

Article

A Curriculum-Embedded Two-Session AI Chatbot-Based History-Taking Practicum in Korean Medicine Diagnostics

In-Young Choi ^{1,†}, Jundong Kim ^{2,†}, Ji-Hwan Kim ³ , Hye-Yoon Lee ⁴ , Won-Hwan Park ¹ , Chang-Eop Kim ^{2,*}  and Dong-Woo Lim ^{1,5,*} 

¹ Department of Diagnostics, College of Korean Medicine, Dongguk University, Goyang 10326, Republic of Korea; azrael123@dongguk.ac.kr (I.-Y.C.); diapwh@dongguk.ac.kr (W.-H.P.)

² Department of Physiology, College of Korean Medicine, Gachon University, Seongnam 13120, Republic of Korea; kimjd@gachon.ac.kr

³ Department of Sasang Constitutional Medicine, College of Korean Medicine, Pusan National University, Yangsan 50612, Republic of Korea; jani77@pusan.ac.kr

⁴ Department of Humanities and Social Medicine, College of Korean Medicine, Pusan National University, Yangsan 50612, Republic of Korea; drlee@pusan.ac.kr

⁵ Institute of Korean Medicine, Dongguk University, Goyang 10326, Republic of Korea

* Correspondence: eopchang@gachon.ac.kr (C.-E.K.); greatwoodong@dongguk.edu (D.-W.L.); Tel.: +82-10-4186-2985 (C.-E.K.); +82-31-961-5833 (D.-W.L.)

[†] These authors contributed equally to this work.

Abstract

Background: History-taking is a core clinical competency in Korean medicine diagnostics, but conventional training methods such as peer role-play and standardized patient-based education have limitations in providing repeated, individualized, and scalable practice opportunities. This study aimed to evaluate the feasibility and educational value of a two-session AI chatbot-based history-taking practicum with automated feedback in Korean medicine diagnostics. **Methods:** This prospective single-arm repeated-measures educational study was conducted with fourth-year students at the College of Korean Medicine, Dongguk University, in May and June 2026. A total of 76 students participated in two chatbot-assisted history-taking sessions using dizziness and shoulder pain scenarios. Students completed surveys on baseline AI familiarity, chatbot experience, usability, and self-efficacy. Self-efficacy was assessed at three time points: before the first session, after the first session, and after the second session. Chatbot-generated feedback scores were compared between session 1 and 2 for each scenario using paired complete-case analyses. Open-ended responses were descriptively categorized. **Results:** Students rated the chatbot-based practicum positively in terms of active participation, perceived usefulness, accessibility, and convenience. Item-level self-efficacy analysis showed significant time effects in two domains: planning the conversation, and closing the conversation appropriately. The overall mean self-efficacy score gradually increased from 3.739 ± 0.546 before the first session to 3.887 ± 0.591 after the second session; however, the overall time effect did not reach statistical significance. Chatbot-generated feedback scores showed scenario-dependent patterns. Scores for the shoulder pain scenario increased from session 1 to session 2 before adjustment, but this change did not remain significant after Holm correction; scores for the dizziness scenario showed a non-significant decreasing trend. Open-ended responses indicated that students valued repeated practice and immediate feedback, while also noting limitations related to feedback accuracy, realism of patient responses, and the lack of physical examination or multimodal diagnostic information. **Conclusions:** The chatbot-assisted practicum was feasible and favorably perceived by students, with selected item-level changes and a modest non-significant upward trend in overall self-efficacy in this exploratory educational study. These findings support the potential role



Academic Editors: Woochun Jun and Pilar García-Díaz

Received: 23 June 2026

Revised: 15 July 2026

Accepted: 17 July 2026

Published: 19 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

of AI chatbot-based simulation as a supplementary, scalable tool for repeated history-taking practice and formative feedback, rather than as a replacement for performance-based clinical skills training. Given the single-arm design and reliance on learner-reported outcomes, these findings should be interpreted as exploratory. Future controlled studies should incorporate objective performance outcomes, expert-validated automated scoring, non-AI comparison groups, and more realistic multimodal clinical scenarios to determine the educational effectiveness of chatbot-assisted history-taking training.

Keywords: artificial intelligence chatbot; history-taking training; Korean medicine education; self-efficacy; medical education

1. Introduction

History-taking is a fundamental competency for clinicians, providing the basis for accurate diagnosis, effective treatment planning, and meaningful patient communication [1]. In Korean medicine education, history-taking skills are typically developed through standardized patient (SP) programs or peer role-play exercises [2]. However, these approaches have inherent limitations, including restricted opportunities for repeated practice, high operational costs, and challenges in ensuring consistency and expertise in patient portrayal [3]. As a result, the development of cost-effective, scalable, and educationally robust alternatives remains an ongoing challenge.

In recent years, artificial intelligence (AI)-driven tools, including large language model (LLM)-based chatbots, have been introduced into medical education [4]. These technologies provide realistic, interactive simulations with virtual patients, offering flexible access, tailored feedback, and the ability to simulate a wide range of clinical scenarios [5]. Chatbots have been reported to enhance learners' communication skills, diagnostic reasoning, and self-efficacy, while reducing resource burdens associated with traditional SP programs. To address the limitations of doctor-patient simulation training in Korean medicine education, recent efforts have been made to develop AI-based patient chatbots [6,7]. Recent reviews of LLM-based virtual patients and AI-assisted clinical skills education have highlighted their potential for scalable practice and individualized feedback, while emphasizing unresolved challenges related to assessment validity, clinical realism, learner safety, and objective demonstration of skill improvement [8–10]. A recent study has also suggested that LLM-based simulated patient interactions combined with structured automated feedback may support clinical decision-making and history-taking practice [11].

Despite these advances, evidence regarding the educational value of AI chatbots in Korean medicine history-taking training remains limited. Previous studies in this area have mainly focused on post-intervention satisfaction, usability, or descriptive educational experiences [6,7]. Recently, our team developed a pilot version of a Clinical Performance Examination (CPX) Practicing Chatbot utilizing prompt engineering, which demonstrated the feasibility of incorporating large language model-based conversational agents into Korean medicine education [12]. According to the report, the pilot-developed CPX Practicing Chatbot demonstrated excellent performance, achieving near-perfect scores across multiple evaluation criteria, including correctness, confidentiality, naturalness, and compliance [12]. However, important challenges remain regarding the validation of automated assessment, objective demonstration of clinical skill improvement, and the limited ability of text-based systems to reproduce non-verbal, physical examination, and multimodal aspects of clinical encounters [13].

More recently, a pilot randomized controlled trial comparing AI chatbot-based simulation with peer role-play for OSCE preparation in Korean medicine suggested that the two methods may have complementary strengths [14]. Peer role-play appeared to better support realistic interaction and procedural performance, whereas AI chatbot-based training showed potential for self-paced, repeatable, and feedback-supported learning, particularly in patient education and clinical reasoning-related domains [14]. These findings indicate that AI chatbots should not be regarded simply as replacements for human-based training, but rather as supplementary tools that may support specific components of clinical skills education.

