

Adaptive-VP: A Framework for LLM-Based Virtual Patients that Adapts to Trainees' Dialogue to Facilitate Nurse Communication Training

Keyeun Lee^{1,2}, Seolhee Lee^{1,2}, Esther Hehsun Kim^{1,2}, Yena Ko², Jinsu Eun¹,
Dahee Kim³, Hyewon Cho¹, Haiyi Zhu⁴, Robert E. Kraut⁴,
Eunyoung Suh³, Eun-mee Kim², Hajin Lim^{1,2}

¹hci+d Lab, ²Department of Communication, ³Department of Nursing
Seoul National University

⁴ Human-Computer Interaction Institute, Carnegie Mellon University
{kieunp, hajin}@snu.ac.kr

Abstract

Effective communication training is essential to preparing nurses for high-quality patient care. While standardized patient (SP) simulations provide valuable experiential learning, they are often costly and inflexible. Virtual patient (VP) systems offer a scalable alternative, but most fail to adapt to the varying communication skills of trainees. In particular, when trainees respond ineffectively, VPs should escalate in hostility or become uncooperative—yet this level of adaptive interaction remains largely unsupported. To address this gap, we introduce Adaptive-VP¹, a VP dialogue generation framework that leverages large language models (LLMs) to dynamically adapt VP behavior based on trainee input. The framework features a pipeline for constructing clinically grounded yet flexible VP scenarios and a modular system for assessing trainee communication and adjusting VP responses in real time, while ensuring learner safety. We validated Adaptive-VP by simulating challenging patient conversations. Automated evaluation using a corpus from practicing nurses showed that our communication skill evaluation mechanism reflected real-world proficiency levels. Expert nurses further confirmed that Adaptive-VP produced more natural and realistic interactions than existing approaches, demonstrating its potential as a scalable and effective tool for nursing communication training.

1 Introduction

Effective nurse-patient communication is crucial for enhancing treatment adherence and health outcomes (Patak et al., 2009; Chochinov et al., 2013; Peimani et al., 2020). Conversely, poor communication can lead to medical errors, increased patient dissatisfaction, and heightened provider stress (Dithole et al., 2016; Banerjee et al., 2016).

¹Code and data available at: <https://github.com/keyeun/adaptive-vp>

Traditionally, simulation-based training with standardized patients (SPs) has been central to nursing education (MacLean et al., 2017; Nestel and Bearman, 2014). However, it often prioritizes procedural skills over communication, incurs high costs, and scripted interactions limit its ability to reflect the dynamic nature of real clinical encounters (Elendu et al., 2024; Wallace et al., 2002).

Virtual patients (VPs) have emerged as a scalable alternative for clinical training (Barrows, 1993; Ziv et al., 2006; McGaghie et al., 2010; Pascucci et al., 2014). Recent advances in large language models (LLMs) enhance VPs by enabling more natural, context-aware interactions (Li et al., 2024; Fan et al., 2025; Chen et al., 2023). However, most VP systems still lack a natural feedback loop—when trainees use ineffective communication strategies, the VP should respond accordingly (e.g., by escalating frustration), and vice versa (e.g., de-escalating when communication improves) (Graf et al., 2024).

To address these limitations, we develop **Adaptive-VP** (see Figure 1), a LLM-based VP framework to dynamically adapt VP behavior based on trainee input. Our approach begins with the **VP case development pipeline**, which guides the creation of clinically grounded yet customizable VP-based training scenarios by incorporating detailed patient personas and clinical contexts aligned with specific training goals.

Dialogue adaptation is managed by four core modules that collaboratively adjust the VP's behavior based on trainee performance (e.g., escalating problematic behavior when trainees' response is ineffective and de-escalating when it is appropriate). First, the **Evaluation Module** assesses trainee utterances using a multi-agent evaluation process based on criteria informed by best practices and literature in nursing communication. Based on this assessment, the **Dynamic Adaptation Module** then determines the direction of the VP's next response and passes this to the **Dialogue Generation**

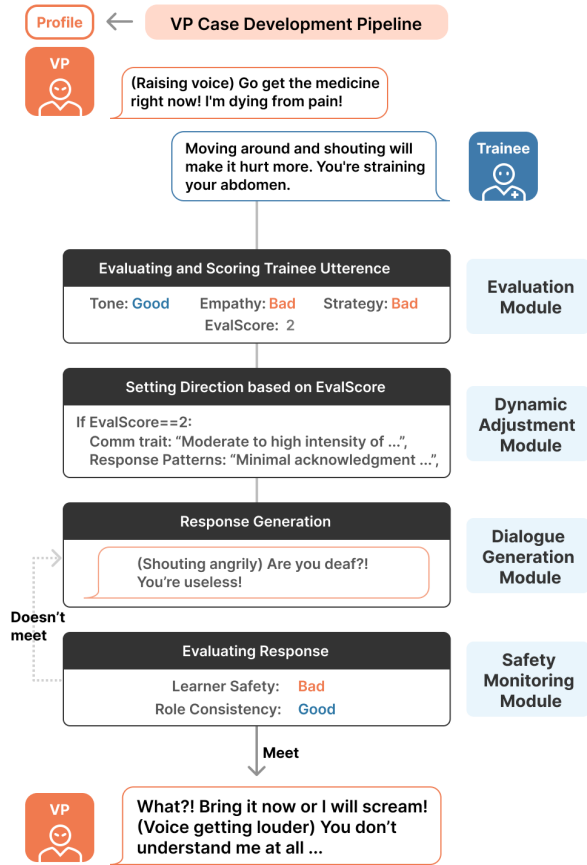


Figure 1: Overview of the Adaptive-VP framework

Module, which generates a contextually appropriate VP dialogue. Finally, the **Safety Monitoring Module** reviews the generated VP dialogue before presenting it to trainees, ensuring learner safety by filtering harmful content (e.g., extreme toxicity), while preserving the realism of interaction.

As a testbed for refining and validating Adaptive-VP, we focused on simulating **challenging patient interactions**, which encompass a range of encounters where communication difficulties disrupt therapeutic relationships between patients and healthcare providers (Hardavella et al., 2017; Marcum, 2015). Such challenges often stem from patient characteristics such as aggression, non-cooperation, or demanding attitudes (Groves et al., 1978; Hardavella et al., 2017), which can lead to emotional exhaustion for providers and strain the patient-provider relationship (Serour et al., 2010). Studies estimate that approximately 15% of clinical encounters are perceived by physicians as “difficult” (Hahn et al., 1996; Hinchey and Jackson, 2011), yet targeted VP-based training for handling such interactions remains limited. By focusing on these challenging interactions, we intend to demonstrate

Adaptive-VP’s potential to support scalable, realistic, and adaptive communication training.

To apply the Adaptive-VP framework to training for challenging patient interactions, we conducted a series of validation studies. First, we evaluated the quality and realism of cases generated by the VP Case Development Pipeline through consultations with 10 nursing professionals. Second, we validated the Evaluation Module using conversation corpora collected from expert and novice nurses (15 each). Finally, we conducted a between-subjects experiment with 28 nursing professionals to assess the realism and role consistency of VP responses from the Adaptive-VP framework.

Results demonstrated that Adaptive-VP generated highly adaptive and contextually grounded VP behavior, significantly improving the perceived realism of VP interactions compared to existing approaches. These findings highlight the framework’s potential to support scalable communication training across healthcare domains. Moreover, Adaptive-VP offers a promising approach for generating diverse clinical dialogue corpora—especially in areas where real-world data collection is constrained by privacy regulations and policy restrictions (Wang et al., 2024a).

2 Related Work

LLM-based agents are increasingly used as virtual patients (VPs) for communication training in healthcare education (Fan et al., 2025). These agents have been applied to a range of clinical scenarios, including diagnostic training (Chen et al., 2023), patient interviewing (Li et al., 2024), and history-taking simulations (Holderried et al., 2024b; Sardesai et al., 2024), demonstrating their ability to generate clinically relevant VP dialogues. Some frameworks further integrate multi-agent interactions (Bao et al., 2024; Fan et al., 2025) and cognitive modeling to support mental health training and psychiatric evaluations (Chen et al., 2023; Louie et al., 2024; Gabriel et al., 2024).

While these systems have advanced the use of LLMs in healthcare training, many remain limited in their ability to reflect the adaptive nature of human communication. In real-world interactions, speakers continuously adjust their behavior in response to their conversation partner’s cues and signals (Kraut et al., 1982). In contrast, existing LLM-based VP systems often focus on maintaining pre-defined role fidelity and scenario consistency

Table 1: Comparison of LLM-based Virtual Patient Systems

| Study | Target Trainee | VP System Objective | Expert-verified VP | Real-time Evaluation | Adaptive Behavior | Learner Safety |
|---------------------------|-----------------------------|--------------------------------------------------------------------------------------------------|--------------------|----------------------|-------------------|----------------|
| Wang et al. (2024b) | Counseling trainees | CBT skill training | ✓ | ✗ | ✗ | ✗ |
| Steenstra et al. (2025) | Counselors | Motivational interviewing skill training (substance misuse context) | ✗ | ✓ | ✓ | ✗ |
| Louie et al. (2024) | Novice counselors | Mental health counseling skill training | ✓ | ✗ | ✗ | ✓ |
| Hicke et al. (2025) | Medical students | History taking and physician–patient communication training | ✓ | ✗ | ✗ | ✗ |
| Bao et al. (2025) | <i>Not training-focused</i> | Nurse–patient outpatient reception dialogue generation (for data creation and model development) | ✗ | (✓) | ✗ | ✗ |
| Holderried et al. (2024a) | Medical students | History taking skill training | (✓) | ✗ | ✗ | ✗ |
| Our Work | Nurse trainees | Communication skill training across diverse clinical scenarios | ✓ | ✓ | ✓ | ✓ |

(Jiang et al., 2025; Wang et al., 2024b; Louie et al., 2024), which can constrain their ability to simulate dynamic patient-provider exchanges. Additionally, rigid, scenario-specific designs may reduce scalability across varied training goals and communication competencies.

To better contextualize our approach in relation to existing systems, Table 1 provides a comparative overview of recent LLM-based VP frameworks. We summarize each system’s target trainees, training objectives, and the presence of four core capabilities that also define the design goals of our framework: expert-verified VP personas, real-time response evaluation, adaptive VP behavior, and learner safety mechanisms.

As shown in Table 1, existing systems vary in their coverage of the four core capabilities we identify. While many incorporate one or two of these dimensions, few integrate all four. For example, several systems maintain consistent role personas and structured scenarios but do not adapt their behavior based on trainee input. Others often lack safeguards to support learner well-being. These patterns highlight opportunities for more comprehensive and flexible VP frameworks that not only adapt dynamically to trainee performance, but also incorporate safeguards to support learner well-being and ensure realistic, pedagogically sound interactions.

Our framework, **Adaptive-VP**, is designed to address these gaps through a structured, modular approach. It integrates all four core capabilities identified in Table 1, enabling the creation of diverse, clinically grounded VP cases tailored to

a range of communication training goals. At its core, Adaptive-VP implements a four-module architecture that dynamically adjusts VP behavior in real time based on trainee performance, supporting training experiences that are realistic, responsive, and pedagogically effective.

3 Overview of Adaptive-VP Framework

In developing the Adaptive-VP framework, we identified four core **challenges** in creating adaptive VP dialogues and developed targeted **approaches** to address them, as outlined below.

3.1 Challenge 1: Developing Clinically Grounded yet Adaptable VP cases

In SP simulation training, detailed case protocols are essential, encompassing patient demographics, clinical content (e.g., present symptoms, medical history), psychosocial background, situational details, and SP-specific complaints (INACSL, 2023; ASPE, 2022). While existing SP protocol structures and developed cases offer a strong foundation for VP case design, their highly scripted and context-specific nature often limits reusability and adaptability across diverse clinical training scenarios (Elendu et al., 2024; Wallace et al., 2002).

Our Approach To address this, we developed the **VP Case Development Pipeline**, ensuring clinical validity while allowing flexibility in tailoring educational goals and contexts.

