

Original Article

Development of an LLM-based CPX Practicing Chatbot for Korean Medicine Education: Implementation of Automated Scoring and Feedback Generation Framework

Jundong Kim¹, Hye-Yoon Lee², Ji-Hwan Kim^{3*}, Chang-Eop Kim^{1*}

¹Department of Physiology, College of Korean Medicine, Gachon University

²Division of Humanities and Social Medicine, School of Korean Medicine, Pusan National University

³Department of Sasang Constitutional Medicine, School of Korean Medicine, Pusan National University

Objectives: This study aimed to develop an AI-based CPX (Clinical Performance Examination) practicing chatbot for Korean medicine education, implementing automated quantitative scoring and qualitative feedback systems to enhance individualized learning experiences.

Methods: Building upon a previously developed CPX practicing chatbot, we integrated a quantitative scoring system and a qualitative text feedback system using Large Language Models (LLMs). Scoring prompts and feedback prompts were designed based on standardized CPX scenarios. We compared the performance of OpenAI's GPT-4 and Anthropic's Claude-3.5-Sonnet models in terms of scoring accuracy, feedback quality, consistency, latency, and fluency. Three sample chat histories representing varying levels of student performance were created for evaluation.

Results: Claude-3.5-Sonnet demonstrated perfect accuracy in quantitative scoring across all samples and faster execution times (average 43.2 seconds) compared to GPT-4. In qualitative feedback generation, GPT-4 provided more specific and actionable feedback, while Claude-3.5-Sonnet produced more natural and supportive language. By combining the strengths of both models, we implemented an optimized system that first uses GPT-4 for detailed feedback generation and then refines the language with Claude-3.5-Sonnet, resulting in accurate scoring and high-quality feedback.

Conclusions: The study successfully developed an LLM-based CPX practicing chatbot that offers automated scoring and individualized feedback for Korean medicine education. The optimized system enhances the learning experience by providing accurate assessments and specific, supportive feedback, addressing limitations in current CPX education practices and resource constraints.

Key Words : Artificial Intelligence; Large Language Model; Clinical Performance Examination; Medical Education; Korean Medicine.

• Received : 31 October 2024 • Accepted : 21 November 2024

• Correspondence to : Chang-Eop Kim

Department of Physiology, College of Korean Medicine, Gachon University

Tel : +82-31-750-5493, E-mail : eopchang@gachon.ac.kr

• Correspondence to : Ji-Hwan Kim

Department of Sasang Constitutional Medicine, School of Korean Medicine, Pusan National University

Tel : +82-55-360-5969, E-mail : jani77@pusan.ac.kr

Introduction

The importance and effectiveness of individualized feedback on learner performance are frequently emphasized in medical education¹⁻³). Individualized feedback allows students to objectively assess their progress, identify and correct mistakes, and strengthen their skills⁴). Individualized feedback approaches can provide significant educational benefits in clinical Korean medicine education, where therapeutic decisions must consider the diversity and individuality of both patients and their conditions. Therefore, there is an emerging need for educational models that can incorporate appropriate feedback processes into the learning journey of students in clinical Korean medicine education⁵).

Clinical Performance Examination (CPX), which is being actively implemented in Korean medicine education, is an ideal educational model that provides practical individualized feedback to students in simulated clinical situations while achieving significant educational outcomes⁶). Since CPX evaluates students' ability to conduct clinical consultations with patients, it serves as a comprehensive assessment tool, examining not only adherence to consistent protocols but also the flexible application of Korean medicine knowledge and effective patient communication skills⁶). Thus, CPX is an ideal model for providing students with practical, individualized feedback in a simulated clinical setting, which can lead to significant educational gains⁷). However, current Korean medicine colleges face significant challenges in providing detailed and in-depth educational feedback to students. These challenges stem from

multiple factors: the time constraints of evaluating various competencies simultaneously, the shortage of qualified professionals and faculties who can objectively assess multiple student practitioners, and insufficient time for individual student evaluation⁸).

The recent emergence of Large Language Models (LLMs), such as ChatGPT and Claude, has opened up possibilities for sophisticated feedback generation in medical education^{9,10}). LLMs are artificial intelligence (AI) systems trained on large volumes of text data that can understand and generate human language, demonstrating particularly strong performance in understanding complex contexts and generating detailed responses⁹). The use of LLMs in CPX education offers several advantages that could overcome existing challenges: students can practice CPX scenarios and receive immediate feedback at any time; comprehensive feedback can be efficiently provided on diverse aspects, such as Korean medicine knowledge, diagnostic reasoning, and communication skills; and individualized feedback can be delivered simultaneously to multiple students, making efficient use of educational resources. These benefits are particularly promising for prospective Korean medicine doctors who need to develop complex and specialized skills¹¹).

Although several studies have recognized and proposed these advantages of LLMs in Korean medicine education, their implementations remain limited in scope. Some LLM-based CPX chatbots have been developed for scenarios like abdominal pain, gastric ulcer, and hypertension, allowing students to practice medical history taking and

physical examinations with AI model^{6,12}). However, these models lack the capability of automated scoring and individualized feedback generation by LLMs for learners. Additionally, while a paper reported on students using LLMs to practice Suicide Risk Assessment for depression patients, it also did not incorporate automated scoring or feedback capabilities for evaluating students' performance¹³).

