Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Large Language Model Based Teaching Pathway Is Feasible for Developing 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

Article

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

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10452795/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

Pharmaceutical care competencies, including pharmacist-patient communication and pharmacy rounds, are core capabilities for clinical pharmacists. However, early-career trainees often lack sufficient real-patient exposure, standardized mentorship, and immediate feedback during skill development. Traditional standardized patients (SPs) are costly and limited in scalability, while early virtual patients (VPs) suffered from rigid, pre-programmed interactions. Large language models (LLMs) offer new opportunities for creating interactive virtual standardized patients (VSPs). However, no study has yet applied an LLM-based VSP teaching pathway specifically to address this early-career training gap in clinical pharmacy.

Objective

This study developed an LLM-based virtual standardized patient teaching pathway – implemented with the open-source DeepSeek-R1 model – designed to provide standardized, replicable communication practice for early-stage clinical pharmacy trainees. We aimed to preliminarily evaluate its feasibility, score consistency, and user acceptance.

Methods

The teaching pathway was built 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. System performance was 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% LoA − 8.17 to 3.40). System 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 anxiety, 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 system 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

Clinical pharmacy is a patient-centered practice discipline. Patient-pharmacist communication, pharmacy rounds, medication education, and medication counselling are the core professional competencies of clinical pharmacists[1–3]. The *Expert Consensus on the Core Competency Framework of Chinese Clinical Pharmacist (2023)* specifically identifies communication and collaboration, patient care, and humanistic literacy as essential core competencies for clinical pharmacists. Its secondary indicator, “communication ability,” requires pharmacists to “ make good use of verbal and non-verbal communication skills for effective communication[4]. However, these competencies must be developed through repeated practice in a real-world scenarios. Junior clinical pharmacists, pharmacists undergoing standardized training, and clinical pharmacy postgraduate generally face practical challenges including insufficient contact with real patients, limited bedside practice opportunities, and lack of training scenarios in their early career development stage[5–7]. Unlike senior pharmacists who have gained experience through years of bedside practice, most novice trainees can only take on an observational role or engage in occasional, limited interactions, remaining in a state of “more watching and less doing.” Their communication skills mainly rely on self-directed exploration, resulting in a long training cycle and low efficiency. This challenge is constrained by limited availability of clinical patients, constraints on teaching time and space, and patient safety considerations[5,8]. Although this consensus establishes competency standards, does not provide a systematic training program. Thus, a key challenge for pharmacy education is to standardize communication training without placing additional burdens on clinical resources.

This challenge can be interpreted through Kolb’s experiential learning theory. According to this theory, which defines learning as “the process whereby knowledge is created through the transformation of experience,” mastering practical skills requires progressing through a cyclical process comprising concrete experience, reflective observation, abstract conceptualization, and active experimentation[9–11]. The concrete experience stage – direct interactions with patients – forms the essential foundation of learning, yet is notably lacking among most junior clinical pharmacists. Virtual patient (VP) simulation-based training can provide trainees with a safe, standardized, and repeatable training scenario, effectively addressing the shortcoming of limited contact with real patients in traditional clinical

training[12]. Therefore, how to safely provide standardized “concrete experiences” without relying on limited clinical resources is the key issue for pharmacy education reform.

Traditionally, standardized patients (SPs) have been employed to simulate doctor-patient interactions, providing an effective substitute for authentic patient encounters in clinical practice settings[13]. SPs use trained actors to simulate clinical communication, creating a realistic clinical teaching environments while supporting real-time interaction and immediate feedback[14]. Nevertheless, their widespread implementation faces considerable limitations, high training and implementation costs, insufficient standardization and consistency (often manifested as script deviations, fatigue-induced performance variability, and subjective bias), and scalability barriers in educational settings[14–17]. A systematic review by Hamilton *et al*.[12] points out that SPs also have inherent limitations such as the inability to simulate certain disease states, and their the training and deployment cost continue to rise, placing a significant financial burden on educational institutions. Internationally, the promotion of simulation-based teaching in pharmacy education also faces multiple obstacles, including time constrains, a shortage of qualified instructors and learner acceptance[18]. Early computer-based virtual patients (VPs) improved accessibility and consistency to some extent, but their pre-programmed, script-based interactions led to highly rigid dialogue. Due to limitations in natural language understanding and generation capabilities, they could not flexibly and naturally respond to learners’ spontaneous inquiries and complex clinical questions[19]. Additionally, early VPs systems were often hampered by prolonged development cycles and high costs[20].

Recent breakthroughs in large language models (LLMs) offer a completely new approach to resolving the aforementioned dilemma. LLMs are sophisticated deep learning systems for natural language processing. Being trained on massive text datasets, they operate through contextual prediction by analyzing preceding linguistic information, enabling them to understand and generate coherent, contextually appropriate human language responses[21,22]. Unlike previous scripted VPs, LLM-driven virtual standardized patients (VSPs) are capable of engaging in open, dynamic conversations, adapting to users’ diverse questions, simulating a range of patient styles, and providing real-time scoring and automated feedback based on evaluation criteria guided by structured prompt engineering. A scope review by Zeng *et al*.[23] systematically reviewed 28 LLM-VSP studies and found that 92.9% were published in the last two years, indicating that this field is currently undergoing rapid development. However, the review also noted that evaluation methods lack standardization, and only 13% of the studies used validated assessment tools, only 21.7% objectively measured learning outcomes, and the vast majority relied on closed-source commercial models (such as ChatGPT), which affects the repeatability and data security of the research. Moreover, current LLM-VSP studies mainly focus on medical history-taking training for medical students, there has been no systematic validation of teaching pathways specifically tailored to clinical pharmacy scenarios (such as medication history-taking, medication counseling, and medication consultation).

