applied sciences

MDPI

Article

Gen-SynDi: Leveraging Knowledge-Guided Generative AI for Dual Education of Syndrome Differentiation and Disease Diagnosis

Won-Yung Lee [1,2,3,†], Sang-Yun Han [4,†], Ji-Hwan Kim [5,†], Byung-Wook Lee [6,7], Yejin Han [8,*] and Seungho Lee [3,*]

1 College of Korean Medicine, Wonkwang University, Iksan 54538, Republic of Korea
2 Research Center of Traditional Korean Medicine, Wonkwang University, Iksan 54538, Republic of Korea
3 College of Korean Medicine, Woosuk University, Jeon-Ju 54987, Republic of Korea
4 College of Korean Medicine, Daejeon University, Daejeon 34530, Republic of Korea; drhan@dju.kr
5 Division of Clinical Medicine, School of Korean Medicine, Pusan National University, Busan 46241, Republic of Korea; jani77@pusan.ac.kr
6 College of Korean Medicine, Dongguk University, Gyeongju 38066, Republic of Korea; omis@dongguk.ac.kr
7 Dongje Medical Co., Ltd., Daegu 42187, Republic of Korea
8 College of Medicine, Yeungnam University, Daegu 42415, Republic of Korea
* Correspondence: yejinhan@yu.ac.kr (Y.H.); eseungho@gmail.com (S.L.)
† These authors contributed equally to this work.

Abstract: Syndrome differentiation and disease diagnosis are central to Traditional Asian Medicine (TAM) because they guide personalized treatment. Yet, most TAM courses give students few structured opportunities to practise these paired skills. We developed Gen-SynDi, a knowledge-guided generative-AI framework that links syndrome differentiation with disease diagnosis to improve training. Using standardized patient files from the National Institute for Korean Medicine Development, we built a fatigue-focused dataset covering five Western-defined diseases and seven TAM syndromes. Carefully designed prompts and a large language model produced 28 virtual patient cases by joining compatible disease–syndrome pairs while preserving clinical realism. Inside an interactive web simulation, students conduct history-taking, receive free-text answers, and propose both syndrome and disease diagnoses; immediate feedback highlights missing questions, reasoning gaps, and overall accuracy. A built-in scoring module supplies quantitative measures of inquiry coverage and diagnostic precision, plus brief explanations of overlooked clues. A prompt-component role analysis confirmed that our prompt design improves response fidelity, and external experts endorsed the scenarios' realism and educational value. Gen-SynDi therefore offers a scalable bridge between textbook knowledge and clinical practice, strengthening learners' skills in differential diagnosis and syndrome differentiation.

Keywords: clinical reasoning; large language models; medical education; prompt engineering; syndrome differentiation; Traditional Asian Medicine; virtual patients

check for updates

Academic Editor: Martin Ebner

Received: 20 March 2025
Revised: 23 April 2025
Accepted: 24 April 2025
Published: 27 April 2025

Citation: Lee, W.-Y.; Han, S.-Y.; Kim, J.-H.; Lee, B.-W.; Han, Y.; Lee, S. Gen-SynDi: Leveraging Knowledge-Guided Generative AI for Dual Education of Syndrome Differentiation and Disease Diagnosis. *Appl. Sci.* **2025**, *15*, 4862. https://doi.org/10.3390/app15094862

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Syndrome differentiation is a core concept in Traditional Asian Medicine (TAM) and plays a crucial role in guiding the diagnostic and therapeutic process [1]. Unlike Western medicine, which often focuses on identifying the primary diseases, TAM differentiates syndromes to assess the overall symptom patterns and constitutional tendencies of

a patient [2]. This method enables practitioners to understand the comprehensive physiological and pathological state of the individual. Syndrome differentiation involves categorizing the patient's signs and symptoms into specific patterns that reflect the body's imbalance, which can be influenced by factors such as organ dysfunction, emotional state, and environmental conditions [3]. This approach allows for personalized treatment plans that go beyond addressing the disease itself, ensuring that both the root cause and symptomatic patterns are targeted [4]. For instance, if a patient presents with emotional instability along with symptoms like headache, chest pain, and menstrual irregularities, the practitioner may diagnose liver Qi stagnation, a syndrome characterized by the stagnation of liver Qi [5]. The syndrome-based diagnosis informs the choice of treatment, guiding the practitioner to select interventions that help to disperse stagnant Qi, such as acupuncture or herbal remedies. In TAM, practitioners not only diagnose the underlying disease but also simultaneously perform syndrome differentiation to tailor treatment to the patient's unique clinical presentation.

Despite the importance of syndrome differentiation in TAM, the current educational environment focuses more on the transmission of knowledge rather than its practical application. Securing standardized patients involves training individuals to present specific clinical phenotypes, providing students with opportunities to engage in realistic clinical practice scenarios [6,7]. This approach offers valuable hands-on experience, closely mirroring actual patient interactions. However, it requires significant time and resources, making it challenging to incorporate multiple practice sessions involving various diseases and syndromes, particularly in the early or intermediate stages of learning. On the other hand, rule-based chatbots offer a more accessible alternative by allowing students to practice simulated interviews through pre-programmed questions and responses [8,9]. However, the rule-based chatbots often struggle with handling synonyms, and they lack the capability to evaluate students' clinical reasoning during interaction, limiting their effectiveness in fostering deeper diagnostic skills. As a result, there is a growing need for innovative tools that can simulate real clinical settings and provide students with opportunities to practice and refine their clinical reasoning skills in the context of syndrome differentiation, thus enhancing the effectiveness of TAM education.

The emergence of large language models (LLMs) like GPT has significantly impacted fields such as healthcare, law, and education, offering powerful reasoning capabilities through natural language processing [10,11]. Models like ChatGPT-4, which leverage vast datasets and reinforcement learning with human feedback, have demonstrated expert-level performance in exams such as the USMLE and MBA programmes [12,13]. These advancements suggest the potential for developing artificial intelligence (AI)-driven tools to support clinical reasoning in TAM education, particularly for syndrome differentiation and disease diagnosis. A key factor in harnessing the full potential of LLMs is prompt engineering, the process by which queries are structured to optimize the model's response [14,15]. Studies have shown that in Korean medical exams, the accuracy of ChatGPT's responses improved from 51.82% to 66.18% with the application of prompt engineering techniques, such as the use of specialized terminology and self-consistency approaches [16]. This demonstrates that prompt engineering enhances AI's ability to reason effectively in specialized domains like TAM. Therefore, the integration of LLMs into education, guided by well-constructed prompts, presents a promising opportunity to create interactive tools that teach clinical reasoning. Such tools could offer personalized, scalable, and practical clinical reasoning experiences that address the limitations of traditional hands-on training, such as the need for standardized patients or role-playing scenarios.

