

Review

Embracing the Future of Medical Education With Large Language Model–Based Virtual Patients: Scoping Review

Jianwen Zeng^{1,2*}, BSc; Wenhao Qi^{1,2*}, BSc; Shiyang Shen¹, BSc; Xin Liu¹, BSc; Sixie Li¹, BSc; Bing Wang¹, BSc; Chaoqun Dong¹, BSc; Xiaohong Zhu¹, BSc; Yankai Shi¹, BSc; Xiajing Lou¹, BSc; Bingsheng Wang¹, BSc; Jiani Yao¹, BSc; Guowei Jiang³, MSc; Qiong Zhang⁴, BSc; Shihua Cao^{1,2}, PhD

¹School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, null, China

²Zhejiang Provincial Research and Evaluation Center for Educational Modernization, Hangzhou, null, China

³Department of Psychiatry and Neuropsychology and Alzheimer Center Limburg, School for Mental Health and Neuroscience (MHeNS), Maastricht University, Maastricht, The Netherlands

⁴Department of Nursing, Zhejiang Provincial People's Hospital, Hangzhou, China

*these authors contributed equally

Corresponding Author:

Shihua Cao, PhD

School of Public Health and Nursing

Hangzhou Normal University

No 2318, Yuhangtang Road, Yuhang District

Hangzhou, 311121

China

Phone: 86 13777861361

Email: csh@hznu.edu.cn

Abstract

Background: In recent years, large language models (LLMs) have experienced rapid development. LLM-based virtual patients have begun to gain attention, offering new opportunities for simulations in medical education.

Objective: This study aims to systematically analyze the current applications, research trends, and challenges of LLM-based virtual patients in medical education and to explore potential future directions for development.

Methods: This study adheres to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Five databases (Web of Science Core Collection, PubMed, IEEE Xplore, Embase, and Scopus) were searched from January 1, 2018, to June 24, 2025, to identify studies related to the application of LLM-based virtual patients in medical education. A comprehensive analysis of LLM-based virtual patients from research design to application and evaluation was conducted.

Results: A total of 28 studies were included in this scoping review. Analysis revealed that 92.9% (26/28) of the studies were published in the past 2 years, indicating that LLM-based virtual patient research is still in its early stages. The research primarily focuses on medical training and spans a wide range of medical disciplines. When using LLMs, advanced technologies such as social robots, virtual reality, and mixed reality are used to present LLM-based virtual patients. Combining these technologies with various supplementary tools enhances the realism of LLM-based virtual patients and improves user interaction. The evaluation of LLM-based virtual patients mainly emphasizes user experience. However, evaluation methods lack standardization, and only 13% (3/23) of studies used validated tools in assessing LLM-based virtual patients, while only 21.7% (5/23) of studies objectively measured learning outcomes facilitated by LLM-based virtual patients. All included studies expressed a positive attitude toward LLM-based virtual patients; however, they overlook privacy and security considerations in practical applications.

Conclusions: LLM-based virtual patients hold significant innovation potential in medical education and are still in the early stages of development. They are primarily applied in medical training and show promise in communication skills training, although they cannot replace real-world interactions. Moreover, the heterogeneity of research designs, the absence of nonverbal cues in interactions, and concerns regarding privacy and security limit their broader implementation. Future research should focus on improving the reliability, realism, safety, and scientific efficacy of LLM-based virtual patients.

Trial Registration: Open Science Framework Registries 10.17605/OSF.IO/DMC9Q; <https://osf.io/DMC9Q/overview>

KEYWORDS

artificial intelligence; virtual patient; large language model; medical education; scoping review

Introduction

In recent years, large language models (LLMs) have made significant progress [1,2]. LLMs are high-performance artificial intelligence (AI) systems capable of understanding and generating natural language [3]. With advancements in AI, LLMs have demonstrated great potential in tasks involving natural language processing [4]. Their applications range from text analysis and summarization to clinical applications, showcasing their flexibility in providing valuable assistance [5-7]. LLMs support user interactions through follow-up questions and are fine-tuned to generate controlled outputs [8,9]. More importantly, they allow developers to create chatbots and virtual assistants with customized behaviors [9]. Given these capabilities, LLMs are expected to become efficient and feasible tools across various domains, including medical education, where traditional virtual patients have also been used.

Virtual patients are computer-based programs that simulate real clinical scenarios, allowing learners to take on the role of health care professionals. The goal is to develop skills and knowledge in specific areas while enabling learners to practice decision-making in a controlled interactive environment [10,11]. Virtual patients have been shown to be effective in teaching, assessment, and clinical reasoning research [12] and have been proposed as a valuable educational tool for practicing clinical reasoning in undergraduate medical education [13]. They play a crucial role in medical education.

However, the application of virtual patients faces logistical challenges and high costs for large-scale implementation [14]. For instance, studies have shown that the technological development cost of a virtual patient is US \$12 per hour, with the average monthly cost of developing and maintaining a virtual patient system being US \$324.75 [15,16]. These limitations often prevent all students from engaging in interactive skill practice or performing multiple exercises, significantly reducing the effectiveness of the practice. However, the emergence of LLMs as a disruptive technology offers unprecedented opportunities to overcome the limitations faced by traditional virtual patients. LLM-based virtual patients combine natural language processing technology with medical knowledge, using LLMs to construct virtual avatars. These avatars are designed to simulate the behavior, symptoms, diagnostic processes, and disease progression of real patients, creating highly realistic and diverse virtual patient models [17-19]. They can present standardized patients in various scenarios, supporting students' clinical reasoning, decision-making, and problem-solving skills while also providing performance analysis and feedback [20,21]. Although LLMs face inherent limitations and negative impacts in practical applications, such as "hallucinations" [22] and influences on independent thinking [23], these challenges do not prevent LLMs from offering new opportunities in medical education simulations.

Recent research has discussed the application of traditional virtual patients, with some systematic reviews highlighting the positive impact of virtual patient simulators on medical communication training. These reviews emphasize the adaptability of virtual patients and their value as a supplement to traditional educational methods [24,25]. However, to date, no study has comprehensively summarized the application of LLM-based virtual patients in medical education. This paper aims to provide a comprehensive overview of the positioning, challenges, and future directions of LLM-based virtual patients in medical education, offering a reference for the better development and application of LLM-based virtual patients.

As an innovative and transformative technology, LLM-based virtual patients demonstrate tremendous potential in medical practice and are expected to drive the field toward greater efficiency, precision, and personalization. To comprehensively analyze their current applications, technological challenges, and future directions, this paper focuses on the following key issues: (1) In which areas of medical education are LLM-based virtual patients primarily applied? Which medical disciplines are involved, and what are the main research directions? (2) What are the primary LLMs currently used? How are the models fine-tuned, and what is the role of prompt engineering? (3) How are LLM-based virtual patients specifically implemented in practical applications? Specifically, how is the instructional design (application scenarios and learning activity design), technological design (integrated technology ecosystem, interaction modes, and auxiliary tools), and assessment design (user experience, learning outcomes assessment, evaluation standards, and evaluation roles) structured? (4) What are the key challenges faced by LLM-based virtual patients, and what are the future research directions?

Methods

Study Design

This study uses a scoping review methodology due to the diversity of the research questions, the heterogeneity of the studies, and the lack of comprehensive previous reviews on this topic [26,27]. The scoping review framework follows the approach proposed by Arksey and O'Malley [26] and is reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist for scoping reviews [27]. The complete PRISMA-ScR checklist is available in [Multimedia Appendix 1](#). This scoping review has been registered in the Open Science Framework (10.17605/OSF.IO/DMC9Q).

Several factors led to deviations from the original protocol regarding inclusion and exclusion criteria. First, given that AI is a rapidly evolving field and high-impact journals frequently publish important peer-reviewed original research in the form of "research letters," we decided not to exclude conference papers and letters. Second, considering the growing attention

on compact LLMs with fewer than 10 billion parameters, we removed the restriction on the model size in the inclusion criteria. Additionally, due to potential difficulties in retrieving non-English literature, and given that English-language publications adequately cover key developments in the fields of natural sciences and medicine, we decided to exclude non-English papers.

Data Sources and Search Strategy

To ensure comprehensive retrieval and consider the interdisciplinary nature of LLM-based virtual patients in medical applications, a literature search was conducted across 5 major databases: PubMed, Web of Science Core Collection, Scopus, Embase, and IEEE Xplore. We collaborated with librarians and medical informatics experts to develop the search strategy. To

ensure thoroughness and minimize the risk of missing relevant literature, the core search terms included 2 categories: one related to “generative AI” and “large language models” and the other related to “virtual patients,” using Boolean operators (eg, AND and OR) for combination. Additionally, we reviewed the reference lists and citations of relevant papers. The search time frame was from June 2018 to April 24, 2025. June 2018 marks the release of the first generative AI model [28]. Literature management and duplicate removal were conducted using EndNote (version 20; Clarivate Analytics) software. A detailed description of the search strategy can be found in Table S1 in [Multimedia Appendix 2](#).

Inclusion and Exclusion Criteria

The inclusion and exclusion criteria are listed in [Textbox 1](#).

Textbox 1. The inclusion and exclusion criteria.

Inclusion criteria

- The literature must focus on research related to large language model–based virtual patients.
- The literature must explicitly address the application of large language model–based virtual patients in medical education, including technical development or practical case studies.
- Eligible types of literature include journal papers, conference papers, and research letters.

Exclusion criteria

- Studies for which the full text is not accessible.
- Duplicate publications.
- Conference abstracts, preprints, books, editorials, reviews, and retracted studies.
- Studies unrelated to medicine, such as research from nonmedical fields like psychology, which does not involve any aspect of medical education.
- Non-English language publications.

Screening and Data Extraction

Before formally determining the inclusion or exclusion of literature, 3 researchers (JZ, SL, and Xin Liu) randomly selected 30 studies for an initial screening to assess the reliability of the screening process. The final calculated Cohen κ value was 0.89, indicating high consistency, and no adjustments were made to the inclusion or exclusion criteria or the researchers involved. In the formal independent screening process, any disagreements were ultimately resolved through intervention by SS. The screening and validation process was completed on April 26, 2025.

To accurately extract data from the included studies, we followed the PRISMA-ScR and created a data extraction form using Microsoft Excel. Two evaluators (JZ and SS) independently completed the form after receiving professional training based on the *Medical Literature Information Retrieval* [29] textbook. Any disagreements between the evaluators were resolved through discussion.

First, we extracted general information from the studies, including the publication year, country, study type, and research objectives. To gain a deeper understanding of the potential applications and challenges of LLMs in medicine, we summarized the key findings and limitations of each study.

Next, the design data extracted consisted of two aspects: (1) general characteristics of LLMs, such as model type, open-source availability, model training (fine-tuning and prompt engineering), and other technologies or tools integrated into LLM-based virtual patients, such as voice assistants and hardware devices, to illustrate the specific design of LLM-based virtual patients; and (2) medical specialty, medical context and tasks, simulated patients, avatars (eg, social robots), participants, and sample sizes to demonstrate the specific application design of LLM-based virtual patients.

Finally, the evaluation data extracted included the evaluation domains (user experience, learning outcomes), evaluation tools (eg, scales and questionnaires), and evaluators (eg, experts) to provide an overview of the overall evaluation details of LLM-based virtual patients use in each study. The data extraction table is provided in Table S2 in [Multimedia Appendix 2](#).