To address these gaps, this study developed a standardized teaching pathway based on LLM technology, tailored to pharmacy round abilities for early clinical pharmacists. For LLM engine selection, we chose DeepSeek-R1 – a free and open-source LLM – ensuring low-cost replication and data security, supporting on-premises deployment, and responding to the field's call for repeatability and privacy protection[24,25]. This pathway consists of three core components: (1) a Structured Case Library (three self-developed standardized cases for pharmacy rounds on the first day of a cancer chemotherapy cycle), covering key aspects of pharmaceutical care such as medication history-taking, medication regimens education, adverse reactions management, and medication consultation; (2) the SPADE framework (Scenario & Setting, Persona, Activity Flow, Discipline Rules, Evaluation), ensuring that the VSP simulates real patients in an anthropomorphic way and adheres to question and answer (Q&A) guidelines; and (3) RTF-based automated scoring prompts (Role, Task, and Format), used to provide instant, standard-based quantitative scores and qualitative feedback. Subsequently, we recruited 21 clinical pharmacy trainees (6 junior pharmacists and 15 graduate students) and conducted a feasibility study involving 30 rounds of simulated pharmacy ward round dialogues. This study aims to answer the following key questions:

How does this LLM-VSP teaching pathway perform in terms of the authenticity, clinical accuracy, and logicality of simulated pharmacy rounds dialogues?

How do the target trainees rate the acceptance, satisfaction, and practicability of this teaching pathway as a skills training tool ?

Does the prompt framework and case design of this teaching path demonstrate good transferability, and can they be applied to other LLM engines or training scenarios ?

By publicly reporting our prompt framework and case design protocols, we aim to prove that LLM-driven VSP teaching pathway is transferable across different LLM engines and clinical communication training scenarios, thereby providing a methodological template for adoption by other teaching institutions.

2 Result

2.1 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%

2.2 Expert Evaluation of LLM-based VSP System Performance

2.2.1 Quantitative Ratings

After the dialogues, the System Testing Expert Panel used the Expert Feedback Evaluation Scale to rate the VSP's performance and provide 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 and evaluation effectiveness 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
System Testing Expert Panel's Evaluation of LLM-based VSP System Performance

Assessment Dimension	No. of Items	Possible Score	Mean Score ± 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
System 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) × 100%.				

2.2.2 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 2.4, provide context for the quantitative ratings above.

2.3 Reliability of the Human Reference Standard and VSP Automated Scoring

2.3.1 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% 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.

2.3.2 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.

2.4 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.

2.5 User Experience Feedback from Trainees

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

2.5.1 Quantitative Ratings

The internal consistency reliability of 18 scale items was excellent (Cronbach's $\alpha = 0.93$), indicating that the reliability of the questionnaire was high. As shown in Table 4 and Fig. 4, participants' evaluations on each dimension of VSP are positive, and the scoring rate is between 81.19% and 89.52%. 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 system stability dimension.

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

Assessment Dimension	No. of Items	Possible Score	Mean Score ± 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
System 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) × 100%.

2.5.2 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”); 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”); lack of flexibility of real patients (“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 to avoid excessive extension (“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

3 Discussion

This feasibility study developed and preliminarily evaluated a standardized, LLM-based virtual patient teaching pathway for pharmaceutical care training, implemented with the open-source DeepSeek-R1 model and the SPADE prompt framework. The automated scoring showed a very strong correlation with expert ratings and trainees reported positive experiences, particularly in reducing communication anxiety. To our knowledge, this is the first study to validate such a pathway in clinical pharmacy education, using DeepSeek as an open-source, replicable LLM, thereby filling a gap in the existing literature.

3.1 Interpretation of key findings.

3.1.1 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 that matches the patients’ style. This is closely linked to the SPADE prompt framework, which strictly restricts DeepSeek to responding solely as an ordinary patient, strictly following the medical cases and avoiding medical terminology or role deviation. Meanwhile, VSP does not show emotional fluctuations or interaction stress, making trainees feel natural and relaxed. This is consistent with the user’s highest-rated item “I feel comfortable without tension during interactions” (4.52/5), suggesting

that the VSP's naturalistic style reduces trainee anxiety. This result aligns with the findings of Ahn and Jeong[9] in nursing virtual patient simulation, namely that high personification can significantly improve psychological safety and promote learners' active participation and repeated practice.

3.1.2 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 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.

Grevisse[26] 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 ($\rho = 0.934$) 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.

Future work should explore few-shot learning with expert-graded edge cases or calibrate the scoring prompt with explicit examples of patient-volunteered information to reduce this bias.

3.1.3 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 3.1.2, 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.

3.1.4 System Stability: Common Technical Challenges

In the participants’ feedback, the system 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 APIs; on-premises deployment can reduce network latency and service queuing.

These problems are common technical limitations among current cloud-based large model platforms, not defects in system design or the prompt framework. Oncu *et al.*[27] also observed technical challenges

such as delayed responses and server overload when using ChatGPT-4o as a VSP. Holderried *et al.*[28] also reported technology stability challenges: 1.4% of the conversation records were lost due to system downtime in their study. Importantly, these fluctuations did not affect the integrity of conversations, clinical accuracy, or scoring consistency. Core teaching functions remained stable in our study. With server optimization and on-premises deployment, these issues may be significantly improved in the future.

3.1.5 Trainee Acceptability: High Usability and Educational Value

Trainees' overall acceptance and satisfaction with LLM-based VSP were both at a high level, and the scores for each dimension ranging from 81% to 89% of the maximum possible score (Table 4). The questionnaire demonstrated excellent internal consistency (Cronbach's α = 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, their learning threshold for conversational learning systems is relatively low such as VSP. Second, virtual environment eliminates the mental stress of interacting with SPs, they can repeatedly practice and get instant feedback without worrying about making mistakes. 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 psychological safety fostered by the VSP, reflected in the highest-rated item "comfortable without tension" (4.52/5), may similarly reduce interpersonal barriers that often impede early-career trainees – with effects similar to the safe, reflective environment created by narrative medicine workshops in Lu *et al.*'s nursing study[29]. The above findings are consistent with the views of Öncü et al.[27], 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.

The results of this study can be explained by Kolb's Experiential Learning Theory[11]. LLM-based VSP provided a safe, standardized, and repeatable simulated environment for trainees to conduct medication history-taking, counseling, and consultation, offering concrete experience. The automated scoring and structured feedback facilitated reflective observation. Repeated practice allowed trainees to refine strategies and transfer skills to new scenarios, completing abstract conceptualization and active experimentation. Thus, LLM-based VSP effectively supports the full experiential learning cycle and is highly valuable for trainees lacking real-patient practice. This is consistent with the findings of Ahn and Jeong [9], who found that virtual patient simulations align well with Kolb's learning theory and promote the development of clinical skills through iterative experiential practice.