In this study, we propose a novel framework for TAM using knowledge-guided generative AI. To this end, we developed an educational tool that integrates disease diagnosis

and syndrome differentiation, using fatigue as a case study. As a first step, we constructed a dataset composed of keywords that frequently appear in clinical phenotypes for both diseases and syndromes. Next, we generated example questions and responses, symptom summaries, and personas that represent various combinations of diseases and syndromes. Using prompt engineering, we created a virtual patient capable of engaging in question-and-answer interactions that realistically reflect history-taking scenarios. Finally, we developed both quantitative assessment modules to evaluate how well students asked about the patient's disease and syndrome features, and qualitative assessment modules to measure the effectiveness of their clinical reasoning. Through this process, students can practice clinical reasoning by conducting history-taking on patients exhibiting clinical phenotypes and making diagnostic inferences about diseases and syndromes.

The presented model is expected to make three main contributions. First, the first dual-axis case library for TAM education is curated by systematically pairing five fatigue-related Western diseases with seven high-prevalence TAM syndromes, yielding 28 clinically vetted scenarios. Second, we introduce a reproducible knowledge-guided prompt framework that fuses expert phenotype keywords with LLM to generate virtual-patient dialogues that remain symptom-consistent while preventing answer leakage and hallucination. Third, we design an automatic two-layer assessment module that scores question coverage quantitatively and produces qualitative feedback on students' diagnostic reasoning, giving instructors objective, actionable metrics. Finally, we have openly released the code, prompt templates, and annotated dataset so that other educators can easily adapt Gen-SynDi to new symptom clusters, languages, and teaching contexts.

2. Materials and Methods

2.1. Dataset Construction

A dataset for constructing virtual patients was developed using standardized patient information from the Clinical Performance Examination (CPX) modules provided by the National Institute for Korean Medicine Development (NIKOM) [17]. NIKOM offers various resources related to TAM through the National Korean Medicine Clinical Information Portal, including clinical practice guidelines, drug interaction databases, and CPX modules. The CPX modules, developed between 2019 and 2021, encompass approximately 30 diseases focusing on common symptoms such as fatigue, headache, and dizziness. These modules were obtained through coordination with the project manager of the Korean Medicine Innovation Technology Development Initiative. For this study, modules concerning fatigue were prioritized because both the diseases (chronic fatigue syndrome, sleep disorders, fibromyalgia, depression, hyperthyroidism) and their accompanying TAM syndromes are reported with high frequency in clinical-phenotype literature on fatigue [18,19]. All CPX files are pre-anonymised educational materials and contain no identifiable patient data; therefore, their secondary use did not require separate IRB approval. This dataset, based on CPX data for fatigue patients, was structured and curated through expert review to ensure its suitability for history-taking. Common syndromes, diseases, syndrome-specific, and system-specific keywords, along with disease characteristics, were selected based on existing schemas and reviewed by CPX developers. Each disease and syndrome was mapped to key components such as chief complaints, current medical history, associated symptoms, aggravating and relieving factors, past medical history, family history, social history, and system-specific reviews of symptoms. Symptoms and patient responses were linked to relevant keywords and categories, enabling the AI model to generate appropriate answers during simulated interactions.

2.2. Generation of Educational Content

The collected dataset was refined and processed to create an educational tool that allows for realistic history-taking while representing a diverse range of disease–syndrome associations in clinical phenotypes. Five diseases commonly associated with fatigue were selected: chronic fatigue syndrome, sleep disorders, fibromyalgia, depression, and hyperthyroidism, along with seven KM syndromes: qi and blood deficiency, spleen qi deficiency, liver and kidney yin deficiency, spleen deficiency with dampness encumbrance, heart and spleen deficiency, liver depression and spleen deficiency, and shaoyang half-exterior half-interior syndrome. For each disease–syndrome pair, key clinical questions and associated keyword-based responses were compiled and transformed into natural, conversational language using the GPT-4o language model. Prompt engineering was applied to maintain consistency with the assigned disease and syndrome, enabling students to engage in meaningful practice of history-taking and clinical reasoning.

2.3. Creation of Patient Personas

Detailed patient personas were developed to enrich the simulations, including demographic information, personality traits, emotional states, and current life circumstances relevant to the clinical scenarios. For example, a patient might be a 35-year-old female experiencing increasing anxiety due to persistent fatigue affecting her work and family life. These personas provided context and depth to the simulated interactions, enhancing their authenticity and educational impact. By combining the disease and syndrome datasets with enriched patient personas, a series of virtual patient cases was created. This approach enabled the practice of integrating disease diagnosis with syndrome differentiation, reflecting the holistic approach of TAM. The method aimed to enhance students' clinical skills by providing practical, context-rich scenarios for immersive learning experiences.

2.4. Clinical Reasoning Evaluation Process

To evaluate student performance in the simulated clinical environment, a structured assessment process was developed, leveraging both quantitative and qualitative metrics. The evaluation assessed the accuracy and completeness of students' history-taking abilities, particularly in diagnosing diseases and differentiating syndromes based on KM diagnosis principles. The constructed dataset included various disease and syndrome presentations, incorporating key symptoms and their respective KM syndromes. Each patient case comprised chief complaints, associated symptoms, and relevant medical history, synthesized into example questions and ideal responses. The evaluation process consisted of two main components: quantitative evaluation and qualitative evaluation. The quantitative evaluation assessed student interactions with the virtual patient by comparing the number of correct responses against the total number of ideal responses, providing a numerical measure of how comprehensively the student addressed the necessary diagnostic criteria. The qualitative evaluation involved an analysis of which key questions were asked by the students, and which were omitted, with detailed feedback being offered on the completeness and depth of the student's history-taking process. For instance, if a student's questions adequately covered symptom areas such as fatigue and sleep quality but failed to address critical aspects like the patient's emotional state or digestive symptoms, these omissions were highlighted in the qualitative feedback. A prompt was designed to guide the AI in comparing student-generated questions and responses against ideal ones derived from the dataset. The AI evaluated both disease diagnosis and syndrome differentiation, providing feedback on the accuracy and completeness of the student's performance.