Therefore, this study aims to develop an LLM-based CPX chatbot that offers both practice opportunities and automated, individualized feedback for students. The feedback system integrates two components: a quantitative scoring system for evaluating performance and a qualitative feedback system that identifies areas for improvement with specific suggestions. Furthermore, this study presents a systematic framework for developing CPX chatbots that can adapt to advancing AI technologies and diverse clinical scenarios. The outcomes of this research are expected to enhance the learning experience in CPX education for Korean medicine.

Method

Our research team previously developed a "CPX Practicing Chatbot" for CPX training of Korean medicine students⁶). In the previous study, a scenario of essential hypertension was obtained from the National Institute for Korean Medicine Development (NIKOM) and modified, and a chatbot acting as a standardized patient was implemented using OpenAI's GPT-4 API. Specifically, role prompting and few-shot prompting techniques were applied to ensure that the chatbot could

consistently play the role of a patient. The model's performance was optimized through efficient data structuring in a chart format and periodic re-prompting. The detailed development process can be found in the paper.

However, the previous CPX Practicing Chatbot had the limitation that it only provided conversation functionality with learners, lacking the ability to evaluate students' performance and provide educational individualized feedback. Therefore, in this study, we aimed to further develop the existing CPX chatbot model by adding a quantitative scoring system and a qualitative feedback generation system, in order to create a more complete educational tool.

1. Development of a Quantitative Scoring System

1) Design of the Scoring Prompt

After the student practitioner completes the CPX session with the chatbot, they can click the "Get Scoring and Feedback" button. This triggers the LLM model to act as an evaluator, using the scoring prompt and the conversation history between the student and the chatbot. The scoring prompt serves as an answer key or rubric, guiding the LLM in assessing the student's performance. It can be thought of as a "whisper from the developer" to the LLM.

The scoring prompt was developed based on "Section 6: Scoring Rubric" from the scenario provided by NIKOM. It includes 15 questions on medical history taking and 4 questions on patient education. For the 15 medical history questions, 1 point is awarded if the standardized patient clearly provides the relevant information; otherwise, 0

points are given. Similarly, for the 4 questions on patient education, 1 point is awarded if the student practitioner clearly delivers the necessary information; otherwise, 0 points are given. The scoring rubric was revised before being provided to the LLM as a prompt to eliminate any ambiguity and ensure that the instructions were as clear as possible for the LLM to understand. The examples of this revision are shown in Table 1.

2) Implementation of the Scoring Process

Since the scoring process involves quantitative evaluation, it is crucial to ensure high accuracy and consistency. To ensure this accuracy and consistency, our research team used a Python ‘for loop’ to make individual LLM API calls for each of the 19 items. For each item, the LLM API was called, provided with the conversation history and scoring prompt, and instructed to generate a score of 0 or 1 along with the rationale, which was then saved. This process was performed individually for items 1 through 19, requiring a total of 19 LLM API calls to ensure that each item was

evaluated accurately and consistently. Finally, an additional LLM API call was made to aggregate the saved scores, calculate the total, and output the result in the specified JSON format. The format for the output after scoring each item was as follows:

```
{
  "score": 1,
  "reason": "Because the patient explicitly
  mentioned resting for 'at least 5 minutes'."
}
```

The format for the final output after aggregating all the scores was as follows:

```
==== Scoring Results ====
Evaluation Results:
Medical History: 10/15 points
Patient Education: 4/4 points
Total Score: 14/19 points
Time Taken for Scoring: 50.61 seconds
"""
```

The scoring process employs the Chain-of-Thought (CoT) prompting technique, which

Table 1. Examples of Original Scenario Scoring Rubric and Corresponding Scoring Prompts Provided to the LLM.

	Original Scenario Scoring Rubric	Scoring Prompt
Medical History Question Example	The patient stated that they had rested sufficiently for at least 5 minutes before measuring their blood pressure. 1 point: Stated 0 points: Not stated	{ "id": "1.1", "prompt": "Award 1 point if the patient explicitly mentions resting for 'at least 5 minutes' before measuring their blood pressure." }
Patient Education Question Example	The patient was informed that they are likely to have essential hypertension. 1 point: Informed 0 points: Not informed	{ "id": "2.1", "prompt": "Award 1 point if the student practitioner mentions the diagnosis or possibility of essential hypertension to the patient. This includes the diagnosis of 'primary hypertension' or 'essential hypertension'." }

ensures that the LLM not only generates an answer but also provides the reasoning behind it. CoT prompting helps the LLM articulate its thought process, making the evaluation more transparent and understandable¹⁴). In this study, each evaluation item was scored along with the reasoning, which was explicitly generated to improve transparency and reliability. This approach aimed to enhance the accuracy of the quantitative evaluation by allowing evaluators to understand the rationale behind each score, thereby increasing the overall credibility and consistency of the assessment.

3) Comparison of Scoring Performance Across Models

The accuracy and reproducibility of the scoring system implemented through the aforementioned procedures were evaluated. For this purpose, three chat history samples were created. The first sample represented an exemplary student, with a score of 10 out of 15 on medical history taking and 4 out of 4 on patient education, resulting in a total score of 14. The second sample represented a student who made an error in pattern identification of Korean medicine, scoring 10 on medical history taking and 3 on patient education, for a total of 13. The third sample represented a student with insufficient skills, scoring 3 on medical history taking and 2 on patient education.

The process of creating these chat history samples was as follows: The web version of the ‘Claude-3.5-Sonnet model’ by Anthropic was provided with a scoring prompt and instructed to generate a conversation history that would achieve a score of 14. The generated content was then

reviewed by the researcher to ensure clarity and eliminate ambiguity in the scoring. Subsequently, a further revision was carried out using web version of the ‘o1-preview model’ by OpenAI, during which a mock scoring was conducted to identify and refine any ambiguous or unclear statements. Each of these three chat history samples was used for repeated evaluations, with each evaluation conducted 10 times. An example of the first sample, which received a score of 14, can be found in Table 2 below.

