

Design and Feasibility of a Large Language Model-Based Teaching Pathway for Clinical Pharmacy Practice Competencies

Jizhong Zhang

The Affiliated Cancer Hospital of Nanjing Medical University

Yuxuan Zhu

The Affiliated Cancer Hospital of Nanjing Medical University

Dengwei Lin

The Affiliated Cancer Hospital of Nanjing Medical University

Xiaoting Yang

The Affiliated Cancer Hospital of Nanjing Medical University

Shuai Yu

The Affiliated Cancer Hospital of Nanjing Medical University

Xin Liu

The Affiliated Cancer Hospital of Nanjing Medical University

Nan Wu

The Affiliated Cancer Hospital of Nanjing Medical University

Min Zhao

zhaomin1618@126.com

The Affiliated Cancer Hospital of Nanjing Medical University

Research Article

Keywords: clinical pharmacy, early-career training, virtual standardized patient, large language model, teaching pathway, simulation-based training

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11023382/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Early-career clinical pharmacists often lack real-patient exposure, standardized mentorship, and immediate feedback. Traditional standardized patients (SPs) are costly and limited in scalability, while early virtual patients (VPs) use rigid, pre-programmed dialogue. Large language models (LLMs) offer interactive virtual standardized patients (VSPs), but few pharmacy-specific teaching pathways integrate structured dialogue generation and automated scoring. This study aimed to design and preliminarily evaluate an LLM-based VSP pathway specifically for clinical pharmacy practice competencies.

Methods

The teaching pathway was implemented using structured prompt frameworks: the SPADE framework for dialogue generation and an RTF (Role-Task-Format)-based prompt for automated scoring. Three standardized oncology pharmacy ward round cases were created. A total of 21 clinical pharmacy trainees (6 junior pharmacists, 15 graduate students) participated in 30 simulation interactions. The pathway's dialogue quality and scoring consistency were assessed via expert panel evaluation (15-item scale), trainee user experience (18-item scale plus open-ended questions), and consistency analysis between automated scores and expert human scores (Spearman correlation, Bland-Altman).

Results

Experts rated anthropomorphism highest (mean 4.44/5 for natural patient-like speech). The VSP automated scores demonstrated an extremely strong rank correlation with expert scores (Spearman's $\rho = 0.934$, $p < 0.001$), but showed systematic underestimation (mean difference -2.38 points, 95% limits of agreement -8.17 to 3.40). Response stability received relatively low ratings (3.62/5 for real-time response). Trainees reported high overall acceptance across dimensions (81–89% of maximum scores), with the highest rating for “comfortable without tension” (4.52/5). Qualitative feedback highlighted realistic details and reduced perceived pressure during interactions, while noting occasional over-answering and technical delays during peak hours.

Conclusions

The LLM-based VSP teaching pathway, guided by SPADE and RTF-based prompt frameworks, can simulate authentic pharmacy ward round dialogues and generate scores that strongly correlate with expert judgments. Although systematic underestimation precludes its use for high-stakes assessment, the teaching pathway shows strong promise as a low-cost, accessible formative training tool for

pharmaceutical care. The prompt engineering frameworks are transferable to other LLMs and clinical communication scenarios.

1 Introduction

Early-career clinical pharmacists typically have limited contact with real patients, restricted bedside practice opportunities, and lack of immediate feedback during communication skill development¹⁻³. Their training mainly relies on self-directed exploration, resulting in a prolonged and inefficient learning cycle. This challenge is constrained by limited clinical resources, teaching time and space, and patient safety considerations^{1,4}. Thus, a key challenge is to provide standardized, repeatable communication training without placing additional burdens on clinical resources.

According to Kolb's experiential learning theory, mastering practical skills requires concrete experience – direct interactions with patients – forming the essential foundation of learning⁵⁻⁷. Yet this foundational stage is notably lacking for most junior clinical pharmacists. Virtual patient (VP) simulation offers a way to provide trainees with a standardized, repeatable experiential learning without relying on limited clinical resources⁸.

Standardized patients (SPs) provide a realistic clinical teaching environments while supporting real-time interaction and immediate feedback⁹. Nevertheless, their widespread implementation faces high implementation costs, insufficient standardization fatigue-induced performance variability, and scalability barriers⁹⁻¹². Early computer-based virtual patients improved accessibility but relied on pre-programmed, script-based dialogues that could not flexibly and naturally respond to learners' spontaneous inquiries and complex clinical questions^{13,14}. These constraints have limited the widespread adoption of simulation-based training in pharmacy education¹⁵.

Large language models (LLMs) offer new possibilities for dynamic, open-ended virtual patient conversations and automated feedback^{16,17}. A recent scoping review of 28 LLM-VSP studies found that 92.9% were published in the last two years but noted a lack of standardized evaluation methods, with only 13% using validated tools and most relying on closed-source commercial models¹⁸. In pharmacy education, artificial intelligence (AI)-based simulated patients have been explored in several ways, including multi-program LLM interview training¹⁹, Generative AI (GenAI)-supported communication during the Objective Structured Clinical Examination (OSCE) under educator guidance²⁰, and rule-based chatbot for medication history practice²¹. These studies have demonstrated the value of AI-based simulation for communication training. However, few studies have integrated structured prompt frameworks for both interaction generation and automated scoring, or examined the agreement between automated and expert scoring.

To address these gaps, this study designed an LLM-based VSP teaching pathway using the cloud-based DeepSeek-R1 model with structured prompt engineering. This pathway incorporates standardized

oncology cases, a SPADE (Scenario & Setting, Persona, Activity Flow, Discipline Rules, Evaluation) prompt framework for dialogue generation and an RTF (Role-Task-Format)-based prompt for automated scoring. We conducted a feasibility study with 21 clinical pharmacy trainees (30 simulated ward round dialogues) to evaluate dialogue quality, scoring consistency, and user acceptance. All prompts and case materials are provided in the supplementary materials to facilitate replication.

We aimed to answer : (1) How does this teaching pathway perform in terms of the authenticity, clinical accuracy? (2) How do trainees rate its usability and educational value?

By publicly reporting our prompt framework and case design protocols, we aim to demonstrate that this teaching pathway is transferable and replicable, providing a methodological template for adoption by other teaching institutions.

2 Methods

2.1 Study Design

This single-center feasibility study aimed to design and preliminarily evaluate a standardized, LLM-based virtual patient teaching pathway for training pharmaceutical care skills. The pathway was implemented using DeepSeek-R1 as an open-source, replicable LLM engine. The study was approved by the Institutional Review Board of the Affiliated Cancer Hospital of Nanjing Medical University (KY-2026-142). An overview of the study workflow is presented in Fig. 1.

2.2 Participants

This study involved three groups of participants:

2.2.1 Case and Prompt Development Expert Panel

Three experienced oncology pharmacy specialists reviewed all core materials, including structured standardized cases, scoring criteria, prompt engineering frameworks and evaluation scales.

2.2.2 Pilot Testing Expert Panel

Three additional specialists (not overlapping with the development panel) served as pilot users, providing quantitative and qualitative feedback to refine the prompts.

2.2.3 Trainee Participants

A convenience sampling method was used to recruit participants. Of the 22 individuals invited, 21 consented and completed the study (95.5% participation rate; 1 declined due to time constraints). The final cohort consisted of six junior pharmacists undergoing clinical pharmacy training and 15 graduate students in clinical pharmacy who had completed 10–12 months of clinical internship. All participants provided informed consent and completed a baseline questionnaire on demographic information, self-

assessed pharmaceutical care experience, and technical background (AI tool usage and prior SP exposure). This baseline data described the sample characteristics and assessed potential influences on subsequent evaluations.

The 30 dialogue records were contributed as follows: 15 trainees completed one scenario each, 1 trainee completed two scenarios each, and 4 trainees completed all three scenarios, totaling 30 valid records. The three standardized cases (VSP-A, VSP-B, VSP-C) were randomly assigned to trainees for each simulation scenario, with the constraint that trainees completing multiple scenarios received different cases each time to minimize learning effects.

2.3 Design of LLM-VSP Teaching Pathway

2.3.1 DeepSeek-R1

DeepSeek-R1 supports ultra-long-context interactions (maximum 128K tokens), facilitating comprehension of lengthy dialogue histories and complex medical records. This enables coherent patient conversations. The choice of DeepSeek-R1 was driven by its open-source, free-to-use nature, as well as our team’s prior validation in pharmacy-related tasks, where it successfully integrated multi-source clinical information to generate structured Subjective, Objective, Assessment, and Plan (SOAP)-format medication logs, confirming its capacity for clinical data integration and structured output²².

2.3.2 Technical Specifications

All VSP interactions were conducted either via the DeepSeek official website (<https://chat.deepseek.com>) on a computer or the DeepSeek mobile APP on an iPad, with the “deep thinking” mode enabled. The model version was DeepSeek-R1-0528 (released May 28, 2025). As the web interface was used, sampling parameters (temperature, top-p) were not user-configurable and remained at platform defaults; DeepSeek recommends temperature = 0.6 for the R1 model. The maximum context window is 128K tokens. All instructions were provided via user prompts, as system prompts are not recommended for the DeepSeek-R1 model. Interactions occurred during July–October 2025.

