Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1560–1568
https://doi.org/10.54941/ahfe1008384

Design and Preliminary Evaluation of a Generative AI Role-Play System for Practicing Cognitive Empathy

Tsubasa Kishiume[1], Taiyo Kojima[2], Toshikazu Kato[1], Toru Nakata[2], and Takashi Sakamoto[2]

[1]Chuo University, Japan
[2]Graduate School of Science and Engineering, Chuo University, Japan

ABSTRACT

When another person's psychological state is not fully verbalized, understanding that person requires active questioning and interpretation. We designed a generative AI role-play system in which the AI portrays a junior colleague who does not readily disclose their feelings. The system combines a predefined persona, rules governing psychological information disclosure, and three dialogue tasks: exploration, interpreting, and paraphrasing. In a preliminary evaluation with 16 participants, illustrative dialogues from two conditions showed both task-aligned communication and instances in which advice or directive statements replaced the intended dialogue behaviors. The mean change in cognitive empathy (CE) was +0.438 in the cognitive-empathy condition and +0.094 in the factual–situational condition. However, baseline scores differed between conditions, and the baseline-adjusted analysis did not show a difference in the same direction. This study presents a concrete design for practicing perspective-taking through dialogue and identifies response-contingent support for users' dialogue behaviors as a direction for future system development.

Keywords: Generative AI, Cognitive empathy, Role-playing, Dialogue training, Human–AI interaction

INTRODUCTION

Understanding other people's feelings and thoughts is important for effective communication in workplaces and educational and research settings. In interpersonal interactions, however, people do not always explicitly describe their psychological states, making it difficult to understand them from their surface-level statements alone. It is therefore necessary to elicit information through questions, infer underlying feelings and thoughts, and compare one's interpretations with the other person's responses.

To support practice in this process of understanding others, we propose a generative AI-based role-play dialogue system. By combining an AI-simulated junior colleague who does not fully disclose their psychological state with three sequential tasks—exploration, interpreting, and paraphrasing—the system provides opportunities for users to actively practice perspective-taking.

This study aims to describe the system design supporting these interactions and to identify directions for improvement through illustrative dialogue cases and preliminary measurements of cognitive empathy.

Received November 7, 2026; Revised November 20, 2026; Accepted November 23, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

RELATED WORK AND POSITIONING OF THIS STUDY

Research has increasingly examined generative AI and conversational agents for communication training. Nouraei et al. (2025) investigated persuasive communication skills training using virtual agents. Sedefoglu et al. (2024) proposed LeadEmpathy, a framework of empathic communication behaviors in leadership settings that includes exploration, interpreting, and paraphrasing. Understanding a junior colleague's perspective and supporting their growth is also relevant to servant leadership (van Dierendonck and Nuijten, 2011). With supportive senior–junior communication as a potential application, the present study focuses on dialogue design for practicing perspective-taking rather than on evaluating leadership outcomes.

Building on this work, we use generative AI not primarily as an agent that expresses empathy toward the user, but as a dialogue partner through which users themselves practice understanding others. In particular, we design an environment for practicing exploration, interpreting, and paraphrasing when the other person's psychological state is not fully disclosed. For preliminary evaluation, we refer to the State Empathy Scale proposed by Shen (2010) and focus on its cognitive empathy (CE) dimension as the primary outcome.

DIALOGUE SYSTEM AND EXPERIMENTAL METHOD

Design of a Dialogue System for Understanding Others

We designed a generative AI role-play system to provide opportunities for users to ask questions, interpret responses, and paraphrase their understanding, rather than merely receive descriptions of another person's psychological state. The design principles are to avoid disclosing all psychological information at the outset, reveal relevant information in response to users' utterances, and provide structured tasks for practicing behaviors involved in understanding others. The system was implemented using Python, Streamlit, and the OpenAI API. Participants took the role of a senior laboratory member responding to a concern raised by a junior colleague, referred to as 'Sato,' portrayed by generative AI.

The scenario involved a junior colleague repeatedly revising presentation slides one week before an interim research presentation. Drawing on the Big Five personality traits, the prompt characterized the junior as high in conscientiousness and low in extraversion and emotional stability. Specifically, the junior was instructed to behave as a responsible, cautious person who worries about failure and does not readily verbalize detailed feelings or thoughts. The background further specified that, although the main slide structure was complete, the junior had difficulty deciding when to stop preparing and was anxious about questions and evaluation. This created a situation in which the underlying psychological state could not be fully understood from initial statements alone.

To prevent the simulated junior from revealing their psychological state all at once, the available information was organized into four categories: Surface, Background, Deep, and Insight. These categories are operational distinctions developed for this study to control the scope of information

disclosed in response to participants' questions and interpretations; they do not represent psychological developmental stages or a mandatory disclosure sequence.