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

Prior LLM-VSP studies[20,28,30,31] 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. Without such decoupling, the fusion of these two roles may cause bias or self-contradiction in the evaluation results. 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.

3.3 Comparison with existing LLM-based VSP studies

Liu and colleagues[32] 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 colloquial, vivid, and accurate responses. Cook *et al*.[20] 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*.[27] 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 and raised data privacy concerns when processing sensitive medical information. 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 developed a VSP based on DeepSeek-R1, a free, open-source LLM that supports on-premises deployment, eliminates per-conversation costs, and ensures data security. The LLM-based VSP 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% of the total score). 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, supporting the utility of the LLM-based VSP for formative assessment.

Wang *et al*.[31] similarly employed prompt engineering to optimize ChatGPT as a VSP and conducted multi-metric assessment. However, their study focused on general medical training rather than pharmaceutical care, and they did not perform quantitative consistency analysis between AI and expert scoring. Our study extends their approach by targeting pharmacy-specific skills and providing rigorous validation of scoring reliability.

Compared with closed-source models such as ChatGPT, DeepSeek-R1 is open-source, free, and locally deployable. These features greatly reduce costs, eliminate per-session fees, and enhance data privacy and security for clinical scenarios involving sensitive medical data. These advantages make it more suitable for widespread use in hospitals and pharmacy education institutions.

3.4 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.

3.5 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) integrating various cues (e.g., vocal tone, facial avatars) to enhance authenticity and non-verbal communication; (4) conducting longitudinal randomized studies comparing VSP versus traditional SP training; and (5) leveraging DeepSeek's open-source architecture for on-premises deployment to eliminate network latency and address data privacy requirements in healthcare institutions.

3.6 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. This teaching pathway directly responds to three key challenges faced by early clinical pharmacists: insufficient contact with real patients, lack of standardized teaching, and lack of immediate feedback. 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, thereby effectively addressing the shortcomings of "more watching and less doing."

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. It can greatly reduce the reliance on SP protocols and clinical bedside resources. In turn, it promotes educational equity and training efficiency in pharmacy education, and may ultimately benefit the quality of pharmaceutical care and patient safety.

Conclusion

This study demonstrates 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 disclosed in this study provide a transferable methodology that other institutions can directly adopt.

Methods

4.1 Study Design

This single-center feasibility study aimed to develop and preliminarily evaluate a standardized, LLM-based virtual patient teaching pathway for training pharmaceutical care skills, including medication history taking, medication counseling, and medication consultation. 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.

4.2 Participants

This study involved three groups of participants:

4.2.1 System Development Expert Panel

Comprising three experienced senior oncology pharmacy specialists, the panel was responsible for the professional review and quality assurance of all core materials developed by the research team, including structured standardized cases, pharmacy ward round scoring criteria, core prompt engineering frameworks and LLM-based VSP evaluation scales. The panel's primary contribution involved providing key recommendations and guidance for improvement, ensuring the clinical accuracy and validity of all developed materials from a clinical standpoint.

4.2.2 System Testing Expert Panel

To independently evaluate the system performance, three additional senior clinical pharmacy specialists (with no overlap with the development expert panel) were invited as pilot users. This panel interacted directly with LLM-based VSP and provided both quantitative and qualitative feedback, which was used for optimization iterations of LLM-based VSP.

4.2.3 Trainee Participants

A convenience sampling method was used to recruit a small cohort of 21 participants for the evaluation study. The group 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 during their undergraduate studies. All participants provided informed consent before taking part in the study and completed a baseline characteristics questionnaire prior to interacting with the LLM-based VSP. The questionnaire was designed to collect demographic information (e.g., age, gender and identity), self-assessed experience in pharmaceutical care core skills (such as medical history taking, medication

counseling) and technical usage background (including frequency of AI tool usage and prior exposure to SPs). The collection of this baseline data serves to describe the overall characteristics of the study sample and to assess its potential influence on subsequent interactions and evaluations.

4.3 Development of LLM-VSP Teaching Pathway

4.3.1 DeepSeek-R1

In this study, DeepSeek-R1 was employed as the carrier LLM engine to implement the VSP-based teaching pathway. DeepSeek-R1 integrates an open-source framework centered on reinforcement learning, incorporating designs such as the Mixture-of-Experts (MoE) architecture and advanced attention mechanisms[33]. This design enables the model to deliver an exceptional performance-to-cost ratio when handling deep reasoning and structured problem-solving tasks[34]. It supports ultra-long-context interactions, with a maximum capacity of 128K tokens *(DeepSeek official documentation, available at*:https://api-docs.deepseek.com/zh-cn/news/news250528*)*, thereby facilitating comprehension and retention of lengthy dialogue histories and complex medical record details. Such a capability allows the LLM-based VSP system to memorize and process the entirety of doctor-patient communication during interactions, which is essential for simulating coherent and realistic patient conversations.The choice of DeepSeek-R1 as the implementation engine was driven by its open-source, free-to-use, and locally deployable nature[24,35], as well as our team's prior validation in pharmacy-related tasks, where it successfully integrated multi-source clinical information (dialogues, medical records, laboratory data) to generate structured SOAP-format medication logs, confirming its capacity for clinical data integration and structured output[36]. These features lower the technical and financial barriers for other institutions to replicate the teaching pathway using their own LLM instances, thereby supporting the goal of educational scalability and reproducibility.

4.3.2 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 systematically reviewed and collected 120 structured medication therapy logs documented by experienced clinical pharmacists over the past two years to ground the cases in real-world practice and enhance their clinical relevance. These logs encompassed multi-dimensional information covering various cancer types, treatment phases, and medication regimens.

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 containing no real patient information. All cases centered on the pharmacy ward round scenario on the first day of chemotherapy for cancer patients and strictly followed a unified interaction protocol: (1) medication history-taking, (2) medication counseling, and (3) medication consultation.

To ensure focused assessment, the cases maintained high consistency in core disease complexity and assessment focus, with an emphasis on evaluating key pharmaceutical care competencies, including medication history-taking, chemotherapy regimen education, adverse reaction management, and medication consultation.