Self-efficacy refers to an individual's belief in their capacity to successfully perform a specific task or handle a given situation [15]. In medical education, self-efficacy is a critical determinant of learners' clinical competence and problem-solving ability [16]. Students with higher self-efficacy are more likely to engage actively in learning, persist in challenging circumstances, and ultimately achieve better clinical performance [17]. Although self-efficacy does not directly represent objective competence, it remains a meaningful educational outcome in exploratory studies of new instructional tools. When combined with usability assessment, chatbot-generated feedback scores, and open-ended learner reflections, self-efficacy can provide useful insight into how students experience and respond to AI-assisted clinical skills training.

Therefore, the present study aimed to evaluate a two-session AI chatbot-based history-taking practicum embedded within a Korean medicine diagnostics course. Unlike previous studies that primarily examined one-time chatbot use or compared chatbot training with peer role-play, the present study focused on repeated chatbot-assisted practice with automated scoring and individualized feedback within a regular Korean medicine diagnostics course. Specifically, we examined students' baseline familiarity with AI chatbots, their perceptions of chatbot usability and learning experience, changes in history-taking self-efficacy across three assessment points, scenario-specific changes in chatbot-generated feedback scores, and qualitative feedback regarding the strengths, limitations, and future improvement of the chatbot-based practicum. Through this approach, we sought to clarify the feasibility and educational value of AI chatbot-assisted history-taking practice as a supplementary learning tool in Korean medicine diagnostics education.

2. Materials and Methods

2.1. Study Design and Participants

This study employed a prospective, single-arm, repeated-measures educational design to evaluate a two-session AI chatbot-based history-taking practicum embedded within a Korean medicine diagnostics course. The study was conducted with fourth-year undergraduate students enrolled in the College of Korean Medicine at Dongguk University in 2026.

A total of 88 students participated in the practicum. Among them, 76 students voluntarily agreed to participate in the study and completed the relevant surveys across the two-session practicum; these students were included in the analysis (46 men and 30 women). Participation in the educational activity was part of the regular curriculum; however, participation in the research survey was voluntary. The survey data were collected anonymously, and no personally identifiable information was included in the analysis.

The study protocol was reviewed and approved by the Institutional Review Board of Dongguk University (DGU IRB 20260022; initial approval). Written consent information was provided to all students before data collection.

2.2. Educational Intervention

The intervention was embedded within the Korean medicine diagnostics practicum and consisted of two chatbot-assisted history-taking sessions conducted on separate class days. The first session was conducted on 22 May 2026, and the second session was conducted on 5 June 2026. Each session was designed to provide students with repeated opportunities to practice medical history-taking and receive immediate feedback using an AI-based virtual patient chatbot.

A CPX chatbot platform was used to simulate virtual patients presenting with common clinical scenarios relevant to Korean medicine diagnostics [12]. The scenarios included two cases each of dizziness and shoulder pain. The chatbot system was developed by adapting standardized patient scenarios into a text-based conversational format and implementing prompt engineering techniques, including role prompting and few-shot prompting [12]. The platform was designed to respond within the boundaries of predefined clinical scenarios while allowing students to conduct history-taking in a free-text conversational format.

During each session, students individually accessed the chatbot through a web-based interface, reviewed the patient's background information, and conducted a history-taking interview with the virtual patient. After completing the interview, students submitted their results and received chatbot-generated feedback. The feedback included scenario-specific scores and comments regarding strengths, missing or insufficiently explored history-taking items, and areas for improvement.

A visual flow diagram of the two-session chatbot-based history-taking practicum is presented in Figure 1.

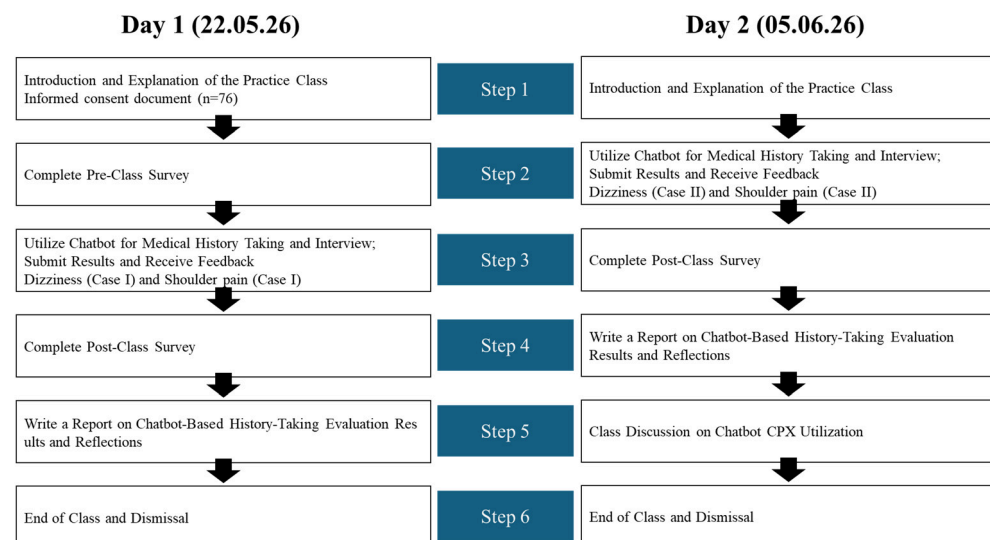


Figure 1. Whole process of chatbot-based medical history-taking class.

2.3. Data Collection: Questionnaire

The survey consisted of four main components: baseline understanding and experience of AI chatbots, evaluation of chatbot experience, evaluation of chatbot utilization, and self-efficacy in medical communication and history-taking. Self-efficacy was assessed using 12 items adapted from the Korean version of the Self-Efficacy Scale for Medical Communication (KSE-12) [18]. The wording of the items was modified to fit the context of chatbot-based history-taking practice in Korean medicine diagnostics.

The KSE-12 items were adapted to reflect chatbot-based history-taking practice in Korean medicine diagnostics while preserving the original focus on self-efficacy in medical communication. The wording was modified by the research team to fit the context of

student–chatbot interaction and Korean medicine diagnostic inquiry. Formal pilot testing or independent psychometric validation of the adapted wording was not conducted.

Chatbot usability was measured using items adapted and refined from previously developed AI usability and satisfaction measures [7]. The Questionnaire for Evaluation of Chatbot Experience was designed from a previous survey conducted with students after role-play in clerkship training [19].

A five-point Likert scale ranging from 1 (“strongly disagree”) to 5 (“strongly agree”) was applied to the survey items, and the responses were evaluated accordingly. For open-ended feedback, students provided qualitative comments on strengths, weaknesses, and suggestions for improvement, based on a modified version of the open-ended question format used in Min et al. (2023) [20].

2.4. Chatbot-Generated Feedback Scores

The chatbot provided scenario-specific feedback scores after each history-taking interaction. Feedback scores were collected for the shoulder pain and dizziness scenarios on both session 1 and 2. For the analysis of chatbot-generated feedback scores, scenario-specific paired complete-case analyses were conducted. Students were included in the paired comparison for a given scenario only if they had available scores for both Session 1 and 2 for that scenario.