The prompt combined the persona, factual information about presentation preparation, psychological background, and condition- and stage-specific response rules. Common rules instructed the AI to maintain a consistent character and scenario, avoid introducing unsupported information, and respond naturally as the junior rather than grade participants' answers. Each response was designed to answer the question directly while generally adding only one central piece of new information to avoid excessive disclosure. When a participant's inference differed from the preset background, the AI was instructed to express the difference as the junior's own experience rather than label the answer correct or incorrect. The junior was also instructed not to immediately settle on a course of action when given advice (Figure 1).

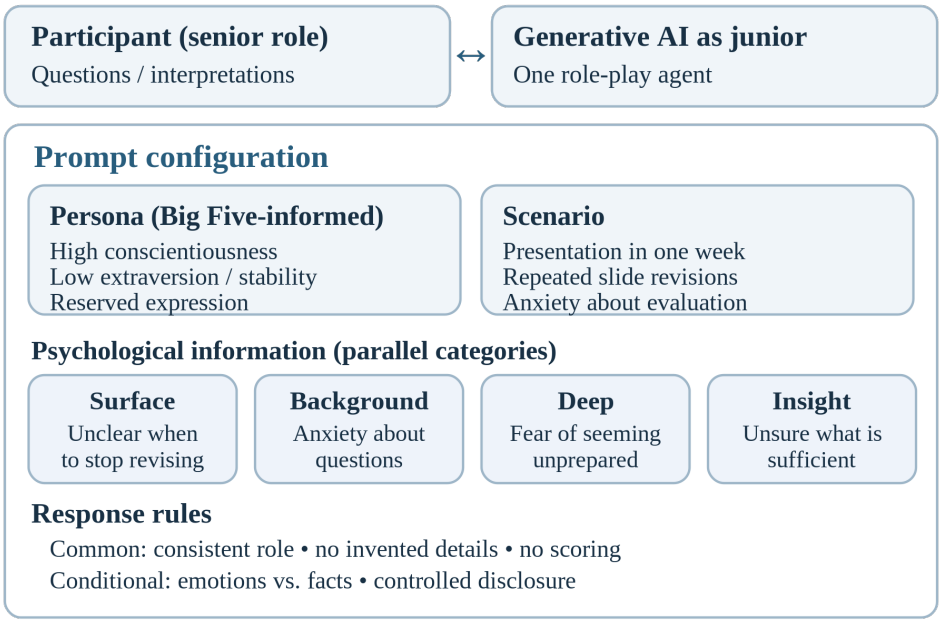


Figure 1: Architecture and prompt configuration of the proposed generative AI role-play system.

Dialogue Tasks and Comparison Conditions

To encourage the practice of understanding another person's psychological state, we defined three dialogue tasks—Exploration, Interpreting, and Paraphrasing—drawing on the empathic communication behaviors organized in LeadEmpathy (Sedefoglu et al., 2024).

During Exploration, participants asked about the junior's feelings and thoughts to elicit information not evident from the initial statement. During Interpreting, they inferred the psychological background from the information obtained and communicated their interpretation to the junior. During Paraphrasing, they summarized the feelings and thoughts they had

understood in their own words and compared this understanding with the junior's response. The sequence was intended to support step-by-step practice, progressing from information gathering to psychological inference and confirmation of understanding.

For the preliminary evaluation, we compared two conditions: a cognitive-empathy condition and a factual–situational condition. The latter focused on events and circumstances rather than psychological states and involved factual exploration, consideration of situational factors, and summarization of facts and circumstances (Table 1).

The scenario, persona, and number of dialogue turns were the same across conditions. Participants were asked to enter 20–80 Japanese characters at each stage. However, both the task instructions shown to participants and the junior AI's response rules differed between conditions; consequently, the comparison concerns the overall combinations of task instructions and AI response rules.

Table 1: Comparison of three-stage dialogue tasks.

Stage	Cognitive Empathy	Factual–Situational
1	Exploration: Feelings and thoughts	Factual Exploration: Facts and circumstances
2	Interpreting: Underlying feelings and thoughts	Situational Reasoning: Contextual factors
3	Paraphrasing: Restating feelings and thoughts	Factual Summarization: Summarizing facts

Experimental Design and Procedure

The study included 16 participants (fourth-year undergraduate students and first- or second-year master's students), with eight participants analyzed in each condition. Participant IDs were assigned in order of arrival, and participants were alternately allocated to the cognitive-empathy and factual–situational conditions according to whether their IDs were odd or even. Assignment was not randomized.