To ensure clinical diversity and training breadth while maintaining comparable difficulty, the three cases were designed with standardized complexity in core tasks, with controlled variations in cancer type, chemotherapy regimen, and consultation focus. This design allowed assessment of generalizability 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+ Etoposide+ Lobaplatin	Serplulimab+ Paclitaxel+ Nedaplatin	Sintilimab+ Oxaliplatin + S-1 (Tegafur, Gimeracil and Oteracil Potassium)
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, ensuring individualized assessment frameworks tailored to the specific content and learning objectives of each case. 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 System Development Expert 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.

4.3.3 Structured Prompt Engineering

The development of the VSP's dialogue interaction and automated scoring capabilities was systematically guided by a structured prompt engineering process. This study drew upon prior research in medical education that employed prompt engineering to build LLM-based VSPs[15,19,28,31,37]. In addition to learning from domain-specific applications, the study also incorporated general design principles derived from established prompt engineering frameworks, including the RTF (Role-Task-Format), CRISPE (Capacity-Role-Insight-Statement-Personality-Experiment), and BROKE (Background-Role-Objectives-Key Results-Evolve)[30,38].

Through multiple rounds of simulated dialogues (n = 20) and iterative refinement, we developed two dedicated prompt systems tailored for pharmacy ward round scenarios: one for dialogue generation and another for automated evaluation. This iterative process aimed to resolve common issues identified in early versions and reported in similar studies[20,28,30,31], including premature answers, excessive length, over-extended responses, non-interactive monologues, irrelevant answers, and disrupted conversation flow. The prompt refinement strategy was guided by our team's prior work on pharmaceutical prompt engineering. Sun *et al*.[39] demonstrated that structured chain-of-thought prompting (RTOG-CoT) 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 systems with substantially enhanced stability and authenticity.

Dialogue Generation Prompt: The SPADE Framework

To systematically guide the VSP's conversational behavior and ensure clinically authentic interactions, we developed the SPADE framework, which comprises five components: Scenario & Setting, Persona, Activity Flow, Discipline Rules, and Evaluation (detailed in Fig. 2).

Automated Scoring Prompt: The RTF-based Prompt

To ensure the reliable execution of the Evaluation task defined in the SPADE, a separate scoring prompt was developed based on the RTF model. This prompt explicitly defines the role of the DeepSeek as an "examiner," mandating it to adhere to precise scoring logic, output formats, and rules such as requiring evidence-based deductions directly quoted from the dialogue text. This design was crucial in guaranteeing the consistency, objectivity, and verifiability of the automated scoring outcomes. This scoring prompt underwent an iterative optimization process similar to that of the dialogue prompt. The final prompts for dialogue generation and automated scoring are provided in Supplementary Material S1 and S2, respectively.

4.3.4 Development and Validation of Assessment Tools

To comprehensively evaluate the performance, acceptability, and replicability of the LLM-VSP teaching pathway, this study developed and employed two dedicated 5-point Likert scales. The design of both scales was informed by the Technology Acceptance Model (TAM) and existing virtual patient evaluation studies, adapting their core dimensions and frameworks with tailored modifications for this study's context[9,22,40,41]. Both scales were reviewed by the expert evaluation panel for content validity, ensuring the suitability of the assessment dimensions and the clarity of each item.

The Expert Feedback Evaluation Scale comprises 15 items and was designed to gather feedback from pilot users – the System Testing Expert Panel. It focused on assessing LLM-based VSP's performance across multiple dimensions, including clinical accuracy, logical consistency, anthropomorphism, system stability, as well as feedback quality and evaluative effectiveness, thereby supporting the optimization of both the prompt engineering and overall system performance.

The User Experience Feedback Questionnaire comprises 20 items, including 18 five-point Likert-scale questions and 2 open-ended questions. It was completed by trainees to systematically gather their subjective feedback across multiple dimensions, such as accuracy, logical consistency, anthropomorphism, system stability, emotional acceptance and trust, as well as teaching/assessment value. The main purpose of the questionnaire is to evaluate the acceptability, practicality, and user satisfaction of the LLM-based VSP as a tool for pharmacy education.

The full item lists and detailed scoring anchors for both the Expert Feedback Evaluation Scale and the User Experience Feedback Questionnaire are provided in Supplementary Material S4 and S5, respectively.

4.4 Expert Evaluation and Refinement

During the late development stage, three experts from the System Testing Expert Panel acted as pilot users engaging in complete pharmacy ward round dialogues with all three VSPs (totaling 9 dialogues). All dialogues were completed and recorded online in a text-based format.

Following each dialogue, the experts used the Expert Feedback Evaluation Scale to rate the VSP's performance and provide qualitative feedback. Thematic analysis of this feedback revealed several recurring issues, which guided a final round of refinements to the prompt engineering framework:

1. **Over-answering**: The VSP occasionally provided excessive information beyond the scope of the expert's questions. 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**: At times, the VSP generated dialogues simulated both pharmacist and patient at the same time, which violated the required question-and-answer format. To address this, we added stricter instructions to the prompt to asked the VSP to respond only as the patient.
3. **Response latency**: Intermittent delays were observed during the study, these delays mainly attributed to the limitations of DeepSeek web service performance at that time, as well as the

model's inherent language processing time for deep thinking and context recalling, which is not a controllable factor nor a defect of our prompt design. In order to reduce the impact of delays on the user experience, we instructed participants to allow a few seconds for VSP to respond. Importantly, this issue did not affect the completeness or integrity of the recorded dialogue data, and has since been largely resolved with subsequent DeepSeek platform updates.

4. **Scoring inaccuracies**: Early versions of the scoring prompt exhibited several problems: (a) it occasionally fabricated unasked questions as score points; (b) it sometimes erroneously used case file information as scoring source; and (c) the scoring criteria are too broad, leading to inconsistent ratings. We therefore (a) explicitly limited the source of scoring to the text of the dialogue record and prohibit the fabrication of questions that the pharmacist did not ask; (b) clearly distinguished between background case information and the conversation to be evaluated and (c) we further refined the Modified Pharmacy Ward Round Scoring Criteria to reduce ambiguity.

These iterative refinements led to the final version of the VSP used in the subsequent participant evaluation stage.

4.5 Data Collection and the Establishment of the Reference Standard

4.5.1 Participant Interaction and Feedback Collection

Following expert validation and optimization, the study proceeded to the participant evaluation stage to examine the effect of LLM-based VSP in an authentic training scenario.

