Original Paper

Large Language Model–Powered AI-Assisted Simulated Patient for Nursing Education: Quasi-Experimental Study

Max Chun Ming Chan, BEng; Danny Chun Kit Hung, MNurs; Tyrone Tai-On Kwok, PhD; Janet Yuen Ha Wong, PhD

School of Nursing and Health Sciences, Jockey Club Institute of Healthcare, Hong Kong Metropolitan University, Hong Kong SAR, China (Hong Kong)

Corresponding Author:

Tyrone Tai-On Kwok, PhD
School of Nursing and Health Sciences, Jockey Club Institute of Healthcare, Hong Kong Metropolitan University
11th Floor, 1 Sheung Shing Street, Homantin, Kowloon
Hong Kong SAR
China (Hong Kong)
Phone: 852 39702944
Email: ttokwok@hkmu.edu.hk

Abstract

Background: Large language model (LLM)–powered virtual patient systems are gaining popularity in nursing education but also face persistent challenges, such as hallucinations and role inconsistency.

Objective: This paper reports the model architecture design, technical considerations, implementation challenges, and evaluation of an LLM-powered AI-assisted simulated-patient system integrated into a nursing triage simulation game.

Methods: The system integrated a Unity-based game client, with GPT-4o hosted through Microsoft AI Foundry to support real-time simulated-patient dialogue in Traditional Chinese using Cantonese syntax. This developmental study comprised exploratory model selection, formative student feedback–informed enhancement from Version 1 to Version 2, development-stage script-fidelity evaluation, and evaluation of system usability and the Net Promoter Score in 2 tutorial cohorts of nursing students.

Results: Key technical contributions include a layered hallucination mitigation architecture that combines grounded prompt inclusion, few-shot conversational exemplars, negative role constraints, and calibrated model parameters. In the predefined script-fidelity evaluation, script-inconsistent responses occurred in 78 of 125 (62.4%) Version 1 outputs and 21 of 125 (16.8%) Version 2 outputs—an absolute difference of 45.6 percentage points. The implementation process also addresses stateless API memory management, cloud-based safety filter calibration for domain-specific clinical prompts, and prompt engineering strategies to prevent role drift and hidden scenario information leakage. GPT-4o was selected as the production model through exploratory development testing that considered script consistency, Cantonese-language suitability, response latency, technical reliability, deployment feasibility, cost considerations, and Unity integration compatibility; 45 of 50 (90%) assessed GPT-4o responses were rated fully script-consistent.

Conclusions: This study demonstrates the development and real-world tutorial deployment of a Cantonese-language, LLM-powered simulated-patient system for nursing triage education. Integrating formative student feedback with structured technical debugging enabled iterative refinement of dialogue fidelity, role consistency, protected-information handling, and game interaction functions. The resulting Version 2 system provides a practical, contextually tailored approach to supporting scalable simulated-patient dialogue in nursing education.

JMIR Nursing 2026;9:e98543; doi: 10.2196/98543

Keywords: large language model; AI; AI simulation; nursing education; virtual simulation

Introduction

Background

Traditional simulated patients have long been recognized as an important pedagogical strategy in nursing education, providing a safe and structured environment in which students can develop clinical judgment and psychomotor skills without compromising patient safety. Scenario-based simulation also supports the development of communication skills, teamwork, and interprofessional collaboration. Despite these advantages, the use of simulated patients is resource-intensive and often constrained by limited availability, logistical demands, and variability in the quality and consistency of feedback [1]. As a result, simulated patients are commonly used as a supplementary rather than primary teaching approach in nursing education.

In recent years, the rapid iteration of AI, specifically in large language models (LLMs) such as OpenAI's GPT Models, is fundamentally reshaping this landscape [2]. These LLMs were trained with massive datasets to understand and synthesize nuanced, human-like dialogue and are now powering a new generation of dynamic virtual patients. This technology provides a highly accessible and scalable alternative to traditional methods, transforming how nursing students develop clinical competencies [3]. LLM-powered apps are powerful enough to simulate patient interactions, provide automated feedback, and support personalized learning experiences that were previously limited by human and material resource constraints [4].

Applications in Health Care Education

LLMs are increasingly applied in health care education to enhance personalization, accessibility, and efficiency, though their implementation remains largely experimental [5]. Concept analyses identify LLMs as innovative and interactive tools supporting student learning, skill development, and teaching quality [6]. Current applications of LLM-powered simulated patients focus mainly on history-taking, nurse-patient communication, and interprofessional training [7]. Systems such as the AI-Powered Medical History-Taking Training and Evaluation System (AMTES) and large language model–based digital patient (LLMDP) enable dynamic, context-sensitive interactions, automated feedback, and performance tracking [8,9], while ChatGPT-driven and virtual reality–based platforms offer realistic, self-paced practice environments [10,11]. Common design features include persona consistency, context-aware responses, structured feedback, and unlimited practice opportunities. Students generally describe these environments as realistic and low-stress, valuing immediate feedback for competency development [10,11].

Evidence of Effectiveness and Limitations

A scoping review of 28 studies found positive attitudes toward LLM-powered virtual patients, with improvements in communication skills, clinical reasoning, and engagement [7]. AI patients often perform comparably to standardized actors, with added benefits of accessibility and scalability. However, most existing studies are small-scale, focusing on perceptions rather than measurable outcomes, limiting generalizability [12]. Effective deployment requires precisely crafted prompts through iterative pilot testing and progressive refinement [13].

Technical and Content Limitations

Current systems are constrained by AI hallucinations undermining content reliability, algorithmic biases, and prompt sensitivity affecting output consistency [14]. Inconsistent interaction patterns include difficulty handling input outliers, medically implausible responses, and fluctuating cross-model performance. These inconsistencies are attributed to restricted patient profile diversity, absent fact-checking mechanisms, and nonstandardized evaluation frameworks [15,16]. Furthermore, the "black box" nature of LLM decision-making complicates clinical safety evaluation [17].

Emerging mitigation strategies show promise. In a systematic review of 39 studies, Li and Lebai Lutfi [18] found that mitigation methods have strong performance in internal medicine and mental health but gaps in multimorbidity and rare disease coverage. Knowledge graph integration (improving top-k accuracy by 16.02%) and targeted fine-tuning (reducing hallucination to 0.31%-5%) enhance realism, though not yet consistently implemented. Most systems are built on OpenAI's GPT family [7], with model temperature frequently adjusted to balance randomness and consistency [19,20]. Aster et al [21] found no significant difference in empathic responses at lower (0.2) vs higher (0.8) temperature settings, suggesting informational accuracy may not be the primary determinant for empathy-focused tasks, unlike history-taking where factual correctness is central. Hua et al [22] emphasized the need to examine how backend model updates and stochastic sampling factors affect output stability over time.

