Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

Beyond Role-Playing: A Gen AI-Driven Health Consultation Training System

Xuewang GENG[a*], Li CHEN[b], Mamiko ETO[c]& Masanori YAMADA[d]
[a] *Faculty of Computer and Information Sciences, Sojo University, Japan*
[b]*Division of Math, Sciences, and Information Technology in Education, Osaka Kyoiku University, Japan*
[c]*Faculty of Health and Welfare Department of Nursing, Seinan jo Gakuin University, Japan*
[d]*Data-Driven Innovation Initiative, Kyushu University, Japan*
* xuewang@cis.sojo-u.ac.jp

Abstract: Recent advances in large language models enable virtual patient simulations to generate natural language dialogue for healthcare education. However, most existing implementations focus solely on dialogue interaction, lacking structured feedback to explain how individual utterances impact the consultation. This study presents an online health consultation training system for school nurse education. The system integrates generative artificial intelligence child avatars, an interaction mechanism based on trust, and structured reflection support. Featuring three child personality profiles, the system evaluates each learner utterance in real time to produce transparent trust score changes and textual rationales. A formative evaluation with five university students demonstrated high overall usability and positive perceptions of the reflection interface. Furthermore, interaction log analysis revealed that empathic strategies consistently produced high trust increases, whereas premature directives caused sharp trust decreases. These findings demonstrate that the system effectively addresses traditional peer role-playing limitations by reproducing diverse child behaviors, capturing dynamic trust building processes, and providing actionable feedback for structured reflection.

Keywords: Health Consultation, Generative AI, School Nursing Education, Learning Analytics

1. Background

In Japan, school nurses, known as yogo teachers, play a critical role in maintaining the physical and mental health of children. When students present with somatic complaints such as headaches or abdominal pain, yogo teachers are often the first point of contact. These physical symptoms frequently mask underlying emotional or social issues, including bullying, school refusal, and abuse (Ministry of Education, Culture, Sports, Science and Technology [MEXT], 2017). Effective health consultation requires yogo teachers to build trust with children through adaptive communication, encouraging students to disclose sensitive concerns (Kikuchi & Ikegawa, 2018). However, developing these competencies during preservice training remains a significant challenge. Current training relies primarily on peer role-playing exercises (Imano, 2009), but this approach has several limitations. First, peers acting as child patients find it difficult to authentically reproduce resistant or reluctant behavior, which restricts the range of consultation scenarios that can be practiced (Nakamura, 2022). Second, role-playing scenarios cannot capture the dynamic nature of trust building. In real consultations, a child's willingness to disclose information depends on the trust that a yogo teacher builds through sustained communication (Kikuchi & Ikegawa, 2018); however, peer actors cannot systematically control this process. Third, when peers evaluate consultation performance, their feedback tends to lack specific justification. Students may receive an overall score but gain little understanding of which communication

strategies were effective or ineffective, making structured reflection on communication skills difficult (Wang et al., 2024).

Immersive technologies such as augmented and virtual reality have demonstrated effectiveness in educational contexts, supporting embodied interaction and situated learning (Geng & Yamada, 2019; Li, Geng, & Yamada, 2025). Building on these developments, recent advances in large language models (LLMs) have opened new possibilities for patient simulation in healthcare education. Virtual patients powered by LLMs can generate contextually appropriate natural language responses, providing scalable and repeatable training opportunities (Holderried et al., 2024). Systematic reviews indicate that simulations based on AI can improve communication skills and clinical reasoning in medical and nursing education (Sengul & Sarıköse, 2025). To address these limitations of both traditional peer role-playing, this study proposes an online health consultation training system that integrates a generative AI virtual child avatar with a trust interaction mechanism and comprehensive reflection features. The system supports real time consultation simulations incorporating multiple child personality profiles and a trust assessment mechanism. This study conducted a formative evaluation of the system, focusing on usability and perceived ease of use for both the simulated consultation and the reflection features.