Additionally, to better understand the differences between the studies, we used 2 separate tools for quality assessment (2 authors [WQ and SS] conducted separate evaluations, and any discrepancies in the final scores were addressed through intervention and discussion by a third author [SC]). Medical Education Research Study Quality Instrument [30], a validated tool for evaluating the quality of quantitative medical education research, has a scoring range from 5=lowest quality to

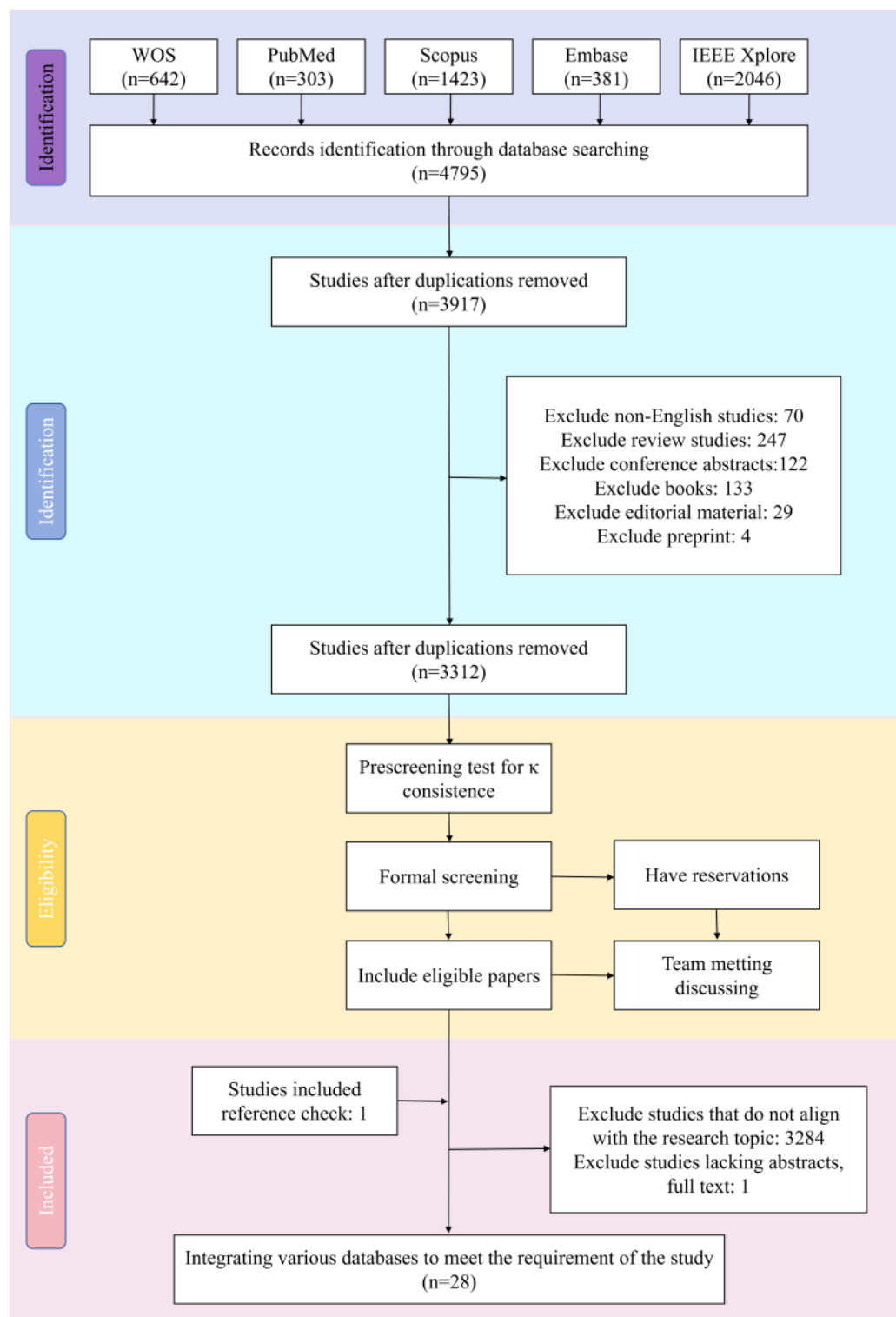
18=highest quality. For qualitative research, we used the QualSyst standard [31], which includes a checklist of 10 criteria for assessing qualitative studies. The score, representing the ratio of obtained points to the maximum possible score, ranges from 0=lowest quality to 1=highest quality. For mixed methods studies, we used both tools simultaneously and reported the scores separately.

For the extracted data, in addition to using the PRISMA-ScR method, we used various other techniques including narrative synthesis, thematic analysis, mapping or data visualization, and descriptive statistical analysis. These methods were used to describe, summarize, and present the application scenarios, research progress, advantages, and limitations of LLM-based virtual patients in medical education in comprehensive formats such as tables, flowcharts, and diagrams.

Results

Study Selection

A preliminary search across the 5 databases identified a total of 4795 papers. After removing duplicates, 3917 papers remained. Non-English language papers, reviews, conference abstracts, editorials, preprints, and similar publications were excluded, leaving 3312 papers. Three researchers (JZ, SS, and SL) screened the titles and abstracts of these papers and conducted further evaluation, resulting in 27 studies that met the inclusion criteria for this review. An additional study that met all the inclusion criteria was identified through citations in the included papers. Therefore, a total of 28 studies were included in this review. Figure 1 shows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study selection. WOS: Web of Science.

Study Characteristics

The included studies (N=28) were published between 2023 and 2025, with the majority concentrated in 2024 and 2025, accounting for 92.9% (26/28) of the total. This field has attracted widespread attention across various countries and regions, with research conducted in 13 different countries or regions. The top 3 countries with the most studies were the United States (6/28, 21.4%), Germany (5/28, 17.9%), and Japan (3/28, 10.7%). The

basic information and quality assessment results for each study can be found in [Multimedia Appendix 3](#).

Development and Design

In total, 7 studies involved the development and design of LLM-based virtual patient programs or platforms [32-38]. Weisman et al [32] provided a detailed description of the steps taken by their team in developing a GPT-4-based virtual simulated patient and communication training platform. Prior

to development, they conducted in-depth interviews to guide design decisions. Other studies provided a more general overview of LLM-based virtual patient applications or platforms. Notably, the virtual patient developed by Shindo et al [36] included a mechanism that identifies and corrects overly detailed initial responses from ChatGPT.

Practical Applications

Overview

In total, 23 studies assessed the actual effectiveness of LLM-based virtual patients, with 2 studies first focusing on design and development before evaluating the practical application of LLM-based virtual patients [32,33]. In contrast, studies evaluating the actual application accounted for over 80% (23/28) of the total, indicating that LLM-based virtual patients in medical education are shifting from the technical development phase to the application phase. This transition highlights the

potential anticipated in improving the quality and efficiency of medical education. This is also the primary focus of our study, and we will present a clear and comprehensive overview of the entire process, from research design to application and evaluation, across 5 key aspects.

Medical Field and Educational Tasks

These studies cover 13 medical specialties, with the highest number in general medicine (n=9), followed by rheumatology and dentistry, each with 2 studies. All these studies are related to medical training in medical education, specifically examining whether interaction with LLM-based virtual patients can help train key medical skills such as history taking and clinical reasoning. For example, Baugerud et al [39] explored the potential of a ChatGPT-3-based child avatar to train participants in interview skills in the context of suspected abuse. Regarding LLM-simulated patients, all but 2 studies explicitly defined the types of simulated patients used [40,41], as detailed in Table 1.

Table 1. Research directions, medical fields, and research content included in large language model-based virtual patient studies.

Reference	Application category	Medical field	Simulated patients
Benfatah et al [42]	Communication skills	Nursing	Patient with respiratory distress
Borg et al [43]	Clinical reasoning	Rheumatology	Patient with rheumatic disease
Brügge et al [20]	Clinical decision-making	Neurology or neurosurgery	Patient with herniated lumbar disc or stroke or meningitis or concussion
Gray et al [44]	N/A ^a	Obstetrics	A mother expecting preterm delivery
Holderried et al [45]	History taking	General medicine	Patient with nausea, weight loss, and chronic fatigue
Holderried et al [21]	History taking	General medicine	Patient with nausea, weight loss, and chronic fatigue
Gutiérrez Maquilón et al [46]	Communication skills	Emergency medicine	A survivor of a traffic accident
Sardesai et al [47]	Anesthesia training	Anesthesiology	Patient with fractured humerus
Yamamoto et al [48]	Interviewing skills	General medicine	Patient with chest pain or abdominal pain or cough or heartburn or fatigue or fever or dizziness or shortness of breath
Aster et al [49]	Empathic expression	Cardiology and emergency medicine	Patient with cardiac conditions
Borg et al [50]	Clinical reasoning	Rheumatology	Patient with rheumatic disease
Cook et al [51]	N/A	Ambulatory medicine	Patient with diabetes or chronic cough
Ko et al [52]	Information gathering skills	Dental	Child survivor of abuse
Öncü et al [53]	Clinical case management	General medicine	Patient with hypertension or brucellosis
Rädel-Abläss et al [54]	History taking	General medicine	Patient with brain hemorrhage
Wang et al [55]	History taking	Gastroenterology	Patient with inflammatory bowel disease
Or et al [56]	History taking	Dental	Patient with dental conditions
Cook [18]	N/A	General medicine	Patient with chronic cough or type 2 diabetes
Yi and Kim [33]	History taking	Urology	Patient with urinary problem
Abou Karam [40]	N/A	General medicine	N/A
Weisman et al [32]	Communication skills	General medicine	Patient with abnormal mammogram results
Baugerud et al [39]	Interviewing skills	Psychology	Child survivor of sexual or physical abuse
Liu et al [41]	N/A	General medicine	N/A

^aN/A: not applicable.

LLM Strategy and Prompting

OpenAI’s GPT series models are the most frequently used, with 95.7% (22/23) of the studies incorporating them, the majority

using more advanced versions, namely, ChatGPT-3.5 or higher. Other LLMs used include HyperCLOVA. Detailed information is provided in Table 2.

Table 2. Information on the use of models in large language model–based virtual patients, language, fine-tuning, and prompts.

LLMs	Language	Fine-tuning	Prompt	Reference
ChatGPT (not specified)	Not specified	No	Scenario	[42]
GPT-3.5-turbo	English	No	Scenario, behavioral, emotional simulation	[43]
GPT-3.5	German	No	Scenario, behavioral, communication style, feedback	[20]
GPT-3.5	English	No	Scenario, communication style, emotional simulation	[44]
GPT-4	German	No	Scenario, behavioral, feedback	[45]
GPT-3.5-turbo	German	No	Scenario, behavioral	[21]
GPT-3.5-turbo	German	No	Scenario, behavioral, communication style, emotional simulation	[46]
Not specified	English	Custom knowledge bank	Not specified	[47]
GPT-4-turbo	Japanese	No	Scenario, emotional simulation	[48]
GPT-3.5	German	No	Scenario, behavioral, communication style	[49]
GPT-3.5-turbo	English	No	Scenario, behavioral, emotional simulation	[50]
GPT-3.5-turbo or 4-turbo	English	No	Scenario, behavioral, communication style, emotional simulation, feedback	[51]
GPT-3	Norwegian	No	Scenario	[52]
GPT-4o	English	No	Scenario	[53]
GPT-4	Not specified	No	Scenario, behavioral, communication style	[54]
GPT-4	English or Chinese	No	Scenario, behavioral, communication style, emotional simulation	[55]
GPT-3.5 Instruct	English	No	Not specified	[56]
HyperCLOVA	Korean	Scripts of medical interviews	Not specified	[33]
GPT-3.5	Not specified	No	Not specified	[40]
GPT-4	English	No	Scenario, behavioral, emotional simulation, feedback	[32]
GPT-3.5-turbo or 4	English	No	Scenario, behavioral, communication style, emotional simulation	[18]
GPT-3	Not specified	741 mock interviews	Not specified	[39]
GPT-3.5-turbo	Not specified	No	Not specified	[41]

In the research process, to better adapt the LLM to specific application needs and improve its performance in simulating patients, some researchers fine-tuned the models. Fine-tuning involves updating the model’s weights using smaller, domain-specific corpora. In total, 13% (3/23) of the studies involved fine-tuning the LLMs used for simulated patients [33,39,47]. For instance, Baugerud et al [39] fine-tuned their model using 741 mock interviews sourced from a forensic interview training program.

To ensure the desired interaction between participants and LLM-simulated patients, carefully designed prompts were used to elicit the required responses from the LLMs. In total, 69.6% (16/23) of the studies described or provided the full prompts used in their research, of which 26.1% (6/23) of the studies used prompt engineering to iteratively optimize the prompts

[20,21,32,45,54,55]. A notable example is the study by Wang et al [55], where they conducted multiple preliminary tests to standardize the evaluation of the model’s responses, generated a list of questions, and then adjusted the prompts until the LLM-based virtual patients performed optimally. Only then, did they finalize the prompt version. A deeper analysis of the 16 studies that described or provided full prompts revealed that the prompts could be classified into five types: (1) contextual prompts: providing background information and setting the interaction scenario, (2) behavioral prompts: restricting or guiding the model’s responses, (3) communication style prompts: adjusting the language style or changing the mode of communication, (4) emotional simulation prompts: incorporating specific emotional responses into the answers, and (5) feedback prompts: offering personalized suggestions based on user

performance. The types of prompts used in each study are provided in [Table 2](#), and typical examples of each prompt type are presented in Table S1 in [Multimedia Appendix 4](#).

Instructional Design Features

A summary of the scenarios presented in each study revealed that the most frequently reported scenario was taking a history inventory with an LLM-based virtual patient (20/23, 87%). Each study's scenarios were reviewed based on the Calgary-Cambridge model of medical communication steps [57], which include (1) gathering information from the patient, (2) building a relationship with the patient, and (3) explaining and planning. Excluding 3 studies [18,32,40], the remaining 87% (20/23) of the studies reported communication skills related to gathering information from the patient. No studies involved building a relationship with the patient or explaining and planning. All studies focused solely on verbal communication skills, without addressing nonverbal behaviors such as gestures or nodding.

In total, 21.7% (5/23) of the studies focused on aspects aimed at improving learning outcomes [20,32,45,51,52], specifically the accuracy and usefulness of feedback automatically generated by LLM-based virtual patients. For instance, Brügge et al [20] demonstrated that the intervention group receiving AI-generated personalized feedback performed better in clinical

decision-making (CDM) than the control group, supporting the effectiveness of feedback. Additionally, 4.3% (1/23) of the study incorporated an additional learning activity, namely, a group-based follow-up workshop [43]. In this workshop, students could discuss and share any issues encountered during the LLM-based virtual patient exercises. Such workshops facilitated better learning outcomes after practicing with LLM-based virtual patients.

Technological Design Features

LLM-based virtual patients are presented through integrated technological ecosystems such as the web, hardware, platforms, software, or applications ([Table 3](#)). The most commonly used platform is OpenAI's public web user interface, used in 7 studies, followed by laptops and social robots, each used in 2 studies. Additionally, to achieve more immersive interactions with LLM-based virtual patients, some studies incorporated advanced technological ecosystems such as mixed reality (MR) and virtual reality (VR). For example, Gutiérrez Maquilón et al [46] examined the integration of GPT-based AI into an MR virtual patient application for communication training of emergency medical services personnel. The system delivered an immersive, sensorially rich MR environment that closely simulated real-world emergency scenarios, maximizing ecological validity.