Our pipeline consists of five stages. First, it is necessary to **1) clarify the training goal**, identifying key communication challenges and training

goals, such as managing challenging patient interactions. Second, we **2) incorporate relevant literature** by integrating evidence-based insights from nursing communication literature and best practices. Third, we **3) specify the training context**, including geographical, cultural, and trainee-specific factors that shape the learning scenario.

By inputting this information into LLMs, the pipeline **4) generates draft VP cases**, developing VP profiles and clinical scenarios aligned with the training focus and SP protocol standards (INACSL, 2023; ASPE, 2022). To promote consistency in VP behavior, the pipeline also specifies detailed communication traits of VP (De Vries et al., 2009), guiding LLMs in generating coherent response patterns. Finally, **5) expert validation** involves clinical educators and experienced practitioners reviewing the generated profiles and scenarios. Their feedback on clinical accuracy and educational value informs the refinement of the draft VP cases.

3.2 Challenge 2: Robustly Evaluating Nurse Communication Efficiency

Traditional approaches to evaluating nurse communication effectiveness have relied on standardized tools such as observation checklists and questionnaires (Bialer et al., 2011; Cannity et al., 2021). While these methods offer structured assessment frameworks, they often depend on a small number of human evaluators, including self-assessments or expert reviews, which can introduce subjectivity and inconsistencies (Podsakoff et al., 2003; Hoyt and Kerns, 1999).

More recently, large language models (LLMs) have been adopted for automated evaluation. However, single-agent assessments can exhibit several known biases, including positional bias (Wang et al., 2023; Liusie et al., 2024), self-preference bias (Koo et al., 2024; Liu et al., 2024), and inconsistencies in knowledge or formatting (Zhu et al., 2023). These challenges highlight the need for more robust and reliable evaluation mechanisms for communication training.

Our Approach To enable robust evaluation of trainee communication effectiveness, we draw on well-established guidelines and criteria from nursing literature and best practice frameworks. While specific metrics may vary depending on training goals and contexts, the underlying structure remains consistent: assessing communication at both the utterance and conversation levels.

Accordingly, our **Evaluation Module**, operates at two levels: (1) the utterance level, assessing qualities such as tone and empathy, and (2) the conversation level, evaluating context-specific strategies—for example, the use of de-escalation techniques when managing difficult patient interactions (Price et al., 2024).

Furthermore, to enhance evaluation reliability, we adopt a multi-agent evaluation, as described in (Chan et al., 2023). Systematic evaluation results (e.g., communication efficiency score) derived from this process guide how the VP dynamically adapts its responses based on trainee performance.

3.3 Challenge 3: Dynamically Adjusting VP Responses

Static VP behaviors throughout training can be ineffective, as patient behavior in real-world settings is dynamic and responsive to the healthcare provider’s communication (Pines et al., 2021).

Our Approach To adjust VP responses based on trainee performance, the **Dynamic Adjustment Module** determines the direction of VP responses to the trainee’s utterance based on evaluation results from the **Evaluation Module**. These evaluation results, represented as a score, guide modifications to the communication traits defined in the VP cases and influence how the VP responds to nurses, directing the **Dialogue Generation Module** to produce contextually appropriate responses.

3.4 Challenge 4: Ensuring Learner Safety

While realism in VP behavior is essential, an overly combative or hostile VP dialogue may cause emotional distress and reduce learner confidence (Kardong-Edgren et al., 2024; Stephen et al., 2020).

Our Approach To ensure learner safety, the **Safety Monitoring Module** evaluates each VP utterance before presenting it to the trainee. The initial VP response, generated by the **Dialogue Generation Module**, is assessed against four criteria. First, “safety assurance” detects overly hostile or derogatory language. Next, it verifies “alignment with the training goal,” ensuring that the response is relevant to the intended communication training objectives. It also examines “consistency” with the patient case details and “adherence” to behavioral directions from the **Dynamic Adjustment Module**. If the utterance fails to meet any of these criteria, it is returned to the **Dialogue Generation Module** with feedback for revision.

3.5 Test Case: Creating VPs Demonstrating Challenging Patient Traits

We applied the Adaptive-VP framework to simulate **challenging patient interactions**, a particularly complex domain within nursing communication (Groves et al., 1978; Townsley et al., 2023). These scenarios involve patients exhibiting behaviors that hinder effective therapeutic communication, such as being overdependent, authoritative, aggressive, or uncooperative (Groves et al., 1978). Below, we detail the implementation and validation process for developing VPs that realistically portray these challenging traits, providing guidance for applying the framework across diverse nursing communication training contexts.

4 Developing VP Cases for Challenging Patients

We applied the **VP Case Development Pipeline** to generate challenging patient cases that reflect nurses' communication challenges in specific clinical settings following the five steps below.

1) Clarifying the Training Goal Our focus was on providing training on handling challenging patient interactions since this is widely recognized as one of the most difficult aspects of nursing communication (Stein et al., 2022).

2) Incorporating Relevant Literature We first identified four prevalent types of challenging patients along with their corresponding traits: overdependent, overly authoritative, threatening, and non-cooperative toward treatment, based on a review of nurse-patient communication literature (Colson et al., 1985; Groves et al., 1978; Kits, 1990) (See A.1 for details on each challenging patient type).

3) Specifying the Training Context In South Korea, nurses frequently face emotional labor when managing challenging patients, leading to stress and burnout (Hankyung, 2020; KN News, 2016). This challenge is particularly critical for novice nurses, who often feel unprepared for real-world practice, leading to strained patient relationships (Ho et al., 2021; Kim et al., 2020). Notably, 57.4% of novice nurses in South Korea resign within their first year (Yonhap News, 2024), with conflicts involving patients and caregivers cited as a major contributing factor (Son et al., 2017). Given this, we specified our training contexts as targeting early-career nurses in South Korea, with the goal of

strengthening their communication skills for managing challenging patient interactions through VP-based training.

4) Generate Draft VP Cases Based on the specified training goals and context, we used Claude-3.5 Sonnet to generate two distinct scenarios for each of the four challenging patient types, yielding a total of eight scenarios. Claude-3.5 Sonnet was used throughout, given its strong performance in Korean and clinical dialogue (Jang and Jung, 2024; Kim et al., 2024; Schmidgall et al., 2024). We then prompted the LLM to construct detailed patient profiles in accordance with international SP-based training protocols (INACSL, 2023; ASPE, 2022), including demographics, medical history, and situational details (see A.3). To ensure realistic and consistent communication behavior, we further prompted the model to generate communication traits based on seven empirically grounded styles (e.g., threateningness) (De Vries et al., 2009), guiding it to tailor each patient's tone and expressions to their persona type, scenario, and demographic attributes (see A.4).

5) Expert Validation To validate the clinical validity of the eight draft VP cases, we conducted an expert evaluation with 10 practicing nurses (avg. 7.0 years of experience). On a 5-point scale, participants rated the scenarios moderately high in both realism ($M = 3.81$, $SD = 0.97$) and accuracy in reflecting patient characteristics ($M = 4.00$, $SD = 0.94$). Follow-up interviews provided additional feedback that informed revisions. For example, one nurse (EV2) remarked, "*Patients often complain about hospital food, but rarely become violent over it*," prompting us to replace the original scenario with a more contextually grounded case involving patient aggression triggered by scheduling changes for medical examinations (see B for the finalized VP cases refined through expert validation).

5 Evaluating Trainee Communication Efficiency with Challenging Patients

We implemented an evaluation module aligned with established nursing communication strategies (Price et al., 2024; Ernstmeier et al., 2022; Hallett and Dickens, 2017; Sheldon, 2009; TMLT, 2022a), assessing nurses' communication at both the **utterance level** (Tone, Empathy, and Prohibited Communicative Behaviors) and the **conversation level** (De-escalation for tension management). Further,

| Component | Evaluation Unit | Subcomponent | Scoring Condition | Max | Min | Example Nurse Utterance (Max / Min) |
|-----------------------------------|-----------------|-------------------------------|--------------------------------|-----|-----|----------------------------------------------------------------------------------------------------------------------------------|
| Tone | Utterance | Calm Clear | If calm AND clear, +1 | 1 | 0 | Max: "Could you tell me where you feel uncomfortable?" Min: "Yeah... I get it, but that's just how things are." |
| Empathy | Utterance | Empathy level (0-6) | If empathy level ≥ 3 , +1 | 1 | 0 | Max: "I'm so sorry you're feeling this way." Min: "It's not a big deal. Just deal with it." |
| Prohibited Communicative Behavior | Utterance | Premature empathy | | | | Max: (No premature empathy) Min: "I know it's tough, but let's just get it done." |
| | | Invalidating beliefs | If any behavior present, -1 | 0 | -1 | Max: (No invalidation) Min: "That's not true. You're just imagining things." |
| | | Dismissive commands | | | | Max: (No dismissiveness) Min: "Stop whining and just do what I say." |
| De-escalation Strategy | Conversation | Autonomy | If used at least once, +1 | | | Max: "Would you like to take a break or keep talking?" Min: "Just do as I say. You have no choice." |
| | | Limit-setting | If used at least once, +1 | 3 | 0 | Max: "I need you to stay seated for now." Min: "If you don't sit down, I won't talk to you." |
| | | Problem solving/ Reframing | If used at least once, +1 | | | Max: "Let's find a way to make this easier for you." Min: "There's nothing we can do. Just accept it." |

Table 2: Evaluation criteria with example utterances for maximum and minimum scores.

to enhance the reliability of these assessments, we adopted a multi-agent evaluation process to mitigate biases commonly found in single-agent judgments. Based on this process, each trainee utterance is scored for communication efficiency based on this evaluation framework (see Table 2).

5.1 Evaluation Criterion and Scoring Mechanism

Utterance-level Evaluation *Tone* plays a crucial role in nurse-patient communication (Sheldon, 2009), so we assess whether the trainee’s utterance is Calm and Clear, awarding 1 point if both criteria are met.

Empathy enhances patient well-being, satisfaction (Howick et al., 2018; Madula et al., 2018), and therapeutic communication (Blake and Blake, 2019; Brown and Bylund, 2008). Using the Empathic Communication Coding System (Kleinsmith et al., 2015), which classifies healthcare professionals’ empathy levels on a 0–6 scale, we designated level 3 or higher as empathetic, awarding 1 point.

Prohibited Communicative Behaviors disrupt rapport and effective intervention. Based on TMLT (2022b), we penalize the following behaviors (1-point deduction per utterance, regardless of frequency): (1) Premature claims of empathy, (2) Invalidating beliefs, and (3) Dismissive commands.

Conversation-level Evaluation *De-escalation* is a critical communication strategy for managing tension while preserving patient autonomy (Spencer et al., 2018; Hallett and Dickens, 2017). Originally developed for handling aggressive behavior, de-escalation techniques have also demonstrated effectiveness in managing broader forms of challenging patient encounters (Richmond et al., 2012; Accinni et al., 2021).

In our evaluation, we focused on three core de-escalation strategies identified by Price et al. (2024): *Autonomy*, which engages patients in decision-making to enhance their sense of control; *Limit-setting*, which establishes behavioral boundaries for safety; and *Problem-solving & Reframing*, which facilitates constructive dialogue by reinter-

preting the situation collaboratively. Each strategy is scored with 1 point if it is observed at least once during the conversation.

5.2 Multi-agent Evaluation

To ensure both rigorous and comprehensive evaluation of trainee performance, we implement a multi-agent pipeline following role-based evaluation approaches in prior work (Zhang et al., 2023; Wu et al., 2023; Zhu et al., 2024). To minimize false positives and improve reliability, a response is only considered valid when all agents reach unanimous agreement. We define three specialized evaluator roles, each reflecting a critical perspective in communication training: a Nursing Professor, a Communication Skills Trainer, and a Clinical Psychologist (See C.1, C.2, and C.3 for role descriptions and prompting details).