2.3.3 Establishment of Structured Standardized Cases

The VSP interactions in this study were based on three self-developed standardized cases. These cases were developed through the following process:

First, we reviewed 120 structured medication therapy logs from the past two years to ground the cases in real-world practice. Subsequently, we cleansed and de-identified the collected data to extract key clinical and educational points. Based on these insights, we created three fully fictional standardized cases for first-day chemotherapy ward rounds, following a unified protocol: history-taking, counseling, and consultation.

To ensure focused assessment, the three cases maintained comparable core complexity with an emphasis on medication history-taking, chemotherapy regimen education, and medication consultation.

To allow generalizability assessment, they varied in cancer type, regimen, and consultation focus while keeping evaluation criteria consistent (Table 1).

Table 1
Standardized Cases Core Information Comparison

Item	VSP-A	VSP-B	VSP-C
Cancer type	Small cell lung cancer	Non-small cell lung cancer	Gastric cancer
Stage	Extensive stage	IIIC,T4N3M0	IV,TxN3M1
Treatment Regimen	Serplulimab+	Serplulimab+	Sintilimab+
	Etoposide+	Paclitaxel+	Oxaliplatin + S-1 (Tegafur, Gimeracil and Oteracil Potassium)
	Lobaplatin	Nedaplatin	
Key Medical History	Hypertension	Coronary Heart Disease	Hypertension
Adverse Reaction	Febrile Neutropenia	Febrile Neutropenia	Thrombocytopenia
Consultation Topic	Management of Neutropenia	Drug Interactions (Statins & Aspirin)	Peripheral Neuropathy Prevention

A separate set of scoring criteria was developed for each of the three cases. All three sets were derived from a common foundation – the “*Assessment Plan for Trainees in Clinical Pharmacist Training Bases*” published by the Pharmaceutical Specialized Committee of the Chinese Hospital Association – and were designated as the “Modified Pharmacy Ward Round Scoring Criteria.” Each version was further optimized to reflect oncology-specific features and the VSP-mediated interaction context, incorporating refined and quantifiable evaluation dimensions (detailed in the Modified Pharmacy Ward Round Scoring Criteria).

All cases and corresponding scoring criteria were examined by the development panel to ensure clinical accuracy, educational relevance, and validity of the assessment tools, resulting in a high degree of consensus. The complete scripts of the three standardized cases are presented in Supplementary Material S3.

2.3.4 Structured Prompt Engineering

This study drew upon prior research in medical education that employed prompt engineering to build LLM-based VSPs^{10,13,23–25}.

Through multiple rounds of simulated dialogues (n = 20) and iterative refinement, we developed two dedicated prompt systems : one for dialogue generation and one for automated evaluation. This iterative process aimed to resolve common issues identified in early versions and reported in similar studies^{14,24–26}. The prompt refinement strategy was guided by our team’s prior work on pharmaceutical prompt

engineering. Sun *et al.*²⁷ demonstrated that RTOG-CoT (Role-Task-Goal-Output chain-of-thought) prompting significantly outperformed zero-shot prompting, raising the semantic F1 score from 0.67 to 0.816, confirming that the prompt quality critically influences LLM performance in pharmacy tasks. Following this approach, we conducted five iterative optimization cycles, resulting in final prompt frameworks with substantially enhanced stability and authenticity.

The SPADE framework guided dialogue generation, which comprises five components: Scenario & Setting, Persona, Activity Flow, Discipline Rules, and Evaluation (Fig. 2). A separate scoring prompt defined the LLM as an “examiner,” mandating precise scoring logic and evidence-based deductions from the dialogue text. Both prompts underwent iterative optimization; final versions are in Supplementary Material S1 and S2.

2.3.5 Development and Validation of Assessment Tools

To comprehensively evaluate the teaching pathway, we developed two 5-point Likert scales, informed by the Technology Acceptance Model (TAM) and prior virtual patient evaluation studies^{5,17,28,29}. Both were reviewed for content validity.

The 15-item Expert Feedback Evaluation Scale gathered feedback from pilot users – Pilot Testing Expert Panel. The 20-item User Experience Feedback Questionnaire (18 Likert-scale and 2 open-ended) was completed by trainees. Both questionnaires were developed for this study. The English-language versions are provided in Supplementary Materials S4 and S5.

2.4 Expert Evaluation and Refinement

During the late development stage, the three pilot experts completed ward round dialogues with all three VSPs (9 dialogues total), recorded in a text format.

After each dialogue, experts rated performance and provided feedback using the Expert Feedback Evaluation Scale. Thematic analysis revealed recurring issues, guiding final refinements to the prompt engineering frameworks:

- (1) Over-answering
- The VSP occasionally provided excessive information. Given that patients in real clinical environment often offer unsolicited information, this over-answering behavior was retained as it simulated real-world patient spontaneous disclosure.
- (2) Role deviation
- The VSP sometimes responded as both pharmacist and patient, which violated the required question-and-answer format. To address this, we added stricter instructions restricting it to patient-only responses.
- (3) Response latency

- Intermittent delays were observed, primarily due to the limitations of the DeepSeek web service at that time and the model's inherent language processing time for context recalling long conversation histories. This was not a defect of our prompt design. We instructed participants to allow a few seconds for response. Importantly, this did not affect data or integrity and the issue has since been largely resolved with subsequent DeepSeek platform updates.
- (4) Scoring inaccuracies
- Early scoring prompts exhibited several problems: (a) fabricating unasked questions as score points; (b) using case file information instead of dialogue text as scoring source; and (c) overly broad criteria leading to inconsistent ratings. We addressed these by : (a) explicitly limiting scoring to the dialogue text record only and prohibiting the fabrication of unasked questions; (b) clearly distinguishing case background information from the conversation being evaluated; and (c) refining the Modified Pharmacy Ward Round Scoring Criteria to reduce ambiguity.

These refinements yielded the final prompt used in the participant evaluation stage.

2.5 Data Collection and the Establishment of the Reference Standard

2.5.1 Participant Interaction and Feedback Collection

Before the interaction, each trainee received a brief oral orientation covering the three-phase ward round structure (medication history-taking → medication counseling → medication consultation). Trainees were informed that the VSP would only respond to questions during the history-taking and counseling phases, with patient questions reserved for the final consultation phase. A brief patient case summary was made available for familiarization prior to each dialogue. No formal training or practice session was provided, as the study aimed to assess baseline performance in an authentic, unscripted interaction.

Following expert validation, each trainee completed 1–3 simulation scenarios based on availability, generating 30 valid dialogue records. Dialogues were conducted via the DeepSeek website (text) or mobile APP (speech-to-text), with all interactions recorded in text. After each dialogue, the scoring prompt was sent to DeepSeek three times independently (no conversation memory carried over). The VSP generated three scores and three feedback reports per dialogue; the mean score was used as the final automated score (30 total, one per dialogue). Each participant then completed the User Experience Feedback Questionnaire. To avoid bias, VSP-generated scores and feedback were released only after all study procedures were completed.

2.5.2 Establishment of the Human Reference Standard

The same three pilot experts manually scored participant performance using the same scoring criteria. Their prior exposure to the VSP during testing helped them distinguish whether information gaps were due to trainee performance versus the VSP's spontaneous disclosures, thereby supporting fairer and more contextualized assessments.

Before formal scoring, three experts completed a brief calibration exercise: they received written instructions, independently scored three pilot dialogues, and discussed results to align understanding.

Inter-rater reliability (IRR) was assessed using intraclass correlation coefficient (ICC) (two-way random-effects, absolute agreement) based on 90 scores (3 experts × 30 dialogues). Both ICC(2,1) (single-measure) and ICC(2,3) (average-measure) were calculated. The mean of the three experts' scores per dialogue served as the benchmark for consistency analysis with VSP automated scores.

2.5.3 VSP Automated Scoring Reliability

For VSP scoring repeatability, ICC was computed using the same model (two-way random-effects, absolute agreement) for the three inferences per dialogue. ICC(2,1) assessed the single inference consistency, and ICC(2,3) assessed the mean of three inferences.

2.6 Statistical Analysis

This study employed a combination of quantitative and qualitative methods to comprehensively evaluate the collected data.

2.6.1 Data Collection and Types

Four types of data were collected: (1) baseline demographics, self-assessed experience and technical usage background (21 participants). (2) expert quantitative ratings and qualitative feedback (15-item scale). (3) trainee quantitative ratings and open-ended feedback (18-item scale + 2 questions, 21 participants); (4) and automated expert scores for 30 dialogue records.