The chatbot-generated scores were interpreted as formative assessment indicators reflecting scenario-specific history-taking performance and feedback, rather than as definitive objective measures of clinical competence.

2.5. Data Analysis and Statistical Analysis

Responses to the survey were collected anonymously using de-identified participant codes via Google Forms before and after the chatbot-based history-taking practice session, and the data were subsequently analyzed.

Categorical variables are presented as frequencies and percentages, and continuous or Likert-scale variables are presented as mean $\pm$ standard deviation.

For item-level self-efficacy analyses, Friedman tests were used to evaluate changes across the three repeated measurements: before the first training session, after the first session, and after the second session. To control for multiple testing across the 12 item-level Friedman omnibus analyses, *p*-values were adjusted using the Benjamini–Hochberg false discovery rate procedure. When appropriate, post hoc pairwise comparisons were conducted using Wilcoxon signed-rank tests with Holm correction for multiple comparisons.

The overall self-efficacy score was calculated as the mean of the 12 self-efficacy items for each student at each assessment point. Changes in the overall mean self-efficacy score across the three assessment points were analyzed using repeated-measures analysis of variance (ANOVA), with assessment of residual normality and sphericity. Greenhouse–Geisser-corrected results were reported when the sphericity assumption was violated. Pairwise comparisons between assessment points were performed using paired *t*-tests, and Holm-adjusted *p*-values were calculated to account for multiple comparisons.

For chatbot-generated feedback scores, scenario-specific paired comparisons between Session 1 and 2 were performed using Wilcoxon signed-rank tests. Holm correction was applied for multiple comparisons across the two scenarios. Because of missing feedback scores at some time points, paired analyses were conducted using complete paired data for each scenario.

All statistical tests were two-sided, and *p*-values less than 0.05 were considered statistically significant. For adjusted analyses, Holm-adjusted *p*-values were used to determine statistical significance. Statistical analysis was performed using GraphPad Prism (version 5.2).

3. Results

3.1. Baseline Understanding of Artificial Intelligence

Table 1 summarizes the participants' baseline understanding and experience of AI chatbots before the chatbot-based history-taking practicum. All participants had previously heard of AI chatbots and had used them at least once (76/76, 100%). Regarding the frequency of chatbot use, 37 students (48.7%) reported using AI chatbots very frequently, almost daily, while 24 students (31.6%) used them frequently, approximately twice a week. The remaining 15 students (19.7%) reported occasional use, approximately once a week. No student reported seldom or no use of AI chatbots.

Table 1. Understanding of artificial intelligence chatbot (Part 1).

Question		Answer (Total n = 76 (%))	
1	Have you ever heard of AI chatbots?	Yes	76 (100)
		No	0 (0)
2	Have you ever used an AI chatbot?	Yes	76 (100)
		No	0 (0)
2-1	How frequently do you use AI chatbots?	Very frequently (Almost daily)	37 (48.7)
		Frequently (Twice a week)	24 (31.6)
		Occasionally (Once a week)	15 (19.7)
		Seldom (Once a month)	0 (0)
		Never or nonresponse	0 (0)
3	Are you familiar with the characteristics of AI chatbots, such as LLMs, conversational AI, and chatbot functionalities?	Yes	58 (76.3)
		No	18 (23.7)
4	Do you subscribe to a paid service of a conversational AI?	Yes	61 (80.3)
		No	15 (19.7)

Most participants also reported familiarity with the characteristics of AI chatbots, including large language models, conversational AI, and chatbot functionalities. Specifically, 58 students (76.3%) answered that they were familiar with these characteristics, whereas 18 students (23.7%) were not. In addition, 61 students (80.3%) reported subscribing to a paid conversational AI service, while 15 students (19.7%) did not. These findings indicate that the participants entered the practicum with a high level of prior exposure to and familiarity with AI chatbot technologies.

Table 2 presents students' baseline perceptions of AI chatbots and their potential role in medical and Korean medicine education. Overall, students showed favorable attitudes toward AI-based education. The highest-rated item was "I believe that AI chatbots have educational benefits" (4.487 ± 0.600), followed by "I believe that clinical education using AI chatbots should be implemented" (4.342 ± 0.664), "I believe that Korean medicine education should incorporate AI chatbots" (4.316 ± 0.752), and "Education based on AI chatbot technology will improve clinical skills" (4.289 ± 0.689). Students also reported that they had frequently observed AI chatbots being used in their immediate environment (4.289 ± 0.727) and showed interest in participating in Korean medicine education using AI chatbots (4.276 ± 0.723).

In contrast, the item "I find information about AI chatbots difficult to understand" showed a relatively low mean score (2.395 ± 0.967), suggesting that most students did not perceive AI chatbot-related information as highly difficult to understand. Collectively, these results suggest that the participants were not only familiar with AI chatbots but also generally receptive to their educational application in Korean medicine training.

Table 2. Understanding of artificial intelligence chatbots (Part 2).

	Question	Mean $\pm$ SD
1	I have read about AI chatbots through various media sources (e.g., internet, articles, news).	3.76 $\pm$ 0.81
2	I find information about AI chatbots difficult to understand.	2.40 $\pm$ 0.97
3	I have seen AI chatbots being used in my immediate environment.	4.29 $\pm$ 0.73
4	I am aware that AI chatbots are used for diagnosis or treatment in medical settings.	3.75 $\pm$ 1.02
5	I am aware that AI chatbots are used in medical education.	4.20 $\pm$ 0.80
6	I believe that AI chatbots have educational benefits.	4.49 $\pm$ 0.60
7	I am interested in participating in lectures involving AI chatbots.	4.20 $\pm$ 0.73
8	I believe that Korean medicine education should incorporate AI chatbots.	4.32 $\pm$ 0.75
9	I am interested in participating in Korean medicine education using AI chatbots.	4.28 $\pm$ 0.72
10	Education based on AI chatbot technology will improve clinical skills.	4.29 $\pm$ 0.69
11	I believe that clinical education using AI chatbots should be implemented.	4.34 $\pm$ 0.66

3.2. Evaluation of Chatbot Experience and Utilization

The chatbot experience was rated positively (Table 3). In the evaluation of chatbot experience, the highest-rated item was active participation in the medical interview practice using a chatbot (4.59 $\pm$ 0.61). Items related to perceived educational usefulness also showed favorable scores, including assessment of diagnostic ability (4.38 $\pm$ 0.69), improvement of clinical knowledge (4.28 $\pm$ 0.73), and improvement of communication skills (4.22 $\pm$ 0.84).

Table 3. Evaluation of chatbot experience.