The scoring performance of OpenAI's latest model, GPT-4o (‘gpt-4o-2024-08-06’), and Anthropic's latest model, Claude-3.5-Sonnet (‘claude-3.5-sonnet-20240620’), was compared using these three chat history samples. Each model was instructed to independently perform repeated scoring 10 times for each of the three samples, and the accuracy of the quantitative scoring results was compared. Based on this comparison, we aimed to determine which of the two models is more suitable for quantitative scoring and apply it accordingly.

2. Development of the Qualitative Text Feedback System

1) Design of the Text Feedback Prompt

After the quantitative scoring is complete, the LLM is called again to begin generating text feedback. At this stage, the conversation history and the scoring results are input into the LLM, along with the ‘textfeedback_prompt’.

The textfeedback_prompt was developed based on not only “Section 6: Scoring Rubric” from the original scenario but also the content from “Section 2: Scoring Criteria.” While “Section 6:

Scoring Rubric” outlines the criteria for scoring the student's responses, it does not provide the rationale for eliciting such answers. In contrast, “Section 2: Scoring Criteria” explains why each item in the scoring rubric needs to be asked. For example, “Section 6: Scoring Rubric” might state, “The patient mentioned that they do not experience significant fatigue, do not feel weakness or tingling in their limbs, and do not drink a lot of water. If the patient provided one or more of these responses, award 1 point; otherwise, 0 points.” Meanwhile, “Section 2: Scoring Criteria” indicates that, considering differential diagnosis for primary aldosteronism,

questions about fatigue, thirst, muscle weakness, and urine volume should be asked. Therefore, the textfeedback_prompt was created based on “Section 2: Scoring Criteria” to maintain consistency between the quantitative scoring and the text feedback, and it also incorporated content from “Section 6: Scoring Rubric,” which was refined further to ensure more accurate and specific feedback.

Additionally, to ensure consistent and stable text feedback generation, a fixed format was provided. According to existing feedback learning methodologies¹⁵⁾, the format was designed to allow the students to first identify what they did

Table 2. An Example of the First Chat History Sample, Who Received a Score of 14.

```
test_conversation_good = [  
    ("Student Practitioner", "Hello, I'm student practitioner Jun-dong Kim. Could you please tell me your name and date of birth?"),  
    ("Min-young Park", "Yes, hello. My name is Min-young Park, and my date of birth is June 3, 1976."),  
    ("Student Practitioner", "What symptoms are you experiencing today?"),  
    ("Min-young Park", "My blood pressure has been high frequently, so I came because I'm worried."),  
    ("Student Practitioner", "You mentioned your blood pressure is high. Since when has it been like this? And have you measured it at home?"),  
    ("Min-young Park", "Yes, it's been between 130 and 140 for about two years. It often reads high when I measure it at home too."),  
    ("Student Practitioner", "Before measuring your blood pressure, did you rest for at least 5 minutes? And have you avoided coffee or cigarettes?"),  
    ("Min-young Park", "Yes, I usually rest for about 10 minutes before measuring it. I don't smoke or drink coffee before measuring my blood pressure."),  
  
    ##### (Scenario details omitted to prevent unintended disclosure of content) #####  
  
    ("Student Practitioner", "To manage hypertension, I recommend a low-salt diet, regular exercise, reducing alcohol and coffee intake, weight management, and self-monitoring of blood pressure."),  
    ("Min-young Park", "Yes, I will try to improve my lifestyle."),  
    ("Student Practitioner", "From a Korean medicine perspective, ○○○○ is suspected. Herbal treatment may also help."),  
    ("Min-young Park", "Okay, thank you."),  
    ("Student Practitioner", "Do you have any other questions?"),  
    ("Min-young Park", "No, I don't. Thank you for the detailed explanation."),  
    ("Student Practitioner", "Alright, take care of your health, and feel free to reach out anytime."),  
]
```

well, then recognize areas needing improvement, and finally receive suggestions for future learning directions. The format is as follows:

“””

Successful consultation.

1. [List of specific examples]
2. [List of specific examples]
3. [List of specific examples]

However, the following points need improvement.

1. [Provide specific examples and suggestions for improvement]
2. [Provide specific examples and suggestions for improvement]
3. [Provide specific examples and suggestions for improvement]

Additional Advice:

[Overall feedback on performance]

[Suggestions for future learning directions]

“””

2) Comparison of Qualitative Text Feedback Generation Performance Across Models

Although the text feedback follows a fixed format, its content can vary significantly depending on the LLM's performance. Therefore, the text feedback generation performance of OpenAI's latest model, GPT-4o ('gpt-4o-2024-08-06'), and Anthropic's latest model, Claude-3.5-Sonnet ('claude-3.5-sonnet-20240620'), were qualitatively compared.

Experiments were conducted by setting the temperature to two different values—0 and 1—for each model, resulting in four configurations: GPT-4o at temperature 0, GPT-4o at temperature

1, Claude-3.5-Sonnet at temperature 0, and Claude-3.5-Sonnet at temperature 1. Temperature is a parameter that adjusts the diversity of the LLM's output; when set closer to 0, the model outputs the most certain response, resulting in higher consistency, whereas setting it closer to 1 leads to more creative answers by considering a broader range of possibilities¹⁶⁾. In this study, by setting the temperature to extreme values of 0 or 1, we aimed to determine which configuration—consistent feedback or varied, creative feedback—was more suitable for educational purposes.