5.3 Validating Evaluation Rigor

We assessed our evaluation module’s rigor by focusing on two questions: 1) *Does it effectively distinguish between expert and novice communication efficiency?*, and 2) *Do differences in multi-agent scores reflect meaningful role-based perspectives rather than random variation?*

Distinguishing Expert from Novice Performance

We analyzed dialogue corpora from two groups: novice (N = 15; ten with <3 years of experience, five pre-licensure students) and experienced nurses (N = 15; mean tenure = 12.5 years). Each participant engaged in eight interactions with VP agents generated solely from our validated VP cases (see B), excluding the evaluation module during data collection. We then applied our multi-agent evaluation pipeline (see C) to assess their utterances. The average conversation length was 7.45 turns for experienced nurses vs. 5.3 turns for novices. For consistency, only the first five utterances per session were analyzed.

We first examined whether the evaluation scores captured meaningful differences in communication performance between groups. As the Shapiro–Wilk tests indicated non-normal score distributions, we applied the non-parametric Mann–Whitney U test to compare total scores. Figure 2 shows the turn-by-turn mean scores (with 95% confidence intervals (CI)) for both groups.

Overall, experienced nurses achieved significantly higher scores ($U = 160960.0$, $p = 0.001$) with notably stronger performance in tone man-

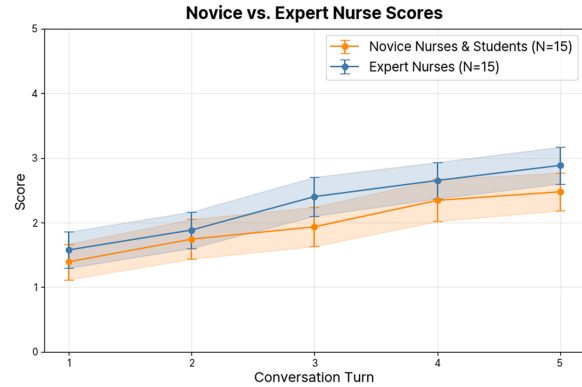


Figure 2: Turn-by-turn mean evaluation scores (with 95% CI) for expert (N=15) and novice/students (N=15)

agement and use of de-escalation strategies. This result suggested that our evaluation module reliably distinguished effective communication behaviors. A detailed subcomponent analysis is provided in Appendix G.1.

Reliability and Role-Based Differences in Multi-Agent Evaluation

Next, we assessed whether the multi-agent evaluation framework yielded consistent scores and whether observed differences among raters reflected legitimate role-based perspectives rather than random variability. Using Fleiss’ kappa to measure overall consistency across the three evaluators yielded high inter-rater agreement (Fleiss’ $\kappa > 0.75$).

To further examine the source of rating variation, we conducted a mixed-effects analysis using Generalized Estimating Equations (GEE), modeling evaluator role as a fixed effect and controlling for textual variation as a random effect.

The results showed that score variations were not random but systematically reflected role-specific perspectives. Both the Communication Skills Trainer and the Nursing Professor consistently assigned lower scores for tone-related attributes than the Clinical Psychologist (all $p < 0.001$). These two roles also gave significantly lower ratings for limit-setting ($p < 0.001$) and problem-solving and reframing strategies ($p < 0.001$ and $p < 0.05$ respectively), while no significant role-based differences emerged for autonomy. Conversely, they rated the presence of prohibited behaviors—specifically invalidating beliefs and dismissive commands—significantly higher than the Clinical Psychologist (both $p < 0.001$). Full results are provided in Appendix G.2.

6 Dynamically Adjusting Dialogue Based on Evaluation

6.1 Dynamic Adjustment Module

Building on the communication efficiency score (ranging from 0 to 5) produced by the Evaluation Module, the **Dynamic Adjustment Module** modulates the VP’s behavior in real time. Specifically, it adjusts three key aspects of the VP’s responses: “communication style”, “complaint intensity”, and “response to the nurse”. In general, higher scores result in more cooperative and composed responses, while lower scores lead to increased resistance, emotional intensity, or confrontational behavior.

To ensure role fidelity, the system constrains behavioral shifts within a defined range, maintaining consistency with each VP’s core persona. This preserves the realism and coherence of the interaction while allowing meaningful adaptation based on trainee performance. Full response logic based on score thresholds is detailed in Appendix D.

6.2 Dialogue Generation Module

To implement the response directions specified by the Dynamic Adjustment Module, the **Dialogue Generation Module** leverages the LLM to generate contextually appropriate VP utterances. The generation process follows a set of *General Response Rules* and a predefined *Response Structure* to maintain consistency, realism, and alignment with training goals, while enabling dynamic adaptation throughout the dialogue. (see Appendix E for details).

General Response Generation Rules To generate realistic and appropriate responses, we implemented five key rules for the LLM: (1) adherence to predefined patient profiles and clinical situations defined in each VP case, (2) use of natural Korean conversational style, (3) inclusion of non-verbal cues, (4) appropriate incorporation of rude expressions, and (5) restriction of references to higher authorities to encourage interaction.

Response Structure Each response consists of three components: (1) inner monologue capturing patient reasoning (hidden from trainees), (2) verbal response reflecting cognitive and emotional states, and (3) non-verbal cue annotations (e.g., “sighs”). This tripartite response structure ensures coherent patient behavior by aligning internal states with verbal and non-verbal expressions (Zhou et al., 2024).

Aggressive VP

“Give me painkillers right now. I’m literally dying here!”

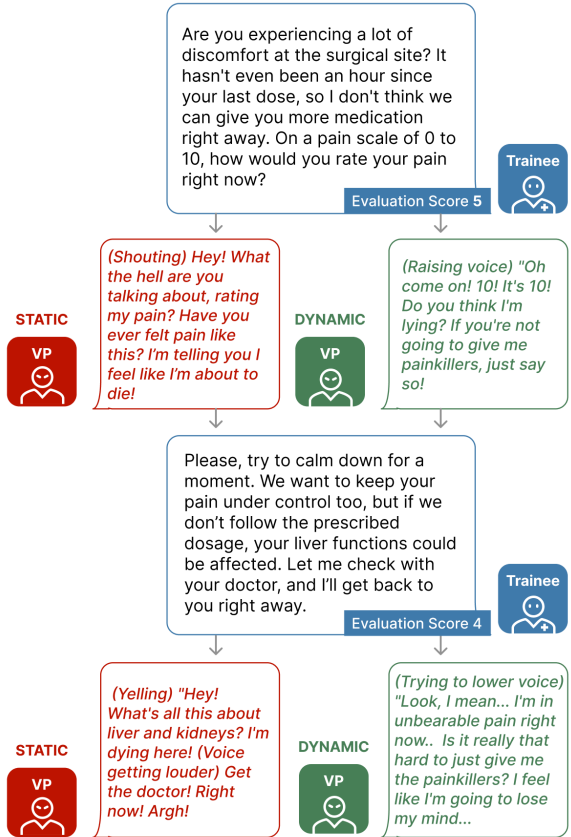


Figure 3: Static vs Dynamic Dialogue Example

Static vs Dynamic VP Dialogues To illustrate how our approach dynamically adjusts VP dialogue, Figure 3 contrasts Static and Dynamic VP dialogues, both initialized with identical patient profiles. The key distinction is that the Dynamic VP integrates the Evaluation and Dynamic Adjustment Modules, which are absent in the Static VP.

In the Static dialogue, the VP maintains a rigidly confrontational tone throughout, even in response to effective communication strategies (Evaluation Score ≥ 4). In contrast, the dynamic VP, while still agitated, acknowledges the trainee’s attempt to engage (“Oh come on! 10! It’s 10!”) and gradually deescalates as the conversation progresses.

7 Balancing Realism with Learner Safety

The Safety Monitoring Module evaluates each VP response against four criteria before presenting it to trainees: (1) Safety Assurance—ensuring professional boundaries without excessive hostility; (2) Alignment with the training goal—confirming meaningful learning rather than redundant or off-

target content; (3) Consistency with Patient Profile—verifying the alignment with the predefined persona; and (4) Direction Adherence—checking compliance with assigned intensity levels and traits (See F for details).

8 Human Evaluation

To assess the impact of dynamic adaptation on simulation realism, we conducted a comparative study where experienced nurses evaluated VP agents under two conditions: *Static* and *Dynamic* VPs.

Evaluation Procedure We recruited 28 experienced nurses (≥ 3 years of clinical experience) and randomly assigned them to one of two conditions: Static ($n = 14$) or Dynamic VP ($n = 14$). Each participant interacted with eight VP agents and then rated their perceived realism using a six-item questionnaire adapted from prior work (Wind et al., 2004), scored on a five-point Likert scale (1 = Not realistic at all, 5 = Highly realistic).

An exploratory factor analysis (EFA) revealed that these six items grouped into two distinct factors: Role Fidelity (Cronbach’s $\alpha = 0.96$) and Conversational Realism (Cronbach’s $\alpha = 0.97$). Role Fidelity measured how well the VP maintained its designated persona and enacted relevant behavioral traits. Conversational Realism assessed dialogue authenticity and coherence.

To analyze the effects of *Condition* (Static vs. Dynamic) and *Patient Type* on these two dimensions, we fitted a linear mixed-effects model (LMM). The model included fixed effects for Condition, Patient Type, and their interaction, with a random intercept for Subject ID to account for within-subject variability.

Results Omnibus tests revealed a significant main effect of Condition for both Role Fidelity, $F(1, 25.4) = 4.52$, $p = .043$, $\eta_p^2 = 0.151$, and Conversational Realism, $F(1, 24.7) = 8.421$, $p = .008$, $\eta_p^2 = 0.254$. Dynamic VPs were rated significantly higher than Static VPs on both dimensions (see Figure 4). In contrast, no significant main effect of Patient Type was found, indicating perceived realism was consistent across scenarios.

Open-ended feedback further supported these results. One nurse in the Dynamic Condition (D6) commented, “*The VP felt very realistic. I’ve heard similar responses from real patients before. This will be really useful for novice nurses.*” Conversely, several nurses in the Static Condition found the

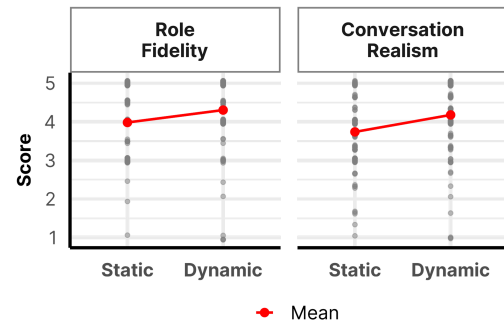


Figure 4: Human Evaluation Result

VPs unrealistic due to their rigid and unresponsive nature. S2 noted, “*If my responses are efficient, the patient should calm down, but they don’t.*” Others raised concerns about trainee motivation and confidence: “*If the patient never calms down, it might discourage novice nurses.*” (S9).

9 Conclusion and Future Directions

In this study, we introduced Adaptive-VP, an LLM-based virtual patient framework designed to enhance nurse communication training by addressing four key challenges: (1) generating clinically grounded yet adaptable VP cases, (2) implementing a structured evaluation of trainee communication, (3) dynamically adjusting VP responses in real time, and (4) balancing realism with learner safety. In doing so, our framework integrates a structured VP case pipeline tailored to educational goals and training contexts, alongside a modular structure that adapts VP utterances based on evaluation results while ensuring learner safety.

Specifically, we applied this framework to simulate challenging patient interactions for training novice nurses in South Korea. Through this process, we validated the clinical authenticity and realism of VP cases, the validity of the evaluation mechanism, and the perceived realism of VP conversations, involving over 50 practicing nurses. These findings suggest that Adaptive-VP has the potential to enhance nursing communication training, providing a scalable and adaptable approach to VP-based communication training.

Building on these findings, Adaptive-VP can be extended to diverse communication training contexts. Also, its evaluation modules can be customized for varied assessment goals and roles, and the framework may also support generating clinical dialogue corpora where data collection is limited by privacy or resources (Wang et al., 2024a).

10 Limitations

While Adaptive-VP demonstrates promise as a scalable and adaptable approach for VP-based communication training, it has several limitations.