2.6.2 Statistical analysis methods

Due to the small sample size and potential clustering effects (30 dialogues from 21 participants), each dialogue was treated as an independent observation. This limitation is addressed in the Discussion. Quantitative analyses were performed using IBM SPSS Statistics (R27.0.1.0). Data are presented as mean ± SD unless otherwise specified.

- (1) Descriptive Statistics
- Continuous variables were described as mean ± SD or median (interquartile range, IQR); categorical variables as frequency (n, %), for baseline data, all scale scores. Normality was tested using Shapiro-Wilk .
- (2) Reliability Analysis
- Inter-Rater Reliability (IRR) and VSP scoring repeatability were assessed using the ICC (two-way random-effects, absolute agreement) based on 90 independent scores (3 experts × 30 dialogues) and the three independent VSP inferences per dialogue, respectively. Both ICC(2,1) (single-measure) and ICC(2,3) (average-measure) were calculated. ICC value ≥ 0.80 was considered indicative of good reliability. Cronbach's α was calculated for the two scales (α > 0.7 acceptable).

- (3) Validity and Consistency Analysis

Correlation between automated and experts scores was assessed using Pearson's or Spearman's coefficient based on normality. Bland-Altman analysis visualized the mean difference and 95% limits of agreement between the two scoring methods. No pre-defined acceptable limits were set, as this feasibility study aimed to describe the agreement patterns rather than make a binary interchangeability judgement.

- (4) Qualitative Data Analysis

Open-ended feedback provided by experts and trainees was analyzed using thematic analysis. Two researchers independently coded the data, identified themes and sub-themes and reached consensus through discussions.

3 Results

3.1 Baseline Characteristics of Trainee Participants

This study recruited 21 trainee participants, the group consisted of six junior pharmacists (28.6%) undergoing clinical pharmacy training and 15 graduate students (71.4%) in clinical pharmacy. All participants had completed baseline characteristics survey, the demographic information, experience in pharmaceutical care core skills and technical usage background of participants are shown in Table 2.

Table 2
Baseline Characteristics of Trainee Participants

Characteristics	number	percentage
Genders		
Men	8	38.1%
Women	13	61.9%
Age		
≤ 25	14	66.7%
26–30	4	19.0%
31–35	3	14.3%
Identity		
Junior pharmacists	6	28.6%
Graduate students	15	71.4%
Self-assessed experience: Conduct medical history taking independently		
Beginner (no/limited experience)	10	47.6%
Intermediate (average experience)	9	42.9%
Advanced (experienced)	2	9.5%
Self-assessed experience: Participate in the pharmacy ward rounding		
Beginner (no/limited experience)	7	33.3%
Intermediate (average experience)	10	47.6%
Advanced (experienced)	4	19.0%
Technical usage: Frequency of using AI tools		
Several times a week or more	17	81.0%
Several times a month or less	4	19.0%
Technical usage: Experiences of contact with SP		
Have personally experienced	15	71.4%
Only observing or without any involvement	6	28.6%
Technical usage: Experiences of Contact with VSP		
Never had any contact	18	85.7%
Had contact before (including observation)	3	14.3%

3.2 Expert Evaluation of LLM-based VSP Performance

3.2.1 Quantitative Ratings

After the dialogues, the expert panel rated the VSP’s performance using the Expert Feedback Evaluation Scale and provided qualitative feedback. As shown in Table 3, the VSP received the highest scores in the anthropomorphism dimension (85.93% of maximum score), with “LLM-based VSP uses natural patient-like speech without artificiality” receiving the highest evaluation (4.44 ± 0.73). The feedback quality dimension scored relatively low, with “after dialogues, LLM-based VSP assesses user proficiency against criteria, delivering objective evaluations” receiving the lowest score of all 15 items (3.11 ± 0.33), the overall dimension percentage was 66.67% (see Table 3 note), indicating the direction for improvement.

Table 3
Pilot Testing Expert Panel’s Evaluation of LLM-based VSP Performance

Assessment Dimension	No. of Items	Possible Score	Mean Score $\pm$ SD	Percentage (%)
Clinical Accuracy	4	20	16.22 ± 1.09	81.11
Logical Consistency	3	15	12.56 ± 1.13	83.70
Anthropomorphism	3	15	12.89 ± 1.17	85.93
Response Stability	3	15	10.67 ± 0.87	71.11
Feedback Quality and Evaluative Effectiveness	2	10	6.67 ± 0.50	66.67
Scores are based on a 5-point Likert scale (1 = Very Poor to 5 = Very Good) for each item. The Possible Score for each dimension is the sum of the highest possible scores for its constituent items. Percentage = (Mean Score / Possible Score) $\times$ 100%.				

3.2.2 Qualitative Feedback

In addition to quantitative ratings, experts provided open-ended feedback. Thematic analysis of their comments revealed four main themes:

Over-answering: Experts noted that the VSP sometimes provided unsolicited information. One commented, “The VSP provided additional information that is not related to the question asked, which felt realistic but occasionally disrupted the pharmacist’s inquiry thinking.”

Role deviation: The VSP occasionally generated dialogues simulating both parties. An expert observed, “Sometimes it directly generated pharmacist-patient dialogue, which was confusing.”

Response latency: Delays were noted. One expert mentioned, “DeepSeek had stalls and delays during peak daytime usage. In addition, there was a noticeable pause before each response, which slightly affected the natural flow.”

Scoring inaccuracies: Issues with the scoring function were reported. For example, an expert stated, “The VSP seemed to make up the score points that the pharmacist had not actually asked.”

These qualitative insights, which informed the system refinements described in Section 4.4, provide context for the quantitative ratings above.

3.3 Reliability of the Human Reference Standard and VSP Automated Scoring

3.3.1 Human reference standard

The IRR among the three expert’s scoring on the 30 dialogue records was excellent. The ICC(2,1) was 0.906 (95% confidence interval (CI): 0.767–0.959, $p < 0.001$), indicating good consistency of individual expert scores. The ICC(2,3) was 0.967 (95% CI: 0.908–0.986, $p < 0.001$), indicating that the mean of the three experts’ scores had high reliability and could serve as the standard reference for subsequent consistency analysis.

3.3.2 VSP automated scoring stability

The repeatability of the VSP automated scoring, based on three independent inferences per dialogue, was also excellent. The ICC(2,1) was 0.939 (95% CI: 0.893–0.969), and the ICC(2,3) was 0.979 (95% CI: 0.961–0.989, $p < 0.001$). These values indicate that the mean of the three VSP inferences was highly stable and comparable to the reliability of the human reference standard.

3.4 Agreement Between LLM-based VSP and Human Scoring

For each dialogue, the VSP automated score was defined as the mean of three independent inferences. Shapiro-Wilk tests revealed that the distribution of the AI-averaged scores significantly deviated from normality ($W = 0.911$, $p = 0.016$), while the expert-averaged scores were normally distributed ($W = 0.977$, $p = 0.735$). Therefore, Spearman’s rank correlation coefficient was used to assess the association between the two scoring methods.

Bland-Altman analysis further evaluated the agreement between the automated scoring and the experts’ scoring. As shown in Fig. 3a (Bland-Altman plot), a systematic underestimation was observed, with a mean difference of -2.38 points (95% CI: -3.44 to -1.32) and 95% limits of agreement (LoA) from - 8.17 to 3.40 (width 11.57 points). Among the 30 data points, only one point (Dialogue 4) fell outside the 95% LoA, with a difference of -9.00 points. The 95% confidence interval of the mean difference did not include zero, confirming a statistically significant systematic underestimation by the VSP. As shown in Fig. 3b (scatter plot), the automated scores showed a strong rank correlation with experts scores (Spearman’s $\rho = 0.934$, $p < 0.001$). This systematic bias (mean difference: -2.38 points) suggests that direct interchangeability without calibration is not supported in the current version. However, the strong rank

correlation indicates potential utility for formative feedback and low-stakes self-assessment. The raw scores for each dialogue are presented in Supplementary Material S11.

3.5 User Experience Feedback from Trainees

21 trainees filled out the User Experience Feedback Questionnaire after completing their all conversations.

3.5.1 Quantitative Ratings

The 18-item scale showed excellent internal consistency (Cronbach’s $\alpha = 0.93$). Participants’ ratings across dimensions were positive, ranging from 81.2% to 89.5% of the maximum score (Table 4, Fig. 4). The single item with the highest score was “I feel comfortable without tension during interactions” (4.52 ± 0.51). The item with the lowest score was “LLM-based VSP responds in real-time without lag” (3.62 ± 1.28), which was consistent with the low score of response stability dimension.

Table 4
Trainee Ratings of LLM-based VSP Across Dimensions (n = 21)

Assessment Dimension	No. of Items	Possible Score	Mean Score $\pm$ SD	Percentage (%)
Clinical Accuracy	4	20	17.24 ± 1.81	86.19
Logical Consistency	2	10	8.38 ± 1.36	83.81
Anthropomorphism	2	10	8.95 ± 1.43	89.52
Response Stability	4	20	16.24 ± 3.19	81.19
Emotional Acceptance and Trust	3	15	12.86 ± 1.56	85.71
Teaching/Assessment Value	3	15	12.67 ± 1.96	84.44
Scores are based on a 5-point Likert scale (1 = Very Poor to 5 = Very Good) for each item. The Possible Score for each dimension is the sum of the highest possible scores for its constituent items. Percentage = (Mean Score / Possible Score) $\times$ 100%.				