The same chat history sample used in Table 2 was utilized for the experiment, representing a good performance with a score of 14 out of 19. Each LLM was instructed to generate feedback five times based on the chat history and the score for this sample. The researchers qualitatively evaluated the generated feedback based on the following five aspects:

- ① Accuracy: Is the feedback accurately generated based on the chat history?
- ② Specificity: Does the feedback offer concrete and actionable suggestions for students?
- ③ Consistency: Does the feedback maintain consistency with the scoring results and the previously generated feedback?
- ④ Latency: Does the feedback generation avoid excessive latency in practical use?
- ⑤ Fluency: Is the feedback written in a natural, supportive tone, especially in Korean?

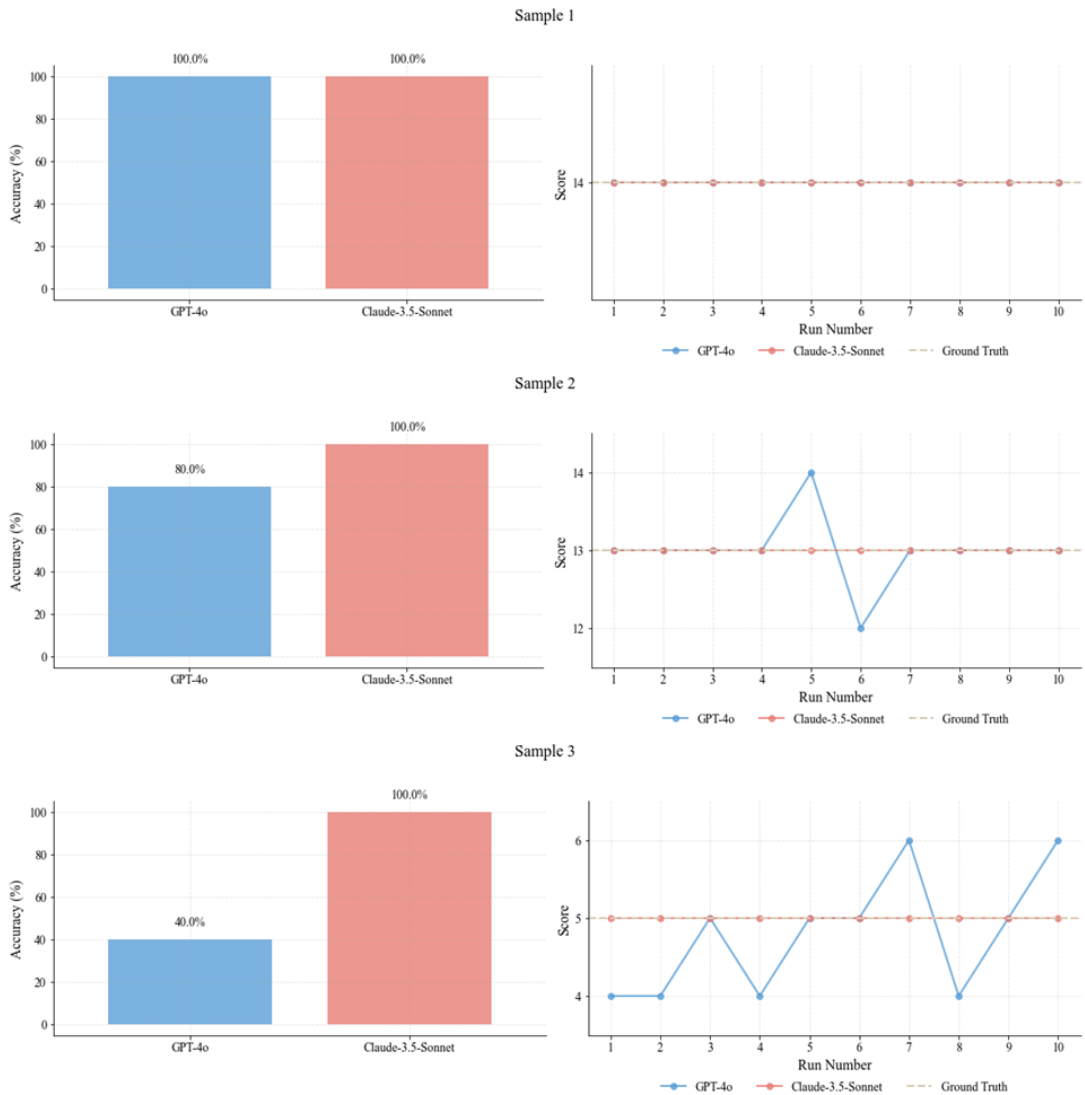


Fig. 1. Comparison of scoring accuracy and consistency between GPT-4o and Claude-3.5-Sonnet model. Left: Scoring accuracy (%) for each sample. Right: Score variations across 10 repeated trials.

Results

1. Evaluation of the Quantitative Scoring System

1) Comparison of Scoring Accuracy Across Models

Scoring accuracy was compared by having each model independently score three sample cases (with scores of 14, 13, and 5) ten times each (Figure 1). The Claude-3.5-Sonnet model demonstrated 100% accuracy across all three samples. In contrast, GPT-4o showed consistent

scoring accuracy only for the first sample (100%), but its accuracy dropped to 80% for the second sample and 40% for the third sample, indicating variability in scoring consistency.