	Question	Mean $\pm$ SD
1	The medical history-taking practice using a chatbot was interesting.	4.42 $\pm$ 0.71
2	I actively participated in the medical history-taking practice using a chatbot.	4.59 $\pm$ 0.61
3	My interest in diagnosing patients increased after the medical interview practice using a chatbot.	4.14 $\pm$ 0.85
4	I believe that role-playing using a chatbot is appropriate for practical training in Korean Medicine Diagnostics.	4.49 $\pm$ 0.73
5	I would recommend applying role-playing (patient–doctor) using a chatbot in other classes.	4.38 $\pm$ 0.72
6	I agree with using chatbot-based role-playing for future students in Korean Medicine Diagnostics practical training.	4.51 $\pm$ 0.62
7	Role-playing with a chatbot helped improve my clinical knowledge.	4.28 $\pm$ 0.73
8	Role-playing with a chatbot was beneficial for assessing my diagnostic ability.	4.38 $\pm$ 0.69
9	Role-playing with a chatbot helped improve my communication skills.	4.22 $\pm$ 0.84
10	I gained confidence in interviewing patients with shoulder pain through role-playing with a chatbot.	4.14 $\pm$ 0.90
11	I gained confidence in interviewing patients with vertigo through role-playing with a chatbot.	3.99 $\pm$ 0.90
12	Role-playing with a chatbot will help improve my ability to respond to real-life situations.	4.39 $\pm$ 0.65
13	Role-playing with a chatbot will help me when treating patients as a future Korean medicine doctor.	4.38 $\pm$ 0.65

The usability evaluation also showed generally high scores (Table 4). The highest-rated item was convenience as a non-face-to-face service (4.55 $\pm$ 0.62), followed by usability without

prior knowledge (4.48 ± 0.62). Regarding interaction quality, students agreed that the chatbot provided appropriate responses (4.21 ± 0.77) and understood their questions well (4.16 ± 0.82). However, the item assessing whether the chatbot's responses were similar to those of real patients showed the lowest score among the utilization items (3.68 ± 1.13).

Table 4. Evaluation of chatbot utilization.

	Question	Mean $\pm$ SD
1	The role-playing activity using a chatbot was useful.	4.28 ± 0.69
2	Participating in role-playing with a chatbot saved time.	4.37 ± 0.71
3	I was able to participate in role-playing with a chatbot regardless of time and location.	4.28 ± 0.70
4	The chatbot's design (prompts) made role-playing with the chatbot easy to perform	4.13 ± 0.82
5	It was interesting to obtain information through role-playing with a chatbot instead of a real patient.	4.26 ± 0.79
6	The conversation process (interaction) with the chatbot was interesting.	4.33 ± 0.77
7	Accessing the chatbot was not difficult.	4.01 ± 1.08
8	The menu and terminology used in the chatbot's responses were not difficult.	4.41 ± 0.61
9	I was able to use the chatbot without prior knowledge.	4.48 ± 0.62
10	The chatbot understood my questions well.	4.16 ± 0.82
11	The chatbot provided appropriate responses to my questions.	4.21 ± 0.77
12	I trust that the chatbot's responses are similar to those of real patients.	3.68 ± 1.13
13	Interacting with the chatbot was not difficult.	4.30 ± 0.67
14	Despite my spelling errors, the chatbot provided appropriate responses to my questions.	4.42 ± 0.68
15	The instructions for using the chatbot were appropriate.	4.46 ± 0.60
16	The chatbot was convenient to use as it was a non-face-to-face service.	4.55 ± 0.62

3.3. Changes in Self-Efficacy After Chatbot Training

The adapted 12-item self-efficacy measure showed high internal consistency at all three assessment points, with Cronbach's alpha values of 0.905, 0.934, and 0.908, respectively. Changes in self-efficacy across the three assessment points are presented in Tables 5 and 6 and Figure 2. In the item-level analysis, self-efficacy scores generally showed increasing trends from the 1st pre-survey to the 2nd post-survey in several items (Table 5). Significant overall time effects were observed in two items after FDR adjustment: making an agenda or plan for the conversation (Q2; 3.51 ± 0.81 , 3.75 ± 0.80 , and 3.82 ± 0.78 ; $p = 0.045$), and closing the conversation by ensuring that the patient's questions had been answered (Q12; 3.79 ± 0.66 , 3.97 ± 0.71 , and 4.09 ± 0.73 ; $p = 0.045$). Q4 showed a non-monotonic pattern across the three assessment points with minor changes. More consistent increasing patterns were observed in Q1, Q2, and Q12. Other items did not show statistically significant overall changes across the three assessment points.

The overall mean self-efficacy score, calculated as the mean of the 12 items for each student, increased gradually across the three assessment points: 3.739 ± 0.546 before the first training session, 3.815 ± 0.625 after the first training session, and 3.887 ± 0.591 after the second training session (Table 6). However, Mauchly's test indicated that the sphericity assumption was violated ($W = 0.882$, $\chi^2(2) = 9.252$, $p = 0.010$); therefore, Greenhouse–Geisser-corrected results were used. The overall time effect was not statistically significant ($F(1.790, 134.224) = 2.720$, $p = 0.076$, partial $\eta^2 = 0.035$).

Table 5. Comparison of Self-Efficacy score across three time points.

Question	Mean $\pm$ SD			Raw <i>p</i> -Value	FDR-Adjusted <i>p</i> -Value
	1st Pre-Survey	1st Post-Survey	2nd Post-Survey		
1	How certain are you that you are able to successfully identify the issues the patient wishes to address during the conversation? 3.62 $\pm$ 0.83	3.82 $\pm$ 0.80	3.86 $\pm$ 0.76	0.030 *	0.097
2	How certain are you that you are able to successfully make an agenda/plan for the conversation with the patient? 3.51 $\pm$ 0.81	3.75 $\pm$ 0.80	3.82 $\pm$ 0.78	0.004 *	0.045 *
3	How certain are you that you are able to successfully urge the patient to expand on his or her problems/worries? 3.64 $\pm$ 0.81	3.78 $\pm$ 0.79	3.87 $\pm$ 0.85	0.168	0.288
4	How certain are you that you are able to successfully listen attentively without interrupting or changing of focus? 4.21 $\pm$ 0.70	4.11 $\pm$ 0.68	4.26 $\pm$ 0.87	0.032 *	0.097
5	How certain are you that you are able to successfully encourage the patient to express thoughts and feelings? 3.72 $\pm$ 0.83	3.76 $\pm$ 0.98	3.86 $\pm$ 0.86	0.589	0.706
6	How certain are you that you are able to successfully structure the conversation with the patient? 3.50 $\pm$ 0.90	3.63 $\pm$ 0.91	3.75 $\pm$ 0.90	0.082	0.196
7	How certain are you that you are able to successfully demonstrate appropriate non-verbal behavior (eye contact, facial expression, placement, posture, and voicing)? 3.86 $\pm$ 0.80	3.78 $\pm$ 0.87	3.95 $\pm$ 0.88	0.230	0.345
8	How certain are you that you are able to successfully show empathy (acknowledge the patient's views and feelings)? 3.82 $\pm$ 0.80	3.74 $\pm$ 0.90	3.89 $\pm$ 0.83	0.409	0.545
9	How certain are you that you are able to successfully clarify what the patient knows in order to communicate the right amount of information? 3.74 $\pm$ 0.79	3.72 $\pm$ 0.78	3.76 $\pm$ 0.88	0.963	0.963
10	How certain are you that you are able to successfully check patient's understanding of the information given? 3.67 $\pm$ 0.72	3.89 $\pm$ 0.78	3.78 $\pm$ 0.87	0.125	0.251
11	How certain are you that you are able to successfully make a plan based on shared decisions between you and the patient? 3.79 $\pm$ 0.68	3.83 $\pm$ 0.82	3.76 $\pm$ 0.83	0.838	0.914
12	How certain are you that you are able to successfully close the conversation by assuring, that the patient's questions have been answered? 3.79 $\pm$ 0.66	3.97 $\pm$ 0.71	4.09 $\pm$ 0.73	0.008 *	0.045 *

Raw *p*-values were obtained from Friedman tests. *p*-values were adjusted across the 12 item-level omnibus tests using the Benjamini–Hochberg false discovery rate procedure. Statistical significance was determined based on FDR-adjusted *p*-values. * indicates *p* < 0.05.