3.5.2 Qualitative Feedback

Thematic analysis of the answers to 2 open-ended questions yielded the following findings.

With regard to the advantages and disadvantages of VSP, the feedback of trainees can be summarized as follows :

- Advantages included : realistic and rich details (“VSP was very vivid in describing his own medical history”); logical and organized response (“clear logic, it is not easy to miss the main points of pharmacist inquiry”); objective and stress-reducing (“VSP was more stable and easier to

communicate than SP emotion, can reduce tension”); and convenient operation (“VSP operation was simple”).

- Disadvantages included : over-answering (“VSP sometimes answered too much, answering questions that had not yet been asked”); technical stability issues (“Occasionally the server is unstable. When the voice input, the drug name recognition is not very clear”); and lack of flexibility (“A bit rigid, lacking the flexibility of real patients”).
- Regarding the aspects mostly needed to be improved to LLM-based VSP, trainees proposed :
- Limiting response scope (“Do not answer non-questions, or over-answer”);
- Optimize technical stability (“Improve network stability”);
- Increase the diversity of diseases (“Include more disease types, such as chronic conditions”);
- Enhance humanity and flexibility (“Need to be more human-like”);
- Strictly follow the pharmacy ward round workflow.

4 Discussion

This feasibility study designed and preliminarily evaluated an LLM-based virtual patient teaching pathway for pharmaceutical care training, implemented with the cloud-based DeepSeek-R1 model and the SPADE prompt framework. The automated scoring showed a strong rank correlation with expert ratings and trainees reported positive experiences.

4.1 Interpretation of key findings.

4.1.1 Anthropomorphism: Strengths of the SPADE Framework

The LLM-based VSP received the highest expert ratings in the anthropomorphism dimension, with a mean of 4.44/5 for “natural patient-like speech”, confirming that DeepSeek can generate natural language matching patient style. This is closely linked to the SPADE prompt framework, which restricts DeepSeek to responding solely as an ordinary patient, following the medical cases and avoiding medical terminology or role deviation. Meanwhile, the VSPs maintain a consistent, non-judgmental interaction style. This aligns with the user’s highest-rated item “I feel comfortable without tension during interactions” (4.52/5), suggesting that the VSP’s predictable responses lower perceived barriers to active participation. This is consistent with Ahn and Jeong⁵, who found that personified virtual patient create a low-stakes environment promoting active learning.

4.1.2 Automated Scoring: High Consistency but Systematic Bias

There was an extremely strong positive rank correlation between the experts’ scores and the VSP’s automated scores (Spearman’s $\rho = 0.934$, $p < 0.001$), indicating that the VSP could accurately rank

trainee performance and demonstrate reliable evaluation capabilities. This result benefits from the RTF-based prompt framework, which imposes strict constraints on grading rules, scoring criteria, and output format. However, we also observed a systematic underestimation: the AI scores were on average 2.38 points lower than the expert ratings.

We attribute this bias to a specific mechanism inherent to pharmaceutical care conversations. In realistic clinical practice, patients usually offer unsolicited information (e.g., medication adherence, adverse reactions or comorbidities) without being clearly asked. We preserved this over-answering behavior in our VSP interactions, because it simulates real-world patient disclosure. However, this causes a scoring dilemma: When the VSP proactively offers key information before being asked, the automated scoring system – which strictly checks explicit questions – may deduct points for “not asked”. In contrast, experts reviewing the same dialogue records award credit based on whether the information was obtained, regardless of who initiated the topic. Although we explicitly emphasized in the scoring prompt that patient-volunteered information – when indirectly elicited by the pharmacist (e.g., through open-ended questions) or confirmed by follow-up – should be credited, DeepSeek occasionally misjudges such cases. This suggests that LLMs tend to interpret criteria literally, paying more attention to explicit question-answer pairs rather than holistic interaction comprehension. Experts may also apply contextual leniency – awarding partial credit based on overall dialogue flow – that LLMs cannot replicate.

Notably, Dialogue 4 was identified as the only outlier that exceeded the 95% consistency limit in the Bland Altman plot. This extreme discrepancy is most likely attributable to the VSP abnormally disclosing excessive unsolicited information, leading the automated scoring system to deduct points because the trainees “did not ask” key information. In contrast, the expert raters awarded full credit, as the information had been successfully obtained during the interaction. This case further highlights the core rating conflicts between AI scoring based on literal rules and human expert judgment based on contextual understanding.

Grevisse³⁰ also reported that GPT-4 produced significantly lower grades than human evaluators in the short-answer responses, with limits of agreement wide enough to question interchangeability. Despite this systematic bias, the strong rank correlation ($p = 0.934$) supports the use of LLM-based VSP for daily practice, self-assessment, and formative feedback where the ability to identify strengths and weaknesses is more important than absolute score accuracy. Taken together, these findings indicate that current LLMs struggle to reliably apply context-sensitive scoring rules that require inferring causal attribution, rather than simply matching surface patterns.

4.1.3 Scoring Accuracy and Hallucination Control

In early expert testing, the feedback quality and evaluation effectiveness dimension received relatively low scores, with a mean rating of 3.11/5 for the item: “After dialogues, LLM-based VSP assesses user proficiency against criteria, delivering objective evaluations.” Experts also noted in their qualitative feedback that the VSP occasionally “made up the score points that the pharmacist had not actually

asked about” and “used case background information instead of dialogue text”. This phenomenon can be attributed to two main factors. First, the early scoring prompts were insufficiently refined, such that the constraint “based solely on the dialogue text” in the scoring prompts was not consistently enforced, leading the LLM to occasionally “borrow” contextual information from medical cases to fill in missing scoring points. Second, the scoring criteria themselves were relatively broad in the early stage, resulting in a lack of precision in evaluations. In response, we refined the RTF-based scoring prompts, more explicitly requiring that scoring be based exclusively on dialogue records rather than cases materials. The scoring criteria for pharmacy ward rounds have also been further refined, with detailed explanations provided for scoring points. In the subsequent participant evaluation phase, these problems markedly diminished, although minor inconsistencies persisted. This suggests that careful prompt engineering can effectively reduce scoring hallucinations. Nevertheless, as detailed in Section 4.1.2, the systematic underestimation of scores (mean difference – 2.38 points) remained, indicating that while prompt engineering can mitigate “hallucinations,” the broader tendency of LLMs to apply stricter criteria than humans experts requires further investigation.

4.1.4 Response Stability and Technical Reliability: Common Technical Challenges

In the participants’ feedback, the response stability dimension received a relatively low score, with the lowest-rated item being “LLM-based VSP responds in real-time without lag” (3.62 ± 1.28). Both the initial feedback from experts and the subsequent qualitative feedback from participants mentioned issues such as stuttering, lag, and network instability. First, during the study period (i.e., daytime peak hours), we used the DeepSeek-R1 website and experienced response delays due to high server load. Second, the DeepSeek-R1 supports 128K-token ultra-long-context interactions, and it requires “deep thinking” when recalling the entire conversation history. Third, the use of the iPad’s speech-to-text feature also increased latency in recognition and transmission. Finally, in this feasibility study, this study utilized public application programming interfaces (APIs).

These problems are common technical limitations among current cloud-based large model platforms, not defects in the prompt framework or teaching pathway design. Oncu *et al.*³¹ also observed technical challenges such as delayed responses and server overload when using ChatGPT-4o as a VSP. Importantly, these fluctuations did not affect the integrity of conversations, clinical accuracy, or scoring consistency. Core teaching interactions remained stable in our study. With server optimization, these issues may be mitigated in the future.

4.1.5 Trainee Acceptability: High Usability and Educational Value

Trainees’ overall acceptance was high, with excellent internal consistency (Cronbach’s $\alpha = 0.93$). This result is primarily due to the following factors: First, the trainees are primarily young pharmacists and graduate students who are widely familiar with different kinds of AI tools, with a low learning curve for the VSP. Second, the virtual environment provides a low-pressure setting where trainees can repeatedly

practice and get instant feedback without the interpersonal evaluative pressure that may accompany face-to-face SP interactions. Third, the VSP's standardized setting is unbound by time and space, trainees can train anytime and anywhere. This highly improves the convenience and accessibility of training. Fourth, the LLM-based VSP can provide automatic instant feedback to help trainees quickly find their own shortcomings and strengths. The VSP's non-judgmental interaction style, reflected in the highest-rated item "comfortable without tension" (4.52/5), may reduce interpersonal barriers that often impede early-career trainees from actively engaging in communication practice – with effects conceptually similar to the supportive, reflective environment described in Lu *et al.*'s narrative medicine training for new nurses³². The above findings are consistent with the views of Öncü *et al.*³¹, who reported that AI-driven VSPs are user-friendly, cost-effective, and highly accessible, and can significantly improve learner satisfaction and willingness to continue using them.