2) Comparison of Latency (Scoring Time)

The average scoring time was faster for Claude-3.5-Sonnet (43.2 seconds) compared to GPT-4o (50.6 seconds). Both models demonstrated response speeds suitable for practical use in educational settings.

2. Evaluation of the Qualitative Text Feedback System

1) Characteristics of Feedback Based on Temperature Settings

With a temperature setting of 0, both models consistently generated similar feedback across five trials. When the temperature was set to 1, more varied feedback was produced, with Claude-3.5-Sonnet often providing more detailed insights into Korean medicine content. However, in a situation where students repeatedly practice the CPX examination and receive feedback after each session, overly varied feedback can make it difficult for them to focus on specific areas for improvement. Consistent feedback allows students to track their progress effectively and clearly understand what aspects need continuous improvement. Therefore, providing feedback with too much variation may reduce the educational effectiveness by causing confusion about which specific skills to work on.

To address this, the temperature was fixed at 0 for both models when comparing feedback quality, ensuring consistency.

2) Comparison of Feedback Quality Across Models

In terms of accuracy, both models performed well, providing feedback that accurately reflected the details of the conversation history without hallucination. However, GPT-4o tended to provide more specific and actionable feedback for improving students' CPX skills. For example, GPT-4o generated feedback like, "You asked about the patient's urine condition, but additional questions regarding frequency or volume of urine are needed beyond checking for foam. This is important for distinguishing kidney disease or diabetes." In contrast, Claude-3.5-Sonnet produced more general feedback, such as, "You asked basic questions about the urine condition but did not inquire about more specific changes." Nonetheless, the Claude-3.5-Sonnet model consistently produced stable feedback, with shorter response times. Moreover, it provided feedback in fluent Korean, delivered in a natural and supportive tone, which is crucial when offering feedback to students. Examples of feedback generated by each model can be found in Table 3 below.

3. Optimization of Scoring and Feedback Generation Based on Quantitative and Qualitative Evaluation Results

Based on the results of both quantitative scoring and qualitative feedback evaluation, our research team implemented the following optimized system. For the quantitative scoring system, the Claude-3.5-Sonnet model was adopted. This decision was made because Claude-3.5-Sonnet demonstrated 100% accuracy across all difficulty levels of the sample and showed faster execution times.

Table 3. Comparison of Qualitative Feedback Generated by GPT-4o and Claude-3.5-Sonnet Under Temperature 0 Setting.

	GPT-4o (temperature : 0)	Claude-3.5-Sonnet (temperature : 0)
Feedback Example in English	<p>Successful Consultation:</p> <ol style="list-style-type: none"> You verified the correct procedure for measuring the patient's blood pressure and inquired thoroughly about factors like coffee or smoking that could affect blood pressure. You appropriately asked about the family history to identify the patient's risk factors for hypertension. You provided a detailed explanation of essential hypertension and the diagnosis (○○○○) and offered specific guidance on further examinations and lifestyle changes. <p>However, the following points need improvement:</p> <ol style="list-style-type: none"> You asked about weight changes, but further questions regarding the reasons for weight gain or related symptoms (e.g., edema, fluid retention) were lacking. Weight change could be a clue to other conditions related to hypertension, so more detailed questioning is required. You inquired about the condition of the patient's urine, but there was a need for additional questions regarding the frequency or volume of urine, beyond checking for foamy urine. This information is important for distinguishing conditions like kidney disease or diabetes. You asked about symptoms such as fatigue, muscle weakness, and thirst, but it would be beneficial to check for other diseases as well (e.g., adrenal or thyroid disease). More comprehensive symptom questioning is needed. <p>Additional Advice:</p> <p>You did a good job of understanding the patient's condition and providing appropriate education. However, a broader and more specific inquiry is required regarding symptoms related to the various causes of hypertension. Specifically, by asking more detailed questions about weight changes, urine condition, and fatigue, you can make more accurate differential diagnoses. For future practice, it would be beneficial to study the question items in more depth related to secondary causes of hypertension and apply these to actual consultations.</p>	<p>Successful Consultation:</p> <ol style="list-style-type: none"> You thoroughly inquired about the patient's blood pressure measurement environment and lifestyle habits. You carefully confirmed the family history and current symptoms. You provided appropriate guidance on necessary examinations and lifestyle changes as part of patient education. <p>However, the following points need improvement:</p> <ol style="list-style-type: none"> The questioning about the medications the patient is currently taking was omitted. This is important for identifying the causes of hypertension and establishing a treatment plan. The questioning about conditions that could be secondary causes of hypertension was insufficient. It is important to confirm the medical history of these conditions. The questioning to identify signs of target organ damage was not sufficient. <p>Additional Advice:</p> <p>You have performed well in basic medical history taking and patient education. However, a more comprehensive approach considering the various causes and complications of hypertension is needed. In the future, deepen your understanding of the causes of secondary hypertension and target organ damage, and use this knowledge to develop a more systematic questioning technique. Also, it would be beneficial to further enhance your ability to integrate both Korean medical diagnoses and Western medical approaches. I encourage your continued growth!</p>
Accuracy	High	High
Specificity	High	Relatively low
Consistency	Moderate	High
Latency	Average of 5 trials: 11.24 seconds	Average of 5 trials: 8.67 seconds
Fluency	Moderate	High, more supportive tone