Each trainee completed 1–3 simulation scenarios according to training availability, generating a total of 30 valid, independent dialogue records for analysis. The dialogues were conducted either via the DeepSeek website on a computer (with text input/output) or the DeepSeek APP on an iPad (with iOS speech-to-text input), with all interactions fully recorded in text format.

After each dialogue, the pre-defined structured scoring prompt was sent to DeepSeek three times independently (with no conversation memory carried over between runs). The VSP generated three quantitative scores and three qualitative feedback reports per dialogue. The mean of the three scores was calculated for each dialogue and used as the final automated score. In total, 30 mean scores (one per dialogue) were collected for consistency analysis. Subsequently, each participant completed the User Experience Feedback Questionnaire after their dialogues to comprehensively assess the acceptability, practicality, and satisfaction of LLM-based VSP among the participating trainees. To avoid performance bias, individual VSP-generated scores and textual feedback were released to participants only after the completion of all study procedures.

4.5.2 Establishment of the Human Reference Standard

The human scoring, performed by the same three experts from the System Testing Expert Panel, served as the reference standard (ground truth) for participant performance. The experts' manual scores were

also based on the Modified Pharmacy Ward Round Scoring Criteria. This intentional design choice was made because the experts' prior exposure to the VSP during the testing phase provided them with a comprehensive understanding of the LLM-based VSP's conversational patterns and limitations – enabling them to better discern whether information gaps stemmed from trainee inadequacy or system constraints.

Before formal scoring, three experts completed a brief calibration exercise. They first received written instructions and key considerations, then independently scored three pilot dialogues (which were different from the 30 study dialogues). The results were compared and discussed to align their understanding, though no formal recalibration of scores was performed.

To assess inter-rater reliability (IRR) among the three experts, the intraclass correlation coefficient (ICC) was calculated using a two-way random-effects model with absolute agreement, based on the 90 independent scores (three experts × 30 dialogues). Both single-measure ICC [ICC(2,1)] and average-measure ICC [ICC(2,3)] were calculated, with ICC(2,1) assessed the consistency of individual expert scores, while ICC(2,3) assessed the reliability of the mean expert score as the reference standard. The arithmetic mean of the three experts' scores for each dialogue record was then taken as the benchmark for consistency analysis with VSP's automated scoring.

4.5.3 VSP Automated Scoring Reliability

For the VSP automated scoring, the same three independent inferences per dialogue were used to calculate the repeatability. The ICC was computed using a two-way random-effects model with absolute agreement, with ICC(2,1) assessing the consistency of a single inference, while ICC(2,3) assessing the reliability of the mean of three inferences. These ICC values quantify the stability of the VSP's automated scoring across repeated inferences.

4.6 Statistical Analysis

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

4.6.1 Data Collection and Types

The following four types of data were collected.

(1) **Baseline data:** Demographic information, self-assessed experience in pharmaceutical care core skills and technical usage background from 21 participants.

(2) **VSP Proficiency Feedback Data:** Quantitative ratings from the 15-item Expert Feedback Evaluation Scale and qualitative textual feedback provided by the System Testing Expert Panel.

(3) **Users Experience Data:** Quantitative ratings from the 18-item User Experience Feedback Questionnaire and qualitative textual feedback from two open-ended questions provided by 21 participants.

(4) **Scoring Consistency Data:** Two sets of scores corresponding to 30 valid dialogue records: the automated scores generated by LLM-based VSP and manual scores assigned by the System Testing Expert Panel.

4.6.2 Statistical analysis methods

All quantitative statistical analyses were performed using IBM SPSS Statistics (R27.0.1.0), with statistical significance level set at $\alpha = 0.05$ (two-tailed). All data are presented as mean ± standard deviation (s.d.) unless otherwise specified. No adjustments for multiple comparisons were required, as all analyses involved single pairwise comparisons or reliability metrics, and no repeated group-wise comparisons were performed.

(1) **Descriptive Statistics**

Continuous variables were described using mean ± standard deviation or median (interquartile range). Categorical variables were described using frequency and percentage (n, %). This analysis was applied for baseline data, all scale scores, and the scores from the VSP and experts. The normality of distribution for continuous variables was assessed using the Shapiro-Wilk test.

(2) **Reliability Analysis**

Inter-Rater Reliability (IRR): To assess the consistency of scores among the three experts, the ICC was calculated using a two-way random-effects model was used to calculate absolute agreement, based on the 90 independent scores (three experts × 30 dialogues). Both single-measure ICC [ICC(2,1)] and average-measure ICC [ICC(2,3)] were calculated. An ICC value≥0.80 was considered indicative of good reliability.

Repeatability of VSP automated scoring: For the three independent scoring runs per dialogue, the ICC was calculated using a two-way random-effects model with absolute agreement. Both single-measure ICC [ICC(2,1)] and average-measure ICC [ICC(2,3)] were reported to evaluate the stability of the VSP scores. An ICC value≥0.80 was considered indicative of good repeatability.

Internal Consistency Reliability of the Scale: For the Expert Feedback Evaluation Scale and the User Experience Feedback Questionnaire developed in this study, Cronbach's α coefficients were calculated to assess internal consistency. An α value > 0.7 was considered acceptable.

(3) **Validity and Consistency Analysis**

Correlation Analysis: The correlation between the mean of the three VSP total scores and the mean of experts’ scores was assessed using either Pearson's or Spearman's correlation coefficient, contingent upon the normality of the data distribution as evaluated by the Shapiro-Wilk test.

Consistency Analysis: Bland-Altman analysis was employed to create Bland-Altman plots, visually displaying the mean difference and 95% consistency limits between the two scoring methods, thereby

assessing the range of consistency and the degree of bias. No pre-defined acceptable limits were set, as this feasibility study aimed to describe the pattern of agreement rather than to make a binary judgment of interchangeability.

(4) **Qualitative Data Analysis**

The open-ended textual feedback provided by experts and trainee participants was processed using thematic analysis. Two researchers independently coded to identify recurring key themes and sub-themes. Subsequently, comparative discussions were held until consensus was reached.

Declarations

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files.

Author contributions

Z.Z. and Y.Z. contributed equally to this work. Z.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

This work was supported by the Jiangsu Pharmaceutical Association Research Project (202564052) and the “Yishan Plan” Project for the High-level Hospital Construction of Jiangsu Cancer Hospital (YSZD202406).