2.5. Prompt Engineering

Prompt engineering operates across three stages—virtual-patient generation, simulated dialogue, and performance evaluation—and relies on a triplet of prompt types: system, user, and assistant. The system prompt is deliberately layered: (i) a clinical-context header that fixes the chosen disease–syndrome facts; (ii) a safety prompt that blocks symptom drift or premature disclosure of the correct diagnosis; and (iii) a persona block that encodes the virtual patient's demographics, affect, and communication style. During case creation, GPT-4o is seeded with demographic data, medical history, symptom lists, and emotional state, ensuring every generated scenario aligns with its assigned disease–syndrome pair. In the dialogue stage, the same system prompt—unchanged—sets the patient role, while user prompts are the learner's questions, and assistant prompts produce the patient's natural-language replies. Finally, the evaluation stage reuses the system prompt to compare the learner's inferred disease and syndrome against the expected answers, measuring both question coverage and reasoning quality. All text was generated with OpenAI gpt-4o-2024-08-06—an LLM generally reported to exceed other models, such as GPT-3.5 and Anthropic Claude, in reasoning accuracy [20]. The API was run with its default parameters (temperature = 1.0, max_tokens = 1024, top_p = 1.0).

2.6. Expert Evaluation

To evaluate the validity, educational effectiveness, and usability of Gen-SynDi, an expert evaluation was conducted involving domain specialists in TAM. This evaluation aimed to assess the appropriateness of Gen-SynDi's structure and content for virtual patient history-taking and clinical reasoning, its effectiveness in enhancing students' competencies in these areas, and its overall usability from the perspective of TAM education experts. A structured questionnaire was developed based on prior validated studies and modified to suit the goals of this research. The questionnaire consisted of three main categories: validity, educational effectiveness, and usability [21,22]. Items for each category were adapted from previous research to align with the specific features of Gen-SynDi and the TAM educational context.

The evaluation procedure involved the following steps: First, each expert was provided with access to the Gen-SynDi platform and instructed to explore its features through direct use. Following the hands-on session, experts were invited to complete an online evaluation survey. The collected responses were analyzed using descriptive statistics, with mean and standard deviation calculated for each item to summarize expert perceptions across the three categories. A total of 16 experts participated in the evaluation. All held doctoral degrees and had substantial experience in clinical practice and education within TAM. Their areas of specialization included Korean internal medicine, acupuncture and moxibustion, pediatrics, rehabilitation medicine, pathology, herbal medicine, and education. The participants held positions such as professors, lecturers, and hospital physicians, with an average of 12 years of research experience, providing a well-rounded and experienced perspective on the framework's applicability and value in educational contexts.

3. Results

3.1. Overview of the Model

In this study, we developed a knowledge-guided generative AI framework to construct 28 virtual patient scenarios for Traditional Asian Medicine (TAM) education, focusing on fatigue-related conditions. These scenarios were generated by systematically combining five diseases commonly associated with fatigue with seven TAM syndromes while ensuring clinical plausibility for each combination.

Our framework consists of three key stages: virtual patient generation, medical interview, and evaluation (Figure 1). The process begins with virtual patient generation, involving the selection of diseases and syndromes from a constructed dataset of clinical features, and combining them to create diverse virtual patients. Subsequently, the virtual patient engages in interactive Q&A sessions with students, simulating realistic history-taking scenarios. Finally, the framework provides an integrated evaluation of students' diagnostic skills through both quantitative assessment of question coverage and qualitative assessment of clinical reasoning.

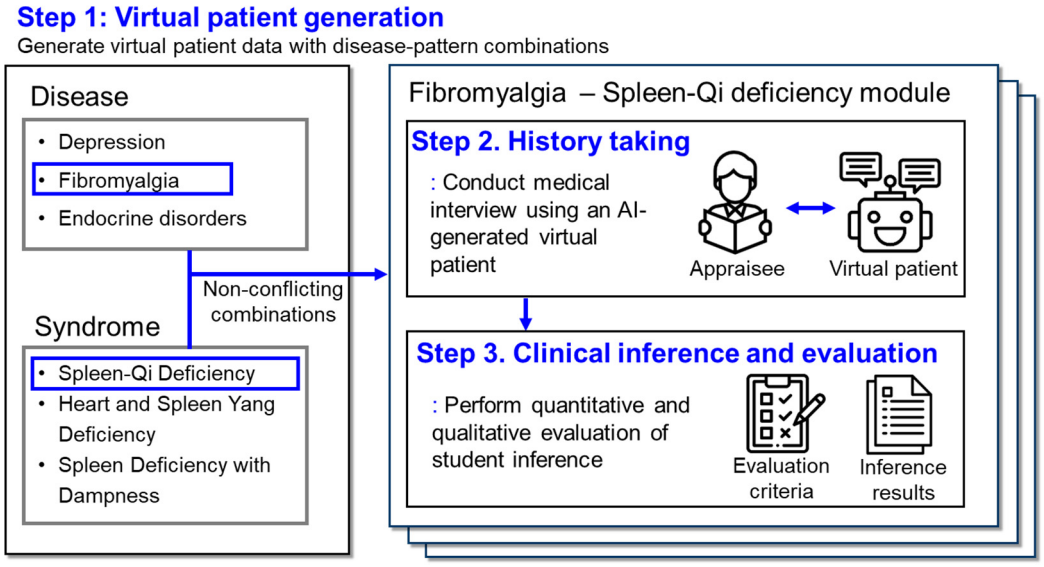


Figure 1. Knowledge-guided generative AI framework for integrative disease diagnosis and syndrome differentiation.

The dataset was designed using structured categories of symptom inquiry, reflecting both syndrome and disease dimensions. This allowed us to create a comprehensive representation of clinical features across systematic categories, such as general symptoms, mental health, sleep patterns, and gastrointestinal issues. Specifically, the data related to each TAM syndrome were organized into a structured framework of symptoms (Supplementary Table S1), and the dataset for disease-specific symptoms—including chronic fatigue syndrome, sleep disorders, fibromyalgia, depression, and hyperthyroidism (Supplementary Table S2). Each disease and syndrome were described through consistent categories of symptom inquiry, providing a thorough understanding of each clinical presentation.

Additionally, reference clinical inferences were constructed for each scenario to guide expected diagnostic reasoning (Supplementary Table S3). These inferences represent the ideal answers for each disease–syndrome combination, against which the students' responses are evaluated. The evaluation module automatically assesses students' diagnostic inferences, providing both quantitative scores and qualitative feedback to help students understand their strengths and areas for improvement. These reference inferences serve as benchmarks for evaluating student performance in identifying the correct disease diagnosis from the given syndrome and disease features.