2. System Design and Development

2.1 Simulation Scenario

The system simulates a health consultation where the learner acts as an elementary school Yogo teacher. The dialogue involves a sixth-grade girl presenting with psychosomatic stomach issues. The underlying cause is a social media peer conflict resulting in class isolation. Since the child hesitates to consult her homeroom teacher or parents, the learner must use appropriate techniques to uncover these psychosocial issues. The child initially reports only physical symptoms, revealing the true problem only after sufficient trust is established. Three child personality profiles share this scenario but feature distinct behavioral patterns and trust thresholds. The active type requires a threshold of 30, deflecting cheerfully but responding well to empathic engagement. The introverted type requires a threshold of 40, providing minimal responses and withdrawing if rushed. The resistant type requires a threshold of 60, acting dismissively and responding primarily to indirect approaches like discussing hobbies. Upon reaching their respective thresholds, each type discloses the core issue in a characteristic manner, ranging from emotional to hesitant or reluctant.

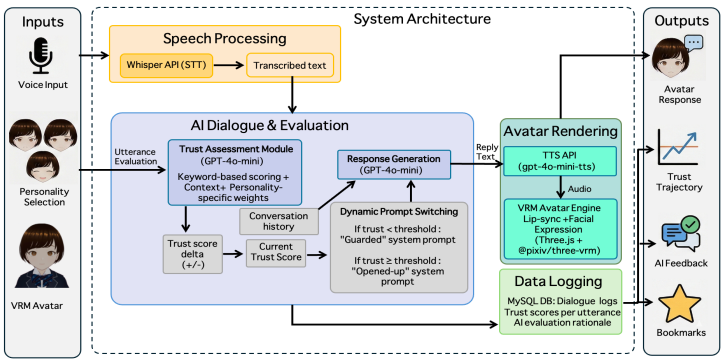

Figure 1. System architecture of the web-based health consultation training system.

2.2 System Architecture

The system comprises four main components: Speech Processing, AI Dialogue and Evaluation, Avatar Rendering, and Data Logging (see Figure 1). Users select a child personality and interact via voice input. The system processes this dialogue to generate a

synchronized avatar response. All interaction data are stored for later reflection. Voice input is captured via MediaRecorder and transcribed by the Whisper API. The transcribed utterance is processed sequentially by two modules. First, the Trust Assessment Module uses GPT-4o-mini to evaluate each utterance by considering keyword relevance, conversational context, and personality-specific weighting, outputting a score change ranging from −20 to +20 along with a textual rationale. This updated score dictates Dynamic Prompt Switching. A score below the threshold triggers a guarded prompt to produce resistant behavior, while reaching the threshold activates an open prompt enabling disclosure. The disclosure thresholds (30, 40, and 60) operate on the cumulative trust score (0–100) and determine when each personality type reveals the core issue. Subsequently, the Response Generation module uses the GPT-4o-mini model to formulate a reply based on the conversation history and the current trust state. The generated reply is synthesized into speech via the TTS API and delivered through a VRM Avatar Engine featuring real time lip sync for Japanese vowels. The facial expressions of the avatar are dynamically modulated across four trust levels. At a low trust ratio below 0.4 (current score ÷ 100), sad and angry expressions convey wariness. These expressions are progressively reduced at medium and high trust levels. Upon reaching the threshold for disclosure, the expression shifts to slight sadness and relaxation to reflect emotional relief. Figure 2 illustrates the training interface where teachers record their dialogue using a designated screen button, which prompts the avatar to reply with synchronized speech and expressions.

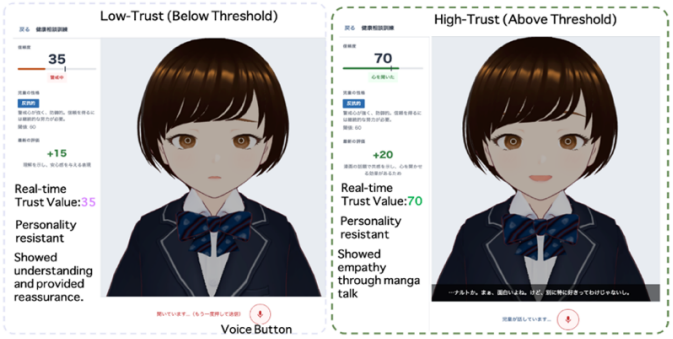

Figure 2. Training interface showing voice-based consultation with the AI-driven child avatar.