4.2 Methodological Contribution: A Transferable Prompt Engineering Paradigm (SPADE and RTF)

Prior LLM-VSP studies^{14,24–26} have reported common technical pitfalls, including premature answers, excessive length, over-extended responses, non-interactive monologues, irrelevant answers, and disrupted conversation flow. Through five iterative optimization cycles (based on 20 rounds of simulated dialogues reported in the Methods section), we substantially mitigated these issues by developing two structured prompt engineering frameworks: SPADE for dialogue generation and a separate RTF-based prompt for automated evaluation.

The SPADE framework systematically addresses these problems as follows. First, role deviation (i.e., VSP occasionally plays the role of both patient and pharmacist) is prohibited by the Persona layer, which strictly restricts the VSP to respond only as a patient, to use colloquial speech, and to obey a predefined information boundary. Second, over-answering and non-interactive monologues (premature answers, excessive length, and irrelevant answers) are constrained by the Discipline Rules layer. The Discipline Rules layer enforces a rigorous question-answer format, limits response length, and prohibits unsolicited advice or teaching. Third, the Activity Flow layer clearly defines the expected interaction phases (medication history-taking → medication counseling → medication consultation) to prevent VSP from skipping steps or moving ahead prematurely. Thus, the problem of disrupted dialogue flow is alleviated. To achieve automated scoring, we further developed the RTF-based prompt to decouple the evaluation task from the dialogue generation task. This separation avoids role confusion, i.e., prevents the LLM-based VSP from acting simultaneously as patient and examiner. The RTF-based scoring prompt explicitly defines the LLM as the "examiner", mandates evidence-based deductions with direct quotations from the dialogue records, and stipulates structured output format. This design mitigates hallucinations to some extent, as reported in our early iterations (e.g., fabricating unasked questions or relying on medical case background rather than the actual dialogue). Importantly, these frameworks are not limited to pharmaceutical care nor to DeepSeek. They are transferable to other clinical communication training scenarios (e.g., history taking, nursing pain assessment) and can be used with other LLMs such as GPT-4 and Gemini, providing a reusable methodology template for future VSP development.

4.3 Comparison with existing LLM-based VSP studies

Liu and colleagues³³ provided early evidence that LLMs could simulate VSPs. They verified that ChatGPT-3.5 could generate realistic patient interactions with lower cost, higher standardization, and no fatigue compared with traditional SPs. Physicians rated the ChatGPT-3.5-based VSP highly (8–10 out of 10) for its accurate and natural responses. Cook *et al.*¹⁴ further evaluated GPT-4.0-Turbo and GPT-3.5-Turbo as VSPs and confirmed high authenticity (mean 4.7/6), user experience (4.9/6), and feedback quality (4.7/6) at low cost (\$0.02–0.51 per conversation).

Öncü *et al.*³¹ recently validated ChatGPT-4o as a feasible VSP for training clinical case management, problem-solving, and crisis management among interns. Their study confirmed high trainee satisfaction and educational value, while also reporting similar technical limitations including response delays and instability.

Together, these studies supported LLM-based VSPs as a scalable, low-cost alternative to traditional SPs. However, these studies relied exclusively on proprietary, closed-source ChatGPT models, which incur usage fees. In addition, their scenarios focused on medical education for diagnosis and chronic disease management rather than pharmaceutical services for pharmacy students.

In contrast, our study directly addresses these gaps. We designed a VSP teaching pathway using DeepSeek-R1, a free, open-source LLM that offers a cost-free alternative to commercial models. This may lower financial barriers for institutions with limited budgets for simulation-based training. The pathway was specifically designed for pharmaceutical care competencies training using two structured prompt frameworks: the SPADE for dialogue generation and the RTF-based framework for automated scoring. Trainees gave favorable ratings across all dimensions (81–90%). Furthermore, we performed a full quantitative consistency analysis, showing an extremely strong correlation between automated and expert scores (Spearman's $\rho = 0.934$) with a systematic bias of -2.38 points.

4.4 Limitations

This study has several limitations. First, the small sample (21 trainees, 30 dialogues) from a single institution limits transferability to other training programs or cultural contexts. Second, only three oncology chemotherapy cases were used, and performance in other disease types (e.g., chronic disease management, outpatient counseling) remains untested. Third, text-based interaction (with optional speech-to-text) does not fully capture non-verbal communication which is essential for pharmaceutical care. Fourth, the expert reference standard, though highly reliable ($ICC(2,3) = 0.967$), was derived from three experts at one institution. Fifth, while we identified a systematic scoring underestimation (-2.38 points), this feasibility study was not designed to fully determine its sources (e.g., expert clinical leniency versus LLM literal interpretation of criteria), limiting our ability to propose a definitive calibration strategy. Sixth, technical delays during peak usage may have influenced user experience ratings. Seventh, the sample size of 30 dialogues was not based on an a priori power calculation for Bland-Altman analysis;

although the correlation was strong, the precision of the 95% limits of agreement may be affected. Finally, we did not assess long-term skill retention or transfer to real patient encounters.

4.5 Future Directions

Future work should focus on: (1) calibrating the scoring model to reduce systematic underestimation, for example, through few-shot expert-graded examples or linear bias correction; (2) expanding case diversity beyond oncology to chronic disease management and ambulatory pharmaceutical care; (3) conducting longitudinal randomized studies comparing VSP versus traditional SP training; and (4) exploring the option of on-premises deployment to potentially reduce network latency and address data privacy requirements in institutional settings.

4.6 Implications for Pharmacy Education

Although the LLM-based VSP cannot replace real patients or expert assessment, it serves as a low-cost, accessible, and standardized supplementary tool for pharmaceutical care training. Using the SPADE dialogue framework and RTF-based automated scoring, trainees can repeatedly practice medication history-taking, medication counseling, and medication consultation in a relaxed atmosphere and obtain instant, objective feedback.

Moreover, this teaching pathway has good replicability and scalability. The SPADE and RTF-based prompts do not rely on a specific LLM, and each training institution can use the open source model or commercial API and rapidly develop similar VSP training tools based on its own resources for common clinical cases. Our prompt framework and case design proposals have been published in the supplementary materials, which greatly reduce the technical threshold and implementation cost. Therefore, this pathway is especially suitable for primary hospitals and training institutions with limited resources.

5 Conclusion

This study suggests that an LLM-based VSP teaching pathway, implemented with the open-source DeepSeek-R1 model and structured through the SPADE and RTF-based prompt frameworks, can simulate pharmacy ward round dialogues and generate scores that strongly correlate with expert judgments (Spearman's $\rho = 0.934$). Systematic underestimation prevents direct substitution for high-stakes use, but the pathway shows strong promise as a formative tool with high user acceptance. With further refinements, it offers a scalable, cost-effective, and replicable solution to address the lack of authentic patient exposure in early-career clinical pharmacy training. Importantly, the SPADE/RTF-based prompt framework and case design proposals described in this study provide a transferable methodology that other institutions can directly adopt.

Abbreviations

AI

artificial intelligence
API
application programming interface
CI
confidence interval
GenAI
generative artificial intelligence
ICC
intraclass correlation coefficient
IQR
interquartile range
IRR
inter-rater reliability
LLM
large language model
LoA
limits of agreement
OSCE
Objective Structured Clinical Examination
RTF
Role-Task-Format
RTOG-CoT
Role-Task-Goal-Output chain-of-thought
SD
standard deviation
SOAP
Subjective, Objective, Assessment, and Plan
SP
standardized patient
SPADE
Scenario & Setting, Persona, Activity Flow, Discipline Rules, Evaluation
TAM
Technology Acceptance Model
VP
virtual patient
VSP
virtual standardized patient

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Jiangsu Cancer Hospital, The Affiliated Cancer Hospital of Nanjing Medical University (Approval No. KY-2026-142). All participants provided written informed consent to participate in this study.

Consent for publication

Not applicable. The cases used in this study were fully fictional, and no identifiable personal data are presented.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the Jiangsu Pharmaceutical Association Research Project (202564087) and the “Yishan Plan” Project for the High-level Hospital Construction of Jiangsu Cancer Hospital (YSZD202406). The funders had no role in study design, data collection, analysis, interpretation, or preparation of the manuscript.

Author Contribution

J.Z. and Y.Z. contributed equally to this work. J.Z. developed the SPADE framework and drafted the Methods and Results. Y.Z. developed the cases, conducted data collection, and drafted the Introduction and Discussion. D.L. performed data cleaning and validated the prompts. X.Y. coordinated the expert panels and performed qualitative analysis. S.Y. performed statistical analyses and prepared figures. X.L. and N.W. critically reviewed and revised the manuscript. M.Z. supervised the study, acquired funding, and provided final approval. All authors reviewed and approved the final manuscript.