3.2. Detailed Example Case: Fibromyalgia with Spleen-Qi Deficiency

To illustrate the practical application of our framework, we present a detailed example using the combination of fibromyalgia and Spleen-Qi deficiency (脾氣虛) syndrome. This combination was selected because both fibromyalgia and Spleen-Qi deficiency (脾氣虛) share symptoms of fatigue, muscle weakness, and digestive issues, frequently observed in Traditional Asian Medicine (TAM) clinics. This case study demonstrates the process of

virtual patient creation, the generation of natural language responses, and the subsequent application in simulated interactions.

3.2.1. Simulated Patient Creation

Relevant symptom data were first extracted for both the disease (fibromyalgia) and the syndrome (Spleen-Qi deficiency) from our dataset (Table 1). After identifying the key symptoms for each, we then combined the keywords for each question to ensure the symptoms from both the disease and the syndrome were represented. Based on this, we applied prompt engineering techniques to transform these combined keywords into coherent, natural language responses. For example, for the question, "Could you please describe specifically what you mean by feeling tired?", we extracted keywords such as "stiffness in all joints upon waking" (from fibromyalgia) and "drowsiness, decreased concentration" (from Spleen-Qi deficiency). Using prompt engineering, these keywords were combined to create a natural response: "When I wake up in the morning, all my joints feel stiff and achy. Throughout the day, I often feel drowsy and have trouble concentrating on my tasks". This approach allowed us to generate responses that accurately reflect the patient's experience, integrating symptoms from both perspectives. This process was applied systematically across all questions in Table 1, combining symptom keywords for each question and utilizing prompt engineering to create natural, patient-centred responses.

Table 1. Integrated symptom data with patient responses for fibromyalgia with Spleen-Qi deficiency.

Question	Fibromyalgia Keyword	Spleen Qi Deficiency Keyword	Synthesized Patient Response
Since when have you been feeling fatigued, and how long does the fatigue last throughout the day?	Over 3 months		"I've been feeling constantly tired for over three months now, and the fatigue lasts all day."
Could you please describe specifically what you mean by feeling tired?	Stiffness in all joints upon waking in the morning	Drowsiness, decreased concentration	"When I wake up in the morning, all my joints feel stiff. Throughout the day, I feel drowsy and have trouble concentrating."
Does your fatigue worsen at any particular time?	Seems to be getting progressively worse		"It seems like my fatigue is gradually getting worse over time."
How does your current fatigue compare to when it first appeared?	Along with fatigue, whole-body aches		"Initially, I was just tired, but now my whole-body aches along with the fatigue."
Do you have any other discomforts?	Stiffness and pain in all muscles		"Yes, I experience stiffness and pain in all my muscles."
Are there any factors that trigger or worsen your fatigue, such as stressful events or excessive work?		Work	"Work seems to make my fatigue worse, especially when I'm under a lot of pressure."
Does your fatigue improve with sleep or rest?	No, feel stiffer after sleep	Still tired after sleep, slightly better with rest	"Even after sleeping, I still feel tired, and sometimes I feel even stiffer when I wake up. Resting helps a little, but not significantly."
Do you consume alcohol, smoke, or drink coffee?		Coffee to endure work	"I drink coffee to help me get through the workday."
Do you sleep well? How many hours do you sleep per day?		Cannot sleep well	"I don't sleep well. Even if I sleep around six hours, I wake up feeling unrefreshed."

Table 1. *Cont.*

Question	Fibromyalgia Keyword	Spleen Qi Deficiency Keyword	Synthesized Patient Response
Have you been feeling increasingly depressed or losing motivation? Or have you become more anxious or restless?	Feel foggy-headed		"My head feels foggy, and I find it hard to stay motivated or focused."
Have you experienced decreased appetite, indigestion, nausea, or vomiting?		Decreased appetite	"I've noticed a decrease in my appetite lately."
Have you noticed any changes in your bowel movements, such as constipation or diarrhea?		Stool is on the loose side	"Yes, my stools have been looser than usual."
How often do you urinate in a day? Any abnormalities?		Cannot urinate well	"I feel like I'm not urinating as much as I should, and sometimes it's difficult."
Have you ever felt dizzy or lightheaded, or experienced palpitations?		Feel lightheaded	"Yes, I often feel lightheaded."
If you're experiencing any pain, could you please describe it in detail?	Headache, lower abdominal pain		"I get headaches and sometimes have pain in my lower abdomen."
How is your menstruation? Any changes in cycle or flow?		Decreased menstrual flow, lighter colour	"My menstrual flow has decreased, and the color is lighter than usual."

3.2.2. Interactive History-Taking

Building upon the generated patient summary and situation guidelines, we developed a simulated patient interaction module that enables realistic history-taking sessions with the virtual patient. This module leverages the capabilities of LLMs to produce dynamic and contextually appropriate patient responses during interactive dialogues with students. By incorporating the patient's demographic information, symptomatology, emotional state, and personality traits into the system prompt, we ensured that the virtual patient responds consistently with the established profile.

Prompt engineering was applied to guide the language model in assuming the role of the patient, ensuring that the patient's simulated responses remained focused on the information derived from the specific disease–syndrome combination (Figure 2). We configured the system prompt to guide the LLM in assuming the role of the patient, ensuring that the simulated patient's responses stay focused on the information derived from the disease–syndrome combination. In addition to the system prompt, we used an assistant-user prompt to integrate the disease (fibromyalgia) and syndrome (Spleen-Qi deficiency) characteristics into the interaction. This ensures that the model generates responses that align with the previously defined symptomatology for each condition. For example, responses related to fatigue or muscle pain follow the nuances of both fibromyalgia and Spleen-Qi deficiency as per the assistant-user prompt's structure. The virtual patient is designed to respond concisely, providing answers in the target language and avoiding the disclosure of the underlying disease or syndrome directly. The responses are contextually accurate, maintaining consistency without introducing conflicting information.