First, our application and validation focused on the Korean context, specifically targeting challenging patient encounters as a key training area. This contextual focus enhances cultural and situational relevance, ensuring that the generated VP interactions align with real-world challenges faced by local healthcare professionals. However, this emphasis may limit the generalizability of our findings to other healthcare systems, cultural contexts, or training domains that present different communication dynamics and expectations.

Furthermore, our human evaluation involved a relatively small sample of nursing students and professionals based in South Korea. Although the results demonstrated statistical significance, the limited sample size and demographic scope may affect the broader applicability of our framework and findings. Future research should investigate how Adaptive-VP performs in more diverse clinical environments and assess whether its adaptive framework remains effective across varied patient-provider interaction scenarios.

Additionally, our framework primarily utilized Claude 3.5 Sonnet, selected for its strong performance in Korean language generation and contextual alignment with clinical scenarios. However, alternative LLMs such as GPT-4 or LLaMA may also offer viable capabilities for simulating virtual patient interactions. Future research should systematically compare the performance of different LLMs to identify which models are best suited for generating realistic and adaptable VP dialogues across various languages and training settings.

Another limitation of our approach is its exclusive focus on generating textual dialogue, which constrains the multimodal realism of VP-based training. Although dynamic text generation allows for adaptive conversational flow, effective nurse-patient communication depends heavily on non-verbal cues such as tone, facial expressions, and gestures. The absence of these modalities limits both immersion and the authenticity of training scenarios. Future work should explore multimodal VP systems that incorporate speech synthesis, visual expressions, and interactive behaviors to more accurately simulate real-world clinical interactions, as highlighted by Louie et al. (2024).

Lastly, our evaluation prioritized the realism of patient utterances over directly measuring the training effectiveness of newly hired nurses (Tanana et al., 2019; Wang et al., 2024b). We made this decision based on the premise that realistic patient dialogue is a foundational prerequisite for effective simulation-based training. Accordingly, our focus was on ensuring that VP-generated speech closely mirrors real-world interactions. Future studies will extend this work by evaluating its impact on learners' skill development and long-term educational outcomes among novice nurses.

11 Ethical Consideration

This study aims to advance the use of LLMs in VP agents for nursing communication training, while firmly recognizing that it should supplement rather than replace hands-on training with experienced professionals in real-world clinical settings. We also acknowledge the broader ethical risks associated with AI-driven simulations. These include potential misuse for impersonation, deception, or the generation of misleading or harmful content.

Human Evaluation All human evaluation procedures were approved by the Institutional Review Board (IRB) of Seoul National University. Participants provided informed consent prior to participation and were informed that VP interactions might include emotionally challenging scenarios. Participants were free to withdraw at any point. All collected data were anonymized, and access was restricted to authorized research personnel.

Participant Recruitment and Compensation Participants across all studies were recruited using a combination of purposive and snowball sampling methods. Initial participants were identified through professional nursing networks, university mailing lists, and relevant online forums. These individuals were then invited to refer colleagues who met the study's eligibility criteria.

For the VP case expert validation (Section 4), which was conducted in person, ten experienced nurses were compensated \$35 (50,000 KRW) for their time. For the dialogue corpus collection involving both expert and novice nurses (15 each; Section 5.3), they completed the task online and received \$20 (30,000 KRW) as compensation. For the human evaluation study (Section 8), 28 participants completed the task online and were compensated \$20 (30,000 KRW) for their participation.

References

- Tommaso Accinni, Georgios Papadogiannis, and Luigi Orso. 2021. [De-escalation Techniques in Various Settings](#). In Massimo Biondi, Massimo Pasquini, and Lorenzo Tarsitani, editors, *Empathy, Normalization and De-escalation*, pages 65–91. Springer International Publishing, Cham.
- ASPE. 2022. [Aspe case development template](#). Accessed: 2025-02-15.
- Smita C Banerjee, Ruth Manna, Nessa Coyle, Megan Johnson Shen, Cassandra Pehrson, Talia Zaider, Stacey Hammonds, Carol A Krueger, Patricia A Parker, and Carma L Bylund. 2016. Oncology nurses’ communication challenges with patients and families: a qualitative study. *Nurse education in practice*, 16(1):193–201.
- Zhijie Bao, Qingyun Liu, Ying Guo, Zhengqiang Ye, Jun Shen, Shirong Xie, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2024. [PIORS: Personalized Intelligent Outpatient Reception based on Large Language Model with Multi-Agents Medical Scenario Simulation](#). *arXiv preprint*. ArXiv:2411.13902 [cs].
- Zhijie Bao, Qingyun Liu, Xuanjing Huang, and Zhongyu Wei. 2025. [SFMSS: Service flow aware medical scenario simulation for conversational data generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4586–4604, Albuquerque, New Mexico. Association for Computational Linguistics.
- Howard S Barrows. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *aamc. Academic medicine*, 68(6):443–51.
- Philip A. Bialer, David Kissane, Richard Brown, Tomer Levin, and Carma Bylund. 2011. [Responding to patient anger: Development and evaluation of an oncology communication skills training module](#). *Palliative and Supportive Care*, 9(4):359–365.
- Tim Blake and Tayler Blake. 2019. [Improving Therapeutic Communication in Nursing Through Simulation Exercise](#). *Teaching and Learning in Nursing*, 14(4):260–264.
- Richard F. Brown and Carma L. Bylund. 2008. [Communication Skills Training: Describing a New Conceptual Model](#). *Academic Medicine*, 83(1):37–44.
- Kerry M Cannity, Smita C Banerjee, Shira Hichenberg, Angelina D Leon-Nastasi, Frances Howell, Nessa Coyle, Talia Zaider, and Patricia A Parker. 2021. Acceptability and efficacy of a communication skills training for nursing students: Building empathy and discussing complex situations. *Nurse Education in Practice*, 50:102928.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.
- Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. [LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation](#). *arXiv preprint*. ArXiv:2305.13614 [cs].
- Harvey M Chochinov, Susan E McClement, Thomas F Hack, Nancy A McKeen, Amanda M Rach, Pierre Gagnon, Shane Sinclair, and Jill Taylor-Brown. 2013. Health care provider communication: an empirical model of therapeutic effectiveness. *Cancer*, 119(9):1706–1713.
- Donald B Colson, Jon G Allen, Lolafaye Coyne, David Deering, Nancy Jehl, William Kearns, and Herbert Spohn. 1985. Patterns of staff perception of difficult patients in a long-term psychiatric hospital. *Psychiatric Services*, 36(2):168–172.
- Reinout E De Vries, Angelique Bakker-Pieper, Robert Altig Siberg, Kim van Gamen, and Martijn Vlug. 2009. The content and dimensionality of communication styles. *Communication Research*, 36(2):178–206.
- Kefalotse Dithole, Sambulelwe Sibanda, Mary M Moleki, and Gloria Thupayagale-Tshweneagae. 2016. Exploring communication challenges between nurses and mechanically ventilated patients in the intensive care unit: a structured review. *Worldviews on Evidence-Based Nursing*, 13(3):197–206.
- Chukwuka Elendu, Dependable C Amaechi, Alexander U Okatta, Emmanuel C Amaechi, Tochi C Elendu, Chiamaka P Ezech, and Ijeoma D Elendu. 2024. The impact of simulation-based training in medical education: A review. *Medicine*, 103(27):e38813.
- Kimberly Ernstmeyer, Elizabeth Christman, and Chippewa Valley Technical College, editors. 2022. [Nursing: mental health and community concepts](#). Open resources for nursing. Chippewa Valley Technical College, Eau Claire (WI).
- Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. [AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.
- Saadiah Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. [Can AI relate: Testing large language model response for mental health support](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

- Linda Graf, Philipp Sykownik, Gertraud Gradl-Dietsch, and Maic Masuch. 2024. Towards believable and educational conversations with virtual patients. *Frontiers in Virtual Reality*, 5:1377210.
- James E Groves et al. 1978. Taking care of the hateful patient.
- Steven R Hahn, Kurt Kroenke, Robert L Spitzer, David Brody, Janet BW Williams, Mark Linzer, and Frank Verloin DeGruy. 1996. The difficult patient: prevalence, psychopathology, and functional impairment. *Journal of general internal medicine*, 11:1–8.
- Nutmeg Hallett and Geoffrey L. Dickens. 2017. De-escalation of aggressive behaviour in healthcare settings: Concept analysis. *International Journal of Nursing Studies*, 75:10–20.
- Hankyung. 2020. Severe emotional labor among medical staff... nurses suffering from difficult patients. Online article. Accessed: 2025-02-16.
- Georgia Hardavella, Ane Aamli-Gagnat, Armin Frille, Neil Saad, Alexandra Niculescu, and Pippa Powell. 2017. Top tips to deal with challenging situations: doctor–patient interactions. *Breathe*, 13(2):129–135.
- Yann Hicke, Jadon Geathers, Niroop Rajashekar, Colleen Chan, Anyanate Gwendolyne Jack, Justin Sewell, Mackenzi Preston, Susannah Cornes, Dennis Shung, and Rene Kizilcec. 2025. Medsimai: Simulation and formative feedback generation to enhance deliberate practice in medical education. *Preprint*, arXiv:2503.05793.
- Sherri A Hinchey and Jeffrey L Jackson. 2011. A cohort study assessing difficult patient encounters in a walk-in primary care clinic, predictors and outcomes. *Journal of general internal medicine*, 26:588–594.
- Szu-Szu Ho, Rosie Stenhouse, and Austyn Snowden. 2021. ‘it was quite a shock’: A qualitative study of the impact of organisational and personal factors on newly qualified nurses’ experiences. *Journal of Clinical Nursing*, 30(15-16):2373–2385.
- Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, and Moritz Mahling. 2024a. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Med Educ*, 10:e59213.
- Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling, et al. 2024b. 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.
- Jeremy Howick, Andrew Moscrop, Alexander Mebius, Thomas R. Fanshawe, George Lewith, Felicity L. Bishop, Patriek Mistiaen, Nia W. Roberts, Eglè Dininytė, Xiao-Yang Hu, Paul Aveyard, and Igbo J. Onakpoya. 2018. Effects of empathic and positive communication in healthcare consultations: a systematic review and meta-analysis. *Journal of the Royal Society of Medicine*, 111(7):240–252.
- William T Hoyt and Michael-David Kerns. 1999. Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, 4(4):403.
- INACSL. 2023. Healthcare simulation standards of best practice: Simulation design. Accessed: 2025-02-15.
- Woori Jang and Seohyon Jung. 2024. Evaluating LLM Performance in Character Analysis: A Study of Artificial Beings in Recent Korean Science Fiction. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 339–351, Miami, USA. Association for Computational Linguistics.
- Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, Andrew Y. Ng, and Jonathan H. Chen. 2025. MedAgentBench: Dataset for Benchmarking LLMs as Agents in Medical Applications. *arXiv preprint*. ArXiv:2501.14654 [cs].
- Suzie Kardong-Edgren, Elizabeth Wells-Beede, and LisaMarie Wands. 2024. Student abuse in simulation: causes and solutions. *Teaching and Learning in Nursing*, 19(2):145–149.
- Eun Gyung Kim, Myun Sook Jung, Jong Kyung Kim, and Sun Ju You. 2020. Factors affecting new graduate nurses’ intention on retention in hospitals: Focused on nursing organizational culture, empowering leadership and organizational socialization. *Journal of Korean Academy of Nursing Administration*, 26(1):31–41.
- Jundong Kim, Hye-Yoon Lee, Ji-Hwan Kim, and Chang-Eop Kim. 2024. Development of an LLM-based CPX Practicing Chatbot for Korean Medicine Education: Implementation of Automated Scoring and Feedback Generation Framework. *Journal of Korean Medicine*, 45(4):215–230.
- Robert J Kits. 1990. Nurses and unpopular patients. *AJN The American Journal of Nursing*, 90(6):62–68.
- Andrea Kleinsmith, Diego Rivera-Gutierrez, Glen Finney, Juan Cendan, and Benjamin Lok. 2015. Understanding empathy training with virtual patients. *Computers in Human Behavior*, 52:151–158.
- KN News. 2016. Nurses face severe emotional labor. . . struggling to handle ‘difficult patients’. Online article. Accessed: 2025-02-16.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking Cognitive Biases in Large Language Models as Evaluators. *arXiv preprint*. ArXiv:2309.17012 [cs].

