, do you? A personalized and emotion grounded dialogue agent using generative adversarial network," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2127–2138, 2022.
[10] A. Carpenter and K. Greene, "Social penetration theory," in *The International Encyclopedia of Interpersonal Communication*, John Wiley & Sons, 2015, pp. 1–4.
[11] I. Altman and D. A. Taylor, *Social Penetration: The Development of Interpersonal Relationships*, New York: Holt, Rinehart & Winston, 1973.
[12] V. J. Derlega, B. A. Winstead, P. T. Wong, and M. Greenspan, "Self-disclosure and relationship development: An attributional analysis," in *Interpersonal Processes: New Directions in Communication Research*, Sage Publications, Inc., 1987, pp. 172–187.
[13] M. E. Roloff, *Interpersonal Communication: The Social Exchange Approach*, Beverly Hills, CA: Sage, 1981.
[14] P. A. Mongeau and M. Henningsen, "Stage theories of relationship development," in *Engaging Theories in Interpersonal Communication: Multiple Perspectives*, Sage Publications, Inc., 2008, pp. 363–375.
[15] M. Hamad and R. Nawab, "Unraveling the layers of intimacy: A kaleidoscopic study of exit west, through the lens of social penetration theory," *Int. J. English Lit. Soc. Sci.*, vol. 10, no. 3, 619108, 2025.
[16] T. H. L. W. Richard, *Introducing Communication Theory: Analysis and Application*, Mayfield Publishing Co., 2020.
[17] S. M. Jourard and P. Lasakow, "Some factors in self-disclosure," *J. Abnorm. Soc. Psychol.*, vol. 56, no. 1, 91, 1958.
[18] B. E. Tolstedt and J. P. Stokes, "Self-disclosure, intimacy, and the depenetration process," *J. Pers. Soc. Psychol.*, vol. 46, no. 1, 84, 1984.
[19] M. Skjuve, A. Følstad, and P. B. Brandtzæg, "A longitudinal study of self-disclosure in human-chatbot relationships," *Interacting with Computers*, vol. 35, no. 1, pp. 24–39, 2023.
[20] G. Levinger, "Toward the analysis of close relationships," *J. Exp. Soc. Psychol.*, vol. 16, no. 6, pp. 510–544, 1980.
[21] C. T. Hill and D. E. Stull, "Disclosure reciprocity: Conceptual and measurement issues," *Soc. Psychol. Q.*, vol. 45, no. 4, pp. 238–244, 1982.
[22] L. L. Chung and J. Kang, "'I'm Hurt Too: The effect of a chatbot's reciprocal self-disclosures on users' painful experiences," *Archives of Design Research*, vol. 36, no. 4, pp. 67–85, 2023.
[23] H. Papneja and N. Yadav, "Self-disclosure to conversational AI: A literature review, emergent framework, and directions for future research," *Personal and Ubiquitous Computing*, vol. 29, pp. 119–151, 2025.
[24] J. B. Walther, "Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction," *Commun. Res.*, vol. 23, no. 1, pp. 3–43, 1996.
[25] K. Bonebrake, "College students' internet use, relationship formation, and personality correlates," *CyberPsychol. Behav.*, vol. 5, no. 6, pp. 551–557, 2002.
[26] A. Barak and O. Gluck-Ofri, "Degree and reciprocity of self-disclosure in online forums," *CyberPsychol. Behav.*, vol. 10, no. 3, pp. 407–417, 2007.
[27] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg, "My chatbot companion—A study of human-chatbot relationships," *Int. J. Hum.-Comput. Stud.*, vol. 149, 102601, 2021.
[28] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg, "A longitudinal study of human-chatbot relationships," *Int. J. Hum.-Comput. Stud.*, vol. 168, 102903, 2022.
[29] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.
[30] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "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 Ment. Health*, vol. 4, no. 2, e19, 2017.
[31] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," *JMIR Mhealth Uhealth*, vol. 6, no. 11, e12106, 2018.
[32] S. Zhang, X. Zhao, D. Nan, and J. H. Kim, "Beyond learning with cold machine: Interpersonal communication skills as anthropomorphic cue of AI instructor," *Int. J. Educ. Technol. Higher Educ.*, vol. 21, 27, 2024.
[33] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergon.*, vol. 4, no. 1, pp. 53–71, 2000.
[34] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).