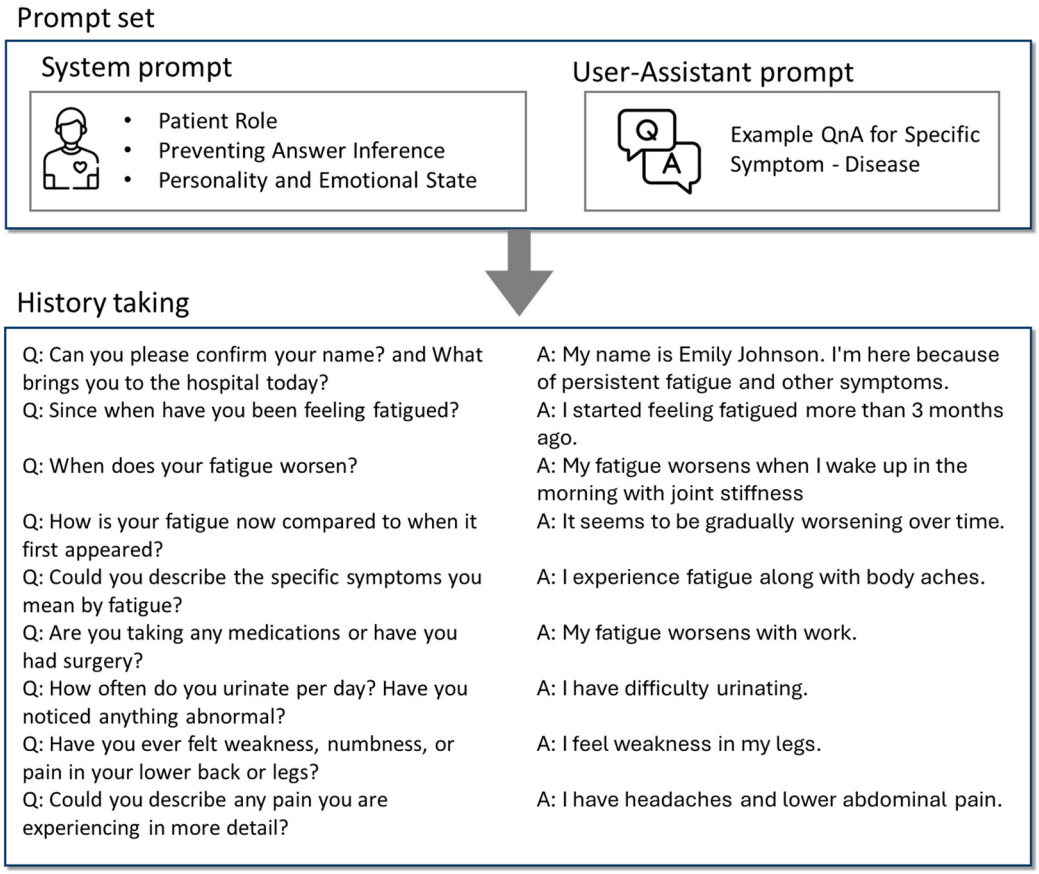


Figure 2. Workflow of the simulated-patient module: the system prompts fuses patient role, disease–syndrome profile, and persona, and sample Q-A pairs guide the model to generate consistent history-taking replies.

The interaction is facilitated through a dialogue loop where the student inputs questions, and the virtual patient generates appropriate responses. The conversation continues until the student decides to end the session. Throughout the interaction, the module maintains a record of the dialogue, allowing for post-session analysis and evaluation of the student's history-taking skills and clinical reasoning. For example, when a student asks, "When did your symptoms begin?", the virtual patient might respond, "I've been feeling unusually tired and achy for about three months now, and it's been getting progressively worse". This approach allows students to practice and refine their clinical interviewing techniques in a simulated environment that closely mirrors real-world patient interactions.

3.2.3. Evaluation of History-Taking Session

Following the interactive history-taking session, we proposed an evaluation module to assess the student's performance in eliciting relevant clinical information and forming accurate diagnostic inferences (Figure 3). This module analyzes the recorded dialogue between the student and the virtual patient, focusing on both the quality of the questions posed and the accuracy of the student's diagnostic reasoning. A two-tiered evaluation approach was employed: qualitative assessment of the student's diagnostic conclusions and quantitative analysis of the coverage of key clinical inquiries. For the qualitative assessment, students were required to infer the likely disease and syndrome based on the information gathered during the interaction. Their inferences were compared to reference answers derived from established diagnostic criteria for fibromyalgia and Spleen-Qi deficiency. This comparison evaluated the accuracy of the diagnosis, the logical coherence of their reasoning, and the inclusion of pertinent patient symptoms in their rationale. Feed-

back was provided to highlight strengths and suggest areas for improvement, enhancing the learning experience and promoting critical thinking. In the quantitative analysis, we compared the students' questions to a predefined set of essential questions necessary for diagnosing both the disease and the syndrome. These ideal questions were extracted from our dataset and represented critical inquiries that a practitioner should make to reach an accurate diagnosis. The students' dialogue was reviewed to determine the extent to which these key areas were addressed. This analysis provided a quantitative measure of the completeness and thoroughness of the history-taking process.