Pairwise comparisons showed that the increase from the 1st pre-training assessment to the 2nd post-training assessment was statistically significant before adjustment (mean difference, +0.148; raw *p* = 0.025), but this result did not remain significant after Holm correction (adjusted *p* = 0.075). Comparisons between the 1st pre-training and 1st post-training assessments and between the 1st post-training and 2nd post-training assessments were not statistically significant.

Figure 2 visualizes the changes in self-efficacy scores across repeated chatbot-based history-taking training. The item-level plot shows that several self-efficacy items tended to increase across the three assessment points, while the overall mean self-efficacy score

showed a gradual upward trend. In the pairwise comparison of overall mean self-efficacy scores, the increase from the 1st pre-training assessment to the 2nd post-training assessment reached statistical significance before adjustment (raw $p = 0.025$).

Table 6. Changes in mean self-efficacy scores across three assessment points.

Assessment Point	Mean	SD	95% CI
1st pre-training	3.739	0.546	3.614–3.864
1st post-training	3.815	0.625	3.672–3.958
2nd post-training	3.887	0.591	3.752–4.022
Analysis	Mean difference	Raw p -value	Adjusted p -value
1st pre vs. 1st post	+0.076	0.294	0.345
1st pre vs. 2nd post	+0.148	0.025	0.075
1st post vs. 2nd post	+0.072	0.172	0.345
Repeated-measures ANOVA	* Greenhouse–Geisser-corrected $F(1.790, 134.224) = 2.720$, $p = 0.076$, partial $\eta^2 = 0.035$		

* Greenhouse–Geisser correction was applied because Mauchly’s test indicated violation of sphericity.

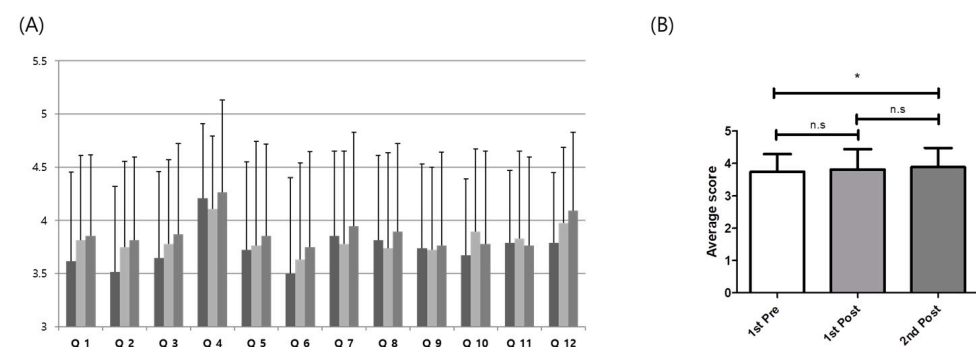


Figure 2. Changes in mean self-efficacy scores across repeated chatbot-based history-taking training. (A) Mean scores for each of the 12 adapted self-efficacy items across three assessment points: before the first training session, after the first training session, and after the second training session. (B) Overall self-efficacy score, calculated as the mean of the 12 items for each student, across the same three assessment points. Data are presented as mean $\pm$ SD. * $p < 0.05$; n.s., not significant.

3.4. Changes in Chatbot-Generated Feedback Scores

Chatbot-generated feedback scores were compared between session 1 and 2 for the shoulder pain and dizziness scenarios (Table 7 and Figure 3). In the shoulder pain scenario, the mean score increased from 8.36 ± 2.37 in session 1 to 8.99 ± 2.35 in session 2. This change was significant before adjustment but did not remain significant after Holm correction (paired $n = 74$; Wilcoxon signed-rank test, raw $p = 0.039$; Holm-adjusted $p = 0.078$).

Table 7. Descriptive comparison of chatbot-generated feedback scores between session-specific cases.

Case Scenario	Paired n (Session 1, 2)	Session 1 Score, Mean $\pm$ SD	Session 2 Score, Mean $\pm$ SD	Mean Difference	Wilcoxon p -Value	Holm Adjusted p -Value
Shoulder pain	74	8.36 ± 2.37	8.99 ± 2.35	+0.61	0.039	0.078
Dizziness	74	8.41 ± 2.05	7.76 ± 2.22	−0.59	0.053	0.078

In contrast, the dizziness scenario showed a non-significant decreasing trend from 8.41 ± 2.05 to 7.76 ± 2.22 (paired $n = 74$; raw $p = 0.053$; Holm-adjusted $p = 0.078$).

Although the Session 1 and Session 2 cases shared the same chief complaint within each condition, they represented different Korean medicine pattern-identification scenarios and therefore differed in the clinically relevant history-taking items required to elicit the corresponding pattern.