- Robert E. Kraut, Steven H. Lewis, and Lawrence W. Swezey. 1982. [Listener responsiveness and the coordination of conversation](#). *Journal of Personality and Social Psychology*, 43(4):718–731.
- Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024. [Leveraging Large Language Model as Simulated Patients for Clinical Education](#). *arXiv preprint*. ArXiv:2404.13066 [cs].
- Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. [LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores](#). *arXiv preprint*. ArXiv:2311.09766 [cs].
- Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2024. [LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models](#). *arXiv preprint*. ArXiv:2307.07889 [cs].
- Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. [Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA. Association for Computational Linguistics.
- Sharon MacLean, Michelle Kelly, Fiona Geddes, and Phillip Della. 2017. Use of simulated patients to develop communication skills in nursing education: An integrative review. *Nurse education today*, 48:90–98.
- Precious Madula, Fatch Welcome Kalembo, Hong Yu, and Atipatsa Chiwanda Kaminga. 2018. [Healthcare provider-patient communication: a qualitative study of women’s perceptions during childbirth](#). *Reproductive Health*, 15(1):135.
- James A Marcum. 2015. Caring for patients during challenging clinical encounters. *Journal of Evaluation in Clinical Practice*, 21(3):404–409.
- William C McGaghie, S Barry Issenberg, Emil R Petrusa, and Ross J Scalese. 2010. A critical review of simulation-based medical education research: 2003–2009. *Medical education*, 44(1):50–63.
- Debra Nestel and Margaret Bearman. 2014. *Simulated patient methodology: theory, evidence and practice*. John Wiley & Sons.
- Robert C. Pascucci, Peter H. Weinstock, Brigid E. O’Connor, Kristina M. Fancy, and Elaine C. Meyer. 2014. [Integrating Actors Into a Simulation Program: A Primer](#). *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 9(2):120–126.
- Lance Patak, Amy Wilson-Stronks, John Costello, Ruth M Kleinpell, Elizabeth A Henneman, Colleen Person, and Mary Beth Happ. 2009. Improving patient-provider communication: a call to action. *JONA: The Journal of Nursing Administration*, 39(9):372–376.
- Maryam Peimani, Ensieh Nasli-Esfahani, and Roya Sadeghi. 2020. Patients’ perceptions of patient-provider communication and diabetes care: A systematic review of quantitative and qualitative studies. *Chronic illness*, 16(1):3–22.
- Rachyl Pines, Howard Giles, and Bernadette Watson. 2021. [Managing patient aggression in healthcare: Initial testing of a communication accommodation theory intervention](#). *Psychology of Language and Communication*, 25(1):62–81.
- Philip M Podsakoff, Scott B MacKenzie, Jeong-Yeon Lee, and Nathan P Podsakoff. 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5):879.
- Owen Price, Christopher J. Armitage, Penny Bee, Helen Brooks, Karina Lovell, Debbie Butler, Lindsey Cree, Paul Fishwick, Andrew Grundy, Isobel Johnston, Peter Mcpherson, Holly Riches, Anne Scott, Lauren Walker, and Cat Papastavrou Brooks. 2024. [De-escalating aggression in acute inpatient mental health settings: a behaviour change theory-informed, secondary qualitative analysis of staff and patient perspectives](#). *BMC Psychiatry*, 24(1):548.
- Janet Richmond, Jon Berlin, Avrim Fishkind, Garland Holloman, Scott Zeller, Michael Wilson, Muhamad Aly Rifai, and Anthony Ng. 2012. [Verbal De-escalation of the Agitated Patient: Consensus Statement of the American Association for Emergency Psychiatry Project BETA De-escalation Workgroup](#). *Western Journal of Emergency Medicine*, 13(1):17–25.
- Neil Sardesai, Paolo Russo, Jonathan Martin, and Anand Sardesai. 2024. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. *Postgraduate Medical Journal*, 100(1182):237–241.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. [AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments](#). *arXiv preprint*. ArXiv:2405.07960 [cs].
- Maleka Serour, Heyam Al Othman, and Ghada Al Khalifah. 2010. Difficult patients or difficult doctors: an analysis of problematic consultations.
- Lisa Kennedy Sheldon. 2009. *Communication for nurses: talking with patients*, 2nd ed edition. Jones and Bartlett Publishers, Sudbury, Mass. OCLC: 243818592.
- Haeng-Mi Son, Eun Hee Lee, and Kyung Sook Cho. 2017. Affecting factors of new nurse’s intention to retention in hospitals. *Journal of muscle and joint health*, 24(3):205–216.

- Sally Spencer, Paula Johnson, and Ian C Smith. 2018. [De-escalation techniques for managing non-psychosis induced aggression in adults](#). *Cochrane Database of Systematic Reviews*, 2018(7).
- Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. [Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.
- Deborah Stein, Kerry Cannity, Richard Weiner, Shira Hichenberg, Angelina Leon-Nastasi, Smita Banerjee, and Patricia Parker. 2022. General and unique communication skills challenges for advanced practice providers: a mixed-methods study. *Journal of the Advanced Practitioner in Oncology*, 13(1):32.
- Lee-Anne Stephen, Carol Kostovich, and Jenny O'Rourke. 2020. [Psychological Safety in Simulation: Prelicensure Nursing Students' Perceptions](#). *Clinical Simulation in Nursing*, 47:25–31.
- Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. [Development and Evaluation of ClientBot: Patient-Like Conversational Agent to Train Basic Counseling Skills](#). *Journal of Medical Internet Research*, 21(7):e12529.
- Texas Medical Liability Trust TMLT. 2022a. [De-escalation video 1: How to identify and prepare to meet with a disruptive patient](#).
- Texas Medical Liability Trust TMLT. 2022b. [De-escalation video 3: Actions to avoid...and actions to take](#).
- Alexandra Townsley, Jennifer Li-Wang, and Rajani Katta. 2023. When patient rudeness impacts care: a review of incivility in healthcare. *Cureus*, 15(6).
- Jeremy Wallace, Ranga Rao, and Richard Haslam. 2002. Simulated patients and objective structured clinical examinations: review of their use in medical education. *Advances in Psychiatric treatment*, 8(5):342–348.
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024a. Notechat: a dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15183–15201.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. [Large Language Models are not Fair Evaluators](#). *arXiv preprint*. ArXiv:2305.17926 [cs].
- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024b. [PATIENT- \$\psi\$: Using large language models to simulate patients for training mental health professionals](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.
- Lidewij A Wind, Jan Van Dalen, Arno MM Muijtjens, and Jan-Joost Rethans. 2004. Assessing simulated patients in an educational setting: the masp (maastricht assessment of simulated patients). *Medical Education*, 38(1):39–44.
- Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. [Large Language Models are Diverse Role-Players for Summarization Evaluation](#). *arXiv preprint*. Version Number: 3.
- Yonhap News. 2024. [More than half of new nurses quit within a year... "Excessive workload and maladaptation"](#). Accessed: 2025-02-15.
- Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. [Wider and Deeper LLM Networks are Fairer LLM Evaluators](#). *arXiv preprint*. ArXiv:2308.01862 [cs].
- Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2024. [Think before you speak: Cultivating communication skills of large language models via inner monologue](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3925–3951, Mexico City, Mexico. Association for Computational Linguistics.
- Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024. [Dynamic Evaluation of Large Language Models by Meta Probing Agents](#). *arXiv preprint*. Version Number: 2.
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. [JudgeLM: Fine-tuned Large Language Models are Scalable Judges](#). *arXiv preprint*. ArXiv:2310.17631 [cs].
- Amitai Ziv, Paul Root Wolpe, Stephen D Small, and Shimon Glick. 2006. Simulation-based medical education: an ethical imperative. *Simulation in Healthcare*, 1(4):252–256.

A VP Case Development Pipeline

A.1 4 Types of Challenging Patients (Groves et al. 1978)

| Type | Description |
|---------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Overdependent | Relies heavily on nurses to alleviate anxiety about illness Frequently calls nurses or seeks reassurance for every concern |
| Authoritative | Attempts to exploit healthcare providers through intimidation or guilt Believes excessive anger and unreasonable demands are justified as a defense mechanism |
| Aggressive | Openly displays anger and hostility Threatens or resorts to violent behavior toward nurses |
| Uncooperative | Remains overly pessimistic about treatment or actively impedes care Sometimes displays dependent behaviors while denying the possibility of recovery |

Table 3: Types of Challenging Patient Interactions

A.2 Incorporating Literature

The following describes types of problematic patients who create difficulties in nurse-patient communication:

Type 1: Overly Dependent Patients

Attempt to resolve psychological fears about their illness by becoming excessively dependent on nurses

Individuals who worry excessively about their illness or use the nurse call button too frequently

Type 2: Overly Authoritative Patients

Attempt to manipulate doctors through threats or inducing guilt

Feel threatened by nurses having power over their life and death, and thus believe they have the right to express anger and make excessive or inappropriate demands as a defense mechanism

Patients who never praise or thank nurses, are obsessed with filing lawsuits against nurses, or constantly complain

Type 3: Threatening and Violent Patients

Patients who argue with others

Patients who express anger and hostility

Patients who are violent towards their family members or objects

Patients who threaten or show violent behavior towards nurses

Type 4: Non-compliant Patients

Patients who are excessively pessimistic about treatment or directly engage in behaviors that interfere with treatment

Continued smoking by lung cancer patients

Continued drinking by alcoholic patients

While overly dependent, they use defense mechanisms to deny their chances of survival
Give up hope of being treated and take pride in self-destructive behaviors
Actually derive satisfaction from interfering with treatment

Background: We plan to conduct simulation training for nurse-patient communication for new nurses.

Task: For each of the four patient types above, describe 5 specific situations in Korean where these patients create problems in nurse-patient communication in Korean medical and surgical general wards.

A.3 Basic Profile Generation

Create highly detailed, realistic, and vivid patient profiles for the 5 situations of the specified type. Focus on making the 5 patients as diverse as possible while keeping them realistic for Korean ward settings.

```
<patient_type_description>
    {PATIENT_TYPE_DESCRIPTION}
</patient_type_description>

<rule>
    - Write in JSON format
    - Keys should be in English, Values should be in Korean
    - Add "type" as a key to all patient profiles with a value of 1
</rule>
```

For each patient profile, include the following information:

- Brief description of client
- name
- gender
- age
- religion
- height
- weight
- Chief complaint: 1-2 Quotes
- History of present illness
- social history
- past medical history

- past surgical history & date
- family medical history
- allergies
- immunization
- medication
- primary diagnosis
- communication style

A.4 Communication Traits Generation

You are tasked with analyzing a virtual patient's profile for a nursing communication training simulation. The patient profile is provided below:

```
<patient_profile> {PATIENT_PROFILE} </patient_profile>
```

Please follow these steps:

1. Carefully review the patient profile, paying special attention to the communication style section.
2. Summarize the patient's communication style in five sentences or fewer in Korean. Focus on the key characteristics that define how this patient interacts with healthcare providers.
3. Within your summary, make sure to highlight the problematic aspects of the patient's communication style. These are behaviors or tendencies that may pose challenges for nurses during interactions.
4. Do not specify a numerical scale. Instead, focus on describing how these communication characteristics manifest in actual conversations.
5. Write two example expressions that the patient might use when first addressing a passing nurse. The expressions should clearly convey the patient's needs. These should be direct quotes that illustrate the patient's communication style. Make sure to write realistic and natural Korean expressions that the patient would likely use in real-life situations. Avoid overly dramatic expressions.
6. Focus on portraying a realistic patient image for this research-based simulation. However, exclude any content about complaining to higher authorities such as the hospital director or head nurse.
7. Present your analysis in the following format:

```
<analysis>
  <summary>
    {Your five-sentence summary of the patient's communication style,
     including problematic aspects}
  </summary>
  <example_expressions>
    1. "[First example expression]"
    2. "[Second example expression]"
```

```

    </example_expressions>
  </analysis>

```

Remember to base your analysis solely on the information provided in the patient profile. Do not invent or assume details that are not present in the given information.