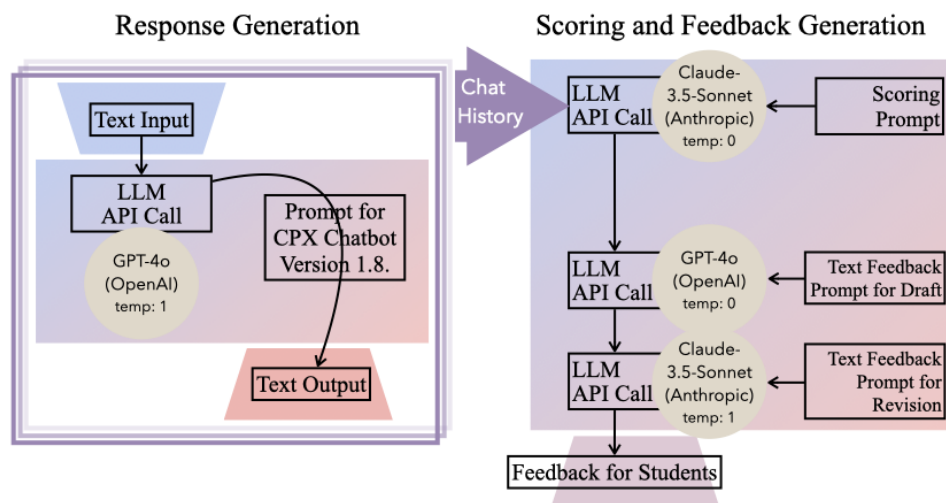


Fig. 2. Workflow of the chatbot's response generation and scoring and feedback system, illustrating the use of different LLM models for specific stages.

Through the qualitative evaluation of feedback generation, we identified the strengths of both models. Thus, a two-step approach combining both models was adopted. GPT-4o showed an advantage in providing more specific and actionable feedback, while Claude-3.5-Sonnet excelled at delivering supportive and natural Korean expressions with faster response times. Therefore, we first used GPT-4o with temperature 0 to generate an initial draft, ensuring in-depth and detailed content. Afterward, we employed Claude-3.5-Sonnet to refine the expressions and tone, making them more natural and supportive. For this step, the temperature for Claude-3.5-Sonnet was set to 1 to utilize richer Korean expressions.

The final optimized scoring and feedback generation system operates as follows:

- ① Perform quantitative scoring and calculate the score using Claude-3.5-Sonnet.

- ② Generate an initial detailed feedback draft using GPT-4o with temperature 0.
- ③ Refine the Korean expression of the feedback using Claude-3.5-Sonnet with temperature 1.
- ④ Present the final scoring and feedback to the student.

This optimized system is expected to provide accurate and consistent scoring along with specific, natural, and supportive educational feedback. The overall structure and accessibility of the chatbot are illustrated in Figure 2, while an example of the user interface and QR code for access are shown in Figure 3.

Discussion

This study developed and evaluated an LLM-based scoring and feedback system for Korean medicine CPX education. Claude-3.5

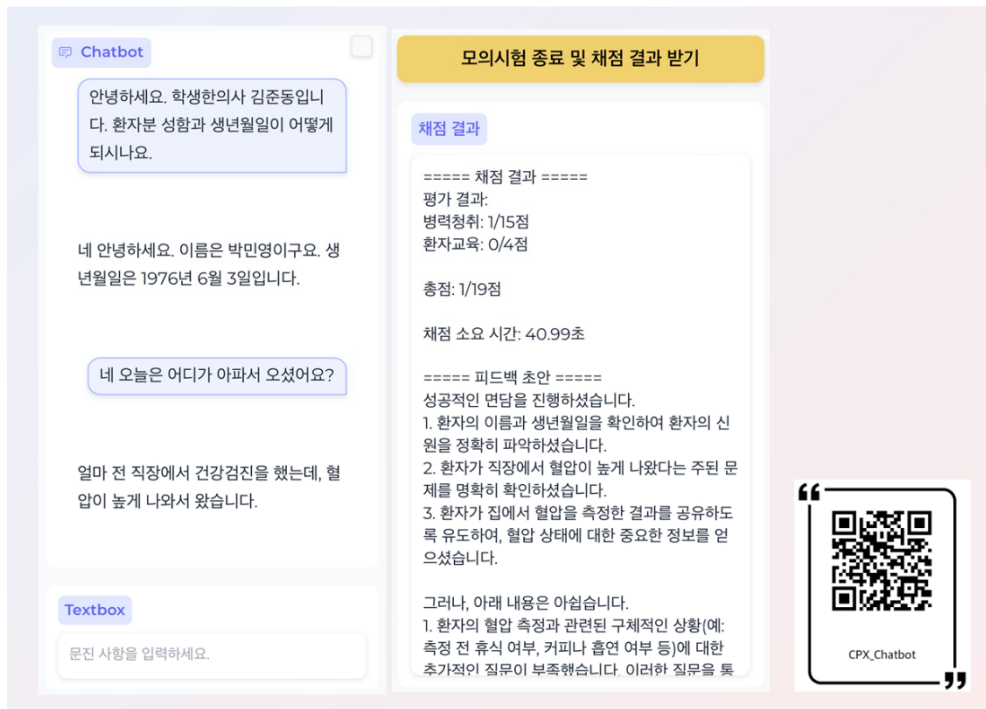


Fig. 3. Screenshot of the chatbot's user interface during a simulated CPX session, showing the scoring result, generated feedback, and QR code for easy access.

-Sonnet demonstrated perfect accuracy in quantitative scoring, with faster execution times (average 43.2 seconds) compared to GPT-4o. In the evaluation of the qualitative text feedback generation system, the strengths of each model were identified, and a method for generating high-quality text feedback by combining both models was proposed.