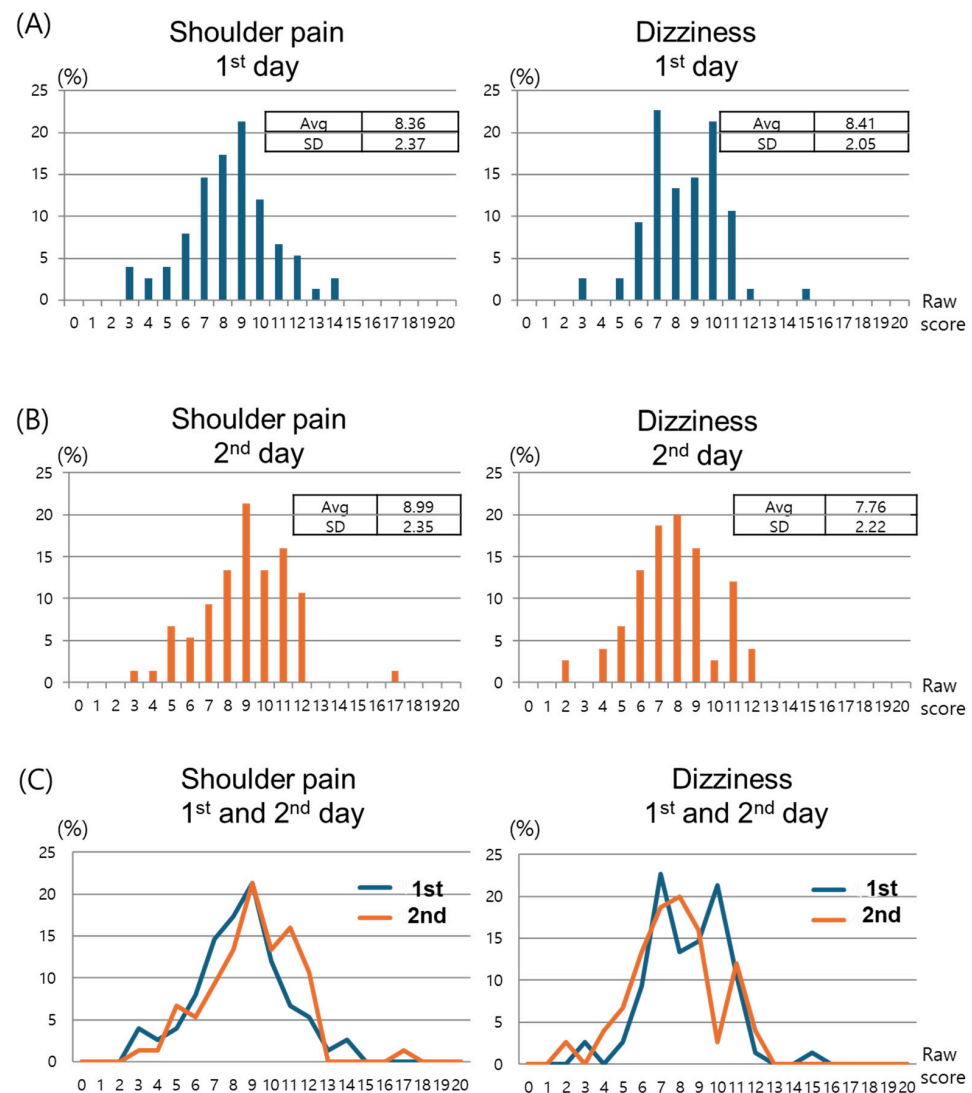


Figure 3. Comparison of Feedback score provided by Chatbot. (A) Distribution of feedback score from Shoulder pain and Dizziness scenario at 1st session. (B) Distribution of feedback score from Shoulder pain and Dizziness scenario at 2nd session. (C) Comparison of score distributions between session 1 and 2 for each scenario.

Overall, chatbot-generated feedback scores showed scenario-dependent patterns rather than consistent improvement across repeated practice sessions. Because the session-specific cases differed in the Korean medicine pattern-identification scenario, required inquiry content, and scoring structure, chatbot-generated score differences were interpreted descriptively rather than as direct evidence of improvement or deterioration in student performance. These findings suggest that chatbot-generated scores may reflect not only students' history-taking performance but also scenario-specific characteristics, case difficulty, and the scoring structure of the chatbot.

3.5. Qualitative Feedback on the Chatbot-Based Training Session

Open-ended responses regarding the chatbot-based history-taking practicum are summarized in Table 8. Students reported several positive aspects of the training, including opportunities for objective self-assessment, detailed and reflective history-taking practice, immediate feedback, and repeated practice without time or place restrictions. Some students also noted that the chatbot feedback helped them recognize their weaknesses and identify areas for improvement.

Table 8. Open-ended responses on the history-taking class.

Theme	Category	Answer
Satisfaction	Objective self-assessment	"I was able to objectively evaluate my own history-taking skills."
	Detailed and reflective practice	"It was helpful for learning because I could think deeply while taking a detailed history."
	Immediate feedback and improvement	"Receiving immediate feedback was helpful because I could identify my weaknesses and areas for improvement, and I could see how I was improving."
	Accuracy and clarity of feedback	"The chatbot feedback was accurate and clear, which was very helpful for learning."
	Accessibility and repeated practice	"It was particularly useful that I could practice repeatedly at any time without restrictions of time or place, and that I could receive scores again based on the feedback."
Dissatisfaction	Overly informative chatbot responses	"The chatbot provided information too readily, even without being asked, which made the history-taking process feel easier than interviewing an actual patient."
	Inaccuracy of chatbot feedback	"The feedback sometimes stated that I had not asked certain questions even though I had. This made the feedback seem slightly inaccurate."
	Limited realism in human interaction	"There are still many aspects that differ from communication with a real human patient, and further improvement is needed."
	Unrealistically precise patient responses	"The virtual patient's answers were too accurate, which made the interaction feel different from an actual patient encounter."
Suggestions for improvement	Inclusion of physical examination	"It would be helpful to add cases involving physical examination."
	Patient education and rapport building	"The assessment of patient education seemed too narrow or overly detailed. It would be better to evaluate patient education and rapport building more broadly."
	Integration of visual diagnostic information	"It would be helpful if additional information, such as images for tongue diagnosis or pulse diagnosis, could also be incorporated."

Students also described several limitations. Some reported that the chatbot provided information too readily or responded too precisely compared with real patients, making the interview feel less realistic. Others noted occasional inaccuracies in chatbot-generated feedback, particularly when questions they had asked were not fully reflected in the feedback. In addition, students pointed out that text-based chatbot interactions differed from communication with real patients.

Suggestions for improvement included expanding the training to include physical examination scenarios, broadening the assessment of patient education and rapport building, and incorporating visual diagnostic information such as tongue diagnosis or pulse-related materials. Overall, the open-ended responses indicated that students valued the chatbot as a flexible and feedback-oriented learning tool, while also recognizing the need for better realism and broader clinical coverage.

4. Discussion

The advancement of artificial intelligence has the potential to significantly transform medical education [6,21]. Incorporating generative AI into the clinical practice curriculum of Korean medicine colleges may enhance students' communication skills [7]. In this prospective single-arm repeated-measures study involving Korean medicine students, we evaluated an AI chatbot-based history-taking practicum embedded within a diagnostics course. Specifically, we implemented a simulation-based diagnostic training method using

a previously developed AI chatbot [12] and evaluated its usability, learner experience, and formative feedback outcomes.

The item-level self-efficacy analysis suggested that chatbot-assisted history-taking practice was associated with selective improvements in specific domains of perceived competence. Significant effects of repeated practicum were observed in items related to identifying the patient's main concerns, preparing an agenda or plan for the conversation, listening attentively, and closing the conversation appropriately (Table 5). These domains are closely related to the structure and procedural organization of a medical interview, which may be well supported by chatbot-based practice because students can repeatedly ask questions, review the flow of the interview, and receive immediate feedback on missing or insufficiently explored items. In contrast, the remaining eight self-efficacy items did not show statistically significant changes. Many of these non-significant items involved more complex communication abilities, such as encouraging patients to elaborate on their concerns, eliciting thoughts and feelings, demonstrating non-verbal behavior (Q7), showing empathy (Q8), checking patient understanding (Q10), and making shared decisions (Q11). These competencies require not only appropriate questioning but also sensitivity to emotional cues, conversational timing, ambiguity, rapport building, and mutual negotiation with the patient [22]. Because the present chatbot interaction was primarily text-based and scenario-bound, it may have provided fewer opportunities to practice the affective, relational, and interactive aspects of clinical communication that typically emerge in encounters with real or standardized patients [23]. Therefore, SP or peer role-play training remains important for non-verbal communication, physical examination, real-time interaction, and exam-like performance demands.