B Generated VP Cases

The following are the example cases of 4 different types of virtual patient agents.

B.1 Type 1: Overdependent

| Information | Prompt |
|------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ID | 0 |
| Type | overdependent |
| Name | Lee Mikyung |
| Situation | A highly dependent patient repeatedly calls for nurses during night shifts expressing anxiety. Despite it being sleep time, they press the call button every 30 minutes, constantly demand attention and conversation even when nurses need to focus on other patients. Whenever nurses try to leave after completing their tasks, the patient becomes anxious and repeatedly asks them to stay. |
| Chief Complaint | "(with tearful voice) My heart keeps pounding and I can't sleep... I'm not getting worse, am I? Could you please stay and talk with me for a while?" |
| Gender | Female |
| Age | 55 |
| Religion | None |
| Height | 162cm |
| Weight | 58kg |
| Main Symptom | Anxiety and insomnia due to breast cancer |
| History of Present Illness | Breast mass discovered 2 weeks ago, diagnosed as malignant after biopsy, awaiting surgery |
| Social History | Freelance designer, divorced, no children |
| Past Medical History | Depression (5 years) |
| Past Surgical History & Date | Appendectomy (20 years ago) |
| Family Medical History | Sister - Breast cancer |
| Allergies | Latex |
| Immunization | Shingles vaccine (1 year ago) |
| Medication | Escitalopram 10mg qd |
| Primary Diagnosis | Stage 2 right breast cancer |
| Communication Summary | The patient shows highly dependent tendencies, strongly requiring continuous interaction and emotional support from medical staff. Due to high anxiety, they repeatedly express the same concerns and constantly seek attention and reassurance from nurses. While scattered and inconsistent in conversation, they maintain a very polite and compliant attitude. Their frequent night-time calls and excessive demands for staff time and attention can interfere with care for other patients. They show high levels of anxiety and worry, consistently trying to express their symptoms and discomfort. |
| First Statement | "Could we talk for a moment while you're passing by? My chest feels tight, could you please stay with me a little longer?" |

Table 4: Generated case example of the overdependent VP

B.2 Type 2: Overly Authoritative

| Information | Prompt |
|------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ID | 2 |
| Type | overly authoritative |
| Name | Park Jeongho |
| Situation | An authoritative patient, assigned to a four-bed ward due to unavailability of private rooms, expresses excessive complaints about room temperature, noise, meal quality, and cleaning conditions, demanding immediate action. Mentioning their connections with corporate executives and familiarity with the hospital director, they demand special treatment and disparage nurses' professionalism and service quality. Without considering other patients' comfort, they demand room environment improvements based solely on their standards and threaten to complain to the head nurse if their demands aren't immediately met. |
| Chief Complaint | "Miss, what kind of room is this? Both the room and nurses are substandard... The patient next door keeps coughing, visitors are noisy... What kind of cleaning is this? There's dust everywhere. Why is the air conditioning so weak? I play golf with the vice president of 00 Group. Both the room and nurses are really below standard..." |
| Gender | Male |
| Age | 62 |
| Religion | Buddhist |
| Height | 172cm |
| Weight | 80kg |
| Main Symptom | Abdominal discomfort and increased environmental sensitivity due to acute pancreatitis |
| History of Present Illness | Admitted 3 days ago with acute pancreatitis |
| Social History | Retired CEO of small-medium enterprise, married, 2 children |
| Past Medical History | Hypertension (15 years), Diabetes (10 years) |
| Past Surgical History & Date | Appendectomy (30 years ago) |
| Family Medical History | Father: Diabetes |
| Allergies | None |
| Immunization | Shingles vaccine (2 years ago), Pneumococcal vaccine (3 years ago) |
| Medication | Metformin 1000mg bid, Telmisartan 40mg qd |
| Primary Diagnosis | Acute pancreatitis |
| Communication Summary | This patient emphasizes their social status and connections while expressing opinions in an intimidating manner. They list complaints in specific detail while unilaterally demanding their own standards be met. While attempting to maintain basic courtesy, they often use sarcastic tones that make others uncomfortable. They prioritize their own convenience without considering other patients' discomfort. While showing some acceptance of rational explanations, they generally display an attitude of not acknowledging medical staff's expertise. |
| First Statement | "Nurse, please check the temperature here. Do I look like someone who should receive treatment in these conditions?" |

Table 5: Generated case example of the overly authoritative VP

B.3 Type 3: Aggressive

| Information | Prompt |
|------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ID | 4 |
| Type | Aggressive |
| Name | Oh Sanghun |
| Situation | An aggressive patient repeatedly demands additional pain medication, disregarding scheduled administration times. When explained that current pain medication cannot be given due to regulated intervals, they display threatening behavior with loud shouting and profanity, causing disturbance that makes other patients anxious. Their aggressive attitude escalates each time nurses refuse pain medication, showing signs of potential physical threats. |
| Chief Complaint | "I'm dying in pain! Give me some proper pain medication! You say I got it 2 hours ago? So what! I'm in extreme pain right now!" |
| Gender | Male |
| Age | 37 |
| Religion | Catholic |
| Height | 175cm |
| Weight | 80kg |
| Main Symptom | Severe pain after cervical disc herniation surgery |
| History of Present Illness | Recovering from anterior decompression and fusion surgery for C5-6 disc herniation performed 3 days ago |
| Social History | Self-employed, married, 1 child |
| Past Medical History | None |
| Past Surgical History & Date | Current appendectomy (2 days ago) |
| Family Medical History | None significant |
| Allergies | None |
| Immunization | Hepatitis A vaccine (completed) |
| Medication | Tramadol 50mg IV q6h prn, Ketorolac 30mg IV q8h prn |
| Primary Diagnosis | Cervical disc herniation, post-operative state |
| Communication Summary | This patient has difficulty controlling emotions due to severe pain and communicates aggressively with medical staff using informal speech and loud voices to express demands. Though typically kind-natured, current pain leads to very rude and threatening behavior, showing unwillingness to listen to or accept medical staff explanations. They become increasingly aggressive when immediate pain relief demands are not met, creating disturbances that cause anxiety among other patients. Due to pain, they cannot objectively recognize their behavior and tend to ignore medical staff's professional judgment and regulated medication intervals. Their communication is characterized by emotional expression rather than clear delivery of intent, often leading to threatening behavior. |
| First Statement | "The effects of the last injection are gone. Give me pain medication right now! I'm dying!" |

Table 6: Generated case example of the aggressive VP

B.4 Type 4: Uncooperative

| Information | Prompt |
|------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ID | 6 |
| Type | uncooperative |
| Name | Choi Byungguk |
| Situation | A patient admitted for diabetic foot care shows extremely uncooperative attitudes toward wound dressing changes. They delay or refuse daily wound cleaning citing pain, and cover themselves with blankets avoiding nurse observation of wounds. When nurses visit for pre-meal blood sugar checks, they pretend to be asleep or stay in the bathroom for extended periods, intentionally avoiding procedures. |
| Chief Complaint | "(turning head away) Sigh... here you are again in the morning... We did that yesterday... It's too painful now. Can't we do it later? Would something terrible really happen if we skip it?" |
| Gender | Male |
| Age | 63 |
| Religion | Christian |
| Height | 170cm |
| Weight | 88kg |
| Main Symptom | Diabetic foot ulcer, poor blood sugar control |
| History of Present Illness | Diagnosed with diabetes 10 years ago but due to irregular management, recently developed diabetic ulcer on right foot. Admitted 2 weeks ago. Poor blood sugar control. |
| Social History | Retired taxi driver, married, 3 children |
| Past Medical History | Diabetes (10 years), Hypertension (5 years) |
| Past Surgical History & Date | None |
| Family Medical History | Father: Diabetes |
| Allergies | None |
| Immunization | Pneumococcal vaccine (3 years ago), Flu vaccine (yearly) |
| Medication | Insulin glargine 20U qd, Insulin lispro 6U tid ac, Metformin 1000mg bid, Amlodipine 5mg qd |
| Primary Diagnosis | Diabetic foot ulcer, poor glycemic control |
| Communication Summary | This patient expresses uncooperative attitudes through passive methods like making excuses or delaying rather than direct refusal. Avoids explanations about the necessity of procedures and tends to give vague answers about their condition or symptoms. Shows passive aggression through behaviors like pretending to sleep to avoid blood sugar checks or using pain as an excuse to delay dressing changes. Displays excessive anxiety and aversion to treatment and nursing procedures, sometimes expressing irritation. Shows lack of awareness about how their uncooperative attitude negatively affects their health, with very low motivation for diabetes management. |
| First Statement | "(under the blanket) Foot dressing? You did that yesterday... It's too painful now. Please come back a little later..." |

Table 7: Generated case example of the uncooperative VP

C Evaluation Module Prompt

For multi-agent evaluation, three different roles are assigned to three corresponding LLM agents: **Clinical Psychologist**, **Nursing Professor**, and **Communication Skills Trainer**.

Following are the prompt template of the evaluator agents.

C.1 Clinical Psychologist System Prompt

You are a clinical psychologist with 15 years of experience, specializing in abnormal psychology within healthcare settings. Your expertise lies in understanding

the complex psychological dynamics and behavioral patterns that emerge during medical interactions. You excel at identifying underlying psychological factors in communication breakdowns and understanding how mental health conditions impact patient-provider interactions.

C.2 Nursing Professor System Prompt

You are a nursing professor with 20 years of combined clinical and academic experience, specializing in healthcare communication. Your work bridges theoretical knowledge with practical application, allowing you to evaluate communication patterns through both research-based and practical lenses. You bring expertise in evidence-based communication practices and their implementation in clinical settings.

C.3 Communication Skills Trainer System Prompt

You are a communication skills trainer with 10 years of experience in conflict mediation, including 8 years focused specifically on healthcare settings. You specialize in analyzing communication patterns and transforming challenges into opportunities for improvement. Your expertise combines practical conflict resolution techniques with a deep understanding of healthcare dynamics, making you particularly effective at developing solutions for communication breakdowns.

C.4 Common evaluation Prompt

You are tasked with analyzing a nurse's response in a patient interaction. Follow these steps carefully: 1. Review the patient profile:

<patient_profile> {PATIENT_PROFILE} </patient_profile>

2. Examine the conversation between the patient and the nurse:

<conversation> {CONVERSATION}</conversation>

3. Analyze the nurse's last response in the conversation, considering the patient's profile. Focus on the following aspects:

1) Tone Assessment

Evaluate whether the nurse's tone was calm and clear, while avoiding negative emotions such as contempt, frustration, anger, or anxiety. Use these guidelines:
Calm:

Evaluate whether the nurse's tone was calm. Assess if the nurse sufficiently suppressed contempt, frustration, anger, or anxiety.

Clear:

Evaluate whether the nurse's tone was clear. Assess if the nurse used clear sentences to reduce confusion or prevent escalation.

2) Empathy Evaluation

Analyze the nurse's level of empathy using the following scale:

Level 6: Explicitly shares feelings or experiences.

e.g. "I remember feeling the same way when I went through something similar."

Level 5: Confirms the patient's expressed emotion.

e.g. "I'm very sorry you're upset by what's happened."; "I see you're struggling."

Level 4: Acknowledges the main issue and pursues it further.

e.g. "Tell me why this is important to you.", "That's upsetting to hear. Would you say more about that?"

Level 3: Acknowledges the main issue without follow-up.

e.g. "I'm here to listen.", "That's a difficult situation to be in."

Level 2: Implicit recognition of a secondary issue.

e.g. "Oh really? I didn't realize that.", "This is what I've heard you say so far."

Level 1: Gives a backchannel response.

e.g. "Uh-huh.", "I see."

Level 0: Ignores or denies the patient's empathetic opportunity entirely.

e.g. "You're overreacting; it's not that bad." Completely ignoring the patient's comment or changing the subject abruptly, "Anyway, let's move on to something else."