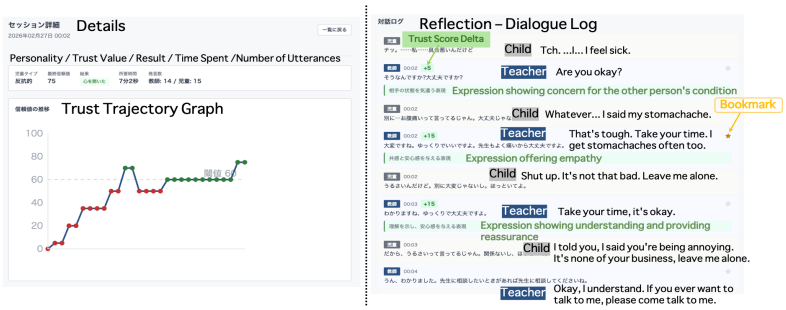

Figure 3. Reflection interface showing dialogue log and trust trajectory graph.

The Data Logging component records dialogue logs, individual trust scores, and AI evaluation rationales to support a comprehensive reflection interface, as illustrated in Figure 3. This interface features a chronological dialogue log detailing both learner and avatar utterances alongside the corresponding trust score changes and feedback generated by AI, clarifying the effectiveness of specific interactions. Additionally, a time series graph visualizes the trust trajectory against a threshold line specific to the selected personality, helping learners identify conversational turning points. Finally, a bookmarking function aggregates selected utterances on a dedicated page to facilitate targeted and selective reflection.

3. Formative Evaluation

Five university students majoring in information sciences participated in the formative evaluation. As this evaluation targeted system usability rather than domain-specific training effectiveness, school nursing expertise was not required. During a 50-minute session, participants completed consultations with at least two child personality types, reviewed the reflection page, and completed the System Usability Scale (SUS; Brooke, 1996), the Learning Analytics Dashboard Success Questionnaire (LADSQ; Park & Jo, 2019), and open-ended questions. The SUS assessed overall usability on a 0–100 scale, where scores above 50.9 are acceptable and above 71.4 are good (Bangor et al., 2009). The LADSQ evaluated the reflection interface on a 5-point Likert scale across five subscales: Visual Attraction (effectiveness of layout and visual design), Usability (ease of navigation), Understanding (learners' comprehension of their consultation performance), Perceived Usefulness (practical value for skill improvement), and Behavioral Changes (motivation to adjust subsequent learning strategies). SUS and LADSQ scores were calculated using standard procedures. Interaction logs were analyzed by categorizing utterances as positive, negative, or neutral based on trust score changes. A sequence heatmap visualized these temporal patterns, with each cell representing a single utterance color-coded by evaluation type and trust score intensity.

4. Results and discussion

4.1 System and Reflection Interface Evaluation

The system’s overall usability and the effectiveness of the reflection interface were evaluated using the SUS and LADSQ instruments. The mean SUS score was 70.0 (SD = 12.91), which significantly exceeds the "OK" threshold of 50.9 and aligns with the "Good" benchmark of 71.4 (Bangor et al., 2009).Item-level analysis highlighted strong positive responses for simplicity (Q3: M = 3.80) and the desire for frequent use (Q1: M = 3.20), though the higher variance in learnability (Q7: M = 3.20, SD = 1.30) suggests some individual differences in the initial onboarding experience.