Pedagogical and Implementation Challenges

Further challenges include faculty readiness, lack of structured pedagogical frameworks, data privacy, ethical concerns, and resource allocation [15,23,24]. Many LLM-driven systems provide rich interactions but lack an instructional feedback loop essential in simulation-based learning [15]. Language accessibility further constrains adoption; most LLMs are trained with English-based datasets and thus not optimized for local linguistic or cultural contexts [18,25].

Research Gap

LLM-powered simulated patients show considerable potential in advancing clinical training; however, existing LLM-driven systems continue to face limitations in content accuracy, clinical depth, and occasional hallucinations, necessitating meticulous prompt design and continuous expert supervision. Previous evaluation studies have largely been small, exploratory, and perception-based, providing limited empirical validation and generalizability. These shortcomings highlight the need for technically robust, pedagogically grounded, and contextually relevant innovations.

The aim of this study was to develop and evaluate a Cantonese-language, LLM-powered simulated-patient system integrated into a nursing-triage simulation game. Specifically, the study aimed to (1) select a feasible model configuration for real-time simulated-patient dialogue, (2) use formative student feedback and technical testing to guide Version 2 enhancement, (3) evaluate script fidelity of the integrated Version 2 configuration, and (4) assess student-reported usability and recommendation outcomes across tutorial cohorts.

Study Design

This 2-phase developmental study used a quasi-experimental design to evaluate student satisfaction and game usability. Phase 1 involved system development from 2024 to February 2025, including clinical scenario development, exploratory model-fit assessment, and selection of a production model configuration. Phase 2 involved Version 1 tutorial deployment in February 2025, formative student feedback and internal technical testing, Version 2 enhancement from February to September 2025, development-stage script-fidelity evaluation, and student usability evaluation during Version 2 tutorial deployment in February 2026.

Sample and Setting

The study sample comprised nursing undergraduates who were enrolled in an emergency nursing theoretical course at an applied university in Hong Kong, possessed a mobile device with access to the game, and were first-time users of AI nursing simulations. Students with prior internship experience in an emergency department were excluded. Based on the effect size reported for AI integration in a similar study design [26], a minimum total sample size of 52 students was required to achieve a statistical power of 0.80 at a significance level of .05. A convenience sample of 226 nursing students in the 2025 academic year and 296 nursing students in the 2026 academic year were voluntarily recruited to participate in the study.

Recruitment and Data Collection

Convenience sampling was applied for recruitment in these 2 cohorts of students via public announcement in class. In the process of informed consent, students were informed about the voluntary nature and the objectives of the study. A written information sheet about the study, questionnaire, and consent form were distributed to the students. Both cohorts of students would complete the online postintervention questionnaire after participating in the 3 cases of AI Triage simulation.

Instruments

Two instruments were used in this study to evaluate the students' experience toward the AI Triage simulation. The first one was the System Usability Scale (SUS) [27], which assesses effectiveness, efficiency, and satisfaction toward the educational tool through 10 statements on a 5-point Likert scale (1="strongly disagree" to 5="strongly agree"). The total SUS score ranged from 0 to 100, with higher scores indicating better usability. The SUS demonstrated high internal consistency, with Cronbach α exceeding 0.90 in previous studies [28]. The second one was the Net Promoter Score (NPS) [29], which was used to evaluate the students' experience and satisfaction. NPS has become common in medication education research to serve as an indicator to assess the value of an educational intervention [30].

Study Analysis

Students' experience was analyzed using IBM SPSS Statistics (version 30; IBM Corp). Demographic variables (sex and employment) were summarized as frequencies and percentages, and group differences were examined using χ^2 tests. Questionnaire data are reported as means and SDs. A quasi-experimental within-participants design was used to evaluate 2 versions among cohort 2025 students, and a between-participants design was used to evaluate the first experience of the 2 versions among 2 cohorts. As normality was assessed using the Kolmogorov-Smirnov test, between-group comparisons were conducted using the Mann-Whitney U test.

Methods

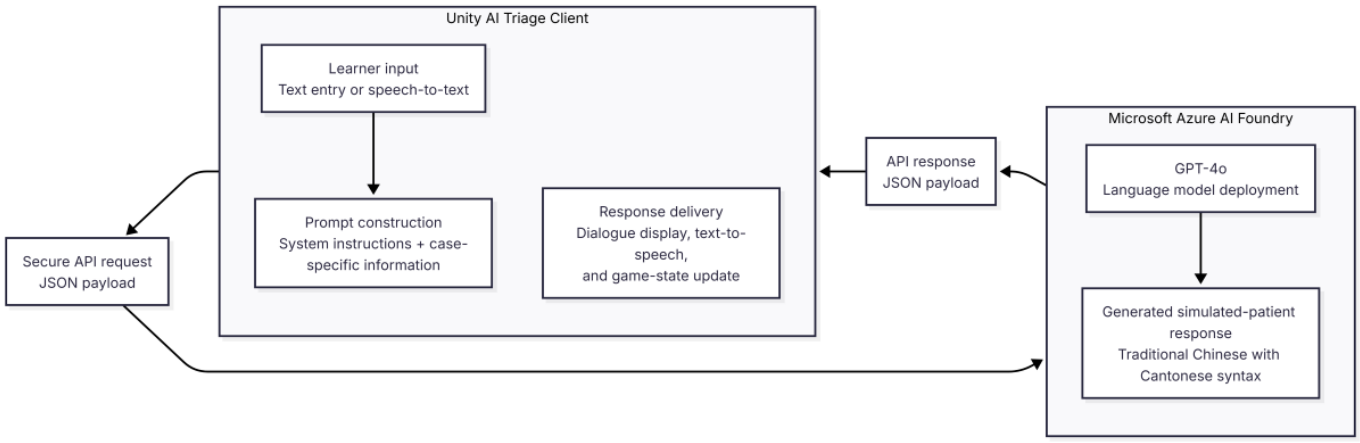
Phase 1: Model Development and Model-Fit Assessment

Overview

The AI Triage system was developed as an LLM-powered simulated-patient component integrated into a Unity-based nursing-triage simulation game. The Unity client captured learner input through text entry or speech-to-text and transmitted requests containing system instructions and case-specific information to GPT-4o hosted through Microsoft AI Foundry on Azure infrastructure. Generated responses were returned to the Unity client for display and use in game-state functions.