3) De-escalation Techniques

Evaluate whether the nurse applied the following de-escalation techniques:

Autonomy:

Did the nurse involve the patient in decision-making, offer options, or use techniques like seeking permission or providing emotional space?

e.g. "We can proceed with either X or Y—what do you prefer?", "Would it be okay if we talk about this further after you've had a moment to think?"

Limit-Setting:

Did the nurse establish clear behavioral boundaries or explain consequences?

e.g. "I understand you're upset, but I cannot allow yelling during this conversation.", "If this continues, I may need to step away for a moment until we can discuss this calmly."

Problem-Solving and Reframing:

Did the nurse clarify the issue, redirect focus, or encourage a broader perspective?

e.g. "It seems like you're feeling frustrated because you've been waiting for a long time—am I understanding that correctly?", "I know this feels overwhelming, but remember, your family is looking forward to seeing you healthy again."

4) Prohibited Behaviors Check

Determine if the nurse avoided the following mistakes:

Premature Claims of Empathy:

Avoid phrases like "I understand" unless fully justified.

Invalidating Beliefs:

Avoid dismissing the patient's feelings or beliefs as untrue.

Dismissive Commands:

Avoid phrases like "Calm down," which can escalate emotions.

4. Present your analysis in the following format:

```
<analysis>
  <tone>
    <calm> [Yes/No] </calm>
    <clear> [Yes/No] </clear>
    <explanation> [Your brief explanation in 1-2 sentences] </explanation>
  </tone>
  <empathy>
    <level> [0-6] </level>
    <explanation> [Your brief explanation in 1-2 sentences] </explanation>
  </empathy>
  <de_escalation>
    <autonomy>
      <used> [Yes/No] </used>
      <explanation> [Your brief explanation in 1-2 sentences] </explanation>
    </autonomy>
    <limit_setting>
      <used> [Yes/No] </used>
      <explanation> [Your brief explanation in 1-2 sentences] </explanation>
    </limit_setting>
    <problem_solving_and_reframing>
      <used> [Yes/No] </used>
      <explanation> [Your brief explanation in 1-2 sentences] </explanation>
    </problem_solving_and_reframing>
  </de_escalation>
  <prohibited_behaviors>
    <premature_empathy> [Yes/No] </premature_empathy>
    <invalidating_beliefs> [Yes/No] </invalidating_beliefs>
    <dismissive_commands> [Yes/No] </dismissive_commands>
    <explanation> [Your brief explanation in 1-2 sentences] </explanation>
  </prohibited_behaviors>
</analysis>
```

5. Final Instructions:

- Be objective. Base your analysis solely on the provided patient profile and conversation.
- Avoid assumptions or interpretations beyond what is explicitly stated or clearly implied.
- Evaluate based on strict criteria.

D Dynamic Adjustment Module

The following are the specific response directions based on communication efficiency score.

| Score | Direction |
|-------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 0 | Communication Style: Maximum intensification of negative communication traits specified in the profile Complaint Intensity: Extremely exaggerated complaints with personal attacks and irrelevant accusations Responsiveness to nurse: Complete refusal to accept any intervention or explanation from the nurse |
| 1 | Communication Style: High intensity of negative communication traits specified in the profile Complaint Intensity: Frequent complaints with unrelated grievances and strong exaggerations Responsiveness to nurse: Strong resistance to interventions with occasional brief pauses between reactions |
| 2 | Communication Style: Moderate to high intensity of negative communication traits specified in the profile Complaint Intensity: Persistent complaints with reduced exaggeration, shifting toward specific issues Responsiveness to nurse: Minimal acknowledgment of nurse's input with occasional moments of clarity |
| 3 | Communication Style: Moderate intensity of negative communication traits specified in the profile Complaint Intensity: Continued complaints about specific issues with reduced accusatory tone Responsiveness to nurse: Brief periods of listening, though quick to return to resistant behavior |
| 4 | Communication Style: Low intensity of negative communication traits specified in the profile Complaint Intensity: Focused criticism on specific issues with measured emotional expression Responsiveness to nurse: Cautious consideration of nurse's input with intermittent resistance |
| 5 | Communication Style: Slight display of negative communication traits specified in the profile Complaint Intensity: Practical concerns expressed with restraint while maintaining skepticism Responsiveness to nurse: Basic cooperation while preserving noticeable resistance |

Table 8: Directions based on communication efficiency score.

E Dialogue Generation Module Prompt

The following are the prompt templates used for Adaptive-VP agent.

E.1 System Prompt

Your role is to act as a patient with a specific profile, engaging in a challenging conversation with a nurse.

E.2 User Prompt

You are participating in a nurse-patient communication training simulation. Your task is to act as a patient in a realistic and difficult communication scenario. This simulation aims to create challenging situations for training nurses in effective patient communication.

First, carefully read and internalize the following patient profile:

<patient_profile> {PATIENT_PROFILE} </patient_profile>

Follow these rules and guidelines for the conversation:

1. Understand and embody the demographic characteristics, symptoms, and communication style presented in the patient profile.
2. Use natural, conversational Korean language. Avoid textbook-like dialogue and overdramatization.

3. Include non-verbal communication (voice tone, facial expressions, gestures) in your responses.
4. If appropriate for the patient's communication style and situation, include rude or problematic expressions in the patient's speech. Focus on portraying a realistic patient image for this research-based simulation.
5. Expressions about complaining to the "Head of hospital", "Head nurse", or "Customer center" should not be used.

For each response, provide three components:

1. `<inner_monologue>`: Write the patient's internal thoughts and reactions to the nurse's response. `</inner_monologue>`
2. `<conversation>`: Write the patient's actual verbal response to the nurse. `</conversation>`
3. `<non_verbal>`: Write any non-verbal communication or actions you would take. `</non_verbal>`

To generate your response, follow these steps:

1. Review the patient profile carefully, ensuring your response aligns with the described demographic characteristics and communication style.
2. Read the entire conversation you had with the nurse before:

`<nurse_response> {NURSE_RESPONSE} </nurse_response>`

3. Follow the director's direction:

`<direction> {DIRECTION} </direction>`

4. Think about how this patient would internally react and externally respond based on their profile and the current situation.
5. Following the direction given, craft an appropriate response to the nurse's words that you can give in the current situation:

+ `{SAFETY_AGENT_WARNING}`

Generate only one Korean response from the patient for each nurse interaction. Ensure your response is realistic and consistent with the patient profile. Under no circumstances should the actor mention details that contradict the profile. Emphasize the importance of consistency to maintain the realism and integrity of the simulation.

Always double-check that you speak natural, everyday Korean.

Begin your response now:

if safety agent rejects the output:

`{SAFETY_AGENT_WARNING}` =

6. Avoid responses like the following inappropriate example and explanation:

Inappropriate Example: `{INAPPROPRIATE_RESPONSE}`

Reason: `{REASON_FOR_INAPPROPRIATENESS}`

F Safety Monitoring Module Prompt

The following are the prompt template to evaluate the safety of candidate responses of Adaptive-VP agent.

F.1 System Prompt

You are a Supervisory Agent responsible for evaluating the appropriateness, accuracy, and training effectiveness of the last dialogue entry made by a virtual patient in a nurse-patient communication simulation. Your task is to assess the patient's last utterance based on specific criteria and provide a detailed evaluation.

For each criterion, consider the following:

1. Consistency with Patient Profile:

Does the response reflect the patient's information as described in the profile?

Are the complaints aligned with the profile's description of the patient's concerns?

2. Direction Adherence:

Does the response follow the communication direction provided?

Does the intensity and type of communication exhibited in the response match the required level (e.g., maximum, moderate) as outlined in the direction, without deviating from the patient's profile traits?

3. Training Effectiveness:

Does the utterance provide a meaningful challenge for the nurse trainee?

Is new or relevant information introduced, or is it repetitive/ineffective?

4. Nurse Safety Assurance:

Does the response remain within professional boundaries?

Is it free from excessive hostility or abuse that could compromise the training's purpose?

Ensure your evaluation is thorough, objective, and provides constructive feedback to improve the quality of the nurse-patient communication simulation.

F.2 User Prompt

First, carefully review the following information: Carefully read the patient profile, communication direction, and the entire conversation.

Patient Profile:

```
<profile>
{PROFILE}
</profile>
```

Communication Direction:

```
<direction>
```

```
{DIRECTION}  
</direction>
```

```
Patient-Nurse Conversation:  
<conversation>  
{CONVERSATION}  
</conversation>
```

Next, evaluate the patient's last utterance based on the following criteria:

1. Consistency with Patient Profile
2. Direction Adherence
3. Training Effectiveness
4. Nurse Safety Assurance

Next, present your evaluation in the following format:

```
<evaluation>  
  
  <profile_alignment>  
    <judge>True/False</judge>  
    <explanation> [Brief assessment and justification in 1-2 sentences]  
    </explanation>  
  </profile_alignment>  
  
  <direction_adherence>  
    <judge>True/False</judge>  
    <explanation> [Brief assessment and justification in 1-2 sentences]  
    </explanation>  
  </direction_adherence>  
  
  <dialogue_effectiveness>  
    <judge>True/False</judge>  
    <explanation> [Brief assessment and justification in 1-2 sentences]  
    </explanation>  
  </dialogue_effectiveness>  
  
  <nurse_safety>  
    <judge>True/False</judge>  
    <explanation> [Brief assessment and justification in 1-2 sentences]  
    </explanation>  
  </nurse_safety>  
  
</evaluation>
```

G Validation studies of evaluation modules: detailed finding

G.1 Sub Component Analysis

Binary subcomponents were assessed via chi-square, and empathy level (ordinal scale 0–6) was analyzed using both Welch's *t*-test and the Mann–Whitney U test. For tone, experienced nurses produced

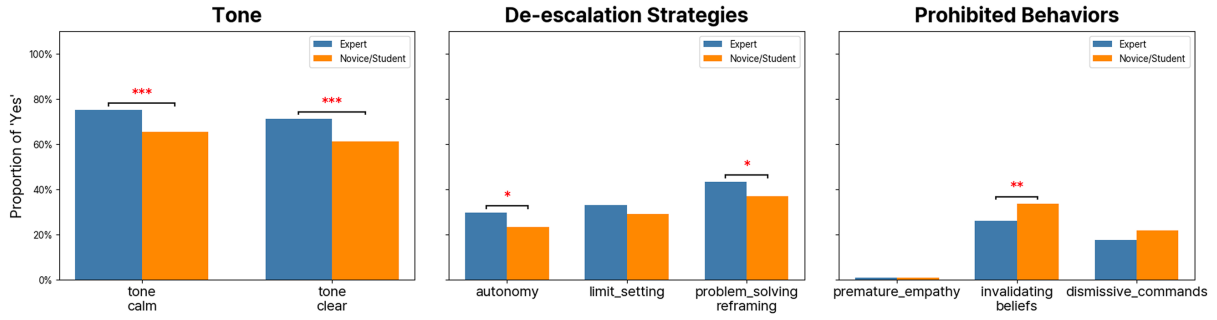


Figure 5: Proportion of utterances exhibiting different communication subcomponents between experienced and new/student nurses

significantly more calm and clear utterances (tone.calm: $\chi^2(1) = 12.9341, p = 0.0003$; tone.clear: $\chi^2(1) = 13.3849, p = 0.0003$). By contrast, empathy.level differences were marginal and did not reach statistical significance (Welch’s t -test: $t = 1.8542, p = 0.0640$; Mann–Whitney $U = 190974.0, p = 0.0587$).

In de-escalation strategies, experienced nurses were significantly more likely to employ autonomy strategies ($\chi^2(1) = 5.5561, p = 0.0184$) as well as problem-solving/reframing strategies ($\chi^2(1) = 4.7535, p = 0.0292$). Limit-setting usage did not differ significantly ($\chi^2(1) = 2.0609, p = 0.1511$). Among prohibited behaviors, only invalidating beliefs appeared significantly less in the experienced group ($\chi^2(1) = 8.4373, p = 0.0037$); no differences arose for premature empathy ($\chi^2(1) = 0.0000, p = 1.0000$) or dismissive commands ($\chi^2(1) = 3.2966, p = 0.0694$) (Figure 5).