Competing interests

The authors declare no competing interests.

References

1. Wang, W., Wang, Y., Shi, X., Zhu, J. & Chen, R. Development of pharmaceutical care competency framework in Chinese hospital pharmacist: a qualitative study in grade a tertiary hospitals. *BMC Med Educ* 25, 1143 (2025). https://doi.org:10.1186/s12909-025-07615-8
2. Bajis, D., Chaar, B., Basheti, I. A. & Moles, R. Pharmacy students' medication history taking competency: Simulation and feedback learning intervention. *Curr Pharm Teach Learn* 11, 1002–1015

(2019). https://doi.org:10.1016/j.cptl.2019.06.007

3. Tadesse, Y. B., Sendekie, A. K., Mekonnen, B. A., Denberu, F. G. & Kassaw, A. T. Pharmacists' Medication Counseling Practices and Knowledge and Satisfaction of Patients With an Outpatient Hospital Pharmacy Service. *Inquiry* 60, 469580231219457 (2023). https://doi.org:10.1177/00469580231219457
4. Chinese Academy of Medical, S., Peking Union Medical College, P. U. M. C. H. & China Medical Board, C. H. A. P. S. C. Expert Consensus on the Core Competency Framework of Chinese Clinical Pharmacist (2023). *Medical Journal of Peking Union Medical College Hospital* 14, 257–265 (2023). https://doi.org:10.12290/xhyxzz.2023-0092
5. 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
6. 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
7. 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
8. 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
9. 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
10. 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
11. 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
12. 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
13. Yi, Y. & Kim, K. J. The feasibility of using generative artificial intelligence for history taking in virtual patients. *BMC Res Notes* 18, 80 (2025). https://doi.org:10.1186/s13104-025-07157-8
14. 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

15. Rädel-Ablass, 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
16. 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
17. 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
18. 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
19. 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
20. 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
21. 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
22. Tu, T. *et al.* Towards conversational diagnostic artificial intelligence. *Nature* 642, 442–450 (2025). https://doi.org:10.1038/s41586-025-08866-7
23. 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
24. Shen, T. *et al.* Rapid deployment of large language model DeepSeek in Chinese hospitals demands a regulatory response. *Nat Med* (2025). https://doi.org:10.1038/s41591-025-03836-y
25. Chen, Y., Shen, J. & Ma, D. DeepSeek's impact on thoracic surgeons' work patterns-past, present and future. *J Thorac Dis* 17, 1114–1117 (2025). https://doi.org:10.21037/jtd-2025b-04
26. 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
27. 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
28. 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
29. 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

30. 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
31. 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
32. 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
33. Jin, I. *et al.* DeepSeek vs. ChatGPT: prospects and challenges. *Front Artif Intell* 8, 1576992 (2025). https://doi.org:10.3389/frai.2025.1576992
34. Peng, Y. *et al.* From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare. *J Biomed Inform* 163, 104791 (2025). https://doi.org:10.1016/j.jbi.2025.104791
35. Błecha, Z. *et al.* Performance of GPT-4o and DeepSeek-R1 in the Polish Infectious Diseases Specialty Exam. *Cureus* 17, e82870 (2025). https://doi.org:10.7759/cureus.82870
36. 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
37. 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
38. Ding, H. *et al.* Evaluation and practical application of prompt-driven ChatGPTs for EMR generation. *NPJ Digit Med* 8, 77 (2025). https://doi.org:10.1038/s41746-025-01472-x
39. 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
40. 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
41. 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