The system was designed to support real-time simulated-patient dialogue in Traditional Chinese characters using Cantonese syntax appropriate to local nursing-triage scenarios. Figure 1 presents the high-level system architecture. Detailed request-processing and deployment specifications are provided in Multimedia Appendix 1.

Figure 1. High-level architecture of the AI Triage simulated-patient system.



Scenario Development and Clinical Review

Eight base nursing-triage scenarios were developed by the research team and clinical educators. The scenarios were then reviewed by external clinical experts and Hong Kong Hospital Authority nursing members for clinical accuracy, relevance to local nursing-triage practice, and alignment of patient scripts with real-world triage processes. The approved scenario scripts specified patient demographics, presenting symptoms, symptom onset and progression, relevant medical and medication history, allergies, travel or exposure history where applicable, and information that should remain unavailable from the patient perspective.

Four scenarios were selected for the exploratory model-selection assessment because they represented the intended range of simulated-patient interactions and triage-history-taking requirements. The remaining scenarios were retained for use in the broader simulation game and subsequent development activities.

Exploratory Model-Selection Assessment

Exploratory model-selection testing was conducted during system development in 2024 to identify a model configuration suitable for Cantonese nursing-triage simulated-patient dialogue. A standardized set of 50 test questions was developed from the approved scenario scripts. Ten questions assessed general triage-history-taking domains applicable across scenarios, including presenting concern, symptom characteristics, symptom timeline, relevant medical history, medication or allergy history, and relevant exposure information. Each of the 4 selected scenarios contributed 10 scenario-specific questions designed to assess maintenance of scripted patient details and responses to clinically relevant triage inquiries.

The same 50-question test set was administered to 6 candidate model configurations: GPT-3.5-turbo, Llama 3.1, GPT-4, GPT-4o, DeepSeek-V2.5, and Amazon Titan. Responses were generated using scenario-specific system instructions and assessed against the corresponding approved case script. During testing, developers documented technical issues relevant to production deployment, including response failures, latency, inappropriate output behavior, deployment feasibility, and Unity integration compatibility.

Selection Criteria and Outcome Definitions

Each model-generated response was assessed against a case-specific reference rubric derived from the approved scenario script. The rubric covered patient identity and history, symptom accuracy and timeline consistency, patient-role behavior, contextual relevance to triage questions, and protection of information that would not normally be known by a patient.

Script adherence was assessed against the approved case script and patient persona using a 3-point scale: 0 for contradictory or unsupported responses, 1 for partially consistent responses, and 2 for fully consistent responses. Role fidelity was classified as maintained or not maintained according to whether the model remained in the patient perspective without adopting a nurse, physician, or general-assistant role. Hidden-information protection was classified as protected or leaked according to whether the response disclosed a diagnosis, triage category, management information, or other scenario metadata unavailable from the patient perspective. Contextual relevance was rated on a 5-point scale from 1 (not relevant) to 5 (fully relevant).

Four development-team members with expertise in nursing-triage simulation and LLM implementation independently assessed each response against the reference rubric. Discrepant ratings were resolved through discussion to reach consensus. Script adherence was summarized as the proportion of responses rated fully script-consistent. Model selection considered descriptive rubric ratings together with Cantonese-language suitability, response latency, technical reliability, deployment feasibility, and compatibility with the Unity environment. The scenario-specific reference rubrics and prompt categories are provided in Multimedia Appendix 2.

Phase 2: Formative Student Feedback–Informed Version 2 Enhancement and Evaluation

Collection and Analysis of Student Feedback

Version 1.0 of the AI Triage game was implemented during tutorial sessions in February 2025 (cohort 1). After gameplay, students participated in recorded group feedback discussions in which they were invited to comment on the simulation experience and suggest improvements.

Eleven tutorial feedback sessions involving 226 students were recorded and reviewed by the study team. The recordings captured group discussions; therefore, individual comment counts were not calculated. The study team transcribed the recordings, extracted the main points, and documented them in a summary matrix. A descriptive content analysis was conducted by grouping feedback into (1) suggested design enhancements and (2) bugs, game interface problems, or technical enhancement needs. The resulting categories were considered alongside system logs and internal technical testing to identify and prioritize Version 2 modifications.

Version 2 was enhanced from February to September 2025. The recorded feedback was used as formative development input and was distinct from the postintervention usability and recommendation outcomes collected from cohort 1 and cohort 2.

Identification of Version 1 Failure Modes

Version 1 failure modes were identified by triangulating formative student feedback with internal development records, including system logs, error observations, and repeated test interactions using the approved triage scenarios. Student feedback was used to identify interaction, interface, workflow, learning-feedback, and perceived-realism issues. Internal testing was used to identify LLM-specific failure modes, including inconsistent responses across sequential questions, role drift, disclosure of protected scenario metadata, inappropriate refusals or nonresponses to selected Cantonese triage prompts, and script-inconsistent responses.

The research and development team reviewed the combined findings and mapped each issue to a potential code, interface, prompt, memory-handling, parameter, or safety-control modification. Modifications incorporated into the completed Version 2 configuration were documented before post-enhancement technical evaluation.

Development of the Version 2 Configuration

The completed Version 2 configuration incorporated integrated modifications to the game interface and workflow, simulated-patient dialogue design, and LLM request-processing configuration. Interface and workflow modifications were prioritized on the basis of student-identified needs related to instructions, case initiation, result review, information display, response timing, audio, scrolling, error recovery, and patient interaction realism.

LLM-related modifications included (1) case-grounded prompt content; (2) 2 to 3 few-shot dialogue examples per scenario; (3) explicit constraints to maintain the simulated-patient role, avoid unsupported clinical details, and prevent disclosure of diagnosis, triage category, management information, or other protected scenario metadata; (4) session-level conversation-history handling, cleared at the end of each session; (5) calibrated generation settings, including reduction of temperature from 0.5 to 0.3, a maximum output setting of 4000 tokens, and frequency and presence penalties of 0; and (6) safety-filter adjustments for Cantonese clinical prompts.

Because these modifications were introduced as an integrated enhancement package, the subsequent evaluation assessed the performance of the completed Version 2 configuration rather than the independent effect of individual changes. Detailed technical implementation and request-processing specifications are provided in Multimedia Appendix 1.

Version 2 modifications were implemented and finalized between February and September 2025 before the February 2026 tutorial deployment to Cohort 2.