Table 1. *Results of SUS and LADSQ questionnaires (n=5)*

Measure	Factors	Mean	Standard Deviation
SUS (1–100)	Total Usability Score	70	12.91
LADSQ (1–5)	Visual Attraction	4.11	0.80
	Reflection Usability	4.05	0.76
	Understanding	4.00	0.88
	Perceived usefulness	4.05	0.78
	Behavioral changes	4.09	0.62

Regarding the reflection interface, Table 1 shows that all five LADSQ subscales surpassed 4.0 on a 5-point scale. Both Visual Attraction (M = 4.11) and Usability (M = 4.05) received high ratings, confirming that the interface was visually comprehensible and easy to navigate. These findings align with prior research indicating that visual design and usability are critical factors influencing learners' engagement (Park & Jo, 2019). Furthermore, the scores for Understanding (M = 4.00), Perceived Usefulness (M = 4.05), and Behavioral Changes (M = 4.09) demonstrate the dashboard's educational relevance. Specifically, the Behavioral Changes dimension was found to be a significant predictor of the regulation of cognition dimension of metacognition (Chen et al., 2020). This aligns with findings that LAD can enhance metacognitive awareness in technology-enhanced learning environments (Geng & Yamada, 2025). Collectively, these results demonstrate that the system achieves high usability and educational relevance, successfully integrating complex real-time AI dialogue and 3D rendering within a user-friendly web browser environment.

4.2 Interaction Log Analysis

In the system, a single utterance is defined as one continuous voice recording initiated and terminated by the learner pressing the recording button on the interface. Each button press produces one transcribed text segment, which is then evaluated as a single unit by the Trust Assessment Module. Interaction logs from twelve training sessions across five participants were analyzed to evaluate how effectively the system simulates realistic consultation dynamics. Figure 4 illustrates the utterance evaluation sequences, with cells color coded by evaluation type and trust score intensity. Distinct patterns emerged across personality types. Active sessions were dominated by positive green cells, demonstrating rapid threshold attainment. Introverted sessions revealed a clear contrast between success and failure. Successful learners transitioned from neutral to predominantly positive evaluations near the midpoint. In contrast, failed sessions remained dominated by neutral gray cells throughout, indicating utterances that neither harmed nor built trust. Notably, one participant conducted two separate sessions with an introverted profile. The first unsuccessful attempt contained nine utterances that were overwhelmingly neutral. The second successful attempt featured fourteen utterances with a significantly higher proportion of positive interactions. This marked improvement likely resulted from reviewing the feedback generated by AI between sessions, which facilitated a strategic shift toward positive communication.

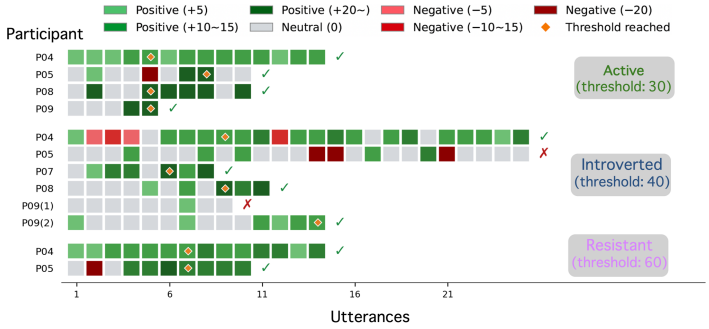

Figure 4. Utterance Evaluation Sequences by Personality Type.

Examination of individual utterances revealed specific techniques that significantly increased trust. First, sharing personal interests built rapport with reluctant children (Dickens et al., 2025); for example, participants stated, "I like reading too, mostly manga and playing games," or asked for anime recommendations. Second, empathic acknowledgment provided deep reassurance, as seen in responses like, "That must be really tough. You have done well. Listen, I am always on your side." Third, affirming courage and promising availability reflected foundational empathic acceptance (Kaluzeviciute, 2020). Conversely, prematurely directing the child to the classroom (e.g., "Go back to the classroom right now, everyone will be worried") caused sharp trust decreases, confirming that directive statements undermine building trust (Ryan, Berry, & Hartley, 2023). Ultimately, these findings address traditional peer role-playing limitations by consistently reproducing diverse child behaviors (Nakamura, 2022), capturing the cumulative nature of building rapport, and providing actionable utterance level feedback to enable structured reflection.