Figures

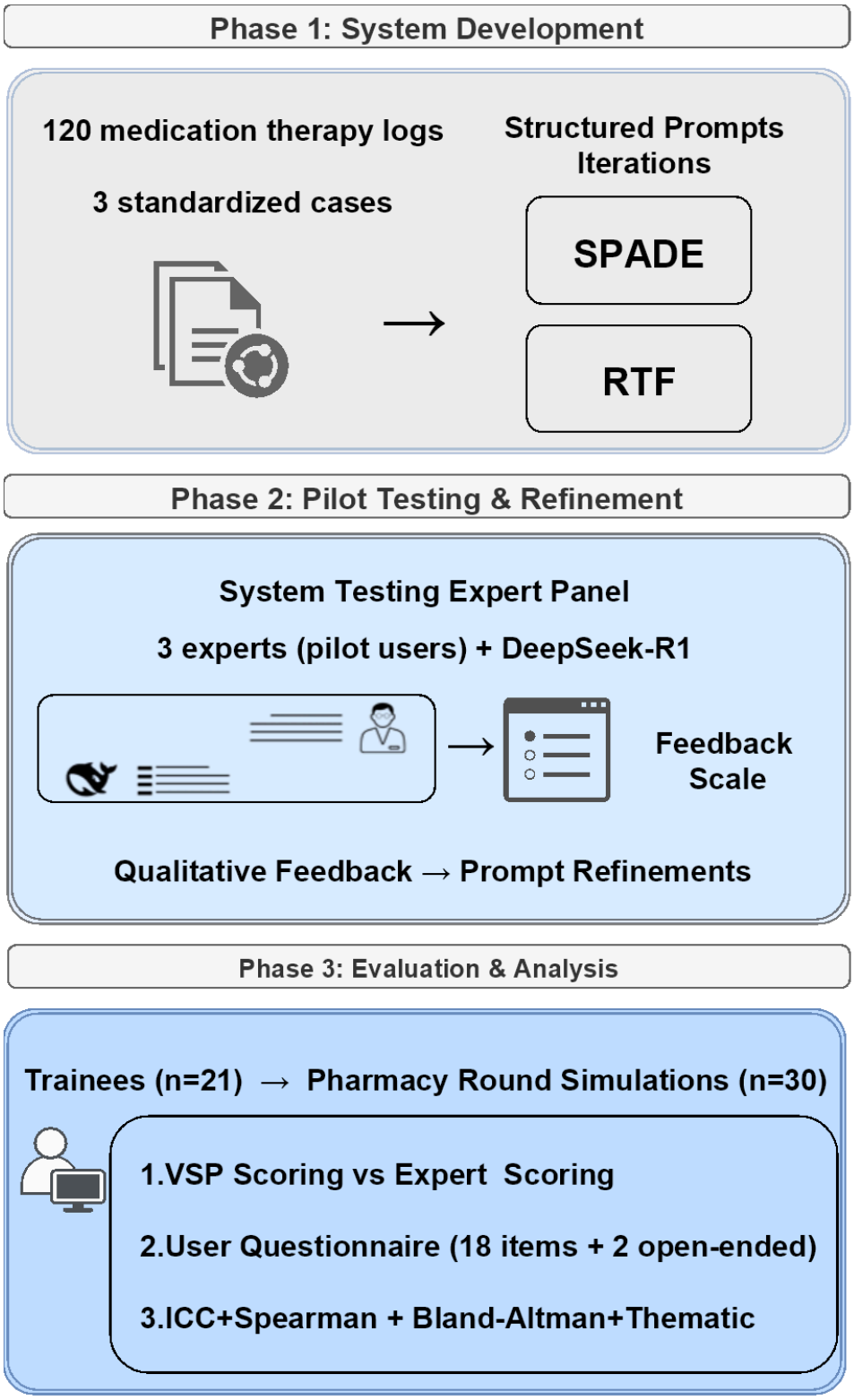


Figure 1

Overview of LLM-based VSP study workflow. The study consisted of three stages: (1) System 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 system; 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.

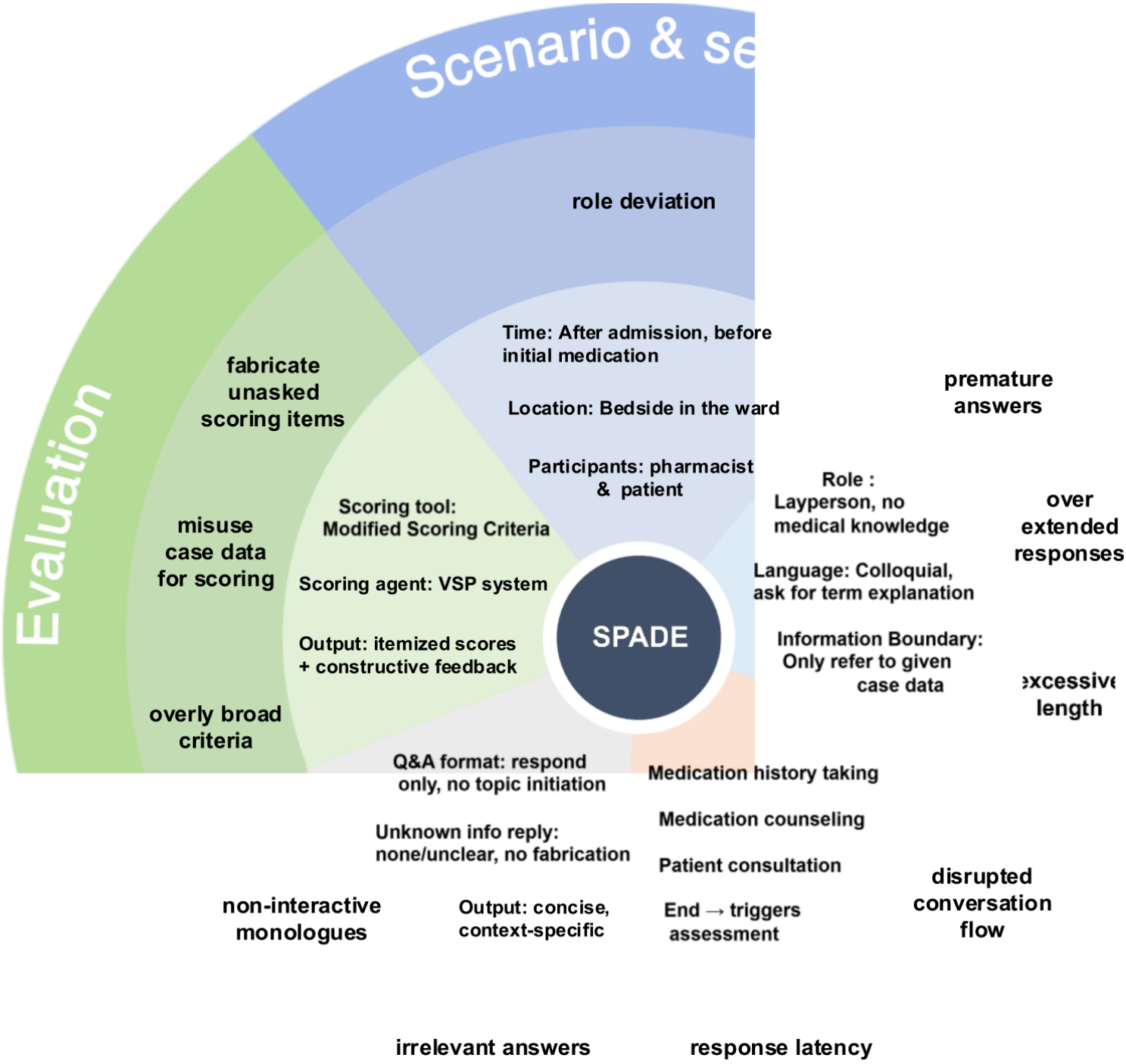


Figure 2

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.