Script-Fidelity Evaluation Procedure

A development-stage script-fidelity evaluation compared the baseline Version 1 configuration with the completed Version 2 configuration. The evaluation used the malaria and neutropenic fever triage scenarios. Prompt categories were derived from the approved scenario scripts and included routine history-taking questions, repeated or rephrased questions to assess within-session consistency, questions about present or absent symptoms, symptom-timeline questions, travel or clinical-history questions, role-challenge prompts, and prompts designed to elicit protected scenario metadata. The scenario-specific reference rubrics and prompt categories are provided in Multimedia Appendix 2.

The same GPT-4o model version, approved case scripts, prompt categories, and planned testing procedure were used for both configurations. Version 1 represented the baseline configuration. Version 2 represented the completed integrated enhancement configuration, including changes to prompt content, conversation-history handling, safety controls, and generation settings. Each model-generated response was treated as one unit of analysis.

A total of 250 responses were generated for assessment: 125 responses under Version 1 and 125 responses under Version 2. For each configuration, 80 responses were generated from the malaria scenario, and 45 responses were generated from the neutropenic fever scenario. The malaria prompt set comprised 16 scenario-specific prompts and the neutropenic fever prompt set comprised 9 scenario-specific prompts. Each prompt was administered in 5 independent sessions per configuration, yielding 80 malaria and 45 neutropenic-fever responses for each configuration. Each independent session began with a new session and did not retain conversation history from a previous test session.

Two raters with expertise in nursing triage simulation and LLM implementation independently assessed all responses against the scenario-specific reference rubrics. Ratings were completed independently before consensus discussion. Disagreements were resolved through consensus. In contrast to phase 1, which used 4 assessors to support exploratory model selection across 6 candidate model configurations and multiple triage scenarios, phase 2 used a narrower, refinement-oriented evaluation focused on a single production model (GPT-4o) and two scenarios using a finalized, scenario-specific rubric. The phase 1 assessment team comprised the lead developer (who selected and configured the models for testing), 2 software developers with experience in applied AI systems (who rated conversational quality from a general AI perspective), and 1 project investigator with expertise in triage workflow and real-world patient responses. Accordingly, script-fidelity assessment in phase 2 was conducted by 2 raters: the lead developer and the same project investigator. The project investigator categorized response issues based on user feedback and triage requirements, while the lead developer implemented prompt and system modifications; the investigator then retested to confirm that revised responses met the approved scenario scripts. This iterative, 2-rater process was sufficient for the narrower, refinement-oriented scope of the phase 2 evaluation.

Outcomes and Analysis

This developmental study with a quasi-experimental student usability evaluation comprised 2 phases. Phase 1, conducted from 2024 to February 2025, involved clinical scenario development, exploratory model-fit assessment, and selection of a production model configuration. Phase 2 began with Version 1 tutorial deployment in February 2025, followed by formative student feedback and internal technical testing. These findings informed Version 2 enhancement from February to September 2025. The completed Version 2 configuration subsequently underwent development-stage script-fidelity evaluation and student usability evaluation during tutorial deployment in February 2026.

The primary technical outcome in phase 2 was the script-inconsistent response rate. A script-inconsistent response, hereafter termed a script-inconsistent hallucination, was defined as a model-generated statement that contradicted a predefined patient characteristic, symptom, clinical history, exposure, or timeline, or introduced clinically material information not supported by the approved scenario script. Each response was classified as fully script-consistent, partially script-consistent, or script-inconsistent. The script-inconsistent response rate was calculated as:

(Number of script-inconsistent responses/Total number of assessed responses) × 100%

Script adherence was calculated as the proportion of responses classified as fully script-consistent. For failure-type analysis, each script-inconsistent response was assigned to one primary, mutually exclusive category: role switching, contradictory symptom, unsupported symptom, incorrect timeline, incorrect travel or clinical history, hidden-information disclosure, unsupported environmental or contextual detail, or misinterpretation of a learner question as a request for protected scenario metadata.

For the development-stage technical evaluation, the results were summarized descriptively as counts and percentages for each configuration and scenario. The absolute difference in script-inconsistent response rates between Version 1 and Version 2 was calculated in percentage points. Because Version 2 incorporated multiple changes as an integrated enhancement package, the comparison was intended to describe the performance of the completed configuration and was not designed to estimate the independent effect of individual modifications.

Student usability outcomes were assessed using the SUS and NPS. Demographic variables, including sex and prior clinical internship experience, were summarized as frequencies and percentages, and cohort differences were examined using χ^2 tests. Questionnaire data were summarized using means and SDs. The 2025 cohort provided a within-cohort comparison of Version 1 and Version 2, whereas first-use experience was compared between the 2025 cohort using Version 1 and the 2026 cohort using Version 2. Normality was assessed using the Kolmogorov-Smirnov test; between-group comparisons were conducted using the Mann-Whitney U test. All analyses were performed using IBM SPSS Statistics (version 30; IBM Corp).

Ethical Considerations

Ethical approval was granted by the Research Ethics Committee of Hong Kong Metropolitan University (reference: HE-SF2024/57) prior to data collection. Written informed consent was obtained from all participants. Voluntary participation and study objectives were explained verbally. All research data were deidentified and pseudonymized at the earliest possible stage.

Results

Phase 1: Model Development and Model-Fit Findings

The final prototype integrated a Unity-based simulation game client, with GPT-4o hosted through Microsoft AI Foundry on Azure infrastructure to generate real-time simulated-patient responses in Traditional Chinese using Cantonese syntax.

During exploratory development testing, GPT-4o was selected for the production prototype based on the combined assessment of script consistency and implementation feasibility. Among 50 assessed responses, GPT-4o produced 45 (90%) fully script-consistent responses. The corresponding results were 43 of 50 (86.0%) responses for GPT-4, 39 of 50 (78.0%) responses for Llama 3.1, 32 of 50 (64.0%) responses for GPT-3.5-turbo, 23 of 50 (46.0%) responses for DeepSeek-V2.5, and 5 of 50 (10%) responses for Amazon Titan (Table 1).

Table 1. Exploratory development-stage model-selection assessment across candidate model configurations.[a]

Model configuration	Fully script-consistent responses, n/N (%)	Selected for production prototype
GPT-3.5-turbo	32/50 (64.0)	No
Llama 3.1	39/50 (78.0)	No
GPT-4	43/50 (86.0)	No
GPT-4o	45/50 (90)	Yes
DeepSeek-V2.5	23/50 (46.0)	No
Amazon Titan	5/50 (10)	No

[a]Fully script-consistent responses were rated against scenario-specific reference rubrics. This exploratory development-stage assessment was not a controlled benchmark comparison because prompt configurations, conversation-history handling, parameter settings, and iterative refinements were not fully standardized across model configurations.