Table 3. Technology ecosystem, interaction modality, and auxiliary tools in the application research of large language model-based virtual patients.

Technology ecosystem	Interaction modality	Auxiliary tools	Reference
Web			
OpenAI's public web UI ^a	Text	No	[20]
OpenAI's public web UI	Text	No	[44]
OpenAI's public web UI	Text	No	[49]
OpenAI's public web UI	Text	No	[55]
OpenAI's public web UI	Text	No	[18]
OpenAI's public web UI	Text	No	[33]
OpenAI's public web UI	Text	No	[41]
Self-developed web interface	Text	No	[21]
Hardware			
Laptops	Text	No	[42]
Social robot	Robot+voice	Furhat software development kit	[43]
Laptops	Text	No	[45]
MR ^b	3D avatar rendering+voice	Microsoft LifeChat LX-3000 headset, OpenAI Whisper, ElevenLabs	[46]
Social robot	Robot+voice	Furhat software development kit	[50]
Tablet	Voice	No	[53]
VR ^c	3D avatar rendering+voice	IBM Watson services	[39]
Platform			
ConVAI	2D avatar+text or voice	No	[47]
Miibo or LINE Corporation	Text	No	[48]
Guided Conversation Designer	Text	No	[54]
Vercel	Text	No	[56]
Software or program			
Python	Text	No	[51]
Hyperskill	Voice	Hyperskill	[32]
Other			
Not specified	Static avatar+text	No	[52]
N/A ^d	N/A	N/A	[40]

^aUI: user interface.^bMR: mixed reality.^cVR: virtual reality.^dN/A: not applicable.

Regarding interaction modes, the majority of studies used natural language, including text (n=14) and voice (n=7), with 1 study supporting both modes. In contrast, the languages used in the studies were more diverse, encompassing 5 different languages (Table 2). One study tested the impact of Chinese and English on LLM-based virtual patient performance, finding no significant performance differences between the test groups using different language combinations [55].

Furthermore, only 26.1% (6/23) of the studies used patient avatars, 4 of which were virtual avatars [39,46,47,52], and 2 were physical embodiments, that is, social robots. In total, 2

studies provided details on the creation and presentation of virtual avatars, both using the Unity game engine for 3D rendering of human models or virtual avatars, which were then presented to participants using head-mounted displays [39,46]. Details are provided in Table 3.

To enhance the realism of virtual patients and improve the user experience, 21.7% (5/23) of the studies used auxiliary tools for LLM-based virtual patients. These tools can be categorized into voice modules, transcription modules, and emotional visualization modules (details and categorization of tools are provided in Table S2 in Multimedia Appendix 4). A notable

example is 2 studies of Borg et al [43,50], where LLM-based virtual patients were presented through social robots, supplemented by the use of FurhatSDK. This setup allowed for voice interaction, as well as the display of subtle facial expressions and emotions, making the virtual patients more anthropomorphic [43,50].

Evaluation Features

Among the 23 studies, a total of 398 participants were clearly identified as experimental participants interacting with LLM-based virtual patients, including various groups such as medical students, clinicians, and emergency medical personnel. Medical students comprised 71.6% (285/398) of the participants, while medical educators made up only 0.5% (2/398).

Two methods were used to evaluate the practical application of LLM-based virtual patients: assessing the user experience of LLM-based virtual patients (23 studies) and evaluating the learning outcomes facilitated by LLM-based virtual patients (5 studies) [20,48,49,52,53]. The characteristics of these 2 types of assessments, including the measurement domains, tools, evaluators, and results, are provided in [Multimedia Appendix 5](#).

Based on Nielsen's usability concepts [58] and existing literature on the assessment criteria for standardized patients and virtual patients [25,59], we reviewed the user evaluations of LLM-based virtual patients used in these studies. The evaluation standards included (1) technological design, (2) realism of simulation, and (3) practicality of learning. In total, 47.8% (11/23) of the studies commonly used at least 1 of the following standards to assess technological design: (1) usability (overall experience or perception after use), (2) satisfaction (acceptance), and (3) errors (technical issues). Regarding the realism of simulation, 60.9% (14/23) of the studies assessed it by asking about authenticity or contextualization. Authenticity (the degree to which LLM-based virtual patients resemble real patients) was the most focused-on area (14/14, 100%). In total, 21.4% (3/14) of the studies assessed contextualization (how closely the simulation resembles real-world scenarios). A total of 39.1% (9/23) of the studies evaluated users' perceptions of the practicality of LLM-based virtual patients in learning, specifically in terms of the learning process (how much they helped achieve target skills) or feedback quality (usefulness or appropriateness of the provided feedback).

For the measurement tools, 3 validated questionnaires were used to assess usability in the technological characteristics [21,46]; 1 validated questionnaire was used to assess authenticity in the simulation realism [50]. Among the self-developed evaluation tools, 3 studies validated and reviewed their scales or questionnaires, which somewhat enhanced the validity, reliability, and applicability of these self-created assessment tools [33,51,54]. Across the 23 studies that evaluated user experience with LLM-based virtual patients, a total of 22 different assessment tools were used. Among these, 18 studies used self-assessment tools completed by users, while 4 studies used expert assessments conducted by clinicians, medical professors, or other domain experts.

Although these studies applied to various skills training ([Table 1](#)), only 21.7% (5/23) of the studies assessed learning outcomes in terms of objective skill measurements. In these studies, the researchers measured changes in learners' skills, such as CDM, and empathetic expression in specific scenarios. One study used a self-created tool for objective skill measurement [53], while the others used validated or reliable tools. For example, the Clinical Reasoning Indicator-History Taking Inventory was used to measure clinical reasoning skills [20]. Of the 5 studies evaluating learning outcomes, 4 used expert assessment, and 1 used self-assessment.

Quality assessment revealed heterogeneity and frequent inconsistencies in the study designs and evaluations, making it challenging to assess the performance of LLM-based virtual patients. Therefore, we provided a general overview of the research findings. Information on the tasks, performance or results, sample sizes, clinical validation methods, and participant demographics for each study can be found in [Multimedia Appendix 6](#). Among these studies, 2 were controlled experiments. The remaining studies were observational in nature. Overall, the included studies demonstrated a positive attitude toward the application of LLM-based virtual patients.

Discussion

Principal Findings

Our review indicates that the application of LLM-based virtual patients has gained considerable momentum in recent years, with many research teams developing diverse and innovative applications. However, the heterogeneity in study designs, evaluation standards, learning outcomes, and their measurements limits the ability to make direct comparisons and draw definitive conclusions, indicating that future research has much ground to cover.

LLMs for Simulated Patients

In these studies, the primary LLM used is the ChatGPT series, indicating that current research on LLM-based virtual patients heavily relies on proprietary models. ChatGPT is a closed-source proprietary model, which raises significant open science issues related to transparency and reproducibility [60,61]. Different OpenAI models, in particular, exhibit notable differences; yet, the source of these discrepancies remains opaque, as OpenAI's closed-source policy makes testing and evaluation impossible. Additionally, OpenAI (and similar systems) continuously updates its models, meaning that research conducted using ChatGPT today may not be directly replicable, or even reproducible, within the next 6 months [62]. This presents challenges to the reproducibility of LLM-based virtual patient research outcomes. In contrast to closed-source (or proprietary) models, open-source models are not affected by undisclosed updates and can be fully deployed locally, avoiding some clinical data privacy issues associated with closed-source models [63]. Furthermore, using open-source LLMs is crucial for reproducibility. With open-source LLMs, researchers can examine the internal structure of the model to understand how it works, customize the code, and flag errors [64]. These details include adjustable parameters and the data on which the model is trained. Currently, many high-performance open-source

LLMs, such as DeepSeek and LLaMA, have emerged, demonstrating strong capabilities in specific domains [65]. Exploring the use of these models could reduce reliance on a single model, greatly improving the universality and reproducibility of research.

To enhance the scientific rigor and scalability of LLM-based simulated patients, the research community must adopt more controlled methodologies. Among studies using ChatGPT, only 2 mentioned setting the “temperature” hyperparameter [49,51], with one of them exploring whether the temperature affects LLM-based virtual patients’ performance. The results indicated that there were no significant differences in performance at different temperatures, but further research is needed to validate this outcome. Temperature is a frequently modified hyperparameter that controls the randomness of the model’s predictions [66]. Some researchers believe that temperature features will play a crucial role in the application of generative AI in medical services, potentially enabling more accurate, empathetic, or creative interactions between AI and health care stakeholders [67]. Currently, there is limited research on the hyperparameters used for LLMs simulating virtual patients. Besides temperature, other parameters, such as the 2 “repetition penalties,” which reduce token repetition and may make responses more diverse, remain unexplored in the context of LLM-based virtual patients. Whether these parameters contribute to more realistic simulations of virtual patients is yet to be determined. Additionally, researchers must address issues related to backend model updates and random factors in the sampling process to ensure the reliability of results [68].

Application of LLM-Based Virtual Patients

The design features of virtual patients, such as interactivity, play an important role in enhancing clinical reasoning skills [69]. Compared to less interactive approaches, highly interactive virtual patients allow educators to better assess students’ clinical reasoning skills by directly observing their abilities [70]. Leveraging advanced technological ecosystems, such as social robots, MR, VR, and auxiliary tools, LLM-based virtual patients can achieve higher levels of interactivity (eg, speech, movement, and eye contact), creating more authentic simulated encounters with potential to enhance learning [71,72]. However, only a minority of studies addressed this dimension.

Despite the potential for realistic simulated interactions, several technical barriers hinder the use of LLM-based virtual patients at a high level. First, LLMs face challenges in emotional understanding and perception. Although LLMs possess subtle capabilities in understanding and managing emotions, they are inefficient in using emotions to facilitate thinking [73]. This results in issues like unrealistic emotional expression and incongruent emotional responses in LLM-based virtual patients. Moreover, training LLMs with multimodal datasets—especially those incorporating speech and video data—can improve the model’s understanding of a patient’s emotional and contextual state, enhancing the naturalness and accuracy of dialogues. However, incorporating these data modalities raises significant privacy concerns, as speech and video data not only threaten patient privacy but also the privacy of clinicians [74], limiting the use of multimodal datasets. Additionally, the embodiment

of virtual patients presents challenges. To achieve realistic interactions and an immersive experience, LLM-based virtual patients typically rely on social robots, VR, and similar hardware. Currently, most robots are designed to express basic emotions and lack sufficient smooth and accurate facial movements, such as eye movements, blinking, eyebrow movements, and particularly lip movements [75,76], hindering more realistic simulation of patient reactions and symptom presentation. VR hardware also faces challenges, including high demands for computational resources and persistent issues with rendering delays [77], which can significantly impact the performance of virtual patients, causing interactions to be sluggish and unnatural, potentially lowering training and learning outcomes.

At present, the application of LLM-based virtual patients primarily focuses on the users, such as medical students and interns, while the core figures driving medical education—such as teachers—have received less attention. Research by Montenegro-Rueda et al [78] shows that integrating ChatGPT into the educational environment can positively impact the teaching process, but its successful implementation depends on the proficiency of the educators, making adequate teacher training key to effective use. Advanced technologies may enhance learning outcomes, but without well-designed curricula or teaching strategies, specific learning results cannot be guaranteed [25]. Lövquist et al [79] argue that establishing and maintaining close relationships between educators, clinicians, and developers are crucial for the development of effective, reliable, and useful VR-based medical training and assessment systems. Similarly, for the development of LLM-based virtual patient medical training and assessment systems, educators’ involvement is essential. Identifying teaching strategies, such as how educators demonstrate the use of virtual patients, explain the medical simulation setup, and provide feedback, plays a significant role in shaping students’ learning outcomes, academic performance, and overall development [80]. Furthermore, this involvement fosters positive teacher-student relationships, which can be reflected in students’ focus and interest in the course content [81]. From the student perspective, experiencing teacher support helps them feel a sense of belonging in the classroom, which enhances their emotional learning, such as their attitude toward the content, thus strengthening their effective learning [81-83]. Additionally, teaching design, often overlooked, is a key factor in optimizing the use of technology and determining its effectiveness. While LLM-based virtual patients indeed have unique and optimal characteristics for medical communication training, their use must be guided by carefully designed teaching interventions to ensure effectiveness. For instance, designing collaborative pair activities or group discussions after interactions with virtual patients can bring added benefits, including increased interactivity, better use of the virtual patient platform, and improved clinical reasoning training [43,84].

Communication is a complex phenomenon that involves not only verbal language but also various nonverbal channels and responses [85]. Nonverbal communication includes conveying information through body signals, such as eye contact, facial expressions, gestures, and acoustic cues (paralanguage) [86].