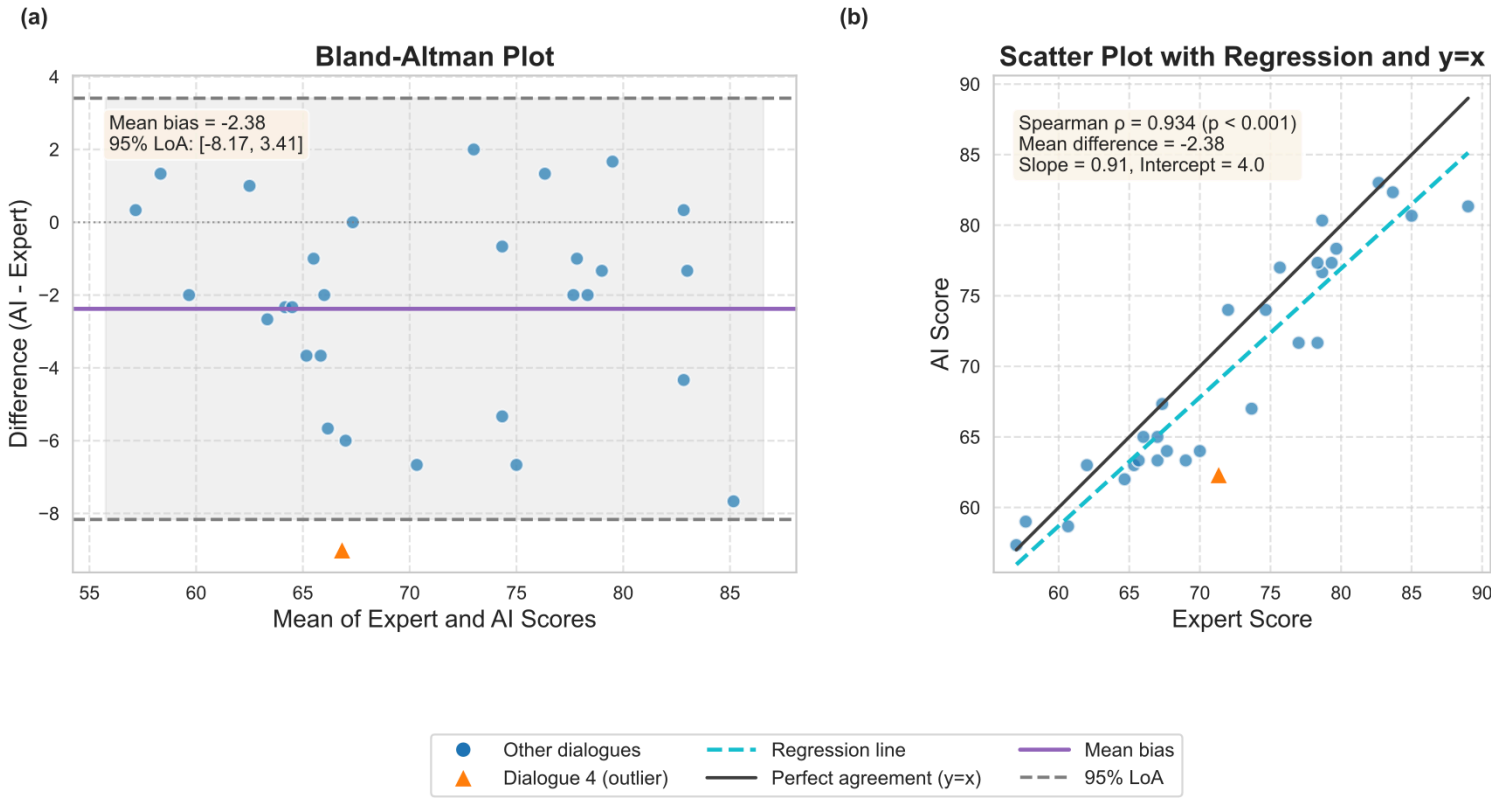


Figure 3

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 ñ = 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.

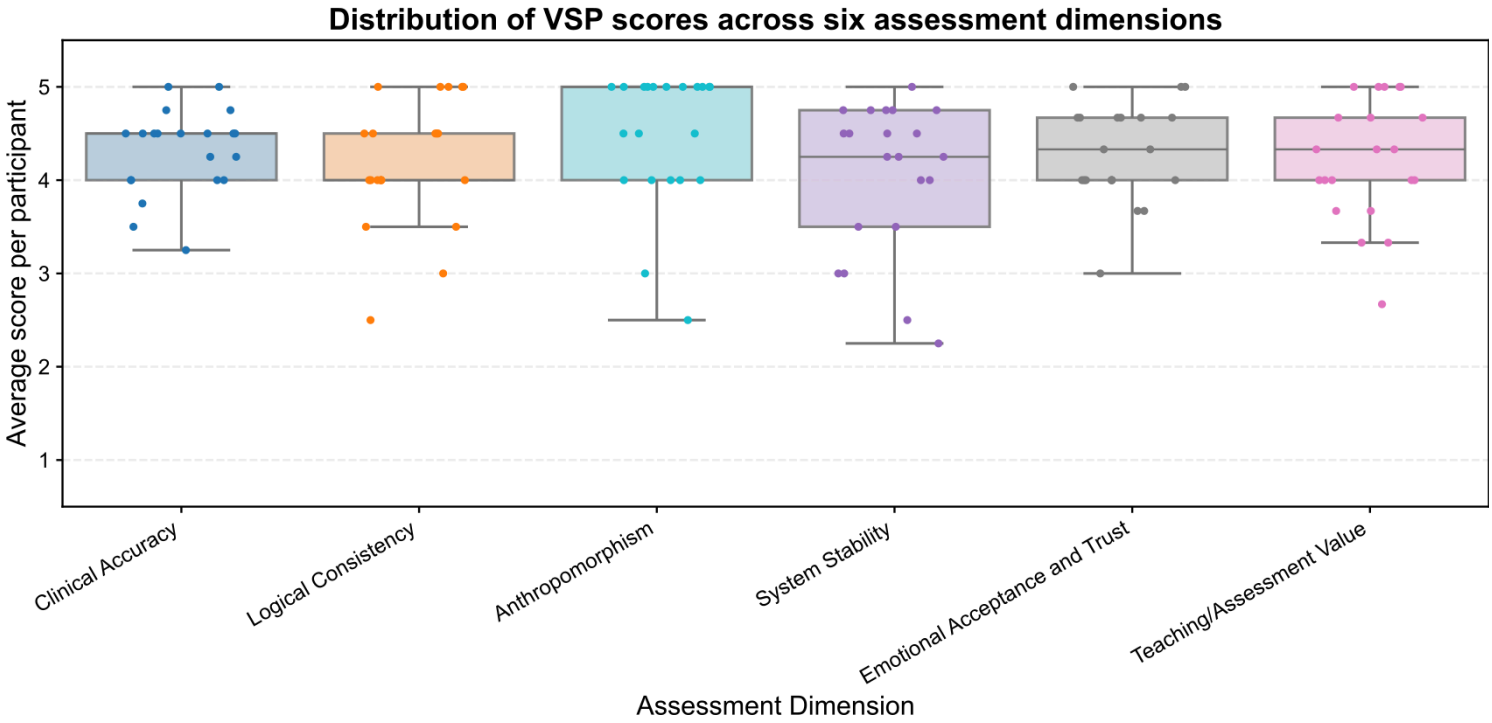


Figure 4

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; System 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