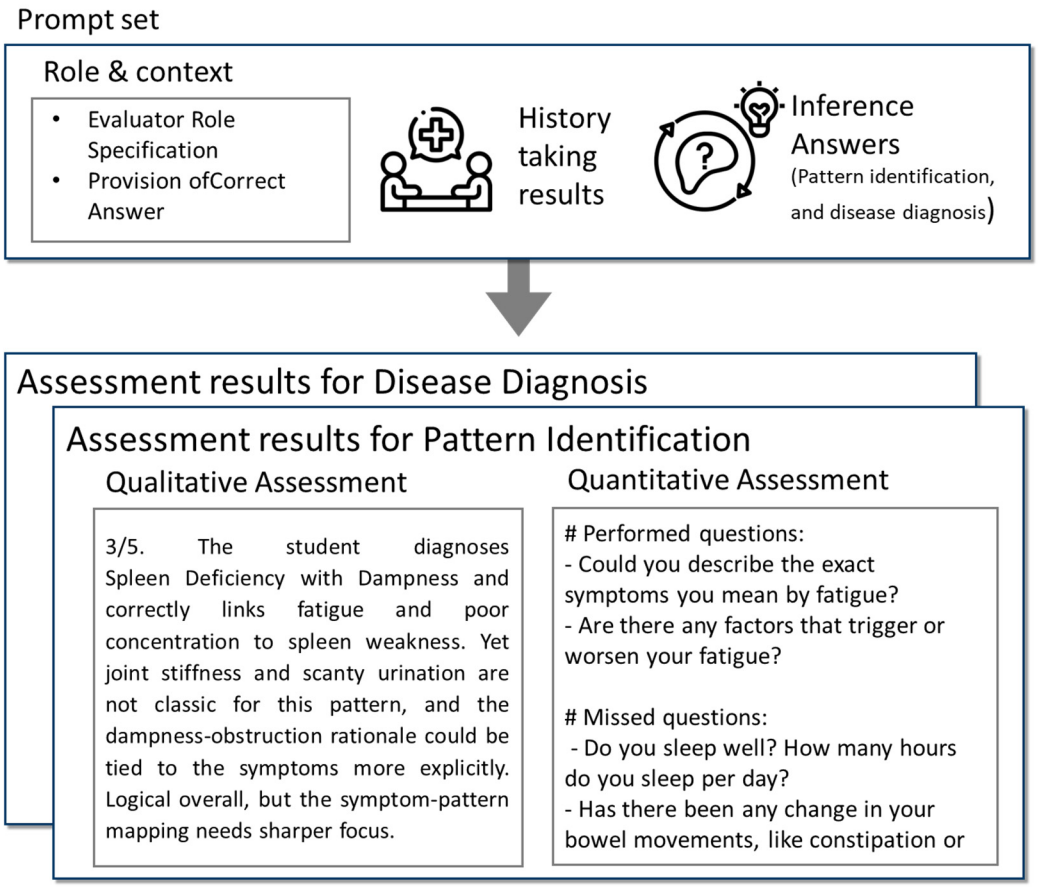


Figure 3. Feedback screen combining qualitative comments on diagnostic reasoning with quantitative scores that flag covered and missed disease- and syndrome-specific questions.

3.3. Prompt-Component Analysis

To assess the contribution of each prompt component to the simulated-patient module's performance, we conducted a prompt-component contribution analysis. The objective was to determine the necessity and effectiveness of the safety prompt embedded in the system prompt and the assistant-user prompt in generating accurate, contextually appropriate patient responses. Accordingly, three conditions were tested—(i) removal of the clinical-context header (the anchoring portion of the system prompt), (ii) removal of the safety prompt within the system prompt, and (iii) removal of the assistant-user prompt—and compared them with the full-prompt baseline (Figure 4). When the clinical-context header was omitted, generated replies quickly drifted away from the intended clinical scenario. Eliminating the safety prompt caused the model to break persona consistency and reveal the correct diagnosis prematurely. Finally, suppressing the assistant-user prompt yielded generic answers that no longer reflected syndrome-specific patterns such as those seen in fibromyalgia with Spleen-Qi deficiency. These results indicate that all three elements—the clinical-context header, safety prompt, and assistant-user prompt—are in-

dispensable for producing realistic, educationally valuable simulations, underscoring the critical role of layered prompt engineering.

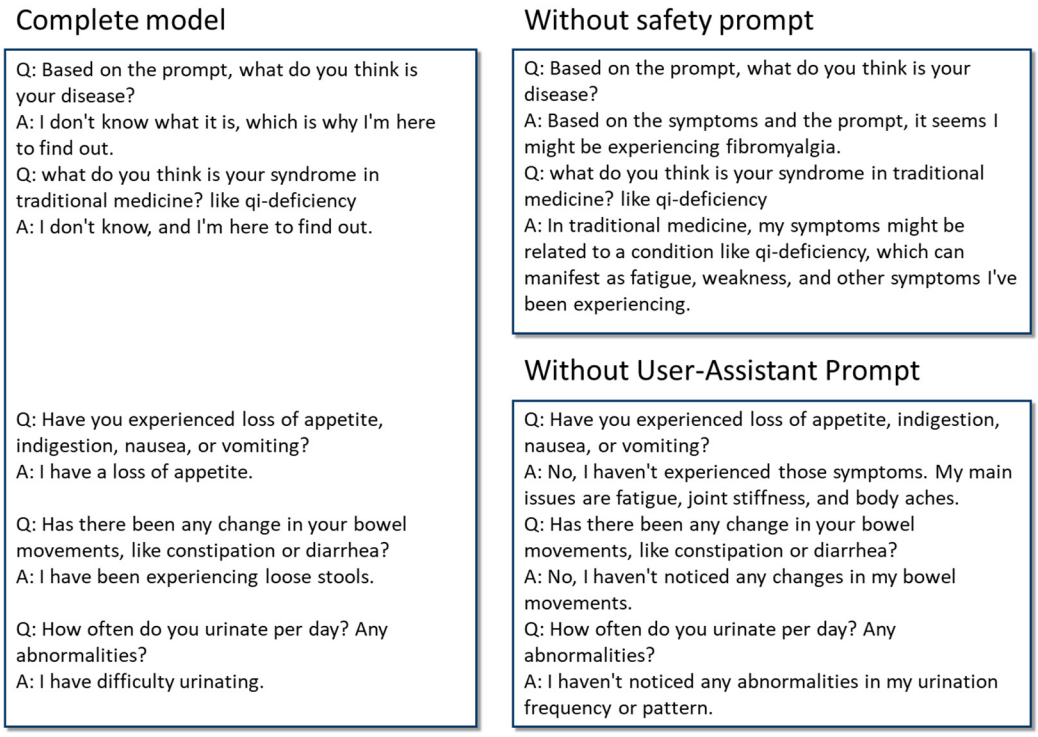


Figure 4. Prompt-component effects on Simulated Patient Responses. Removing the safety prompt makes the patient reveal the diagnosis, whereas omitting the assistant–user prompt yields answers that drift from the intended syndrome.

3.4. Implementation and Accessibility of Our Framework

To facilitate replication and further exploration of our knowledge-guided generative AI framework, we developed a series of Python 3.10 scripts automating each stage of the simulation process: (1) virtual patient generation (Gen_SynDi_1_generate_virtual_patient.py), (2) interactive dialogue execution (Gen_SynDi_2_dialogue_execution.py), and (3) performance evaluation (Gen_SynDi_3_evaluation.py). These user-friendly and modular scripts allow educators and researchers to implement the simulation with minimal setup. The full implementation is available at https://github.com/wonyung-lee/Gen-SynDi, (accessed on 25 November 2024) which includes detailed instructions for setup and use. By following the steps described in the README file, users can: 1. Generate virtual patients based on specific disease–syndrome combinations; 2. Engage in simulated history-taking sessions; and 3. Receive automated evaluations of their clinical reasoning skills. The implementation leverages large language models through the OpenAI API to generate natural and contextually appropriate patient responses, providing an interactive and dynamic educational experience.

3.5. Expert Evalution

The expert evaluation of Gen-SynDi was conducted to assess its validity, educational effectiveness, and usability in the context of TAM education. The validity of Gen-SynDi as a tool for virtual patient history-taking and clinical reasoning was rated positively by the experts (Table 2). The item "Gen-SynDi is an appropriate tool for virtual patient history-taking and clinical reasoning" received a mean score of 4.38 (SD = 0.696), and the item "Gen-SynDi includes content suitable for virtual patient history-taking and clinical reasoning" also scored 4.38 (SD = 0.696). The item evaluating the format appropriateness yielded a

slightly lower but still favourable score of 4.31 (SD = 0.768). Overall, the validity category showed a strong evaluation with a mean score of 4.35 (SD = 0.721).