These findings were descriptive development-stage results and informed selection of the production prototype. Although all model configurations were assessed using the same scenario scripts and 50-question test set, prompt configurations, conversation-history handling, parameter settings, and iterative refinements were not fully standardized across all model tests. Accordingly, these findings do not establish general superiority of GPT-4o over alternative models. The selected prototype subsequently underwent further refinement and Version 2 technical evaluation in phase 2.

Phase 2: Formative Student Feedback, Version 2 Enhancement, and Technical Evaluation

Student Feedback and Version 1 Issues

Student review discussions and internal testing identified Version 1 limitations in 4 areas: learning-feedback functions, interface and workflow, simulated-patient realism, and LLM response reliability. Students requested clearer instructions, improved result review and explanation, simultaneous display of question history and assessment information, and more flexible practice conditions. Students also described the patient dialogue as overly linear and suggested more emotionally expressive or dynamic patient behavior.

Technical and interface concerns included occasional nonresponses or display of code-like content, unexpected chat scrolling, absent audio output, slow assessment interactions, unnecessary blocking pop-ups, and restricted exit options after errors. Internal testing additionally identified inconsistent answers to repeated questions, occasional role drift, potential disclosure of protected scenario metadata, inappropriate refusals or nonresponses to selected Cantonese triage questions, and script-inconsistent clinical information.

These findings informed the completed Version 2 enhancement package (Table 2).

Table 2. Student feedback and development findings informing completed Version 2 enhancements.

Evidence source	Version 1 issue	Completed Version 2 modification	Intended function
Student feedback and internal testing	Dialogue was overly linear; patient responses were too passive; occasional role drift	Added few-shot patient-dialogue examples, explicit patient-role constraints, and refined dialogue prompting	Improve role fidelity, dialogue variation, and contextual response quality
Student feedback and internal testing	Inconsistent sequential answers; occasional absent or code-like AI responses	Added session-level conversation-history handling, refined request/error handling, and adjusted Cantonese safety-filter settings	Improve within-session continuity and response reliability
Internal testing	Script-inconsistent facts and unsupported clinical information	Added case-grounded prompt content, explicit negative constraints, and reduced model temperature from 0.5 to 0.3	Improve script consistency and reduce unsupported output
Internal testing	Protected scenario metadata could be disclosed	Added explicit restrictions on disclosure of diagnosis, triage category, management information, and protected metadata	Improve hidden-information protection
Student feedback	Instructions and game initiation were unclear	Updated tutorial screen, and enlarged UI[a] components	Improve onboarding and navigation
Student feedback	Limited result review and insufficient explanation	Implemented new game history system and AI analysis system for reference	Improve postcase learning feedback
Student feedback	Slow, disruptive, or confusing interface behavior	Implemented concurrent audio input, and changed AI model API call regions to reduce latency	Improve usability and gameplay flow

[a]UI: User Interface.

Completed Version 2 Enhancements

Version 2 incorporated completed game interface, workflow, and LLM-related enhancements in response to the identified usage and technical issues. Game interface and workflow modifications improved the start sequence and onboarding instructions through an updated animated tutorial screen and enlarged user-interface components, strengthened postcase review through a new game-history system and AI analy-

sis function, and enabled concurrent audio input to reduce conversation latency.

The LLM enhancement package incorporated case-grounded prompt content, few-shot patient-dialogue examples, explicit patient-role and protected-information constraints, session-level conversation-history handling, calibrated generation settings, and safety-filter adjustments for Cantonese clinical prompts. The few-shot examples demonstrated the expected patient role, response style, and boundaries of the approved scenario script. For example, the malaria-case examples guided the simulated patient to affirm the approved symptoms, including cough and fever, when asked about respiratory symptoms.

The completed changes are summarized in Table 2. As multiple modifications were introduced during the same development cycle, their individual contribution to subsequent technical outcomes could not be determined.

Postenhancement Technical Evaluation of Script Fidelity

A total of 250 model-generated responses were assessed across the malaria and neutropenic fever scenarios, including 125 responses generated using Version 1 and 125 responses generated using Version 2. Script-inconsistent responses occurred in 78 of 125 (62.4%) Version 1 responses and 21 of 125 (16.8%) Version 2 responses, representing an absolute reduction of 45.6 percentage points (Table 3).

Table 3. Script-inconsistent response outcomes by scenario and system configuration.

Scenario	Assessed responses, n	Baseline: script-inconsistent, n (%)	Final: script-inconsistent, n (%)	Absolute difference, percentage points
Malaria	80	54 (67.5)	15 (18.8)	48.7
Neutropenic fever	45	24 (53.3)	6 (13.3)	40.0
Overall	125	78 (62.4)	21 (16.8)	45.6

In the malaria scenario, script-inconsistent responses decreased from 54 of 80 (67.5%) Version 1 responses to 15 of 80 (18.8%) Version 2 responses—an absolute reduction of 48.7 percentage points. In the neutropenic fever scenario, the rate decreased from 24 of 45 (53.3%) Version 1 responses to 6 of 45 (13.3%) Version 2 responses—an absolute reduction of 40.0 percentage points.

The malaria scenario provides an illustration of the classification procedure. The approved case script specified fever, chills, muscle pain, and cough. When asked, “你有冇咳呀?” (“Do you have a cough?”), the response “有啊，咳咗幾日” (“Yes, I have been coughing for several days”) was classified as script-consistent. In contrast, “我冇咳喎” (“I do not have a cough”) was classified as script-inconsistent because it contradicted the approved symptom profile.

Residual Failure Modes and Remaining Limitations

Under Version 1, role switching was the most frequent script-inconsistent failure type, occurring in 17 of 78 (21.8%) script-inconsistent responses, followed by contradictory symptom descriptions in 16 of 78 (20.5%) responses. Under Version 2, unsupported symptoms were the most frequent residual failure type, occurring in 6 of 21 (28.6%) script-inconsistent responses, followed by incorrect timelines in 5 of 21 (23.8%) responses (Table 4).

Table 4. Primary categories of script-inconsistent responses by system configuration.