Acknowledgements

The authors thank all trainee participants and expert panel members for their involvement in this study.

Data Availability

All data supporting the findings of this study are available within the article and its Supplementary Information, except for the original dialogue transcripts, which are not publicly available due to participant privacy and confidentiality. De-identified data are available from the corresponding author on reasonable request.

References

1. 1. Gitti, S. A., Al-den, S. S. B. & Abbas, A. F. When Patients Say No: Rethinking Refusal in Bedside Medical Education. *Clin Teach* **22**, e70235 (2025). <https://doi.org/10.1111/tct.70235>
2. 2. Sarli, G. et al. Using Virtual Patients to Assess and Improve Clinicians' Emotional Self-awareness: a Randomized Controlled Study. *Acad Psychiatry* **48**, 18–28 (2024). <https://doi.org/10.1007/s40596-023-01909-z>
3. 3. Jalal, Z. et al. Communications Skills in the Pharmacy Profession: A Cross Sectional Survey of UK Registered Pharmacists and Pharmacy Educators. *Pharmacy (Basel)* **6** (2018). <https://doi.org/10.3390/pharmacy6040132>
4. 4. Gharib, A. M., Bindoff, I. K., Peterson, G. M. & Salahudeen, M. S. Exploring global perspectives on the use of computer-based simulation in pharmacy education: a survey of students and educators. *Front Pharmacol* **15**, 1494569 (2024). <https://doi.org/10.3389/fphar.2024.1494569>
5. 5. Ahn, S. & Jeong, H. W. Exploring nursing students' experiences with virtual patient-based health assessment simulation program: A qualitative study. *Nurse Educ Today* **153**, 106826 (2025). <https://doi.org/10.1016/j.nedt.2025.106826>
6. 6. Mitsugi, N., Ijuin, K., Oshiyama, C., Hara, E. & Nishimura, T. An AI-mediated framework for recursive learning: transforming individual experiences into organizational knowledge and autonomous engagement in elderly care. *Front Digit Health* **7**, 1529072 (2025). <https://doi.org/10.3389/fdgth.2025.1529072>
7. 7. Eisenberg, S., Walles, B. & Long, E. Theory in action for graduate nurse educator students: An experiential learning project. *Nurse Educ Today* **143**, 106391 (2024). <https://doi.org/10.1016/j.nedt.2024.106391>
8. 8. Hamilton, A., Molzahn, A. & McLemore, K. The Evolution From Standardized to Virtual Patients in Medical Education. *Cureus* **16**, e71224 (2024). <https://doi.org/10.7759/cureus.71224>
9. 9. Sezer, B., Sezer, T. A., Teker, G. T. & Elcin, M. Developing a virtual patient: design, usability, and learning effect in communication skills training. *BMC Med Educ* **23**, 891 (2023). <https://doi.org/10.1186/s12909-023-04860-7>
10. 10. Rädcl-Abläss, K. et al. Teaching opportunities for anamnesis interviews through AI based teaching role plays: a survey with online learning students from health study programs. *BMC Med Educ* **25**, 259 (2025). <https://doi.org/10.1186/s12909-025-06756-0>
11. 11. Stamer, T., Steinhäuser, J. & Flägel, K. Artificial Intelligence Supporting the Training of Communication Skills in the Education of Health Care Professions: Scoping Review. *J Med Internet Res* **25**, e43311 (2023). <https://doi.org/10.2196/43311>

12. 12. Laverde, N., Grévisse, C., Jaramillo, S. & Manrique, R. Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training. *Comput Struct Biotechnol J* **27**, 2481–2491 (2025). <https://doi.org/10.1016/j.csbj.2025.05.025>
13. 13. Luo, M. J. et al. A large language model digital patient system enhances ophthalmology history taking skills. *NPJ Digit Med* **8**, 502 (2025). <https://doi.org/10.1038/s41746-025-01841-6>
14. 14. Cook, D. A., Overgaard, J., Pankratz, V. S., Del Fiol, G. & Aakre, C. A. Virtual Patients Using Large Language Models: Scalable, Contextualized Simulation of Clinician-Patient Dialogue With Feedback. *J Med Internet Res* **27**, e68486 (2025). <https://doi.org/10.2196/68486>
15. 15. Katoue, M. G. & Ker, J. Pharmacists' experiences and perceptions about simulation use for learning and development of clinical skills in Kuwait. *Int J Clin Pharm* **41**, 1451–1461 (2019). <https://doi.org/10.1007/s11096-019-00904-5>
16. 16. Ferreira Santos, J., Ladeiras-Lopes, R., Leite, F. & Dores, H. Applications of large language models in cardiovascular disease: a systematic review. *Eur Heart J Digit Health* **6**, 540–553 (2025). <https://doi.org/10.1093/ehjdh/ztaf028>
17. 17. Tu, T. et al. Towards conversational diagnostic artificial intelligence. *Nature* **642**, 442–450 (2025). <https://doi.org/10.1038/s41586-025-08866-7>
18. 18. Zeng, J. et al. Embracing the Future of Medical Education With Large Language Model-Based Virtual Patients: Scoping Review. *J Med Internet Res* **27**, e79091 (2025). <https://doi.org/10.2196/79091>
19. 19. Walter, J. M. et al. Utilization of Artificial Intelligence as Simulated Patients to Enhance Student Interviewing Skills in the Pharmacy Curriculum Across Multiple Programs. *J Am Coll Clin Pharm* **9**, e70238 (2026). <https://doi.org/10.1002/jac5.70238>
20. 20. Lim, A. et al. Using an educator-guided generative artificial intelligence (GenAI) tool for developing communication skills in undergraduate pharmacy students. *Currents in Pharmacy Teaching and Learning* **18**, 102675 (2026). <https://doi.org/https://doi.org/10.1016/j.cptl.2026.102675>
21. 21. Hall, S. et al. Evaluation of the use of chatbots as simulated patients in capstone clinical pharmacy education: A pilot study. *Currents in Pharmacy Teaching and Learning* **18**, 102588 (2026). <https://doi.org/https://doi.org/10.1016/j.cptl.2026.102588>
22. 22. Zhu, Y. et al. Generating subjective, objective, assessment, and plan (SOAP)-structured medication logs using DeepSeek-R1 through prompt engineering and multi-source clinical information. *Digit Health* **12**, 20552076261428234 (2026). <https://doi.org/10.1177/20552076261428234>
23. 23. Peralta Ramirez, A. A. et al. Clinical Simulation with ChatGpt: A Revolution in Medical Education? *J cme* **14**, 2525615 (2025). <https://doi.org/10.1080/28338073.2025.2525615>
24. 24. Holderried, F. et al. A Language Model-Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study. *JMIR Med Educ* **10**, e59213 (2024). <https://doi.org/10.2196/59213>

25. Wang, C. et al. Application of Large Language Models in Medical Training Evaluation-Using ChatGPT as a Standardized Patient: Multimetric Assessment. *J Med Internet Res* **27**, e59435 (2025). <https://doi.org/10.2196/59435>
26. Weisman, D. et al. Development of a GPT-4-Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study. *JMIR Form Res* **9**, e65670 (2025). <https://doi.org/10.2196/65670>
27. Sun, Y. et al. Prompt-guided knowledge-augmented large language model for reliable clinical medication review. *J Biomed Res*, 1–5 (2026). <https://doi.org/10.7555/jbr.39.20250515>
28. Mudenda, S., Mufwambi, W., Mwale, R. S., Kathewera, B. & Lubanga, A. F. Attitudes and usage of ChatGPT among pharmacy students in a Sub-Saharan African country, Zambia: findings and implications on the education system. *BMC Med Educ* **25**, 1237 (2025). <https://doi.org/10.1186/s12909-025-07833-0>
29. Pham, L., Klaus, T. & Bateh, J. Key factors influencing intention to use ChatGPT: An empirical study of U.S. students. *Acta Psychol (Amst)* **260**, 105516 (2025). <https://doi.org/10.1016/j.actpsy.2025.105516>
30. Grévisse, C. LLM-based automatic short answer grading in undergraduate medical education. *BMC Med Educ* **24**, 1060 (2024). <https://doi.org/10.1186/s12909-024-06026-5>
31. Öncü, S., Torun, F. & Ülkü, H. H. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Med Educ* **25**, 278 (2025). <https://doi.org/10.1186/s12909-025-06877-6>
32. Lu, N. et al. A narrative medicine-based training program increases the humanistic care quality of new nurses in cancer hospital. *Precision Medical Sciences* **12**, 23–28 (2023). <https://doi.org/10.1002/prm2.12091>
33. Liu, X. et al. ChatGPT: when the artificial intelligence meets standardized patients in clinical training. *J Transl Med* **21**, 447 (2023). <https://doi.org/10.1186/s12967-023-04314-0>
34. List of Figures Figure 1. Overview of LLM-based VSP study workflow. Figure 2. Schematic diagram of the SPADE prompt framework and corresponding flawed items in iterative training. Figure 3. Comparison of VSP automated scores with expert scores. Figure 4. Distribution of dimension-averaged scores across six assessment dimensions.
35. List of Tables Table 1. Standardized Cases Core Information Comparison Table 2. Baseline Characteristics of Trainee Participants Table 3. Pilot Testing Expert Panel's Evaluation of LLM-based VSP Performance Table 4. Trainee Ratings of LLM-based VSP Across Dimensions