However, the reviewed studies mostly focused on verbal communication, with limited attention to nonverbal behaviors. These nonverbal elements often carry as much, if not more, information than verbal communication itself. For instance, in interactions with patients, doctors primarily rely on facial expressions, body language, vocal tone, and other subtle cues to interpret meaning and make clinical decisions [87]. Some studies have shown that medical students' communication skills, particularly in nonverbal communication and empathy toward patients, are insufficient [88-90], highlighting the need to focus on improving learners' nonverbal communication skills. To train nonverbal communication skills using LLM-based virtual patients, more advanced LLMs, such as multimodal LLMs, must be used, in combination with more advanced technological ecosystems. However, limited by technological and cost-effectiveness issues, especially technical limitations like system failures, language processing challenges, and system overloads [91], as well as the inability of virtual patients to fully simulate real patient responses, achieving the goal of training nonverbal communication skills remains challenging.

To enhance the realism of LLM-based virtual patients, fine-tuning LLMs with training data relevant to target outcomes is necessary. Bui et al [92] fine-tuned 3 open-source models using existing datasets, data scraped from Vietnamese medical online forums, and data extracted from Vietnamese medical textbooks. The results showed that the fine-tuned models performed better than their base versions on evaluation metrics such as BertScore, Rouge-L, and the "LLM as Judge" method, confirming the effectiveness of the fine-tuning process. Currently, only a few studies mention fine-tuning, and no detailed investigations have been conducted. Future research should explore whether fine-tuning LLM-based virtual patients using different specific types of data, such as medical dialogues, can optimize their performance. Previous studies have fine-tuned LLMs using doctor-patient dialogue datasets, showing significant improvements in the model's ability to understand patient needs and provide targeted suggestions [93]. Furthermore, only a few studies have explicitly addressed prompt engineering in the design of LLM-based virtual patient prompts. By optimizing the input structure, prompt engineering plays a crucial role in refining AI and LLM outputs [94]. Modifying and optimizing prompts to make them more specific lead to more accurate and focused LLM outputs [95], thus improving the performance of LLM-based virtual patients and enabling more realistic and accurate simulated education.

Evaluation of LLM-Based Virtual Patients

The design and evaluation of the 23 studies on the practical application of LLM-based virtual patients are heterogeneous and often inconsistent, making it difficult to accurately assess the task performance and application effectiveness of LLM-based virtual patients and masking the potential of LLM-based virtual patients in medical education. Particularly in terms of evaluation, only a few studies used validated tools to assess LLM-based virtual patients, indicating a lack of standardization in the evaluation process. Furthermore, evaluations have largely focused on subjective indicators—users' experiences—which means that the heterogeneity of these outcome measures may prevent

cross-study comparisons. While these studies show enthusiasm for the use of LLM-based virtual patients in medical training, whether they can be further applied to medical education needs careful consideration. Due to the lack of standardization, LLM-driven virtual standardized patient training is unlikely to be used as part of summative clinical examinations or assessments of learner communication skills. The limitations of current evaluation standards highlight the need for broader evaluations of simulation robustness.

Currently, several effective and reliable evaluation methods or frameworks could be considered for future research. To assess user experience, tools like the Subjective Assessment of Speech System Interfaces (SASSI) questionnaire [96] and Witmer's Presence Questionnaire [97] could be used. The SASSI questionnaire is an effective, reliable, and sensitive measure of users' subjective experience with speech recognition systems, including dimensions such as system response accuracy, likability, cognitive demand, annoyance, habitability, and speed. It includes 39 Likert items, scored from 1=strongly disagree to 7=strongly agree. Using this questionnaire to assess the usability of LLM-based virtual patients in voice interaction could provide insights into speech recognition accuracy, interaction smoothness, and users' experiences. However, the SASSI has been used only in limited speech recognition systems. Witmer's Presence Questionnaire, consisting of 22 self-report items, each with a 7-point Likert scale, assesses the sense of immersion in virtual environments, with higher scores indicating stronger immersion.

From an educational training perspective, models like Kirkpatrick's 4-level training evaluation model and Kolb's experiential learning theory could be introduced. Kirkpatrick's model, introduced in 1959, is one of the most widely used and well-known frameworks for evaluating training and development programs. It has 4 levels: reaction, learning, behavior, and results. Due to its robustness, adaptability, and applicability, using this model could lead to more effective evaluation of training outcomes [98]. Since only 5 (17.9%) studies measured learning outcomes, applying this model to assess the training effectiveness of LLM-based virtual patients in medical training could enable researchers to provide a more comprehensive demonstration of virtual patient systems' impact on medical education, showcasing learners' performance after virtual patient training and analyzing improvements in knowledge mastery, skill application, and CDM, rather than merely focusing on learners' immediate reactions or academic performance. Kolb's experiential learning theory emphasizes the process of learning through experience as an integrated cycle of 4 stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. Each stage is interrelated, guiding learners from direct experience to critical reflection, conceptual understanding, and the application of new knowledge [99]. Using this framework to guide the evaluation of LLM-based virtual patient systems could not only provide students with an immersive learning experience but also help medical students combine theory with practice through continuous feedback, reflection, and experimentation. Additionally, frameworks for automated interaction assessment and AI-structured clinical

examinations to assess LLM performance in clinical tasks could also be adapted [100,101].

Only 5 studies involved the measurement of objective skills, making it unclear how effective LLM-based virtual patients are in training users' skills. This not only reflects the intent of clinical educators but also highlights the characteristics of the simulated training environment, which often faces time constraints and limited resources, making long-term sustainability difficult [102]. Overall, the current evaluation practices for LLM-based virtual patients in communication training lack rigor. More controlled approaches are needed to improve scalability and scientific rigor, such as using validated tools and systematically examining changes in student behavior or clinical outcomes, rather than simply focusing on students' attitudes or performance in the simulated environment. Additionally, researchers could reduce the limitations associated with self-reported data by collecting psychophysiological data from digital sensors (eg, electroencephalography, extreme energy ratio, and heart rate) during students' interactions with virtual patients and providing real-time feedback [103]. In conclusion, the insufficient standardized assessment of learning outcomes means we must remain cautious in judging the practical value of LLM-based virtual patients, though all 23 studies show positive user attitudes toward LLM-based virtual patients. While this attitude does not equate to scientifically validated educational effectiveness, it can help enhance learners' motivation, undeniably indicating the promising potential of LLM-based virtual patients in medical training.

Privacy and security, key aspects often not addressed or discussed in the studies, are important considerations in LLM-based virtual patient research. Most of the studies used cloud-based LLMs, but cloud-based LLMs typically require users to upload explicit requests during inference, which inevitably raises concerns about data security and user privacy [104-107]. Specifically, the process of guiding LLMs to simulate patients through prompts inevitably includes patient-related information. The popular privacy protection method for LLMs is to encrypt user medical requests to prevent LLM service providers' servers from accessing private user data. Common methods, such as solid-state encryption technology [108], significantly mitigate the risk of LLM operators or potential attackers leaking or misusing patient data for commercial or other purposes [109]. While these privacy protection methods are effective, challenges remain for medical LLMs, such as resource consumption and potential impacts on model accuracy and reliability [110]. A new method, adaptive compressed-based privacy-preserving LLM, has been proposed, which avoids the aforementioned issues while demonstrating strong privacy protection capabilities and high response accuracy [110]. Local deployment of LLMs is also a reliable method for addressing data leakage and privacy concerns. This approach ensures that user data stay within the organization, significantly enhancing data security and privacy protection [111]. Researchers have proposed an innovative compact LLM framework for local deployment of electronic health record data, which not only addresses privacy concerns in medical environments but also overcomes challenges related to limited computational resources [112]. Furthermore, from a patient data perspective, using

synthetic patient data can help resolve privacy issues [113]. Synthetic data do not pose the same privacy concerns as real patient data because they are not linked to any specific individual [114].

The Ethics of Using LLM-Based Virtual Patients

The use of LLM-based virtual patients in medical education involves several ethical dimensions, including data ownership and consent for use, data representativeness and bias, and privacy [115]. These dimensions reflect the relationships, responsibilities, and moral obligations between virtual patients as a technological tool and actual patients, health care professionals, and educators.

Data Ownership and Consent for Use

Using LLM-based virtual patients with patient data raises core questions about ownership, consent, and anonymization. When fine-tuning models or supplying prompts, patients may need explicit consent or at minimum clear notice that their data are used. Safeguarding informed consent and data rights is central to the ethics of virtual patient simulations.

Data Representativeness and Bias

LLMs may present potential algorithmic biases that lead to discriminatory behaviors and stereotypes, potentially resulting in unfair treatment of certain groups [116-119]. If these biases are not identified and corrected in a timely manner, virtual patients may contribute to incorrect diagnoses or treatment plans in certain populations, thus exacerbating inequalities in health care services. It is imperative that researchers and developers of LLM-based virtual patients proactively address these biases to prevent harmful consequences and ensure equitable health care training environments.

Privacy

The application of LLM-based virtual patients may involve the processing of real patient-related data. Even when using synthetic or virtual patient data, it is crucial to ensure that the data are deidentified, anonymized, and fully protected to prevent potential personal information leaks. Adequate safeguards must be implemented to protect patient privacy and ensure that virtual patient data are handled ethically and securely.

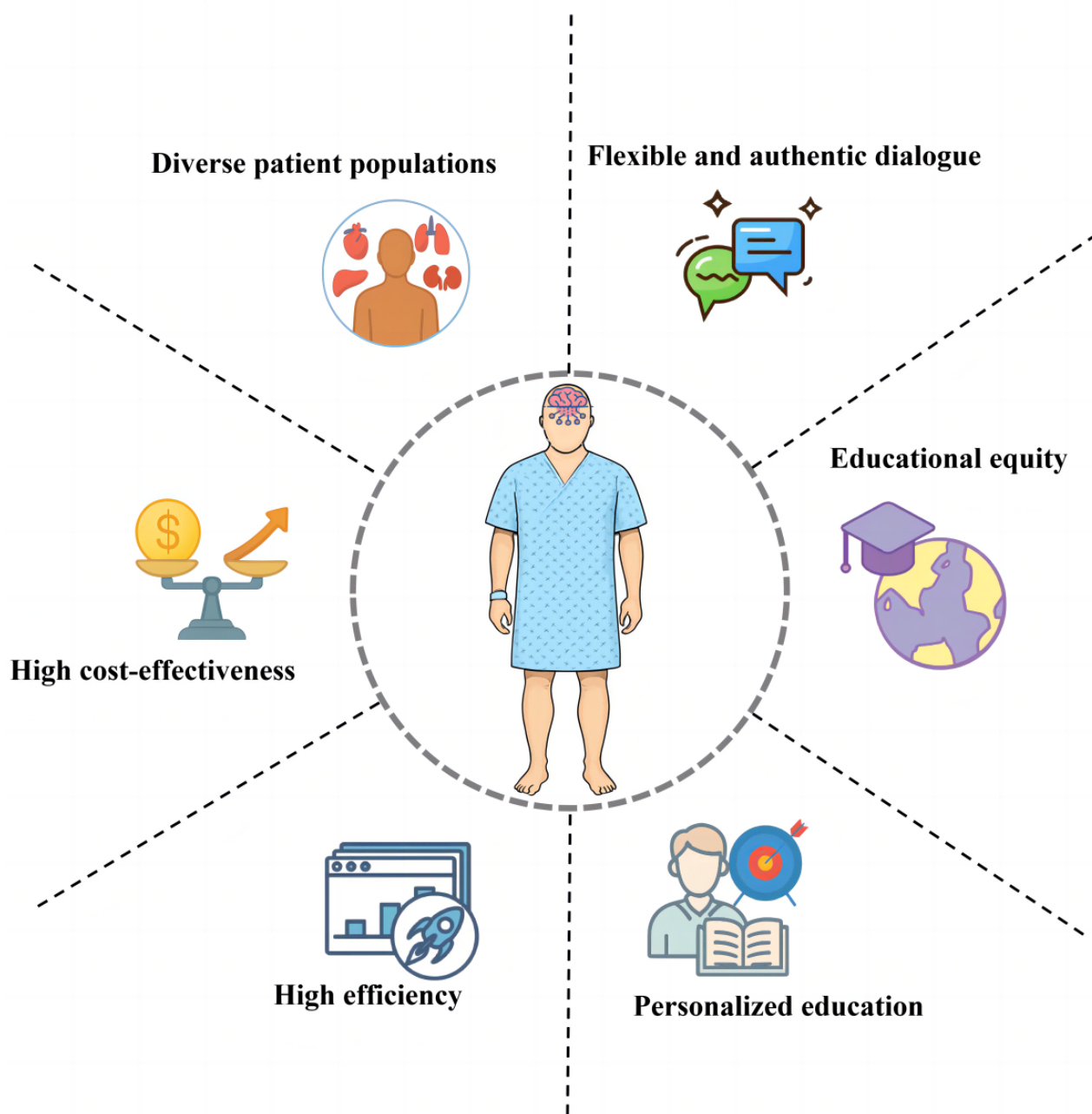
Situation of LLM-Based Virtual Patients

Integrating AI into medical education is crucial for equipping health care professionals with the key skills needed to provide optimal patient care in the future [120], and the use of LLM-based virtual patients undoubtedly aligns with this trend. Compared to most existing virtual patients, LLM-based virtual patients demonstrate unscripted, responsive dialogues that exhibit realism and flexibility. This realism is advantageous for training, assessment, and research on shared decision-making [121-124] and other management reasoning processes [125-127]. LLMs can simulate a diverse range of patients. For rare diseases, medical students often find it difficult to encounter real patients experiencing these conditions during clinical rotations. LLM-based virtual patients offer high cost-efficiency, such as reducing the resources required for interaction with real patients and specialized facilities. Through this LLM-based approach, thousands of preference-sensitive virtual patients can be created

with greater efficiency and even higher realism than current labor-intensive methods. Each virtual patient can be “created” as a single document page, with different variants added by changing just a few sentences [51]. Furthermore, the application of LLM-based virtual patients can help mitigate educational inequities. AI-driven simulations can be accessed by an

unlimited number of students across different geographic locations. The system enables anyone to train at any time, unrestricted by time or spatial limitations, thereby democratizing access to high-quality educational experiences. The advantages of LLM-based virtual patients are shown in Figure 2.