Failure type	Operational definition	Baseline configuration, n	Final configuration, n
Role switching	The simulated patient responded as a nurse, physician, or general assistant rather than as the patient	17	2
Contradictory symptom	The response denied, altered, or contradicted a symptom specified in the approved script	16	4
Unsupported symptom	The response introduced a symptom not specified in the approved script	11	6
Incorrect timeline	The response altered the approved onset, duration, sequence, or progression of symptoms	9	5
Incorrect travel or clinical history	The response contradicted the approved travel, exposure, medication, or medical history	7	4
Hidden-information disclosure	The response disclosed diagnosis, triage category, management information, or other protected metadata	7	0
Unsupported environmental or contextual detail	The response introduced a material setting or contextual detail not specified in the scenario	6	0
Misinterpretation of question as request for scenario metadata	The response treated a learner’s question as a request to disclose scenario-specific technical or hidden information	5	0
Total script-inconsistent responses	Sum of all mutually exclusive failure categories	78	21

No hidden-information disclosures, unsupported environmental or contextual details, or misinterpretations of learner questions as requests for protected scenario metadata were observed among the 125 Version 2 responses assessed. The results describe the performance of the integrated Version 2 configuration in the predefined test set and do not isolate the effects of individual enhancement components or establish performance across clinical scenarios beyond those evaluated.

Student Usability and Recommendation Outcomes

Overview

In addition to the development-stage technical evaluation, student usability and recommendation outcomes were evaluated during tutorial deployment. These student outcomes were distinct from the predefined script-fidelity assessment and were assessed using the SUS and NPS among the 2025 and 2026 tutorial cohorts. The following results present participant characteristics and usability outcomes across the 2 cohorts (Table 5).

Table 5. Characteristics of study participants.

Characteristics	Cohort 2025 (n=111), n (%)	Cohort 2026 (n=282), n (%)	*P* value
Sex (male)	50 (45)	136 (48)	.57
Temporary undergraduate nursing student (TUNS) experience	37 (33)	78 (28)	.26

A total of 393 nursing students participated in the study (n=111 in the 2025 cohort; n=282 in the 2026 cohort). Overall, 47.3% (n=186) of participants were male, and 29.0% (n=115) had prior clinical internship experience. Baseline comparisons revealed no statistically significant differences between the 2 cohorts regarding sex distribution (P=.57) or internship experience (P=.26), indicating baseline comparability.

SUS Outcomes

System usability of the AI Triage simulation game was evaluated using both within-group and between-group comparisons. A within-group analysis of the 2025 cohort (n=111) demonstrated statistically significant improvements in perceived system usability and interface quality when transitioning from Version 1 (mean 58.31, SD 9.54) to the enhanced Version 2 (mean 62.65, SD 11.66; U=7324; Z=3.39; P <.001).

To assess first-time user experience, a between-group comparison was conducted between the 2025 cohort (Version 1; n=111) and the 2026 cohort (Version 2; n=282). The enhanced Version 2 (mean 65.57, SD 12.36) showed significantly higher system usability scores than Version 1 (mean 58.31, SD 9.54; U=10,566; Z=5.57; P<.001).

NPS Outcomes

User satisfaction for enhanced Version 2 was evaluated using the NNPS across both cohorts. Both cohorts reported positive outcomes, with NPS ranging from +36.6% (2025 cohort; n=111) to +41.1% (2026 cohort; n=282), demonstrating student favorability toward the enhanced simulation platform.

Discussion

Principal Findings

This study developed and deployed a Cantonese-language, LLM-powered simulated-patient system for nursing-triage education through an iterative process. GPT-4o was selected through an exploratory model-fit assessment, and formative student feedback from Version 1 tutorial deployment, together with internal technical testing, informed an integrated Version 2 enhancement package. In the predefined script-fidelity evaluation, Version 2 generated fewer script-inconsistent responses than Version 1. Student usability and recommendation outcomes were collected in 2 tutorial cohorts. These findings illustrate the value of integrating user feedback with structured technical evaluation during development of LLM-enabled nursing simulations.

Technical and Pedagogical Implications

The development and deployment of the LLM-powered AI-assisted simulated-patient system described in this study revealed several considerations, recurring problems, and practical suggestions that are broadly relevant to educators seeking to implement similar systems in nursing curricula. However, the substantial difference in response rates between the 2025 and 2026 cohorts (n=111, 49.1% vs n=282, 95.3%), respectively) may have introduced selection bias and limits the generalizability of the findings.

A primary consideration is the trade-off between clinical safety and conversational realism. LLM-generated responses can fabricate plausible but script-inconsistent clinical details, a hallucination problem documented extensively in the present implementation. This issue is not unique to our system—a comprehensive review of 40 LLM-based virtual patient studies confirmed that factual accuracy, role consistency, and ethical accountability remain persistent challenges

across the field [31]. In fact, recent evidence demonstrates that even flagship commercial LLMs with advanced safety mechanisms are highly vulnerable to prompt-injection attacks, with success rates exceeding 90% in simulated clinical dialogues [32]. Our findings suggest that layered controls, including case grounding, few-shot examples, explicit role and protected-information constraints, generation-setting calibration, and platform-level safety adjustments, may reduce script-inconsistent responses while preserving the intended dialogue function.

Another related problem is the overcautious behavior of cloud-based safety filters. As documented in our debugging process, Microsoft AI Foundry's content safety classifiers intermittently blocked legitimate Cantonese clinical questions, such as "Why do you come to see a doctor?" by misinterpreting certain phrasings as sensitive content. This highlights a broader challenge: safety guardrails designed for general-purpose applications may be poorly calibrated for domain-specific clinical simulations, particularly in non-English languages [14,17]. We suggest that developers conduct systematic testing using structured sets of clinical prompts in the target language during deployment to identify and iteratively tune safety thresholds, ensuring essential triage dialogue is not inadvertently suppressed.

Linguistic equity represents another significant barrier. Most LLM training corpora are English-dominant, and models frequently process non-English languages through English intermediary encoding, which compromises generative quality in non-Anglophone contexts [33]. Our implementation required extensive testing and iterative prompt refinement to achieve clinically natural spoken Cantonese, and even the best-performing model (GPT-4o, at the time of development) demanded continuous adjustment. Addressing this disparity requires developing region-specific clinical corpora for fine-tuning, collaborating with local clinicians and linguists for validation, and benchmarking across diverse linguistic populations [18,25].

The absence of standardized evaluation frameworks presents a further obstacle to advancing this field. Current evidence for LLM-powered virtual patients relies predominantly on student satisfaction and self-reported confidence, with few studies using objective measures of clinical competence or behavioral change [7,12]. We recommend that future research adopt validated clinical assessment instruments, implement multisite controlled trials with adequate sample sizes, and use longitudinal designs that track transfer of simulation-acquired skills to actual patient care settings [31].