Figures

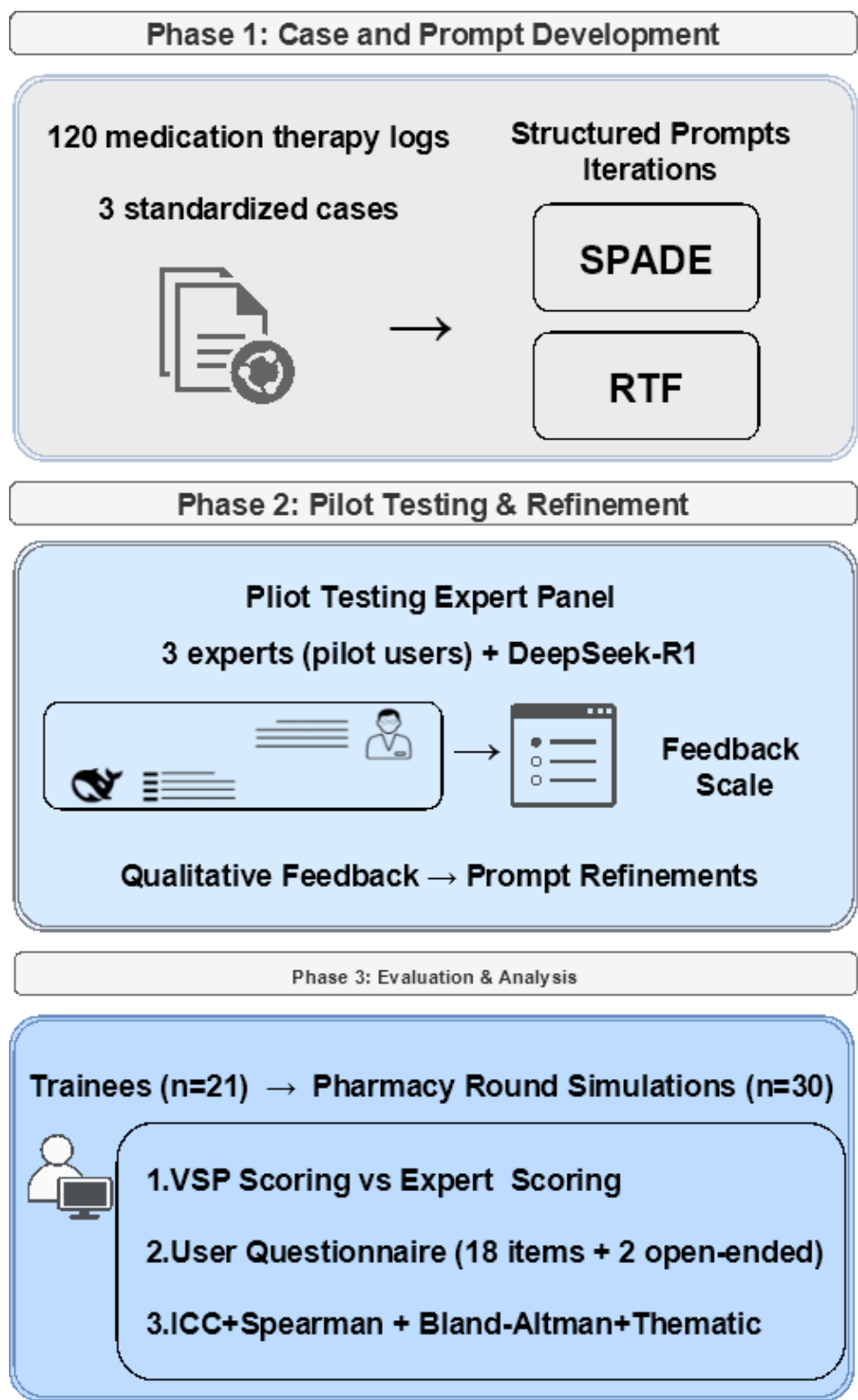


Figure 1

Overview of LLM-based VSP study workflow. The study consisted of three stages: (1) Case and Prompt Development, during which structured medication therapy logs were used to create fictional cases and structured prompts (SPADE for dialogue and RTF for scoring). The structured prompts were developed and refined through 20 rounds of simulated conversations to ensure stability and authenticity; (2) Pilot Testing & Refinement, in which three experts interacted with VSP and provided quantitative and

qualitative feedback to refine the prompt frameworks; and (3) Evaluation and Analysis, in which 21 trainees completed 30 simulated pharmacy rounds dialogues, followed by automated scoring by the VSP, experts blinded scoring, and analysis of user experience questionnaire. Data analysis included ICC, Spearman correlation, Bland-Altman plots, Cronbach's α , and thematic analysis.

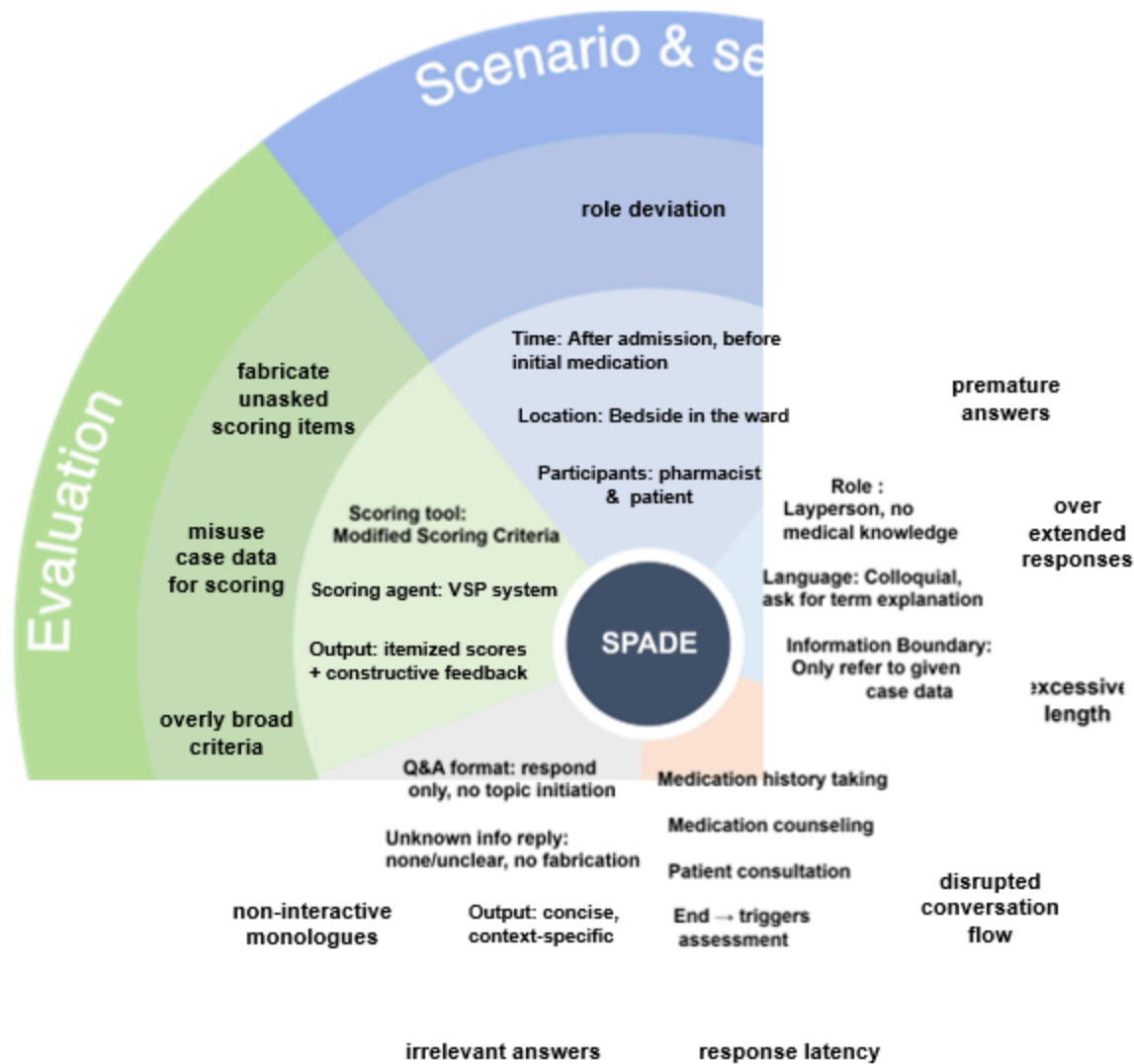


Figure 2

Schematic diagram of the SPADE prompt framework and corresponding flawed items in iterative training. This concentric annular chart is divided into five equal sectors corresponding to the five core modules of the SPADE framework: Scenario & Setting, Persona, Activity Flow, Discipline Rules, and Evaluation. The central circle is marked with the core concept SPADE prompt; the innermost ring summarizes the detailed specification of each SPADE component; the middle ring lists typical defects of model output encountered during prompt iteration and training; the outermost ring labels the five constituent dimensions of the SPADE framework respectively.