The experiment consisted of Phase 1, involving dialogue with generative AI, and Phase 2, involving responses to differing opinions (Figure 2). State empathy and humility were measured before the experiment. In Phase 1, participants completed the three-stage dialogue tasks corresponding to their assigned condition, and state empathy was measured again immediately afterward.

Following a 30-second break, Phase 2 presented three fixed-response situational judgment scenarios involving disagreement with the same junior colleague. State empathy and humility were measured again after Phase 2. Phase 2 did not involve generative AI response generation and served as an additional exploratory task. The main analyses reported here concern the dialogue cases and changes in cognitive empathy before and after Phase 1.

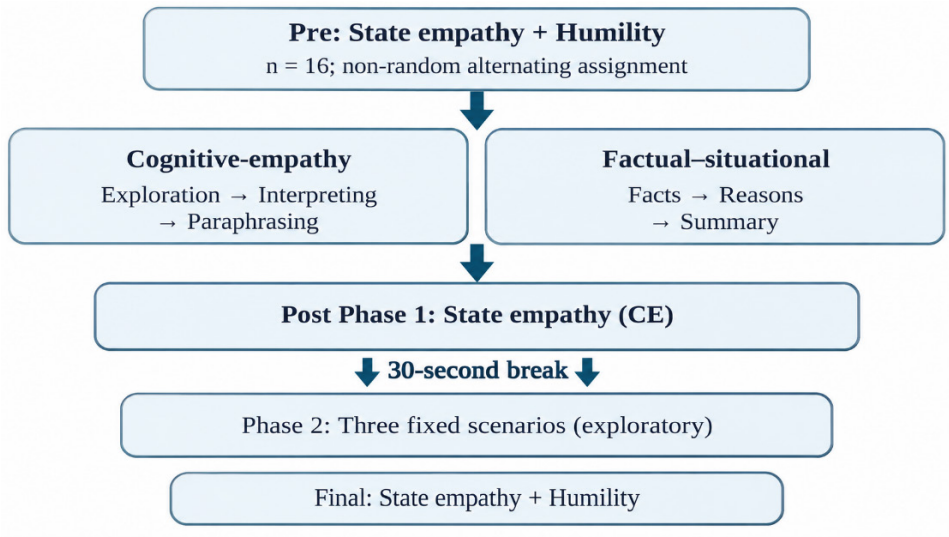


Figure 2: Experimental design and measurement timeline.

Measures and Analysis

We focused on cognitive empathy (CE), a dimension of state empathy related to inferring and understanding another person's feelings and thoughts, as the primary outcome.

For state empathy measurement, we referred to Shen's (2010) State Empathy Scale. Four items each for cognitive empathy (CE), affective empathy (AE), and associative empathy (AS) were translated into Japanese and adapted to refer to the simulated junior. Each item was rated from 0 to 4, and mean scores were calculated for each subscale.

For the primary descriptive evaluation, we calculated changes in CE scores (ΔCE) from before the experiment to immediately after Phase 1 and compared the average changes and their between-condition difference. We also examined illustrative examples from actual dialogue logs to consider how participants' utterances corresponded to the assigned task at each stage.

RESULTS

Characteristics of the Observed Dialogues

Figure 3 presents illustrative dialogue sequences observed in the two conditions. In the cognitive-empathy condition, participants asked about and inferred the junior's anxiety, while the junior AI elaborated on concerns about answering presentation questions and expressing explanations. In the factual–situational condition, dialogues primarily addressed the progress of slide preparation and remaining revisions. These cases illustrate different conversational focuses and do not demonstrate that all four categories of psychological information were disclosed in sequence.

Figure 4 shows cases in which participants in the cognitive-empathy condition offered advice about rehearsal and preparing extra slides during the interpreting and paraphrasing stages. In the factual–situational condition, participants also encouraged consultation with a supervisor and independent decision-making. Thus, in both conditions, advice or directive statements sometimes took precedence over the intended tasks. These illustrative cases were not used to quantify the frequency of dialogue behaviors or differences between conditions.

Cognitive-empathy	Factual–situational
Stage 1: Exploration **PARTICIPANT** What is worrying you most? **JUNIOR** I worry about freezing when questions come.	**Stage 1: Factual exploration** **PARTICIPANT** What remains unfinished? **JUNIOR** The outline is ready; I still need revisions and rehearsal.
Stage 2: Interpreting **PARTICIPANT** Explaining it in your own words may matter. **JUNIOR** I had not considered that. My reasoning matters.	**Stage 2: Situational reasoning** **PARTICIPANT** Identify the details that need revising. **JUNIOR** New issues arise each time. I cannot decide when to stop.
Stage 3: Paraphrasing **PARTICIPANT** You now value explaining in your own words. **JUNIOR** Yes. I want to explain things in my own words.	**Stage 3: Factual summarization** **PARTICIPANT** The outline is ready, but the stopping point is unclear. **JUNIOR** Yes. I cannot decide when revisions are complete.