The analysis of the overall mean self-efficacy score further supports a cautious interpretation of the educational effect (Table 6). Although the mean self-efficacy score gradually increased from the first pre-training assessment to the first post-training assessment and then to the second post-training assessment, the overall time effect did not reach statistical significance ($p = 0.076$). The comparison between the first pre-training and second post-training assessments showed a significant difference before adjustment ($p = 0.025$), but this result was no longer significant after Holm correction ($p = 0.075$). Therefore, the findings should not be interpreted as evidence that the chatbot-based practicum significantly improved overall history-taking self-efficacy. Rather, the results suggest a modest upward trend in perceived competence across repeated sessions, with improvement signals concentrated in selected components of the interview process. This pattern may indicate that repeated chatbot-assisted practice is more effective for supporting the cognitive and structural organization of history-taking than for developing nuanced interpersonal communication skills.

The analysis of chatbot-generated feedback scores provides an additional perspective that was not examined in previous works (Figure 3, Table 7). In the shoulder pain scenario, feedback scores increased from session 1 to session 2, showing a significant change before adjustment but not after Holm correction. In contrast, the dizziness scenario showed a non-significant decreasing trend. From an educational perspective, the absence of consistent score improvement across both scenarios does not necessarily undermine the value of chatbot-assisted practice. The changes in these scores may reflect not only students' history-taking performance but also differences in case difficulty, scenario structure, expected key history items, and the scoring criteria embedded in the chatbot system. If the second dizziness case required more specific or discriminating questions than the first case, a lower score may reflect increased case complexity or stricter scenario-specific expectations rather than deterioration in students' learning. For future implementation, automated scoring systems should be further calibrated across scenarios, reviewed by clinical educators, and

validated against expert-rated history-taking performance. Such refinement would allow chatbot-generated feedback to serve as a more reliable formative assessment tool within Korean medicine diagnostics education.

The usability results indicate that the AI chatbot was technically accessible and educationally acceptable for Korean medicine students (Table 4). Students reported high levels of active participation, perceived usefulness, ease of use, and convenience, suggesting that the chatbot-based practicum was feasible within a regular diagnostics course. This favorable response may have been facilitated by the participants' high baseline familiarity with AI chatbots and frequent prior use of conversational AI (Tables 1 and 2). In this context, students' AI literacy may have lowered the technical barrier to participation and supported the implementation of a new chatbot-assisted learning format. The generally high baseline familiarity with conversational AI limited the ability to examine differential effects by prior AI experience and may restrict generalizability to AI-naïve learners. However, the relatively lower rating for similarity to real patients (Q12, Table 4) indicates that usability and accessibility alone are insufficient to ensure authentic clinical simulation. The naturalness of the AI chatbot's responses, which had been regarded as a weakness by the developer [12], was also recognized by students as an ongoing concern. The chatbot was originally developed and tested using ChatGPT-4, and several of the limitations observed in this study are likely to be addressed by adopting more advanced versions of GPT, which offer improved contextual understanding and conversational naturalness. The quality of chatbot-assisted learning, however, remains vulnerable to substantial variations depending on service environments or policy changes in generative AI models [24].

The open-ended responses further clarified these limitations (Table 8). Students valued the opportunity for repeated practice and immediate feedback, but they also noted that the chatbot sometimes provided information too readily, responded more precisely than real patients, or failed to fully reflect questions that had already been asked in its feedback. In addition, some students suggested that future systems should include physical examination scenarios and provide visual diagnostic information, tongue diagnosis, pulse diagnosis information, or other examination findings. The current platform could not deliver certain essential visual or sensory information related to the symptom/disease, which takes a great part in diagnostics in Korean Medicine [25].

To overcome these limitations, future chatbot-based educational systems should move beyond text-only interaction toward more realistic and multimodal simulation. Recent advances, such as a doctor–patient interaction simulator supporting both text-based and voice-based modes [26], suggest promising directions. Furthermore, the integration of immersive technologies including virtual reality (VR) and augmented reality (AR)—already applied in diverse educational contexts—could help address these shortcomings [27,28]. Incorporating such features may substantially enhance the quality of diagnostic education and mitigate the limitations of current chatbot-based models.

In addition, this single-institution study relied primarily on voluntary, self-reported survey responses; therefore, response-related selection bias, novelty-related engagement with the chatbot, and limited generalizability to other educational settings cannot be excluded.

At present, the chatbot program remains in the development stage and lacks a diverse range of patient cases. Further refinement is required to better align the content with the competency level of fourth-year students. In addition, external evaluation of chatbot performance has not yet been conducted, highlighting the need for assessment by educational experts. From an educational perspective, future studies should include comparative analyses between AI-assisted and traditional approaches, as well as the development and evaluation of diverse AI-integrated, project-based learning models to broaden the applicability and effectiveness of such innovations.

5. Conclusions

This exploratory study evaluated a two-session AI chatbot-based history-taking practicum embedded within a regular Korean medicine diagnostics course. The study addressed its primary aim by demonstrating that repeated chatbot-assisted practice with automated scoring and individualized feedback was feasible to implement in a real course setting and was generally well accepted by students. Students reported favorable usability and perceived educational value, while overall self-efficacy scores showed a numerical but statistically non-significant increase across the assessment points. Chatbot-generated feedback scores were interpreted as descriptive, case-specific formative outputs rather than validated measures of clinical competence. Although the adapted KSE-12 showed high internal consistency in the present sample, the modified instrument was not fully validated for chatbot-based Korean medicine diagnostics training; therefore, the self-efficacy findings should be interpreted as exploratory learner-reported perceptions.

These findings suggest that AI chatbot-assisted training may serve as a supplementary tool for providing accessible, repeatable history-taking practice and formative feedback in Korean medicine education. However, the single-arm design, reliance on learner-reported outcomes, absence of expert-rated performance assessment, and limited validation of automated scoring restrict conclusions regarding educational effectiveness. Future studies should use controlled or randomized designs, psychometrically matched cases, expert-rated clinical performance outcomes, validated automated scoring systems, and more realistic multimodal clinical scenarios to determine whether chatbot-assisted training improves objective clinical skills.

Author Contributions: Conceptualization, C.-E.K. and D.-W.L.; methodology, J.K. and C.-E.K.; software, J.K., C.-E.K. and J.-H.K.; validation, J.-H.K. and H.-Y.L.; formal analysis, H.-Y.L.; investigation, D.-W.L., I.-Y.C. and W.-H.P.; data curation, I.-Y.C. and J.-H.K.; writing—original draft preparation, D.-W.L. and I.-Y.C.; writing—review and editing, C.-E.K. and J.K.; supervision, W.-H.P. and C.-E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the 2022 Gachon University Research Fund (GCU-202206720001).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of Dongguk University (DGU IRB 20260022; initial approval).