The quantitative scoring system of our CPX Practicing Chatbot offers several critical advantages for Korean medicine education. Firstly, Claude-3.5-Sonnet's perfect accuracy and consistency enable a standardized evaluation process, effectively mitigating the subjectivity commonly found in current assessment practices especially in Korean

medicine CPX. Second, the system's rapid processing time of 43.2 seconds confirms its practicality for actual educational settings. Most importantly, the system promotes self-directed learning by allowing students to practice clinical scenarios repeatedly, while objectively tracking their progress through clear, numerical feedback. This ability to monitor improvement addresses the inherent cost, time, and physical constraints of conventional CPX education, ultimately enhancing the learning experience.

Also, the qualitative text feedback system proposed in this study, with its two-stage approach, has several educational implications. By setting the temperature value to 0 during draft

generation, the system was designed to provide concise, focused feedback consistently. This aligns with recent research suggesting that providing limited amounts of feedback more frequently is significantly more effective than delivering comprehensive feedback all at once¹⁵). Frequent, targeted feedback helps maintain learner focus and ensures that key areas for improvement are clearly understood, which ultimately enhances learning outcomes. Additionally, several key principles have been established for effective feedback¹⁵). These include: providing specific and actionable improvements, basing feedback on learner's actual behaviors and decisions, and maintaining a supportive and constructive attitude. Our two-stage approach fully embodies these principles by using GPT-4o to generate highly accurate and specific initial drafts, followed by Claude-3.5-Sonnet to refine them into supportive and natural expressions.

Notably, clinical practice in Korean medicine involves a complex dual diagnostic process: practitioners must conduct differential diagnosis for the patient's presenting symptoms while simultaneously performing pattern identification that considers the patient's overall condition. The feedback generation system of our CPX Practicing Chatbot provides comprehensive guidance by not only identifying strengths and weaknesses but also actively suggesting specific learning directions for future improvement. This systematic approach to feedback represents a potentially transformative solution for Korean medicine CPX education, where students have struggled with the complexity of clinical skill development without adequate guidance¹⁷).

Another significant contribution of this study is

the establishment of a systematic evaluation framework. As AI technology rapidly evolves with continuous emergence of new LLMs, and CPX education requires ongoing development of diverse clinical scenarios, a structured evaluation system becomes crucial. Building upon our previous research that proposed methodology for evaluating chatbot response performance⁶), this study further developed methods for assessing both quantitative scoring and qualitative feedback systems. The framework provides objective criteria for evaluating performance and selecting the most suitable models whenever new LLMs or clinical scenarios are developed (Figure 4). Moreover, its applicability is not limited to Korean medicine education but extends to other fields such as general medicine, dentistry, and nursing. This broad applicability represents a sustainable methodology for integrating AI technology into medical education.

However, this study has several limitations. First, the chatbot has been developed using only one essential hypertension scenario, limiting its coverage of diverse clinical situations. Second, the inherent possibility of hallucination in LLMs cannot be completely eliminated. Third, the actual educational effectiveness of the chatbot when applied to students has not yet been empirically validated.

Future studies are planned to address these limitations. We plan to expand the educational scope of the CPX Practicing Chatbot by developing a comprehensive scenario library. Additionally, we will collaborate with medical education departments to objectively evaluate the system's educational effectiveness. In the long

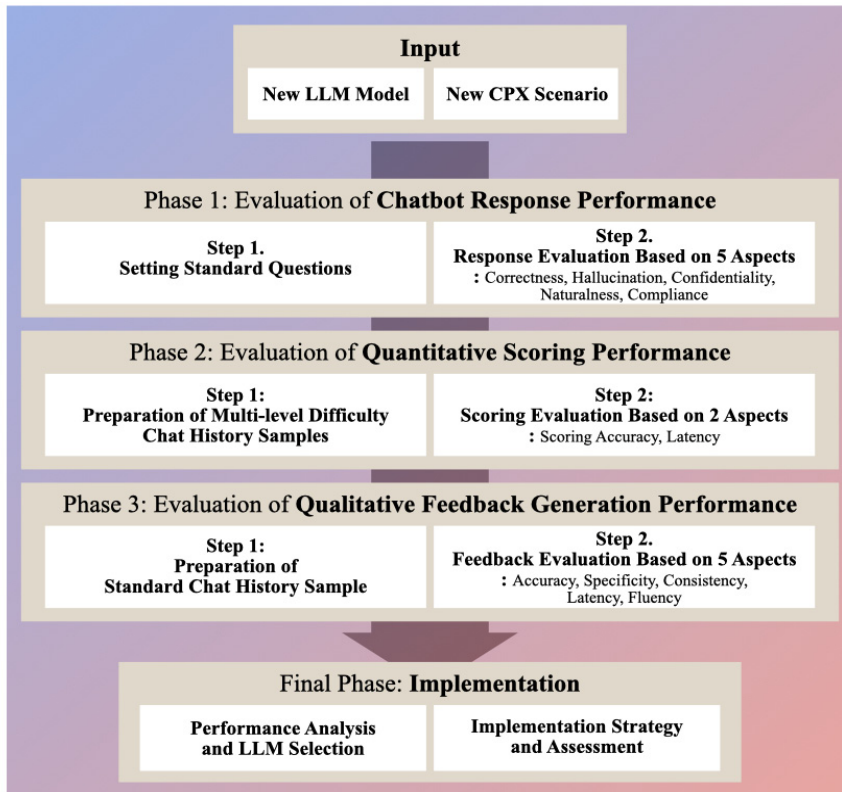


Fig. 4. Implementational framework for LLM-based CPX chatbot for a newly developed LLMs and CPX scenarios.

term, we aim to enhance the realism of CPX practice by incorporating audiovisual features such as voice input/output and virtual patient implementation.