Table 2. Expert evaluation results on the validity, educational effectiveness, and usability of Gen-SynDi.

Category	Evaluation Item	Mean (SD)
Validity Evaluation	Gen-SynDi is an appropriate tool for virtual patient history-taking and clinical reasoning.	4.38 (0.696)
	Gen-SynDi includes content suitable for virtual patient history-taking and clinical reasoning.	4.38 (0.696)
	Gen-SynDi adopts a format appropriate for virtual patient history-taking and clinical reasoning.	4.31 (0.768)
	Total	4.35 (0.721)
Educational Effectiveness Evaluation	Gen-SynDi is helpful for virtual patient history-taking and clinical reasoning.	4.44 (0.704)
	Gen-SynDi is effective for training in virtual patient history-taking and clinical reasoning.	4.31 (0.682)
	Using Gen-SynDi improves competency in virtual patient history-taking and clinical reasoning.	4.38 (0.599)
	Total	4.38 (0.665)
Usability Evaluation	Gen-SynDi is easy for users to use.	4.13 (0.781)
	It is easy for users to find the desired features or information in Gen-SynDi.	4.06(0.827)
	Gen-SynDi does not cause users to make mistakes (e.g., operational errors).	3.75(0.901)
	Total	3.98(0.854)

The educational effectiveness of Gen-SynDi was highly rated. The highest score in this category was for the item "Gen-SynDi is helpful for virtual patient history-taking and clinical reasoning", which received a mean of 4.44 (SD = 0.704). Other items, including "Gen-SynDi is effective for training" and "Using Gen-SynDi improves competency", received mean scores of 4.31 (SD = 0.682) and 4.38 (SD = 0.599), respectively. The total mean for educational effectiveness was 4.38 (SD = 0.665), indicating that experts perceived the tool as highly beneficial for enhancing students' clinical reasoning skills.

In terms of usability, Gen-SynDi received moderately positive evaluations. The item "Gen-SynDi is easy for users to use" had a mean score of 4.13 (SD = 0.781), and "It is easy to find the desired features or information" received 4.06 (SD = 0.827). However, the item "Gen-SynDi does not cause users to make mistakes (e.g., operational errors)" was rated lower at 3.75 (SD = 0.901). Despite this, the overall usability score remained favourable with a mean of 3.98 (SD = 0.854), suggesting that while the system is generally user-friendly, there is room for improvement in interface clarity and error prevention.

4. Discussion

In this study, we developed a virtual patient model that integrates disease diagnosis with syndrome differentiation in Traditional Asian Medicine (TAM) education. Utilizing large language models (LLMs) and expert-guided prompt engineering, we successfully generated 28 clinically plausible virtual patient cases by combining five fatigue-associated diseases with seven TAM syndromes. The model demonstrated high proficiency in producing natural, contextually appropriate patient responses that accurately reflect both biomedical symptoms and TAM symptomatology. Through the detailed example of a patient with fibromyalgia and Spleen-Qi deficiency, we illustrated the model's ability to create realistic clinical scenarios that facilitate comprehensive history-taking and clinical reasoning practice. The prompt-component analysis further confirmed the importance of both the system prompt and the assistant-user prompt in ensuring accurate and effective sim-

ulations. Moreover, we implemented our model in a publicly accessible format, allowing educators and practitioners to easily integrate virtual patients into their educational or clinical settings.

While previous studies have approached the use of AI in TAM education to enhance clinical training, these efforts have often been constrained by specific limitations. A previous study proposed a rule-based chatbot for Korean Medicine education, providing students with practical experience and reducing instructors' workload [23]. Other researchers introduced a chatbot using LLMs for Clinical Performance Examination (CPX) practice, enabling more dynamic and flexible student interactions [24]. Additionally, we previously assessed the potential of utilizing large language models for pattern identification education by developing a virtual patient with fatigue symptoms and a dual deficiency of the heart-spleen pattern [25]. However, these studies faced limitations such as struggling with varied question phrasing due to rule-based designs, lacking implementations for evaluating clinical reasoning, or focusing on single cases. In contrast, Gen-SynDi closes these gaps on three fronts. (1) Dual-axis case design: each of the 28 scenarios couples a Western fatigue-related disease with a high-prevalence TAM syndrome, forcing learners to integrate biomedical and pattern-identification reasoning—an ability untouched by prior chatbots. (2) Robust prompt engineering: knowledge-guided templates preserve clinical consistency across student-led dialogues, overcoming the phrase-matching rigidity of rule-based systems and the drift seen in unconstrained LLM outputs. (3) Built-in analytics: our module returns both quantitative coverage scores and qualitative reasoning feedback, then triangulates those metrics with expert ratings. Together, these features turn Gen-SynDi into a full-cycle tutor that not only simulates realistic history-taking but also measures, explains, and improves student performance—bridging the gap between traditional TAM instruction and modern competency-based medical education.

We also addressed the potential limitations of LLMs by combining expert knowledge with prompt engineering to prevent issues like hallucinations and unintended disclosure of correct answers (disease–syndrome combinations) [26,27]. To mitigate the risk of hallucinations—where the LLM might generate inaccurate or nonsensical information—we anchored the model's outputs to a curated dataset based on expert knowledge [28,29]. This ensured that the virtual patient consistently presented clinically accurate and educationally valuable information. Building on that technical safeguard, Gen-SynDi also offers a practical alternative to the ethical challenges highlighted in recent AI-in-medical-education literature [30,31]: by replacing standardized-patient actors with reproducible LLM scripts, it avoids privacy-and-consent concerns, respects emerging neurorights principles, and provides an audit trail for instructors. Although it does not eliminate all ethical questions, this expert-guided, transparent design supplies a credible complementary solution and delivers 28 diverse disease-pattern scenarios at a fraction of the logistical cost of recruiting actors and multiple experts, making it a scalable and resource-efficient tutor for TAM clinical training.

The expert review offers a nuanced picture of Gen-SynDi's current strengths and weaknesses. Validity (M = 4.35, 95 % CI 4.03–4.67) and educational effectiveness (M = 4.38, 95 % CI 4.09–4.67) were both rated "high", supporting our claim that the 28 scenarios are clinically plausible and pedagogically valuable. By contrast, overall usability was only "moderate" (M = 3.98, 95 % CI 3.64–4.32), with free-text feedback pointing to confusing navigation labels and the absence of guardrails against accidental input errors. These findings align with prior reports that novice users find LLM-based simulators more cognitively demanding than rule-based interfaces.