5. Conclusion

This study presented an online health consultation training system integrating generative artificial intelligence child avatars, a trust evaluation mechanism, and structured reflection support. Formative evaluation demonstrated good usability and positive perceptions of the reflection interface. Interaction logs confirmed the system differentiates consultation techniques, revealing that empathic strategies consistently increased trust while premature directives caused sharp decreases, aligning with established principles. Future research will

involve preservice Yogo teacher participants within formal courses to assess training effectiveness using performance-based measures such as consultation quality rubrics. Additionally, the system's pedagogical design will be strengthened by incorporating proactive scaffolding strategies, such as guided prompts and adaptive feedback during consultations, grounded in clinical communication frameworks. Expanding the scenario library beyond a single social media conflict to other prevalent issues is also planned.

Acknowledgements

This study is funded by Japan Society for the Promotion of Science (JP26K16815, JP25K17079, JP22H00552) and the Telecommunications Advancement Foundation.

References

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Chen, L., Lu, M., Goda, Y., Shimada, A., & Yamada, M. (2020). Factors of the use of learning analytics dashboard that affect metacognition. *In Proceedings of the 17th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2020)* (pp. 295–302).

Dickens, R., Leroy, P., Eppich, W., & Brenner, M. (2025). Trustful relationships between healthcare professionals and children: A concept analysis using Rodgers' evolutionary approach. *European Journal of Pediatrics*, *184*(7), 464.

Geng, X., & Yamada, M. (2019). Development and design of a compound verb AR learning system employing image schemas. In Proceedings of the International Conference on Mobile Learning 2019 (pp. 73–80).

Geng, X., & Yamada, M. (2025). The role of dashboards in augmented-reality-based language learning: Enhancing language learning and metacognitive awareness. SAGE Open, 15(2), 21582440251341675.

Holderried, F., Stegemann-Philipps, C., Herschbach, L., Moldt, J.-A., Nevins, A., Griewatz, J., Holderried, M., Herrmann-Werner, A., Festl-Wietek, T., & Mahling, M. (2024). A generative pretrained transformer (GPT)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR Medical Education*, *10*(1), e53961.

Imano, Y. (2009). Effectiveness of a seminar on health counseling activities for developing practical skills of Yogo teachers. *Human Welfare Studies*, *12,* 61–73.

Kaluzeviciute, G. (2020). The role of empathy in psychoanalytic psychotherapy: A historical exploration. Cogent Psychology, 7(1), 1748792.

Kikuchi, M., & Ikegawa, N. (2018). The initial stage of the continuous support process of health consultation conducted by Yogo teachers. *Japanese Journal of School Health, 60*(1), 26–40.

Li, T., Geng, X., & Yamada, M. (2025). Embodied learning in virtual reality: Enhancing Japanese psychomimetic word acquisition through emotional experiences. IEEE Access.

MEXT. (2017). Support for children with modern health issues: *Focusing on the role of Yogo teachers.* https://www.mext.go.jp/a_menu/kenko/hoken/1384974.htm

Nakamura, S. (2022). Learning through role-playing in health counseling training. *Nursing Journal of Osaka Aoyama University*, *5*, 29–33.

Park, Y., & Jo, I. (2019). Factors that affect the success of learning analytics dashboards. *Educational Technology Research and Development, 67*, 1547–1571.

Ryan, R., Berry, K., & Hartley, S. (2023). Therapist factors and their impact on therapeutic alliance and outcomes in child and adolescent mental health: A systematic review. *Child and Adolescent Mental Health*, *28*(2), 195–211.

Sengul, T., & Sarıköse, S. (2025). Enhancing learning outcomes through AI-driven simulation in nursing education: A systematic review. *Clinical Simulation in Nursing*, *106*, 101797.

Wang, J., Wang, B., Liu, D., Zhou, Y., Xing, X., Wang, X., & Gao, W. (2024). Video feedback combined with peer role-playing: A method to improve the teaching effect of medical undergraduates. *BMC Medical Education*, *24*(1), 73.