Figure 3: Illustrative dialogue sequences under the cognitive-empathy and factual–situational conditions.



Figure 4: Illustrative dialogue sequences showing departures from the assigned tasks under both conditions.

Note. Dialogue excerpts are condensed English translations from one participant per condition; stage labels are assigned tasks.

Preliminary Cognitive Empathy Measurements

Table 2 presents the mean CE scores and changes before and after Phase 1. The mean change was +0.438 in the cognitive-empathy condition and +0.094 in the factual–situational condition, yielding a descriptive difference of +0.344. However, baseline CE scores were higher in the factual–situational condition, and the analysis adjusting for baseline scores did not produce a difference in the same direction. These descriptive results therefore do not establish an empathy-enhancing effect of the intervention.

Table 2: Changes in cognitive empathy scores following Phase 1.

Measure	Cognitive Empathy	Factual–Situational
Pre	3.063	3.531
Post	3.500	3.625
Change (ΔCE)	+0.438	+0.094
Between-condition difference (ΔΔCE)	**+0.344**	

Note. n = 8 per condition. CE scores range from 0 to 4. ΔCE = Post − Pre. ΔΔCE = ΔCE (Cognitive Empathy) − ΔCE (Factual–Situational).

DISCUSSION

This study developed a dialogue environment for practicing exploration, interpreting, and paraphrasing by combining a character who does not readily disclose psychological information with response-contingent information disclosure and control (Figure 1). Figure 3 illustrates dialogue focusing on the junior's psychological state as well as dialogue centered on factual and situational information, showing how different interaction tasks can be enacted through generative AI role-play. These selected cases should be understood as illustrative rather than representative of all participants' behaviors.

Figure 4 also reveals instances in both conditions in which advice or directive statements took precedence over the assigned tasks. This observation suggests a direction for developing systems that support users' efforts to understand others in response to their utterances.

For example, when advice is offered prematurely, a future system could prompt users to ask about the junior's feelings; when directive statements precede factual summarization, it could encourage users to organize the information gathered first. These are proposed extensions to the system developed in this study.

Although the mean CE change was greater in the cognitive-empathy condition, baseline scores differed between groups, and the baseline-adjusted result was not in the same direction. Future studies with larger samples and better-balanced baseline scores will examine potential effects on empathy. We will also investigate the influence of the familiar interim research presentation scenario by including different participants and contexts.

Future work will systematically classify dialogue logs to evaluate participants' utterances and patterns of psychological information disclosure, while separately examining the contributions of task instructions and AI response control. We will also assess the applicability of the approach to human–human communication and the effectiveness of response-contingent support features.

CONCLUSION

We designed a generative AI role-play system in which an AI-simulated junior colleague does not readily disclose their psychological state, enabling users to practice exploration, interpreting, and paraphrasing. The preliminary

evaluation identified dialogue cases broadly aligned with the intended tasks as well as cases in both conditions where advice or directive statements took precedence.

These observations suggest that environments for practicing perspective-taking may benefit not only from opportunities for dialogue but also from support that responds to users' utterances. Future work will systematically evaluate dialogue behaviors and the effectiveness of such support features.

ETHICAL APPROVAL

This study was conducted with the approval of the Research Ethics Committee for Studies Involving Human Participants at Chuo University (Approval No. 2021-026(2)).

ACKNOWLEDGMENT

This work was supported by a Collaborative Research Project of the Institute of Science and Engineering, Chuo University.

REFERENCES

Nouraei, F., Rebello, K., Fallah, M., Murali, P., Matuszak, H., Jap, V., Parker, A., Paasche-Orlow, M., and Bickmore, T. (2025). Virtual Agent-based Communication Skills Training to Facilitate Health Persuasion Among Peers. Proceedings of the ACM on Human-Computer Interaction, 9(2), Article CSCW203, 1–24. https://doi.org/10.1145/3711101

Sedefoglu, D., Lahnala, A., Wagner, J., Flek, L., and Ohly, S. (2024). LeadEmpathy: An Expert Annotated German Dataset of Empathy in Written Leadership Communication. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 10237–10248. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.894/

Shen, L. (2010). On a Scale of State Empathy During Message Processing. Western Journal of Communication, 74(5), 504–524. https://doi.org/10.1080/10570314.2010.512278

van Dierendonck, D., and Nuijten, I. (2011). The Servant Leadership Survey: Development and Validation of a Multidimensional Measure. Journal of Business and Psychology, 26(3), 249–267. https://doi.org/10.1007/s10869-010-9194-1