Figure 2. Advantages of using large language model–based virtual patients in medical education and training.



Although LLM-based virtual patients offer numerous advantages, they should be seen as a useful and cost-effective supplementary tool rather than a replacement for real-life interactions. Their greatest limitation lies in their inability to handle nonverbal communication, which is crucial for developing important skills such as empathy communication. Research has shown that while medical students interact with virtual patients with empathy, these interactions, both quantitatively and qualitatively, are insufficient to replace real-life interactions, such as with standardized patients [128].

In conclusion, while the application of LLM-based virtual patients holds great promise, integrating them into medical education is inevitable. However, educators must remain vigilant and consider both the positive and negative impacts that this integration may bring [129].

Future Development Directions

The rapid development of LLMs has led to the emergence of many high-performance models. Researching the performance of other models in patient simulation can better address the

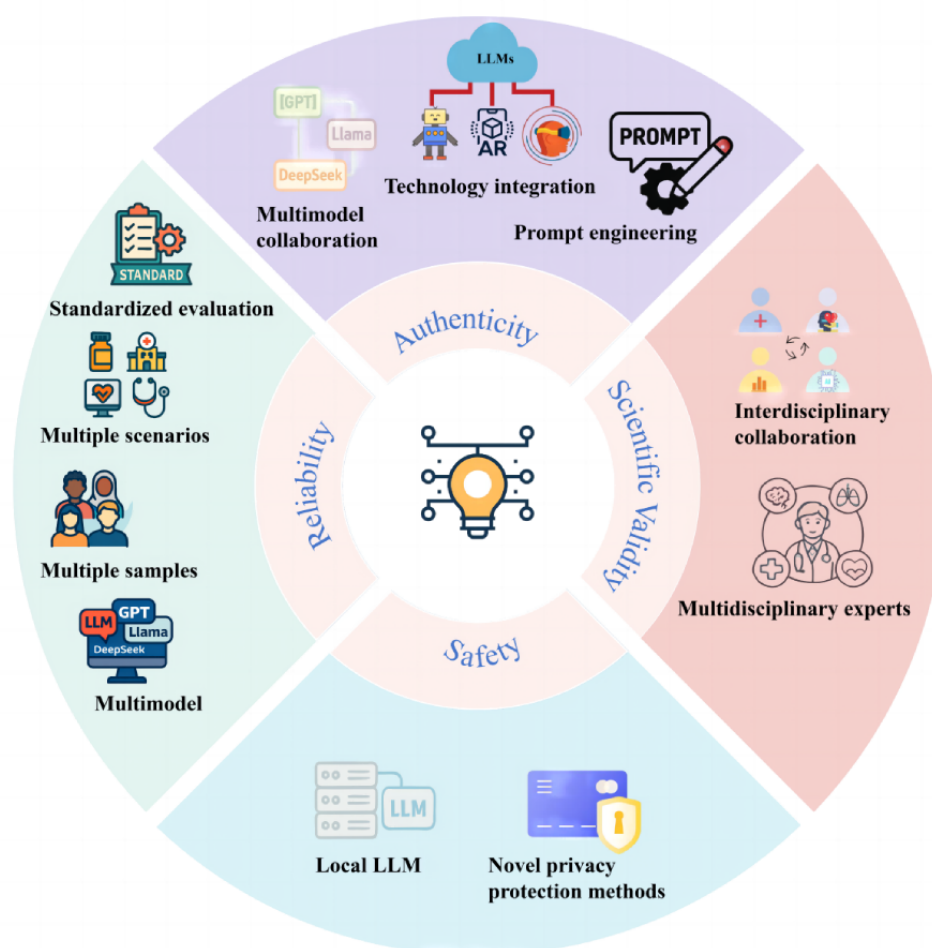
diverse needs of medical training, technological iterations, and security issues. Using representative and diverse samples, such as those from different geographic regions, cultures, and backgrounds, will help ensure the broad applicability and accuracy of the results. Additionally, creating diverse scenarios and establishing standardized evaluation methods to study and assess LLM-based virtual patients are equally important.

The realism of LLM-based virtual patients remains a key area for improvement. Compared to using a single model, collaborative multimodel systems have demonstrated superior performance [130]. Exploring the potential of multiple LLMs working together to reduce errors and enhance realism in simulated patients holds promise. When combined with more advanced technological ecosystems, such as social robots, augmented reality, VR, or MR, these systems can facilitate multisensory interactions, offering significant benefits for medical skills training. Furthermore, further exploration of prompt engineering related to simulating virtual patients can play a critical role in improving the realism of LLM-based virtual patients' performance.

Currently, the research on LLM-based virtual patients has overlooked safety concerns, particularly data privacy and protection. Using secure local LLMs can reduce data privacy risks, although it requires additional computational resources. Methods for privacy protection still need to be explored and developed in future research.

Finally, ensuring the scientific rigor of LLM-based virtual patient design remains an area of further study. Early interprofessional education training has the potential to enhance leadership, collaboration, and communication among health care teams, ultimately improving patient safety [131]. Incorporating the knowledge and experience of multidisciplinary medical experts (such as clinicians, pharmacists, and rehabilitation specialists) to optimize LLM-based virtual patients will allow for more comprehensive and scientific applications, achieving the goal of interprofessional education training for users. Moreover, collaboration among interdisciplinary teams should also be explored. Integrating expertise from medical professionals, data scientists, AI researchers, psychologists, and other fields in the design and development of LLM-based virtual patients will significantly enhance their scientific foundation. The 4 future research directions are illustrated in Figure 3.

Figure 3. Future development directions in LLM-based virtual patient research. LLM: large language model.



Strengths and Limitations

To the best of our knowledge, this is the first comprehensive review focused on LLMs simulating virtual patients. We have

summarized the entire process of using LLM-simulated patients for medical training, covering aspects from experimental design to outcome evaluation. This review highlights the development, design, and application processes, providing valuable references

for the higher-quality, more effective, and scientific application of LLM-based virtual patients in medical education.

However, this study has several limitations. First, our review was limited to English-language literature, potentially overlooking high-quality research published in other languages. Second, the inclusion of studies relied on the subjective judgment of the researchers, which may introduce selection bias. Additionally, as research on LLM-based virtual patients is still in its early stages, the limited number of relevant studies may affect the comprehensiveness and representativeness of the analysis.

Conclusions

This scoping review adopts a rigorous methodology to summarize and discuss the current state of applications of

LLM-based virtual patients in medical education. The findings indicate that research on LLM-based virtual patients has gradually increased over the past 2 years. They provide learners with opportunities to repeatedly practice communication skills and receive timely, appropriate feedback, offering significant economic benefits. As a result, they hold promising prospects in delivering effective medical skills training. However, further improvements are needed in areas such as research design, model implementation, humanization, privacy and security, and evaluation criteria. Additionally, it is important to clarify that LLM-based virtual patients should serve as a valuable supplement to traditional simulation-based education rather than a replacement. Future research is essential to further investigate their reliability, authenticity, safety, and scientific rigor.

Acknowledgments

This study is supported by Zhejiang Province Traditional Chinese Medicine Science and Technology Project (2023ZF134); First-Class Course of Zhejiang Province; 2024 Research Project of Engineering Research Center of Mobile Health Management System, Ministry of Education; 2022 Zhejiang Province First-Class Undergraduate Courses, Zhejiang Provincial Department of Education; and 2024 Higher Education Research Project of Zhejiang Higher Education Society.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

JZ conceptualized the study, set the research methodology, and edited the manuscript. WQ conceptualized the study and revised the manuscript. SS, Xin Liu, and SL organized the data and set the research methodology. SC conceptualized the study, reviewed and edited the manuscript, acquired funding, managed the project, and performed formal analysis. Bing Wang, XZ, CD, YS, Bingsheng Wang, Xiajing Lou, JY, GJ, and QZ conducted formal analysis and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR checklist.

[\[DOC File , 77 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Search strategy and data extraction form.

[\[DOCX File , 21 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Basic information for each study.

[\[XLSX File \(Microsoft Excel File\), 21 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Typical examples of each prompt and auxiliary tool categories and descriptions.

[\[DOCX File , 21 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Evaluation characteristics of large language model-based virtual patients.

[\[XLSX File \(Microsoft Excel File\), 13 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Summary of large language model-based virtual patient applications in medical training.

[\[DOCX File, 25 KB-Multimedia Appendix 6\]](#)

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
2. Xiao H, Zhou F, Liu X, Liu T, Li Z, Liu X, et al. A comprehensive survey of large language models and multimodal large language models in medicine. *Inf Fusion*. 2025;117:102888. [doi: [10.1016/j.inffus.2024.102888](https://doi.org/10.1016/j.inffus.2024.102888)]
3. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307(2):e230163. [doi: [10.1148/radiol.230163](https://doi.org/10.1148/radiol.230163)] [Medline: [36700838](https://pubmed.ncbi.nlm.nih.gov/36700838/)]
4. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*. 2023;56(2):1-40. [doi: [10.1145/3605943](https://doi.org/10.1145/3605943)]
5. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: comparative analysis. *JMIR Med Educ*. 2024;10:e63430. [FREE Full text] [doi: [10.2196/63430](https://doi.org/10.2196/63430)] [Medline: [39504445](https://pubmed.ncbi.nlm.nih.gov/39504445/)]
6. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. *NPJ Digit Med*. 2024;7(1):205. [FREE Full text] [doi: [10.1038/s41746-024-01208-3](https://doi.org/10.1038/s41746-024-01208-3)] [Medline: [39112822](https://pubmed.ncbi.nlm.nih.gov/39112822/)]
7. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239. [doi: [10.1056/NEJMs2214184](https://doi.org/10.1056/NEJMs2214184)] [Medline: [36988602](https://pubmed.ncbi.nlm.nih.gov/36988602/)]
8. Roumeliotis KI, Tselikas ND. ChatGPT and open-AI models: a preliminary review. *Future Internet*. 2023;15(6):192. [doi: [10.3390/fi15060192](https://doi.org/10.3390/fi15060192)]
9. Wu T, He S, Liu J, Sun S, Liu K, Han Q, et al. A brief overview of chatGPT: the history, status quo and potential future development. *IEEE/CAA J Autom Sin*. 2023;10(5):1122-1136. [doi: [10.1109/jas.2023.123618](https://doi.org/10.1109/jas.2023.123618)]
10. Baumann-Birkbeck L, Florentina F, Karatas O, Sun J, Tang T, Thaug V, et al. Appraising the role of the virtual patient for therapeutics health education. *Curr Pharm Teach Learn*. 2017;9(5):934-944. [doi: [10.1016/j.cptl.2017.05.012](https://doi.org/10.1016/j.cptl.2017.05.012)] [Medline: [29233327](https://pubmed.ncbi.nlm.nih.gov/29233327/)]
11. Posel N, Mcgee JB, Fleischer DM. Twelve tips to support the development of clinical reasoning skills using virtual patient cases. *Med Teach*. 2015;37(9):813-818. [doi: [10.3109/0142159X.2014.993951](https://doi.org/10.3109/0142159X.2014.993951)] [Medline: [25523009](https://pubmed.ncbi.nlm.nih.gov/25523009/)]
12. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. *Acad Med*. 2010;85(10):1589-1602. [doi: [10.1097/ACM.0b013e3181edfe13](https://doi.org/10.1097/ACM.0b013e3181edfe13)] [Medline: [20703150](https://pubmed.ncbi.nlm.nih.gov/20703150/)]
13. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. *Med Educ*. 2009;43(4):303-311. [doi: [10.1111/j.1365-2923.2008.03286.x](https://doi.org/10.1111/j.1365-2923.2008.03286.x)] [Medline: [19335571](https://pubmed.ncbi.nlm.nih.gov/19335571/)]
14. Huang G, Reynolds R, Candler C. Virtual patient simulation at US and Canadian medical schools. *Acad Med*. 2007;82(5):446-451. [doi: [10.1097/ACM.0b013e31803e8a0a](https://doi.org/10.1097/ACM.0b013e31803e8a0a)] [Medline: [17457063](https://pubmed.ncbi.nlm.nih.gov/17457063/)]
15. Isaza-Restrepo A, Gómez MT, Cifuentes G, Argüello A. The virtual patient as a learning tool: a mixed quantitative qualitative study. *BMC Med Educ*. 2018;18(1):297. [FREE Full text] [doi: [10.1186/s12909-018-1395-8](https://doi.org/10.1186/s12909-018-1395-8)] [Medline: [30522478](https://pubmed.ncbi.nlm.nih.gov/30522478/)]
16. O'Neil T, Sewall J, Marchand R. Time-shared computer-assisted preclinical instruction: a short trial and evaluation. *J Med Educ*. 1976;51(09):765-767. [doi: [10.1097/00001888-197609000-00009](https://doi.org/10.1097/00001888-197609000-00009)] [Medline: [784998](https://pubmed.ncbi.nlm.nih.gov/784998/)]
17. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. *ArXiv*. Preprint posted online on April 25, 2024. [FREE Full text] [doi: [10.48550/arXiv.2404.13066](https://doi.org/10.48550/arXiv.2404.13066)]
18. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. *Med Teach*. 2025;47(1):40-42. [doi: [10.1080/0142159X.2024.2376879](https://doi.org/10.1080/0142159X.2024.2376879)] [Medline: [38992981](https://pubmed.ncbi.nlm.nih.gov/38992981/)]
19. Voigt H, Sugamiya Y, Lawonn K, Zarriß S, Takanishi A. LLM-powered virtual patient agents for interactive clinical skills training with automated feedback. *ArXiv*. Preprint posted online on August 19, 2025. [FREE Full text] [doi: [10.48550/arXiv.2508.13943](https://doi.org/10.48550/arXiv.2508.13943)]
20. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med Educ*. 2024;24(1):1391. [FREE Full text] [doi: [10.1186/s12909-024-06399-7](https://doi.org/10.1186/s12909-024-06399-7)] [Medline: [39609823](https://pubmed.ncbi.nlm.nih.gov/39609823/)]
21. Holderried F, Stegemann-Philipp C, Herschbach L, Moldt J, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT) powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. *JMIR Med Educ*. 2024;10:e53961. [FREE Full text] [doi: [10.2196/53961](https://doi.org/10.2196/53961)] [Medline: [38227363](https://pubmed.ncbi.nlm.nih.gov/38227363/)]
22. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med*. 2024;177(2):210-220. [doi: [10.7326/M23-2772](https://doi.org/10.7326/M23-2772)] [Medline: [38285984](https://pubmed.ncbi.nlm.nih.gov/38285984/)]
23. Mu Y, He D. The potential applications and challenges of ChatGPT in the medical field. *Int J Gen Med*. 2024;17:817-826. [FREE Full text] [doi: [10.2147/IJGM.S456659](https://doi.org/10.2147/IJGM.S456659)] [Medline: [38476626](https://pubmed.ncbi.nlm.nih.gov/38476626/)]
24. Kelly S, Smyth E, Murphy P, Pawlikowska T. A scoping review: virtual patients for communication skills in medical undergraduates. *BMC Med Educ*. 2022;22(1):429. [FREE Full text] [doi: [10.1186/s12909-022-03474-9](https://doi.org/10.1186/s12909-022-03474-9)] [Medline: [35659213](https://pubmed.ncbi.nlm.nih.gov/35659213/)]

25. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. *Med Educ*. 2020;54(9):786-795. [doi: [10.1111/medu.14152](https://doi.org/10.1111/medu.14152)] [Medline: [32162355](https://pubmed.ncbi.nlm.nih.gov/32162355/)]
26. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol Theory Pract*. 2005;8(1):19-32. [doi: [10.1080/1364557032000119616](https://doi.org/10.1080/1364557032000119616)]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467-473. [FREE Full text] [doi: [10.7326/M18-0850](https://doi.org/10.7326/M18-0850)] [Medline: [30178033](https://pubmed.ncbi.nlm.nih.gov/30178033/)]
28. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. California. OpenAI; 2018.
29. Aijing L, Shuangcheng Y, Lu M. Medical Literature Information Retrieval. Beijing. People's Medical Publishing House; 2005.
30. Cook DA, Reed DA. Appraising the quality of medical education research methods: the Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. *Acad Med*. 2015;90(8):1067-1076. [doi: [10.1097/ACM.0000000000000786](https://doi.org/10.1097/ACM.0000000000000786)] [Medline: [26107881](https://pubmed.ncbi.nlm.nih.gov/26107881/)]
31. Kmet LM, Cook LS, Lee RC. Standard quality assessment criteria for evaluating primary research papers from a variety of fields. University of Alberta. 2004. URL: <https://ualberta.scholaris.ca/items/b8ff8755-6efb-4fb2-941a-def0d418fd07> [accessed 2025-10-28]
32. Weisman D, Sugarman A, Huang YM, Gelberg L, Ganz PA, Comulada WS. 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*. 2025;9:e65670. [FREE Full text] [doi: [10.2196/65670](https://doi.org/10.2196/65670)] [Medline: [40246299](https://pubmed.ncbi.nlm.nih.gov/40246299/)]
33. Yi Y, Kim K. The feasibility of using generative artificial intelligence for history taking in virtual patients. *BMC Res Notes*. 2025;18(1):80. [FREE Full text] [doi: [10.1186/s13104-025-07157-8](https://doi.org/10.1186/s13104-025-07157-8)] [Medline: [39994780](https://pubmed.ncbi.nlm.nih.gov/39994780/)]
34. Thesen T, Alilonu NA, Stone S. AI Patient Actor: an open-access generative-AI app for communication training in health professions. *Med Sci Educ*. 2025;35(1):25-27. [doi: [10.1007/s40670-024-02250-2](https://doi.org/10.1007/s40670-024-02250-2)] [Medline: [40144116](https://pubmed.ncbi.nlm.nih.gov/40144116/)]
35. Takata T, Yamada R, Oliveira NRA, Xu K, Fujimoto M. Development of a virtual patient model for Kampo medical interview: new approach for enhancing empathy and understanding of Kampo medicine pathological concepts. 2024. Presented at: Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS); November 9-12, 2024; Himeji, Japan. [doi: [10.1109/scisisis61014.2024.10759962](https://doi.org/10.1109/scisisis61014.2024.10759962)]
36. Shindo N, Uto M. ChatGPT-based virtual standardized patient that amends overly detailed responses in objective structured clinical examinations. 2024. Presented at: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky; July 8-12, 2024:263-269; Recife, Brazil. [doi: [10.1007/978-3-031-64315-6_22](https://doi.org/10.1007/978-3-031-64315-6_22)]
37. Ng H, Koh A, Foong A, Ong J. Real-time hybrid language model for virtual patient conversations. 2023. Presented at: International Conference on Artificial Intelligence in Education; July 3-7, 2023:780-785; Tokyo, Japan. [doi: [10.1007/978-3-031-36272-9_71](https://doi.org/10.1007/978-3-031-36272-9_71)]
38. Grévisse C. RasPatient Pi: a low-cost customizable LLM-based virtual standardized patient simulator. In: Communications in Computer and Information Science. Germany. Springer Science and Business Media Deutschland GmbH; 2025:125-137.
39. Baugerud G, Johnson MS, Dianiska R, Røed RK, Powell MB, Lamb ME, et al. Using an AI-based avatar for interviewer training at children's advocacy centers: proof of concept. *Child Maltreat*. 2025;30(2):242-252. [FREE Full text] [doi: [10.1177/10775595241263017](https://doi.org/10.1177/10775595241263017)] [Medline: [38889731](https://pubmed.ncbi.nlm.nih.gov/38889731/)]
40. Abou Karam G. Revolutionizing medical education: ChatGPT3.5 ability to behave as a virtual patient. *Med Sci Educ*. 2024;34(6):1559-1564. [doi: [10.1007/s40670-024-02121-w](https://doi.org/10.1007/s40670-024-02121-w)] [Medline: [39758479](https://pubmed.ncbi.nlm.nih.gov/39758479/)]
41. Liu X, Wu C, Lai R, Lin H, Xu Y, Lin Y, et al. ChatGPT: when the artificial intelligence meets standardized patients in clinical training. *J Transl Med*. 2023;21(1):447. [FREE Full text] [doi: [10.1186/s12967-023-04314-0](https://doi.org/10.1186/s12967-023-04314-0)] [Medline: [37415217](https://pubmed.ncbi.nlm.nih.gov/37415217/)]
42. Benfatah M, Marfak A, Saad E, Hilali A, Nejari C, Youlyouz-Marfak I. Assessing the efficacy of ChatGPT as a virtual patient in nursing simulation training: a study on nursing students' experience. *Teach Learn Nurs*. 2024;19(3):e486-e493. [doi: [10.1016/j.teln.2024.02.005](https://doi.org/10.1016/j.teln.2024.02.005)]
43. Borg A, Jobs B, Huss V, Gentline C, Espinosa F, Ruiz M, et al. Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology. *Rheumatol Int*. 2024;44(12):3041-3051. [doi: [10.1007/s00296-024-05731-0](https://doi.org/10.1007/s00296-024-05731-0)] [Medline: [39412574](https://pubmed.ncbi.nlm.nih.gov/39412574/)]
44. Gray M, Baird A, Sawyer T, James J, DeBroux T, Bartlett M, et al. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of chatGPT: exploratory observational study. *JMIR Med Educ*. 2024;10:e50705. [FREE Full text] [doi: [10.2196/50705](https://doi.org/10.2196/50705)] [Medline: [38300696](https://pubmed.ncbi.nlm.nih.gov/38300696/)]
45. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. *JMIR Med Educ*. 2024;10:e59213. [FREE Full text] [doi: [10.2196/59213](https://doi.org/10.2196/59213)] [Medline: [39150749](https://pubmed.ncbi.nlm.nih.gov/39150749/)]

46. Gutiérrez Maquilón R, Uhl J, Schrom-Feiertag H, Tscheligi M. Integrating GPT-based AI into virtual patients to facilitate communication training among medical first responders: usability study of mixed reality simulation. *JMIR Form Res*. 2024;8:e58623. [FREE Full text] [doi: [10.2196/58623](https://doi.org/10.2196/58623)] [Medline: [39661979](https://pubmed.ncbi.nlm.nih.gov/39661979/)]
47. Sardesai N, Russo P, Martin J, Sardesai A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. *Postgrad Med J*. 2024;100(1182):237-241. [doi: [10.1093/postmj/qgad137](https://doi.org/10.1093/postmj/qgad137)] [Medline: [38240054](https://pubmed.ncbi.nlm.nih.gov/38240054/)]
48. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, et al. Enhancing medical interview skills through AI-simulated patient interactions: nonrandomized controlled trial. *JMIR Med Educ*. 2024;10:e58753. [FREE Full text] [doi: [10.2196/58753](https://doi.org/10.2196/58753)] [Medline: [39312284](https://pubmed.ncbi.nlm.nih.gov/39312284/)]
49. 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](https://doi.org/10.1007/s40670-025-02342-7)] [Medline: [40625933](https://pubmed.ncbi.nlm.nih.gov/40625933/)]
50. Borg A, Georg C, Jobs B, Huss V, Waldenlind K, Ruiz M, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. *J Med Internet Res*. 2025;27:e63312. [FREE Full text] [doi: [10.2196/63312](https://doi.org/10.2196/63312)] [Medline: [40053778](https://pubmed.ncbi.nlm.nih.gov/40053778/)]
51. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. *J Med Internet Res*. 2025;27:e68486. [FREE Full text] [doi: [10.2196/68486](https://doi.org/10.2196/68486)] [Medline: [39854611](https://pubmed.ncbi.nlm.nih.gov/39854611/)]
52. Ko H, Hovden EAS, Köpp UMS, Johnson MS, Baugerud GA. Using an AI-driven child chatbot avatar as a training tool for information gathering skills of dental and medical professionals: a pilot study. *Appl Cogn Psychol*. 2025;39(1):e70022. [doi: [10.1002/acp.70022](https://doi.org/10.1002/acp.70022)]
53. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Med Educ*. 2025;25(1):278. [FREE Full text] [doi: [10.1186/s12909-025-06877-6](https://doi.org/10.1186/s12909-025-06877-6)] [Medline: [39979969](https://pubmed.ncbi.nlm.nih.gov/39979969/)]
54. Rädcl-Ablass K, Schliz K, Schlick C, Meindl B, Pahr-Hosbach S, Schwendemann H, 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*. 2025;25(1):259. [FREE Full text] [doi: [10.1186/s12909-025-06756-0](https://doi.org/10.1186/s12909-025-06756-0)] [Medline: [39966894](https://pubmed.ncbi.nlm.nih.gov/39966894/)]
55. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, et al. Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: multimetric assessment. *J Med Internet Res*. 2025;27:e59435. [FREE Full text] [doi: [10.2196/59435](https://doi.org/10.2196/59435)] [Medline: [39742453](https://pubmed.ncbi.nlm.nih.gov/39742453/)]
56. Or AJ, Sukumar S, Ritchie HE, Sarrafpour B. Using artificial intelligence chatbots to improve patient history taking in dental education (pilot study). *J Dent Educ*. 2024;88(Suppl 3):1988-1990. [doi: [10.1002/jdd.13591](https://doi.org/10.1002/jdd.13591)] [Medline: [38783404](https://pubmed.ncbi.nlm.nih.gov/38783404/)]
57. Silverman J, Kurtz S, Draper J. *Skills for Communicating with Patients*. Boca Raton, FL. CRC Press; 2016.
58. Nielsen J. *Usability Engineering*. Burlington, MA. Morgan Kaufmann; 1994.
59. Wind LA, Van Dalen J, Muijtjens AMM, Rethans J. Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). *Med Educ*. 2004;38(1):39-44. [doi: [10.1111/j.1365-2923.2004.01686.x](https://doi.org/10.1111/j.1365-2923.2004.01686.x)] [Medline: [14962025](https://pubmed.ncbi.nlm.nih.gov/14962025/)]
60. Ollion E, Shen R, Macanovic A, Chatelain A. ChatGPT for text annotation? Mind the hype. *Open Science Framework*. 2023. URL: https://osf.io/preprints/socarxiv/x58kn_v1 [accessed 2025-10-25]
61. Wu X, Duan R, Ni J. Unveiling security, privacy, and ethical concerns of ChatGPT. *J Inform Intell*. 2024;2(2):102-115. [doi: [10.1016/j.jiixd.2023.10.007](https://doi.org/10.1016/j.jiixd.2023.10.007)]
62. Kristensen-McLachlan RD, Canavan M, Kárdos M, Jacobsen M, Aarøe L. Are chatbots reliable text annotators? Sometimes. *PNAS Nexus*. 2025;4(4):pgaf069. [FREE Full text] [doi: [10.1093/pnasnexus/pgaf069](https://doi.org/10.1093/pnasnexus/pgaf069)] [Medline: [40171238](https://pubmed.ncbi.nlm.nih.gov/40171238/)]
63. Larson DB, Koirala A, Cheuy LY, Paschali M, Van Veen D, Na HS, et al. Assessing completeness of clinical histories accompanying imaging orders using adapted open-source and closed-source large language models. *Radiology*. 2025;314(2):e241051. [doi: [10.1148/radiol.241051](https://doi.org/10.1148/radiol.241051)] [Medline: [39998369](https://pubmed.ncbi.nlm.nih.gov/39998369/)]
64. Spirling A. Why open-source generative AI models are an ethical way forward for science. *Nature*. 2023;616(7957):413. [doi: [10.1038/d41586-023-01295-4](https://doi.org/10.1038/d41586-023-01295-4)] [Medline: [37072520](https://pubmed.ncbi.nlm.nih.gov/37072520/)]
65. Gibney E. What are the best AI tools for research? Nature's guide. *Nature*. Feb 17, 2025;578:123-125. [doi: [10.1038/d41586-025-00437-0](https://doi.org/10.1038/d41586-025-00437-0)] [Medline: [39962251](https://pubmed.ncbi.nlm.nih.gov/39962251/)]
66. 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. [FREE Full text] [doi: [10.3389/frai.2025.1504805](https://doi.org/10.3389/frai.2025.1504805)] [Medline: [39931218](https://pubmed.ncbi.nlm.nih.gov/39931218/)]
67. Davis J, Van Bulck L, Durieux BN, Lindvall C. The temperature feature of ChatGPT: modifying creativity for clinical research. *JMIR Hum Factors*. 2024;11:e53559. [FREE Full text] [doi: [10.2196/53559](https://doi.org/10.2196/53559)] [Medline: [38457221](https://pubmed.ncbi.nlm.nih.gov/38457221/)]
68. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, et al. A scoping review of large language models for generative tasks in mental health care. *NPJ Digit Med*. 2025;8(1):230. [FREE Full text] [doi: [10.1038/s41746-025-01611-4](https://doi.org/10.1038/s41746-025-01611-4)] [Medline: [40307331](https://pubmed.ncbi.nlm.nih.gov/40307331/)]
69. Huwendiek S, Reichert F, Bosse H, de Leng BA, van der Vleuten CPM, Haag M, et al. Design principles for virtual patients: a focus group study among students. *Med Educ*. 2009;43(6):580-588. [doi: [10.1111/j.1365-2923.2009.03369.x](https://doi.org/10.1111/j.1365-2923.2009.03369.x)] [Medline: [19493183](https://pubmed.ncbi.nlm.nih.gov/19493183/)]

70. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. *BMC Med Educ.* 2022;22(1):365. [FREE Full text] [doi: [10.1186/s12909-022-03410-x](https://doi.org/10.1186/s12909-022-03410-x)] [Medline: [35550085](https://pubmed.ncbi.nlm.nih.gov/35550085/)]
71. Kalet AL, Song HS, Sarpel U, Schwartz R, Brenner J, Ark TK, et al. Just enough, but not too much interactivity leads to better clinical skills performance after a computer assisted learning module. *Med Teach.* 2012;34(10):833-839. [FREE Full text] [doi: [10.3109/0142159X.2012.706727](https://doi.org/10.3109/0142159X.2012.706727)] [Medline: [22917265](https://pubmed.ncbi.nlm.nih.gov/22917265/)]
72. Maicher KR, Stiff A, Scholl M, White M, Fosler-Lussier E, Schuler W, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity. *Med Teach.* 2022;45:1-7. [doi: [10.1080/0142159X.2022.2130216](https://doi.org/10.1080/0142159X.2022.2130216)] [Medline: [36346810](https://pubmed.ncbi.nlm.nih.gov/36346810/)]
73. Vzorin GD, Bukinich AM, Sedykh AV, Vetrova II, Sergienko EA. The emotional intelligence of the GPT-4 large language model. *Psychol Russ.* 2024;17(2):85-99. [doi: [10.11621/pir.2024.0206](https://doi.org/10.11621/pir.2024.0206)] [Medline: [39552777](https://pubmed.ncbi.nlm.nih.gov/39552777/)]
74. Qiu J, Yuan W, Lam K. The application of multimodal large language models in medicine. *Lancet Reg Health West Pac.* 2024;45:101048. [FREE Full text] [doi: [10.1016/j.lanwpc.2024.101048](https://doi.org/10.1016/j.lanwpc.2024.101048)] [Medline: [38524685](https://pubmed.ncbi.nlm.nih.gov/38524685/)]
75. Al Moubayed S, Beskow J, Skantze G, Granström B, Furhat: a back-projected human-like robot head for multiparty human-machine interaction. 2012. Presented at: International Training School on Cognitive Behavioural Systems, COST 2102; February 21-26, 2012:114-130; Dresden, Germany. [doi: [10.1007/978-3-642-34584-5_9](https://doi.org/10.1007/978-3-642-34584-5_9)]
76. Minh Trieu N, Truong Thinh N. A comprehensive design for biomimetic behavior of robotic head. *Mech Based Des Struct Mach.* 2024;53(5):3203-3224. [doi: [10.1080/15397734.2024.2420766](https://doi.org/10.1080/15397734.2024.2420766)]
77. Albert R, Patney A, Luebke D, Kim J. Latency requirements for foveated rendering in virtual reality. *ACM Trans Appl Percept.* 2017;14(4):1-13. [doi: [10.1145/3127589](https://doi.org/10.1145/3127589)]
78. Montenegro-Rueda M, Fernández-Cerero J, Fernández-Batanero JM, López-Meneses E. Impact of the implementation of ChatGPT in education: a systematic review. *Computers.* 2023;12(8):153. [doi: [10.3390/computers12080153](https://doi.org/10.3390/computers12080153)]
79. Löwquist E, Shorten G, Aboulafia A. Virtual reality-based medical training and assessment: the multidisciplinary relationship between clinicians, educators and developers. *Med Teach.* 2012;34(1):59-64. [doi: [10.3109/0142159X.2011.600359](https://doi.org/10.3109/0142159X.2011.600359)] [Medline: [22250676](https://pubmed.ncbi.nlm.nih.gov/22250676/)]
80. Zhang H, Yang J, Liu Z. Effect of teachers' teaching strategies on students' learning engagement: moderated mediation model. *Front Psychol.* 2024;15:1475048. [FREE Full text] [doi: [10.3389/fpsyg.2024.1475048](https://doi.org/10.3389/fpsyg.2024.1475048)] [Medline: [39691665](https://pubmed.ncbi.nlm.nih.gov/39691665/)]
81. Jennings PA, Greenberg MT. The prosocial classroom: teacher social and emotional competence in relation to student and classroom outcomes. *Rev Educ Res.* 2009;79(1):491-525. [doi: [10.3102/0034654308325693](https://doi.org/10.3102/0034654308325693)]
82. Frymier AB, Houser ML. The teacher - student relationship as an interpersonal relationship. *Commun Educ.* 2000;49(3):207-219. [doi: [10.1080/03634520009379209](https://doi.org/10.1080/03634520009379209)]
83. Titsworth S, McKenna TP, Mazer JP, Quinlan MM. The bright side of emotion in the classroom: do teachers' behaviors predict students' enjoyment, hope, and pride? *Commun Educ.* 2013;62(2):191-209. [doi: [10.1080/03634523.2013.763997](https://doi.org/10.1080/03634523.2013.763997)]
84. Edelbring S, Parodis I, Lundberg IE. Increasing reasoning awareness: video analysis of students' two-party virtual patient interactions. *JMIR Med Educ.* 2018;4(1):e4. [FREE Full text] [doi: [10.2196/mededu.9137](https://doi.org/10.2196/mededu.9137)] [Medline: [29487043](https://pubmed.ncbi.nlm.nih.gov/29487043/)]
85. Zerbini G, Schneider P, Reicherts M, Roob N, Jung-Can K, Kunz M, et al. A novel multi-measure approach to study medical students' communication performance and predictors of their communication quality—a cross-sectional study. *BMC Med Educ.* 2025;25(1):685. [FREE Full text] [doi: [10.1186/s12909-025-07018-9](https://doi.org/10.1186/s12909-025-07018-9)] [Medline: [40346504](https://pubmed.ncbi.nlm.nih.gov/40346504/)]
86. Aziz A, Farhan F, Hassan F, Qaiser A. Words are just noise, let your actions speak: impact of nonverbal communication on undergraduate medical education. *Pak J Med Sci.* 2021;37(7):1849-1853. [FREE Full text] [doi: [10.12669/pjms.37.7.4180](https://doi.org/10.12669/pjms.37.7.4180)] [Medline: [34912406](https://pubmed.ncbi.nlm.nih.gov/34912406/)]
87. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Educ Couns.* 2009;74(3):295-301. [doi: [10.1016/j.pec.2008.11.015](https://doi.org/10.1016/j.pec.2008.11.015)] [Medline: [19150199](https://pubmed.ncbi.nlm.nih.gov/19150199/)]
88. Baugh AD, Vanderbilt AA, Baugh RF. Communication training is inadequate: the role of deception, non-verbal communication, and cultural proficiency. *Med Educ Online.* 2020;25(1):1820228. [FREE Full text] [doi: [10.1080/10872981.2020.1820228](https://doi.org/10.1080/10872981.2020.1820228)] [Medline: [32938330](https://pubmed.ncbi.nlm.nih.gov/32938330/)]
89. Graf J, Loda T, Zipfel S, Wosnik A, Mohr D, Herrmann-Werner A. Communication skills of medical students: survey of self- and external perception in a longitudinally based trend study. *BMC Med Educ.* 2020;20(1):149. [FREE Full text] [doi: [10.1186/s12909-020-02049-w](https://doi.org/10.1186/s12909-020-02049-w)] [Medline: [32393242](https://pubmed.ncbi.nlm.nih.gov/32393242/)]
90. Vogel D, Meyer M, Harendza S. Verbal and non-verbal communication skills including empathy during history taking of undergraduate medical students. *BMC Med Educ.* 2018;18(1):157. [FREE Full text] [doi: [10.1186/s12909-018-1260-9](https://doi.org/10.1186/s12909-018-1260-9)] [Medline: [29970069](https://pubmed.ncbi.nlm.nih.gov/29970069/)]
91. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ.* 2019;5(1):e13930. [FREE Full text] [doi: [10.2196/13930](https://doi.org/10.2196/13930)] [Medline: [31199295](https://pubmed.ncbi.nlm.nih.gov/31199295/)]
92. Bui N, Nguyen G, Nguyen N, Vo B, Vo L, Huynh T, et al. Fine-tuning large language models for improved health communication in low-resource languages. *Comput Methods Programs Biomed.* 2025;263:108655. [FREE Full text] [doi: [10.1016/j.cmpb.2025.108655](https://doi.org/10.1016/j.cmpb.2025.108655)] [Medline: [39987667](https://pubmed.ncbi.nlm.nih.gov/39987667/)]

93. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus*. 2023;15(6):e40895. [FREE Full text] [doi: [10.7759/cureus.40895](https://doi.org/10.7759/cureus.40895)] [Medline: [37492832](https://pubmed.ncbi.nlm.nih.gov/37492832/)]
94. Dietrich N, Bradbury NC, Loh C. Prompt engineering for large language models in interventional radiology. *AJR Am J Roentgenol*. 2025;225(2):e2532956. [doi: [10.2214/AJR.25.32956](https://doi.org/10.2214/AJR.25.32956)] [Medline: [40334089](https://pubmed.ncbi.nlm.nih.gov/40334089/)]
95. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res*. 2023;25:e50638. [FREE Full text] [doi: [10.2196/50638](https://doi.org/10.2196/50638)] [Medline: [37792434](https://pubmed.ncbi.nlm.nih.gov/37792434/)]
96. HONE KS, GRAHAM R. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Nat Lang Eng*. 2001;6(3-4):287-303. [doi: [10.1017/s1351324900002497](https://doi.org/10.1017/s1351324900002497)]
97. Witmer BG, Singer MJ. Measuring presence in virtual environments: a presence questionnaire. *Presence Teleoper Virtual Environ*. 1998;7(3):225-240. [doi: [10.1162/105474698565686](https://doi.org/10.1162/105474698565686)]
98. Alsalamah A, Callinan C. The Kirkpatrick model for training evaluation: bibliometric analysis after 60 years (1959–2020). *Ind Commer Train*. 2021;54(1):36-63. [doi: [10.1108/ict-12-2020-0115](https://doi.org/10.1108/ict-12-2020-0115)]
99. Kolb DA. *Experiential Learning: Experience as the Source of Learning and Development*. New Jersey. FT Press; 2014.
100. Liao Y, Meng Y, Wang Y, Liu H, Wang Y, Wang Y. Automatic interactive evaluation for large language models with state aware patient simulator. *ArXiv*. Preprint posted online on July 15, 2024. [doi: [10.2139/ssrn.4890649](https://doi.org/10.2139/ssrn.4890649)]
101. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. *NPJ Digit Med*. 2024;7(1):84. [FREE Full text] [doi: [10.1038/s41746-024-01083-y](https://doi.org/10.1038/s41746-024-01083-y)] [Medline: [38570554](https://pubmed.ncbi.nlm.nih.gov/38570554/)]
102. Delisle M, Pradarelli JC, Panda N, Haynes AB, Hannenberg AA. Methods for scaling simulation-based teamwork training. *BMJ Qual Saf*. 2020;29(2):98-102. [doi: [10.1136/bmjqs-2019-009515](https://doi.org/10.1136/bmjqs-2019-009515)] [Medline: [31597783](https://pubmed.ncbi.nlm.nih.gov/31597783/)]
103. Schembre SM, Liao Y, Robertson MC, Dunton GF, Kerr J, Haffey ME, et al. Just-in-time feedback in diet and physical activity interventions: systematic review and practical design framework. *J Med Internet Res*. 2018;20(3):e106. [FREE Full text] [doi: [10.2196/jmir.8701](https://doi.org/10.2196/jmir.8701)] [Medline: [29567638](https://pubmed.ncbi.nlm.nih.gov/29567638/)]
104. Song C, Huang R, Hu S. Private-preserving language model inference based on secure multi-party computation. *Neurocomputing*. 2024;592:127794. [doi: [10.1016/j.neucom.2024.127794](https://doi.org/10.1016/j.neucom.2024.127794)]
105. Wang T, Zhai L, Yang T, Luo Z, Liu S. Selective privacy-preserving framework for large language models fine-tuning. *Inf Sci*. 2024;678:121000. [doi: [10.1016/j.ins.2024.121000](https://doi.org/10.1016/j.ins.2024.121000)]
106. Wang W, Niyato D, Xiong Z, Yin Z. Detection and authentication for cross-technology communication. *IEEE Trans Veh Technol*. 2025;74(2):3157-3171. [doi: [10.1109/tvt.2024.3467109](https://doi.org/10.1109/tvt.2024.3467109)]
107. Yao D, Li B. Is split learning privacy-preserving for fine-tuning large language models? *IEEE Trans Big Data*. 2024;1-12. [doi: [10.1109/tbdata.2024.3524101](https://doi.org/10.1109/tbdata.2024.3524101)]
108. Munjal K, Bhatia R. A systematic review of homomorphic encryption and its contributions in healthcare industry. *Complex Intell Systems*. 2022;9:1-28. [FREE Full text] [doi: [10.1007/s40747-022-00756-z](https://doi.org/10.1007/s40747-022-00756-z)] [Medline: [35531323](https://pubmed.ncbi.nlm.nih.gov/35531323/)]
109. Li X, Liu S, Lu R, Khan MK, Gu K, Zhang X. An efficient privacy-preserving public auditing protocol for cloud-based medical storage system. *IEEE J Biomed Health Inform*. 2022;26(5):2020-2031. [doi: [10.1109/JBHI.2022.3140831](https://doi.org/10.1109/JBHI.2022.3140831)] [Medline: [34990371](https://pubmed.ncbi.nlm.nih.gov/34990371/)]
110. Gong X, Gao J, Sun S, Zhong Z, Shi Y, Zeng H, et al. Adaptive compressed-based privacy-preserving large language model for sensitive healthcare. *IEEE J Biomed Health Inform*. 2025:PP. [doi: [10.1109/JBHI.2025.3558935](https://doi.org/10.1109/JBHI.2025.3558935)] [Medline: [40198302](https://pubmed.ncbi.nlm.nih.gov/40198302/)]
111. Montagna S, Ferretti S, Klopfenstein L, Florio A, Pengo M. Data decentralisation of LLM-based chatbot systems in chronic disease self-management. 2023. Presented at: GoodIT '23: Proceedings of the 2023 ACM Conference on Information Technology for Social Good; September 6-8, 2023:205-212; Lisbon, Portugal. [doi: [10.1145/3582515.3609536](https://doi.org/10.1145/3582515.3609536)]
112. Qu Y, Dai Y, Yu S, Tanikella P, Schrank T, Hackman T. A novel compact LLM framework for local, high-privacy EHR data applications. *ArXiv*. Preprint posted online on December 3, 2024. [FREE Full text]
113. Katalinic M, Schenk M, Franke S, Katalinic A, Neumuth T, Dietz A, et al. Generation of a realistic synthetic laryngeal cancer cohort for AI applications. *Cancers (Basel)*. 2024;16(3):639. [FREE Full text] [doi: [10.3390/cancers16030639](https://doi.org/10.3390/cancers16030639)] [Medline: [38339389](https://pubmed.ncbi.nlm.nih.gov/38339389/)]
114. Chen A, Chen DO. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. *Sci Rep*. 2022;12(1):17917. [FREE Full text] [doi: [10.1038/s41598-022-23011-4](https://doi.org/10.1038/s41598-022-23011-4)] [Medline: [36289292](https://pubmed.ncbi.nlm.nih.gov/36289292/)]
115. Cohen IG. What should ChatGPT mean for bioethics? *Am J Bioeth*. 2023;23(10):8-16. [doi: [10.1080/15265161.2023.2233357](https://doi.org/10.1080/15265161.2023.2233357)] [Medline: [37440696](https://pubmed.ncbi.nlm.nih.gov/37440696/)]
116. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. *JMIR Med Educ*. 2023;9:e48163. [FREE Full text] [doi: [10.2196/48163](https://doi.org/10.2196/48163)] [Medline: [37279048](https://pubmed.ncbi.nlm.nih.gov/37279048/)]
117. Sahu PK, Benjamin LA, Singh Aswal G, Williams-Persad A. ChatGPT in research and health professions education: challenges, opportunities, and future directions. *Postgrad Med J*. 2023;100(1179):50-55. [doi: [10.1093/postmj/qgad090](https://doi.org/10.1093/postmj/qgad090)] [Medline: [37819738](https://pubmed.ncbi.nlm.nih.gov/37819738/)]
118. Singh S, Ramakrishnan N. Is ChatGPT biased? A review. *Open Science Framework*. 2023. URL: <https://osf.io/preprints/osf/9xkbu> [accessed 2025-10-07]

119. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. *J Educ Eval Health Prof.* 2024;21:6. [FREE Full text] [doi: [10.3352/jeehp.2024.21.6](https://doi.org/10.3352/jeehp.2024.21.6)] [Medline: [38486402](https://pubmed.ncbi.nlm.nih.gov/38486402/)]
120. Nagi F, Salih R, Alzubaidi M, Shah H, Alam T, Shah Z, et al. Applications of artificial intelligence (AI) in medical education: a scoping review. *Stud Health Technol Inform.* 2023;305:648-651. [doi: [10.3233/SHTI230581](https://doi.org/10.3233/SHTI230581)] [Medline: [37387115](https://pubmed.ncbi.nlm.nih.gov/37387115/)]
121. Bomhof-Roordink H, Gärtner FR, Stiggelbout AM, Pieterse AH. Key components of shared decision making models: a systematic review. *BMJ Open.* 2019;9(12):e031763. [FREE Full text] [doi: [10.1136/bmjopen-2019-031763](https://doi.org/10.1136/bmjopen-2019-031763)] [Medline: [31852700](https://pubmed.ncbi.nlm.nih.gov/31852700/)]
122. Cook DA, Hargraves IG, Stephenson CR, Durning SJ. Management reasoning and patient-clinician interactions: insights from shared decision-making and simulated outpatient encounters. *Med Teach.* 2023;45(9):1025-1037. [doi: [10.1080/0142159X.2023.2170776](https://doi.org/10.1080/0142159X.2023.2170776)] [Medline: [36763491](https://pubmed.ncbi.nlm.nih.gov/36763491/)]
123. Elwyn G, Durand MA, Song J, Aarts J, Barr PJ, Berger Z, et al. A three-talk model for shared decision making: multistage consultation process. *BMJ.* 2017;359:j4891. [FREE Full text] [doi: [10.1136/bmj.j4891](https://doi.org/10.1136/bmj.j4891)] [Medline: [29109079](https://pubmed.ncbi.nlm.nih.gov/29109079/)]
124. Hargraves IG, Fournier AK, Montori VM, Bierman AS. Generalized shared decision making approaches and patient problems. Adapting AHRQ's SHARE approach for purposeful SDM. *Patient Educ Couns.* 2020;103(10):2192-2199. [FREE Full text] [doi: [10.1016/j.pec.2020.06.022](https://doi.org/10.1016/j.pec.2020.06.022)] [Medline: [32636085](https://pubmed.ncbi.nlm.nih.gov/32636085/)]
125. Cook DA, Durning SJ, Sherbino J, Gruppen LD. Management reasoning: implications for health professions educators and a research agenda. *Acad Med.* 2019;94(9):1310-1316. [doi: [10.1097/ACM.0000000000002768](https://doi.org/10.1097/ACM.0000000000002768)] [Medline: [31460922](https://pubmed.ncbi.nlm.nih.gov/31460922/)]
126. Cook DA, Sherbino J, Durning SJ. Management reasoning: beyond the diagnosis. *JAMA.* 2018;319(22):2267-2268. [doi: [10.1001/jama.2018.4385](https://doi.org/10.1001/jama.2018.4385)] [Medline: [29800012](https://pubmed.ncbi.nlm.nih.gov/29800012/)]
127. Cook DA, Stephenson CR, Gruppen LD, Durning SJ. Management reasoning: empirical determination of key features and a conceptual model. *Acad Med.* 2023;98(1):80-87. [doi: [10.1097/ACM.0000000000004810](https://doi.org/10.1097/ACM.0000000000004810)] [Medline: [35830267](https://pubmed.ncbi.nlm.nih.gov/35830267/)]
128. Deladisma AM, Cohen M, Stevens A, Wagner P, Lok B, Bernard T, et al. Do medical students respond empathetically to a virtual patient? *Am J Surg.* 2007;193(6):756-760. [doi: [10.1016/j.amjsurg.2007.01.021](https://doi.org/10.1016/j.amjsurg.2007.01.021)] [Medline: [17512291](https://pubmed.ncbi.nlm.nih.gov/17512291/)]
129. Qadir J. Engineering education in the era of ChatGPT: promise and pitfalls of generative AI for education. 2023. Presented at: IEEE Global Engineering Education Conference (EDUCON); May 1-4, 2023; Kuwait, Kuwait. [doi: [10.1109/educon54358.2023.10125121](https://doi.org/10.1109/educon54358.2023.10125121)]
130. Shang K, Chang C, Yang C. Collaboration among multiple large language models for medical question answering. ArXiv. Preprint posted online on May 22, 2025. [FREE Full text] [doi: [10.48550/arXiv.2505.16648](https://doi.org/10.48550/arXiv.2505.16648)]
131. van Diggele C, Roberts C, Burgess A, Mellis C. Interprofessional education: tips for design and implementation. *BMC Med Educ.* 2020;20(Suppl 2):455. [FREE Full text] [doi: [10.1186/s12909-020-02286-z](https://doi.org/10.1186/s12909-020-02286-z)] [Medline: [33272300](https://pubmed.ncbi.nlm.nih.gov/33272300/)]

Abbreviations

AI: artificial intelligence

CDM: clinical decision-making

LLM: large language model

MR: mixed reality

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

SASSI: Subjective Assessment of Speech System Interfaces

VR: virtual reality

Edited by T Leung, A Coristine; submitted 15.Jun.2025; peer-reviewed by A Altozano, L Zhu, N Zhang, P Sharma, N Acharya, H Maheshwari, E Popa; comments to author 15.Jul.2025; revised version received 04.Aug.2025; accepted 29.Sep.2025; published 13.Nov.2025

Please cite as:

Zeng J, Qi W, Shen S, Liu X, Li S, Wang B, Dong C, Zhu X, Shi Y, Lou X, Wang B, Yao J, Jiang G, Zhang Q, Cao S
Embracing the Future of Medical Education With Large Language Model-Based Virtual Patients: Scoping Review
J Med Internet Res 2025;27:e79091

URL: <https://www.jmir.org/2025/1/e79091>

doi: [10.2196/79091](https://doi.org/10.2196/79091)

PMID:

©Jianwen Zeng, Wenhao Qi, Shiyang Shen, Xin Liu, Sixie Li, Bing Wang, Chaoqun Dong, Xiaohong Zhu, Yankai Shi, Xiajing Lou, Bingsheng Wang, Jiani Yao, Guowei Jiang, Qiong Zhang, Shihua Cao. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 13.Nov.2025. 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 the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.