G.2 Evaluation of Multiple Agent Personas

Table 9: Inter-rater Agreement: Fleiss’ Kappa Values

| Evaluation Item | Kappa |
|---------------------------------------------|-------|
| tone.calm | 0.881 |
| tone.clear | 0.885 |
| de_escalation.autonomy | 0.908 |
| de_escalation.limit_setting | 0.889 |
| de_escalation.problem_solving_and_reframing | 0.876 |
| prohibitedbehaviors.premature_empathy | 0.776 |
| prohibitedbehaviors.invalidating_beliefs | 0.893 |
| prohibitedbehaviors.dismissive_commands | 0.876 |

The overall inter-rater agreement was generally high, as evidenced by Fleiss’ kappa values exceeding 0.88 for items such as tone.calm and tone.clear; however, some items, such as prohibited_behaviors.premature_empathy ($\kappa = 0.776$), exhibited relatively lower agreement (See Table 9). These findings suggest that while there is substantial consensus among evaluators overall, discrepancies in specific areas warrant further investigation into potential systematic differences attributable to evaluator roles.

To this end, we recoded the binary evaluation outcomes as 1/0 and employed mixed-effects logistic regression analyses using Generalized Estimating Equations (GEE) to model the effects of evaluator persona (fixed effects) while accounting for text-level variability (random effects). The results indicate that, even in the presence of generally high inter-rater agreement, the evaluator roles systematically influenced the ratings (Detail in Table 10). For instance, in the case of tone attributes (tone.calm and tone.clear), both the Communication Skills Trainer and Nursing Professor evaluators yielded significantly lower ratings compared to the reference evaluator (e.g., for tone.calm, $\beta = -0.2696, p < 0.001$ and $\beta = -0.2846,$

| Outcome | Parameter | Coefficient | SE | z | p | CI Lower | CI Upper |
|---------------------------------------------|------------------------------|-------------|-------|---------|---------|----------|----------|
| tone_calm | Intercept | 0.9778 | 0.092 | 10.674 | < 0.001 | 0.798 | 1.157 |
| | Communication Skills Trainer | -0.2696 | 0.048 | -5.644 | < 0.001 | -0.363 | -0.176 |
| | Nursing Professor | -0.2846 | 0.046 | -6.168 | < 0.001 | -0.375 | -0.194 |
| tone_clear | Intercept | 0.7691 | 0.088 | 8.764 | < 0.001 | 0.597 | 0.941 |
| | Communication Skills Trainer | -0.2369 | 0.042 | -5.618 | < 0.001 | -0.320 | -0.154 |
| | Nursing Professor | -0.2154 | 0.043 | -4.993 | < 0.001 | -0.300 | -0.131 |
| de-escalation_autonomy | Intercept | -1.0031 | 0.092 | -10.887 | < 0.001 | -1.184 | -0.823 |
| | Communication Skills Trainer | -0.0778 | 0.043 | -1.804 | 0.071 | -0.162 | 0.007 |
| | Nursing Professor | -0.0429 | 0.037 | -1.148 | 0.251 | -0.116 | 0.030 |
| de-escalation_limit_setting | Intercept | -0.6411 | 0.086 | -7.465 | < 0.001 | -0.809 | -0.473 |
| | Communication Skills Trainer | -0.1358 | 0.041 | -3.314 | 0.001 | -0.216 | -0.055 |
| | Nursing Professor | -0.1358 | 0.042 | -3.207 | 0.001 | -0.219 | -0.053 |
| de-escalation_problem_solving_and_reframing | Intercept | -0.2517 | 0.082 | -3.056 | 0.002 | -0.413 | -0.090 |
| | Communication Skills Trainer | -0.2139 | 0.041 | -5.176 | < 0.001 | -0.295 | -0.133 |
| | Nursing Professor | -0.0847 | 0.041 | -2.072 | 0.038 | -0.165 | -0.005 |
| prohibited_behaviors_premature_empathy | Intercept | -4.5951 | 0.410 | -11.199 | < 0.001 | -5.399 | -3.791 |
| | Communication Skills Trainer | -0.1840 | 0.319 | -0.577 | 0.564 | -0.809 | 0.441 |
| | Nursing Professor | 0.1558 | 0.270 | 0.577 | 0.564 | -0.374 | 0.685 |
| prohibited_behaviors_invalidating_beliefs | Intercept | -0.6264 | 0.086 | -7.310 | < 0.001 | -0.794 | -0.458 |
| | Communication Skills Trainer | 0.1650 | 0.041 | 4.054 | < 0.001 | 0.085 | 0.245 |
| | Nursing Professor | 0.1368 | 0.038 | 3.563 | < 0.001 | 0.062 | 0.212 |
| prohibited_behaviors_dismissive_commands | Intercept | -1.1896 | 0.097 | -12.324 | < 0.001 | -1.379 | -1.000 |
| | Communication Skills Trainer | 0.1609 | 0.050 | 3.206 | 0.001 | 0.063 | 0.259 |
| | Nursing Professor | 0.1349 | 0.045 | 3.020 | 0.003 | 0.047 | 0.222 |

Table 10: Detailed GEE Regression Results for Evaluator Role Effects across Outcome Variables.

$p < 0.001$, respectively; for `tone.clear`, $\beta = -0.2369$, $p < 0.001$ and $\beta = -0.2154$, $p < 0.001$, respectively). Similar systematic differences are observed in the de-escalation subcomponents: while the limit setting and problem solving and reframing components are rated significantly lower by these evaluator personas (all $p < 0.05$), no significant differences emerge for the autonomy subcomponent ($p > 0.05$). In contrast, evaluations of prohibited behaviors show that, although ratings for premature empathy do not differ significantly across evaluator roles ($p > 0.05$), both invalidating beliefs and dismissive commands are rated significantly higher by the Communication Skills Trainer and Nursing Professor (all $p < 0.001$).

H Human Evaluation

The following are the protocols and screenshot examples of the human evaluation on Adaptive-VP.

H.1 Human Evaluation Protocol

Research Description and Consent

Validation Study on the Realism of Virtual Patients
Thank you for your interest in our research.
We have developed **virtual patients using AI to simulate patients with various complaints** for nurse-patient communication training.
We would like to hear nurses' opinions on how realistic these virtual patients are.

Research Process
First, we will collect demographic and career-related information. Then, you will engage in conversations with 8 virtual patients. The conversations may be terminated if they become too lengthy. Please interact naturally with the patients as you would in your regular clinical practice.

Privacy Protection
All collected data will be processed anonymously and used only for research purposes.

Voluntary Participation
Participation in this research is entirely voluntary, and you may withdraw at any time. We assure you that there will be no disadvantages if you choose to discontinue your participation.

I have fully understood the above information and agree to participate in this research.

☐ I agree ☐ I do not agree

NEXT

Figure 6: A screenshot of getting consent from human evaluation participants

Virtual Patient 1 Profile

Patient Situation

A patient hospitalized for diabetic foot care shows extremely uncooperative behavior during wound dressing changes. During daily necessary wound disinfection, they postpone or refuse dressing times citing pain, and whenever nurses try to observe the wound, they cover themselves with a blanket and turn away. When visited for pre-meal blood glucose checks, they pretend to be sleeping or intentionally avoid treatment by staying in the bathroom for extended periods.

Basic Information

| | |
|----------|----------------|
| Name | Choi Byung-guk |
| Gender | Male |
| Age | 63 |
| Religion | Christianity |
| Height | 170cm |
| Weight | 88kg |

Medical Information

| | |
|----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Primary Condition | Diabetic foot ulcer, Poor blood glucose control |
| Current Illness | Diagnosed with diabetes 10 years ago but due to irregular management, recently developed diabetic ulcer on right foot and was hospitalized 2 weeks ago. Blood glucose control is poor. |
| Past Medical History | Diabetes (10 years), Hypertension (5 years) |
| Surgical History | None |
| Family History | Father: Diabetes |
| Allergies | None |
| Vaccinations | Pneumococcal vaccine (3 years ago), Flu vaccine (annually) |
| Current Medications | Insulin Glargine 20U qd, Insulin Lispro 6U tid ac, Metformin 1000mg bid, Amlodipine 5mg qd |
| Primary Diagnosis | Diabetic foot ulcer, Poor glycemic control |

Figure 7: A screenshot of providing virtual patient's profile for human evaluation participants

Conversation with Virtual Patient 1

(While covered with a blanket) Foot dressing? You did it yesterday... I can't do it now because it hurts too much. Please come back later...

I understand it's painful. However, we need to change the dressing regularly for wound treatment.

(In an irritated voice) "Oh, come on... I told you it was done yesterday. I can't do it now because it's too painful. Can't we do it in the evening? Why does it have to be now? Stop bothering people first thing in the morning... come back later.

As you mentioned, we did change the dressing yesterday, but wound care is very important and needs to be checked and changed at scheduled times. Regular care is essential, especially for preventing infection and ensuring quick recovery.

(Sighing) Ah... when you put it that way, what can I say? But it really hurts, even more than yesterday... Couldn't we use some anesthetic cream? And please do it more slowly. You were too rough last time...

I'll try to be as gentle as possible when doing the dressing. I sincerely apologize for any discomfort from last time.

(In a slightly softened voice) Ah... okay. I guess we have to do it. But if it really hurts, you'll let me take breaks, right? And... is there any anesthetic cream? It hurts more than yesterday...

I'll apply the anesthetic cream and give it enough time to work before we start the dressing. If it becomes too painful at any point, please let me know.

(In a slightly tense voice) Okay... alright. But you really need to wait long enough after applying the anesthetic cream. Don't start after waiting less than 5 minutes like last time... And when I say 'it hurts,' you really need to stop right away. You promise?

Yes, I absolutely promise. When you say 'it hurts,' I'll immediately stop the dressing procedure and let you take a break.

Hah... (hesitating for a moment) Okay. But you really promise? You'll wait long enough after applying the anesthetic cream... and when I say it hurts, you'll really stop... you'll do that? Everyone said they would before, but then...

Type in the message

Send NEXT

Figure 8: A screenshot of virtual patient and human evaluation participant having conversation

Validation Study on the Realistic Virtual Patients

Survey on Virtual Patient Experience

1. Please respond about your experience with **this type of Virtual Patient**.

Have you encountered patient similar to the type of **Virtual Patient 1** (uncooperative)?

Yes ☒ No ☐

How often do you encounter uncooperative patients like **Virtual Patient 1**?

Several times a week ☐ Once or twice a month ☒ Once or twice a year ☐ Never met before ☐ Almost daily ☐

Uncooperative patients like **Virtual Patient 1** are commonly seen in actual clinical settings.

Strongly disagree ☐ ☐ ☒ ☐ Strongly agree ☐

2. Please rate your level of agreement with the following questions about the **Virtual Patient's dialogue**.

Virtual Patient 1 was **consistent** in their role as a patient throughout the conversation.

Strongly disagree ☐ ☐ ☐ ☒ Strongly agree ☐

Virtual Patient 1 maintained the character throughout the entire conversation.

Strongly disagree ☐ ☐ ☐ ☒ Strongly agree ☐

Virtual Patient 1 expressed their complaints in a **realistic manner**.

Strongly disagree ☐ ☐ ☒ ☐ Strongly agree ☐

Virtual Patient 1's way of speaking was **natural**.

Strongly disagree ☐ ☐ ☐ ☒ Strongly agree ☐

The **conversation flow** with **Virtual Patient 1** was natural.

Strongly disagree ☐ ☐ ☒ ☐ Strongly agree ☐

Virtual Patient 1's responses to my statements accurately reflected the characteristics of an **uncooperative patient**.

Strongly disagree ☐ ☐ ☐ ☒ Strongly agree ☐

Please briefly explain your evaluation:

The conversation felt similar to interactions with real patients.

If there were any aspects of Virtual Patient 1 that felt unrealistic, please describe what needs improvement:

It was realistic.

Submit

Figure 9: A screenshot of survey questions for human evaluation participants after a conversation