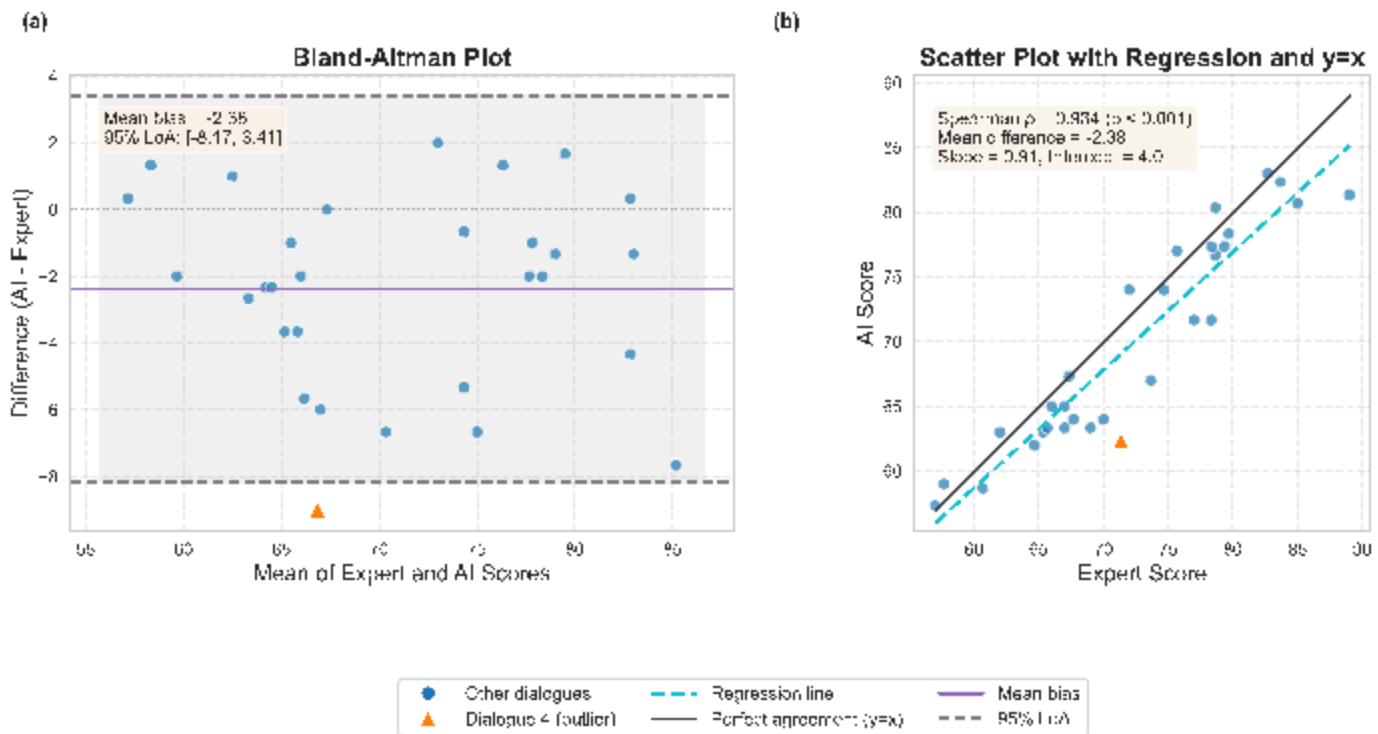


Figure 3

Comparison of VSP automated scores with expert scores.

This figure presents a comprehensive agreement analysis between the automated VSP scoring system and human expert ratings using two complementary visualizations. No error bars are displayed in either panel, as the data points represent individual paired measurements rather than aggregated summary statistics.

(a) Bland–Altman plot illustrating the agreement between the two scoring methods. The x-axis shows the mean of the expert and AI scores for each dialogue, while the y-axis shows the difference (AI score minus expert score). Blue circles represent the 29 non-outlier dialogues. The orange triangle highlights Dialogue 4, which is the single outlier falling outside the 95% limits of agreement. The solid purple horizontal line indicates the mean bias (−2.38 points), and the two dashed grey horizontal lines mark the 95% limits of agreement (lower: −8.17; upper: 3.40). The light grey shaded area visually emphasizes the region between the two limit lines.

(b) Scatter plot comparing the expert scores (x-axis) against the AI scores (y-axis). The dashed cyan line represents the linear regression fit (slope = 0.91, intercept = 4.0), and the solid black line indicates the line of perfect agreement ($y = x$). Spearman’s rank correlation coefficient is $\rho = 0.934$ ($p < 0.001$). Again, the outlier Dialogue 4 is distinctly marked as an orange triangle, while all other dialogues are shown as blue circles.

Symbols and line definitions: Blue circles, non-outlier dialogues; orange triangles, Dialogue 4 (outlier); solid purple line, mean bias; dashed grey lines, 95% limits of agreement; light grey shaded area, interval

between the limits; dashed cyan line, linear regression; solid black line, perfect agreement.

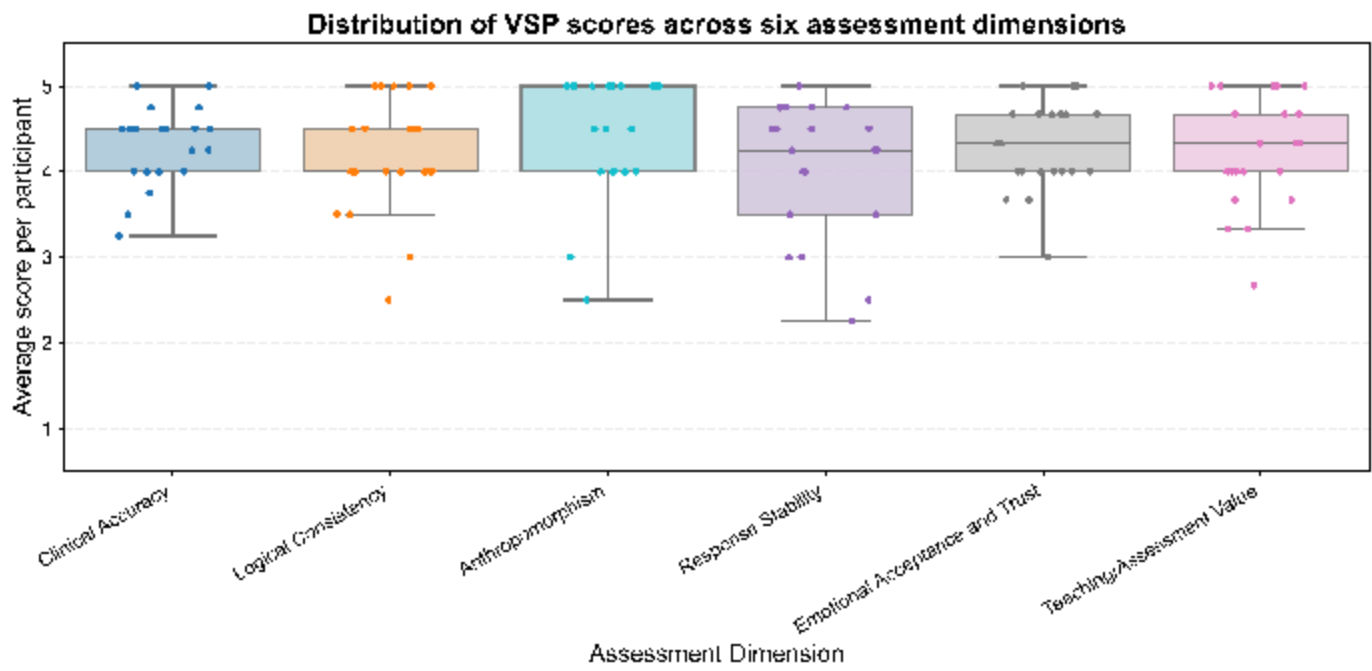


Figure 4

Distribution of dimension-averaged scores across six assessment dimensions.

Boxes represent interquartile ranges (IQR), with horizontal lines indicating medians. Whiskers extend to the most extreme data points within 1.5×IQR from the box edges; dots indicate individual responses (n=21 per dimension). Higher scores indicate better user-perceived performance. Clinical Accuracy, 4 items; Logical Consistency, 2 items; Anthropomorphism, 2 items; Response Stability, 4 items; Emotional Acceptance and Trust, 3 items; Teaching/Assessment Value, 3 items.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials.pdf](#)