Finally, institutional readiness and ethical governance deserve attention. Health care organizations deploying AI-powered educational tools should establish cross-functional governance committees encompassing compliance, clinical operations, informatics, and data privacy to oversee system evaluation, approval, and ongoing monitoring [34]. Equally important, nurse educators require structured professional development in AI literacy, prompt engineering principles, and ethical AI use, so they can serve as informed stewards of these technologies and critically evaluate their pedagogical outputs rather than relying on them uncritically [23,25,35].

Further Development

LLM-powered virtual patients represent a paradigm shift from traditional rigid, scripted simulations to dynamic, open-ended conversations that adapt to student inputs [1]. Building upon the existing AI Triage game framework, AI Triage future Version 3.0 will evolve the system from a primarily instructor-authored simulation into a learner-generated scenario platform. This enhancement will allow students to design their own triage cases, including patient profiles, presenting complaints, and contextual details, which the LLM will then transform the custom-made scenario into realistic, clinically coherent interactions in Cantonese. By enabling students to self-author and iteratively refine scenarios with feedback from peers and teachers, the app aims to deepen clinical reasoning, foster creativity, and strengthen ownership of learning.

To support accurate and safe learner-created content, Version 3.0 will incorporate structured scenario templates, embedded clinical guardrails, and automated checks to reduce hallucinations and maintain alignment with evidence-based practice. These features will be complemented by configurable difficulty levels and feedback modes, allowing educators to scaffold tasks while giving students greater flexibility in pacing, repetition, and focus areas. In combination, these design choices position the upgraded system as a scalable environment for self-directed, practice-intensive learning that extends beyond scheduled class time.

Conclusions

This study developed and deployed a Cantonese-language, LLM-powered simulated-patient system integrated into a nursing-triage simulation game. GPT-4o was selected through an exploratory development-stage assessment, and formative student feedback from Version 1 tutorial deployment, together with internal technical testing, informed an integrated Version 2 enhancement package.

In a predefined script-fidelity evaluation, the Version 2 configuration produced fewer script-inconsistent responses than the Version 1 configuration across the 2 evaluated triage scenarios. The study illustrates a practical development approach that combines clinically reviewed case scripts, local-language dialogue design, student feedback, and structured technical testing to support simulated-patient interactions in nursing education.

Further research should evaluate the system across additional clinical scenarios and institutions, assess robustness after model or platform updates, and determine whether use of the system improves clinical reasoning, triage performance, or transfer of learning to practice.

Acknowledgments

The authors thank the clinical experts, university colleagues, and Hong Kong Hospital Authority nursing members who reviewed the triage scenarios for clinical accuracy, relevance to local nursing-triage practice, and alignment with real-world triage processes. The authors also thank the participating nursing students for their time and feedback.

Generative AI Use

Generative AI was not used in the preparation, writing, or editing of this manuscript.

Funding

This study was supported by Hong Kong Metropolitan University Quality Enhancement Measures (QEM) funding (QEM-1125; principal investigator: DCKH).

Data Availability

The deidentified data supporting the findings of this study are not publicly available because they contain participant-derived educational data and scenario materials that could compromise assessment integrity. Deidentified data may be available from the corresponding author upon reasonable request, subject to approval by the Research Ethics Committee of Hong Kong Metropolitan University, and in accordance with applicable institutional and data-protection requirements.

Authors' Contributions

MCMC contributed to technical methods, AI Triage system development and implementation, technical evaluation and analysis, and the initial manuscript draft. DCKH contributed to study conceptualization, project and quasi-experimental study design, the *Introduction* section, and interpretation of findings. TTOK verified technical implementation and contributed to manuscript review and revision. JYHW contributed to study design, ethics-approval coordination, and manuscript review. All authors approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Technical implementation and request-processing workflow of the AI Triage simulated-patient system.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Scenario-specific reference rubrics and prompt categories used for model-fit and script-fidelity evaluation.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 2]

References

1. Hamilton A, Molzahn A, McLemore K. The evolution from standardized to virtual patients in medical education. Cureus. Oct 2024;16(10):e71224. [doi: 10.7759/cureus.71224] [Medline: 39525234]
2. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res. Jan 7, 2025;27:e59069. [doi: 10.2196/59069] [Medline: 39773666]
3. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore). Jul 5, 2024;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
4. 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. 2025;27:2481-2491. [doi: 10.1016/j.csbj.2025.05.025] [Medline: 40547455]
5. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare (Basel). Mar 10, 2025;13(6):603. [doi: 10.3390/healthcare13060603] [Medline: 40150453]
6. Harrington J, Booth RG, Jackson KT. Large language models in nursing education: concept analysis. JMIR Nurs. Aug 22, 2025;8:e77948. [doi: 10.2196/77948] [Medline: 40845300]
7. Zeng J, Qi W, Shen S, et al. Embracing the future of medical education with large language model–based virtual patients: scoping review. J Med Internet Res. Nov 13, 2025;27:e79091. [doi: 10.2196/79091] [Medline: 41232097]
8. Liu Y, Shi C, Wu L, et al. Development and validation of a large language model–based system for medical history-taking training: prospective multicase study on evaluation stability, human-AI consistency, and transparency. JMIR Med Educ. Aug 29, 2025;11:e73419. [doi: 10.2196/73419] [Medline: 40882613]
9. Luo MJ, Bi S, Pang J, et al. A large language model digital patient system enhances ophthalmology history taking skills. NPJ Digit Med. Aug 4, 2025;8(1):502. [doi: 10.1038/s41746-025-01841-6] [Medline: 40760042]