This study illustrates the potential for transforming Korean medicine CPX education using AI technology, providing practical solutions for longstanding challenges in evaluator availability and feedback quality. By implementing a system that integrates both quantitative scoring and qualitative feedback, we have proposed a practical solution to overcome the human and material resource constraints in Korean medicine education. Furthermore, the significance of this

study lies in proposing a systematic framework that can effectively evaluate and implement rapidly evolving AI technologies and various clinical scenarios. The methodology presented in this study is expected to contribute to improving both the quality and efficiency of Korean medicine education.

Conclusion

This study has demonstrated the potential of AI technology to revolutionize CPX education in Korean medicine through a novel evaluation and feedback system. By successfully integrating

quantitative scoring and qualitative feedback mechanisms, our chatbot system offers a practical solution to overcome conventional resource limitations while ensuring consistent, objective assessment. The proposed systematic framework for evaluating and implementing LLM-based educational technologies not only addresses current challenges in Korean medicine education but also provides a sustainable methodology applicable across various medical education fields. Through continued development and refinement, this approach has the potential to significantly enhance both the quality and accessibility of clinical education in Korean medicine.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00339889). And this research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2024-00413256).

References

1. Ende, J. (1983). Feedback in clinical medical education. *Jama*. 250(6). 777-781.
2. Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of educational research*. 77(1). 81-112.
3. Fullerton, P. D.Sarkar, M.Haque, S. & McKenzie, W. (2022). Culture and understanding the role of feedback for health professions students: Realist synthesis protocol. *BMJ Open*. 12(2). e049462. 10.1136/bmjopen-2021-049462
4. Anderson, P. A. (2012). Giving feedback on clinical skills: Are we starving our young? *Journal of graduate medical education*. 4(2). 154-158.
5. Woo, H. (2024). A case study on chuna manual medicine education using various teaching, learning and assessment methods. *Journal of Korean Medicine Education*. 2(1). 6-13. <https://doi.org/10.23215/JKME.PUB.2.1.6>
6. Kim, J., Lee, H.-Y., Kim, J.-H., Kim, C.-E., Kim, J., Lee, H.-Y., et al. (2024). Pilot development of a'clinical performance examination (cpx) practicing chatbot'utilizing prompt engineering. *Journal of Korean Medicine*. 45(1). 200-212.
7. Lee, Y. H.Lee, Y.-M. & Kim, B. S. (2010). Content analysis of standardized-patients' descriptive feedback on student performance on the cpx. *Korean Journal of Medical Education*. 22(4). 291-301.
8. Jo, H. & Min, S. (2020). The current status and future operations of clinical performance evaluation (cpx) in the nationwide colleges (graduate schools) of traditional korean medicine. *The Journal of Korean Medical History*. 33(2). 9-21. <http://dx.doi.org/10.15521/jkmh.2020.33.2.009>
9. Bae, H.Park, S.-Y. & Kim, C.-E. (2024). A practical guide to implementing artificial intelligence in traditional east asian medicine research. *Integrative Medicine Research*. 13(3). 101067.
10. Kang, B.Lee, S.Bae, H. & Kim, C. (2024). Current status and direction of generative

- large language model applications in medicine - focusing on east asian medicine. *Journal of Physiology & Pathology in Korean Medicine*. 38(2). 49-58. <http://dx.doi.org/10.15188/kjopp.2024.04.38.2.49>
11. Lim, C.Han, H.Hong, J. & Kang, Y. (2016). Competency modeling for doctor of korean medicine & application plans. *Jkm*. 2016; 37 (1): 101-13.
 12. Han, Y. (2024). Usability and educational effectiveness of ai-based patient chatbot for clinical skills training in korean medicine. *Korean Journal of Acupuncture*. 41(1). 27-32. [10.14406/acu.2024.001](https://doi.org/10.14406/acu.2024.001)
 13. Kwon, C.-Y. (2024). Utilization of generative artificial intelligence chatbot for training in suicide risk assessment of depressed patients: Focusing on students at a college of korean medicine. *Journal of Oriental Neuropsychiatry*. 35(2). 153-162.
 14. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*. 35(24824-24837).
 15. Bienstock, J. L., Katz, N. T., Cox, S. M., Hueppchen, N., Erickson, S. & Puscheck, E. E. (2007). To the point: Medical education reviews—providing feedback. *American journal of obstetrics and gynecology*. 196(6). 508-513.
 16. Yu, C. X.James, C. S. Y. & David, P. H.-L. P. Can llms have a fever? Investigating the effects of temperature on llm security.
 17. Jeong, S.-H.Kim, J.-P.Kang, Y.-J.Jeong, H. I. & Kim, K. H. (2020). A survey of recognitions and satisfaction with education in traditional korean medicine. *Journal of Society of Preventive Korean Medicine*. 24(3). 49-56.

ORCID

Jundong Kim <https://orcid.org/0000-0002-3919-1508>
Hye-Yoon Lee <https://orcid.org/0000-0002-9486-1703>
Ji-Hwan Kim <https://orcid.org/0000-0001-7270-0987>
Chang-Eop Kim <https://orcid.org/0000-0001-8281-9148>