Informed Consent Statement: Written informed consent was obtained from all participants. Participation in the research survey was voluntary, and only data from students who agreed to participate were included in the analysis.

Data Availability Statement: Dataset available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KSE-12	Korean Self-Efficacy Scale-12
SP	Standardized Patient
AI	Artificial Intelligence
LLM	Large Language Model
CPX	Clinical Performance Examination
VR	Virtual Reality
AR	Augmented Reality

References

1. Peart, P. Clinical history taking. *Clin. Integr. Care* **2022**, *10*, 100088. [\[CrossRef\]](#)
2. Cho, E.; Jung, H.-J.; Leem, J. Peer role-play in a College of Korean Medicine to improve senior students' competencies in patient care and communication: A case analysis and proposal for a model. *J. Korean Med.* **2022**, *43*, 49–64. [\[CrossRef\]](#)
3. Lee, Y.H.; Lee, Y.-M.; Kim, B.S. Content analysis of standardized-patients' descriptive feedback on student performance on the CPX. *Korean J. Med. Educ.* **2010**, *22*, 291–301. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Abd-Alrazaq, A.; AlSaad, R.; Alhuwail, D.; Ahmed, A.; Healy, P.M.; Latifi, S.; Aziz, S.; Damseh, R.; Alrazak, S.A.; Sheikh, J. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Med. Educ.* **2023**, *9*, e48291. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Öncü, S.; Torun, F.; Ülkü, H.H. AI-powered standardised patients: Evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Med. Educ.* **2025**, *25*, 278. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Han, Y. Usability and educational effectiveness of ai-based patient chatbot for clinical skills training in korean medicine. *Korean J. Acupunct.* **2024**, *41*, 27–32. [\[CrossRef\]](#)
7. Kwon, C.-Y. Utilization of generative artificial intelligence chatbot for training in suicide risk assessment of depressed patients: Focusing on students at a college of korean medicine. *J. Orient. Neuropsychiatry* **2024**, *35*, 153–162.
8. Zeng, J.; Qi, W.; Shen, S.; Liu, X.; Li, S.; Wang, B.; Dong, C.; Zhu, X.; Shi, Y.; Lou, X. Embracing the Future of Medical Education With Large Language Model-Based Virtual Patients: Scoping Review. *J. Med. Internet Res.* **2025**, *27*, e79091. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Thind, B.S.; Javidi, D.; Schwartz, L.M. Artificial intelligence in undergraduate medical education clinical skills curricula: A scoping review of implementations since 2022. *Front. Digit. Health* **2026**, *8*, 1830254. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Li, D.; Lebai Lutfi, S. Large Language Model-Based Virtual Patient Systems for History-Taking in Medical Education: Comprehensive Systematic Review. *JMIR Med. Inform.* **2026**, *14*, e79039. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Brügge, E.; Ricchizzi, S.; Arenbeck, M.; Keller, M.N.; Schur, L.; Stummer, W.; Holling, M.; Lu, M.H.; Darici, D. 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*, 1391. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Kim, J.; Lee, H.-Y.; Kim, J.-H.; Kim, C.-E. Pilot development of a 'clinical performance examination (CPX) practicing chatbot' utilizing prompt engineering. *J. Korean Med.* **2024**, *45*, 200–212. [\[CrossRef\]](#)
13. Preiksaitis, C.; Rose, C. Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review. *JMIR Med. Educ.* **2023**, *9*, e48785. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Lee, H.-Y.; Kim, J.; Choi, H.; Bae, H.; Jeong, A.; Choi, S.; Kim, J.-H.; Kim, C.-E. Comparing AI chatbot simulation and peer role-play for OSCE preparation: A pilot randomized controlled trial. *BMC Med. Educ.* **2025**, *25*, 1755. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Heslin, P.A.; Klehe, U.-C.; Keating, L.A. *The SAGE Encyclopedia of Industrial and Organizational Psychology*, 2nd ed.; SAGE Publications, Inc.: Thousand Oakes, CA, USA, 2017.
16. Kang, H.-S.; Kim, Y.-Y.; Lee, H.-S. Predictive factors influencing clinical competence in nursing students. *J. Korea Acad.-Ind. Coop. Soc.* **2018**, *19*, 389–398.
17. Demirören, M.; Turan, S.; Öztuna, D. Medical students' self-efficacy in problem-based learning and its relationship with self-regulated learning. *Med. Educ. Online* **2016**, *21*, 30049. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Gil, C.-R.; Sung, K.-M. Validity and reliability of the Korean version of self-efficacy questionnaire (KSE-12). *J. Digit. Converg.* **2020**, *18*, 337–345.
19. Song, M.-K.; Hong, S.-U. A survey of students' satisfaction on participation learning using role-play in clerkship. *J. Korean Med. Ophthalmol. Otolaryngol. Dermatol.* **2012**, *25*, 65–77. [\[CrossRef\]](#)
20. Min, Y.; Kwak, S.-J. A Comparative Study on the Usability by the Platform of Artificial Intelligence Chatbot Service in Library. *J. Korean BIBLIA Soc. Libr. Inf. Sci.* **2023**, *34*, 183–203.
21. Sun, L.; Yin, C.; Xu, Q.; Zhao, W. Artificial intelligence for healthcare and medical education: A systematic review. *Am. J. Transl. Res.* **2023**, *15*, 4820. [\[PubMed\]](#)
22. Roter, D.L.; Frankel, R.M.; Hall, J.A.; Sluyter, D. The expression of emotion through nonverbal behavior in medical visits: Mechanisms and outcomes. *J. Gen. Intern. Med.* **2006**, *21*, 28–34. [\[CrossRef\]](#)
23. Laymouna, M.; Ma, Y.; Lessard, D.; Schuster, T.; Engler, K.; Lebouché, B. Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *J. Med. Internet Res.* **2024**, *26*, e56930. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Chen, L.; Zaharia, M.; Zou, J. How is ChatGPT's behavior changing over time? *Harv. Data Sci. Rev.* **2024**, *6*. [\[CrossRef\]](#)
25. Ko, M.; Lee, J.; Yun, K.; You, S.; Lee, M. Perception of pattern identification in traditional medicine: A survey of Korean medical practitioners. *J. Tradit. Chin. Med.* **2014**, *34*, 369–372. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Hicke, Y.; Geathers, J.; Rajashekar, N.; Chan, C.; Jack, A.G.; Sewell, J.; Preston, M.; Cornes, S.; Shung, D.; Kizilcec, R. MedSimAI: Simulation and formative feedback generation to enhance deliberate practice in medical education. *arXiv* **2025**, arXiv:2503.05793.

27. Chuang, C.-H.; Lo, J.-H.; Wu, Y.-K. Integrating chatbot and augmented reality technology into biology learning during COVID-19. *Electronics* **2023**, *12*, 222. [\[CrossRef\]](#)
28. Dweik, H.; Nawahdah, M. Bridging reality and AI: The role of augmented reality chatbots in interactive learning. In Proceedings of the INTED2025, 19th International Technology, Education and Development Conference, Valencia, Spain, 3–5 March 2025; pp. 4136–4141.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.