10. Chen PJ, Liou WK. ChatGPT-driven interactive virtual reality communication simulation in obstetric nursing: a mixed-methods study. Nurse Educ Pract. May 2025;85:104383. [doi: 10.1016/j.nepr.2025.104383] [Medline: 40315616]
11. Liaw SY, Tan JZ, Lim S, et al. Artificial intelligence in virtual reality simulation for interprofessional communication training: mixed method study. Nurse Educ Today. Mar 2023;122:105718. [doi: 10.1016/j.nedt.2023.105718] [Medline: 36669304]
12. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model–powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ. Aug 16, 2024;10:e59213. [doi: 10.2196/59213] [Medline: 39150749]
13. Wang C, Li S, Lin N, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. J Med Internet Res. Jan 1, 2025;27:e59435. [doi: 10.2196/59435] [Medline: 39742453]
14. Chow JCL, Li K. Large language models in medical chatbots: opportunities, challenges, and the need to address AI risks. Information. 2025;16(7):549. [doi: 10.3390/info16070549]
15. Elhilali A, Ngo ASH, Reichenpfader D, Denecke K. Large language model-based patient simulation to foster communication skills in health care professionals: user-centered development and usability study. JMIR Med Educ. Dec 12, 2025;11:e81271. [doi: 10.2196/81271] [Medline: 41385781]
16. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond). Dec 19, 2025;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
17. Asaad A, Chy A, Kabir AS, Nakib A, Tabassum N. Navigating the labyrinth: a review of explainability and trustworthiness in large language model-powered systems for sensitive decision-making. Sci Technol Sci Soc. 2025;2(7):5-19. [doi: 10.59324/stss.2025.2(7).02]
18. Li D, Lebai Lutfi S. Large language model-based virtual patient systems for history-taking in medical education: comprehensive systematic review. JMIR Med Inform. Jan 2, 2026;14:e79039. [doi: 10.2196/79039] [Medline: 41481915]
19. Davis J, Van Bulck L, Durieux BN, Lindvall C. The temperature feature of ChatGPT: modifying creativity for clinical research. JMIR Hum Factors. Mar 8, 2024;11:e53559. [doi: 10.2196/53559] [Medline: 38457221]
20. Workum JD, van de Sande D, Gommers D, van Genderen ME. Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare. Front Artif Intell. 2025;8:1504805. [doi: 10.3389/frai.2025.1504805] [Medline: 39931218]
21. Aster A, Ragaller SV, Raupach T, Marx A. ChatGPT as a virtual patient: written empathic expressions during medical history taking. Med Sci Educ. 2025;35(3):1513-1522. [doi: 10.1007/s40670-025-02342-7] [Medline: 40625933]
22. Hua Y, Na H, Li Z, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med. Apr 30, 2025;8(1):230. [doi: 10.1038/s41746-025-01611-4] [Medline: 40307331]
23. Chan MMK, Wan AWH, Cheung DSK, et al. Integration of artificial intelligence in nursing simulation education: a scoping review. Nurse Educ. 2025;50(4):195-200. [doi: 10.1097/NNE.0000000000001851] [Medline: 40168182]
24. El Arab RA, Al Moosa OA, Abuadas FH, Somerville J. The role of AI in nursing education and practice: umbrella review. J Med Internet Res. Apr 4, 2025;27:e69881. [doi: 10.2196/69881] [Medline: 40072926]
25. Kim Y, Shin S, Yoo SH. Performance of large language models in non-English medical ethics-related multiple choice questions: comparison of ChatGPT performance across versions and languages. BMC Med Ethics. Dec 9, 2025;26(1):168. [doi: 10.1186/s12910-025-01316-z] [Medline: 41366422]
26. Han JW, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Med Educ. Dec 1, 2022;22(1):830. [doi: 10.1186/s12909-022-03898-3] [Medline: 36457086]
27. Brooke J. SUS: a “quick and dirty” usability scale. In: Jordan PW, Thomas PB, McClellandIL, WeerdmeesterB, editors. Usability Evaluation in Industry. CRC Press; 1996:189-194. [doi: 10.1201/9781498710411-35]
28. Vlachogianni P, Tselios N. Perceived usability evaluation of educational technology using the System Usability Scale (SUS): a systematic review. J Res Technol Educ. May 27, 2022;54(3):392-409. [doi: 10.1080/15391523.2020.1867938]
29. Baquero A. Net Promoter Score (NPS) and customer satisfaction: relationship and efficient management. Sustainability. 2022;14(4):2011. [doi: 10.3390/su14042011]
30. Takahashi Y, Agatsuma N, Oguro F, Oura M, Abo-Rass F, Nishikawa Y. Usage of Net Promoter Score in medical education literature: a scoping review. Cureus. Nov 2025;17(11):e95919. [doi: 10.7759/cureus.95919] [Medline: 41333511]
31. Jo YW, Lee M, Yang HJ. Large language model-based virtual patient simulations in medical and nursing education: a review. Appl Sci. 2025;15(22):11917. [doi: 10.3390/app152211917]

32. Lee RW, Jun TJ, Lee JM, Cho SI, Park HJ, Suh J. Vulnerability of large language models to prompt injection when providing medical advice. JAMA Netw Open. Dec 1, 2025;8(12):e2549963. [doi: 10.1001/jamanetworkopen.2025.49963] [Medline: 41632124]
33. Tierney AA, Reed ME, Grant RW, Doo FX, Payán DD, Liu VX. Health equity in the era of large language models. Am J Manag Care. Mar 2025;31(3):112-117. [doi: 10.37765/ajmc.2025.89695] [Medline: 40053403]
34. The responsible use of AI in healthcare (RUAIH). The Joint Commission; 2025. URL: https://digitalassets.jointcommission.org/api/public/content/dcfcf4f1a0cc45cdb526b3cb034c68c2 [Accessed 2026-08-30]
35. Hardie P, Darley A, Derwin R, et al. Applications, attitudes and ethical considerations of generative artificial intelligence (Gen AI) in nursing education: a scoping review. BMC Nurs. Jan 16, 2026;25(1):148. [doi: 10.1186/s12912-025-04253-9] [Medline: 41545825]

Abbreviations

AMTES: AI-Powered Medical History-Taking Training and Evaluation System
LLM: large language model
LLMDP: large language model–based digital patient
NPS: Net Promoter Score
SUS: System Usability Scale

Edited by Elizabeth Borycki; peer-reviewed by Mei Chen, Takeshi Koike; submitted 20.Apr.2026; initial comments to author 16.Jul.2026; final revised version received 13.Sep.2026; accepted 21.Oct.2026; published 19.Nov.2026

Please cite as:
Chan MCM, Hung DCK, Kwok TTO, Wong JYH
Large Language Model–Powered AI-Assisted Simulated Patient for Nursing Education: Quasi-Experimental Study
JMIR Nursing 2026;9:e98543
URL: https://nursing.jmir.org/2026/1/e98543
doi: 10.2196/98543

© Max Chun Ming Chan, Danny Chun Kit Hung, Tyrone Tai-On Kwok, Janet Yuen Ha Wong. Originally published in JMIR Nursing (https://nursing.jmir.org), 19.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Nursing, is properly cited. The complete bibliographic information, a link to the original publication on https://nursing.jmir.org/, as well as this copyright and license information must be included.