Gen-SynDi has several potential limitations that frame our next research steps. Methodological transparency awaits full confirmation through independent replications; likewise,

our expert panel evaluation offers only short-term evidence, with no head-to-head trials against conventional CPX or rule-based chatbots. The current case library is confined to fatigue presentations and text-only history-taking, so key differentiators such as laboratory or immunologic markers—crucial for teasing apart fibromyalgia from ME/CFS—remain outside the simulation, and cross-cultural validity has yet to be tested beyond Korean TAM nosology. Finally, while safety prompts and audit logs curb hallucinations and privacy risks, broader questions about cognitive autonomy and over-reliance on AI still warrant scrutiny. To address these limitations, a human-in-the-loop design strategy could be adopted, whereby instructors review and adjust AI-generated evaluation feedback before it is provided to learners. Alternatively, a participatory design approach could be employed, involving both instructors and learners in the development process to ensure that the AI model is co-designed in a way that preserves learners' cognitive autonomy. Even with these constraints, Gen-SynDi is, to our knowledge, the first openly accessible tutor that pairs disease diagnosis with TAM pattern identification and automatically provides qualitative and quantitative feedback, both questioning coverage and reasoning quality, providing an affordable bridge between phenotype-based teaching and competency-based assessment. Future work will address this by integrating multimodal inputs—laboratory panels, imaging summaries, and wearable-sensor data—to create even more realistic and diagnostically challenging scenarios.

5. Conclusions

Gen-SynDi shows that expert-guided LLM prompts, grounded in curated phenotypes, can generate 28 fatigue-focused cases that blend biomedical diagnosis with TAM syndrome differentiation. Expert reviewers rated validity and educational effectiveness high, confirming its capacity to train students' clinical reasoning, while the built-in analytics module automatically evaluates question coverage and diagnostic logic. Moderate usability scores signal the need for clearer navigation and stronger error-prevention, but the system already offers a cost-efficient, privacy-respecting resource that supplements traditional methods such as standardized-patient actors and live tutorials.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app15094862/s1, Table S1: History Taking Dataset of Fatigue-Related Syndrome; Table S2: History Taking Dataset of Fatigue-Related Disease; Table S3: Example Responses for Clinical Reasoning in Fatigue-Related Syndromes and Diseases.

Author Contributions: Conceptualization, B.-W.L.; data curation, B.-W.L.; formal analysis, Y.H.; investigation, W.-Y.L., S.-Y.H., J.-H.K. and S.L.; resources, W.-Y.L. and S.-Y.H.; software, J.-H.K.; supervision, S.L.; visualization, Y.H. and B.-W.L.; writing—original draft, W.-Y.L., S.-Y.H. and Y.H.; writing—review and editing, J.-H.K. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by Wonkwang University in 2025.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets supporting the conclusions of this article are included within the article.

Acknowledgments: The first author (W.Y. Lee) would like to thank the Ph.D. programme of Woosuk University for providing the thesis completion through this work. The authors also gratefully acknowledge the anonymous reviewers whose constructive comments substantially strengthened the manuscript. The grammar and spelling of this work were refined with the help of AI.

Conflicts of Interest: Author Byung-Wook Lee was employed by Dongje Medical Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Jin, H.-J.; Lee, J.; Jang, E. Research trends of pattern identification of korean medicine using the network analysis. *J. Korea Contents Assoc.* **2014**, *14*, 1037–1046. [CrossRef]
2. Kim, J.W.; Kim, H.J.; Jang, E.S.; Jung, H.J.; Hwang, M.W.; Nam, D.H. Survey on pattern identification and treatment of chronic fatigue in Korea medicine. *J. Physiol. Pathol. Korean Med.* **2018**, *32*, 126–133. [CrossRef]
3. Birch, S.; Alraek, T.; Bovey, M.; Lee, M.S.; Lee, J.A.; Zaslawski, C.; Robinson, N.; Kim, T.-H.; Bian, Z.-x. Overview on pattern identification–history, nature and strategies for treating patients: A narrative review. *Eur. J. Integr. Med.* **2020**, *35*, 101101. [CrossRef]
4. Ferreira, A.S.; Lopes, A.J. Chinese medicine pattern differentiation and its implications for clinical practice. *Chin. J. Integr. Med.* **2011**, *17*, 818–823. [CrossRef]
5. Li, X.J.; Qiu, W.Q.; Da, X.L.; Hou, Y.J.; Ma, Q.Y.; Wang, T.Y.; Zhou, X.M.; Song, M.; Bian, Q.L.; Chen, J.X. A combination of depression and liver Qi stagnation and spleen deficiency syndrome using a rat model. *Anat. Rec.* **2020**, *303*, 2154–2167. [CrossRef]
6. Huang, M.; Yang, H.; Guo, J.; Fu, X.; Chen, W.; Li, B.; Zhou, S.; Xia, T.; Peng, S.; Wen, L. Faculty standardized patients versus traditional teaching method to improve clinical competence among traditional Chinese medicine students: A prospective randomized controlled trial. *BMC Med. Educ.* **2024**, *24*, 793. [CrossRef]
7. Chen, Y.-L.; Hou, M.C.; Lin, S.-C.; Tung, Y.-J. Educational efficacy of objective structured clinical examination on clinical training of traditional Chinese medicine–a qualitative study. *Complement. Ther. Clin. Pract.* **2015**, *21*, 147–153. [CrossRef] [PubMed]
8. 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]
9. Baglivo, F.; De Angelis, L.; Casigliani, V.; Arzilli, G.; Privitera, G.P.; Rizzo, C. Exploring the possible use of AI chatbots in public health education: Feasibility study. *JMIR Med. Educ.* **2023**, *9*, e51421. [CrossRef]
10. Tian, S.; Yang, W.; Le Grange, J.M.; Wang, P.; Huang, W.; Ye, Z. Smart healthcare: Making medical care more intelligent. *Glob. Health J.* **2019**, *3*, 62–65. [CrossRef]
11. Bourke, J.; Roper, S.; Love, J.H. Innovation in legal services: The practices that influence ideation and codification activities. *J. Bus. Res.* **2020**, *109*, 132–147. [CrossRef]
12. Brin, D.; Sorin, V.; Vaid, A.; Soroush, A.; Glicksberg, B.S.; Charney, A.W.; Nadkarni, G.; Klang, E. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci. Rep.* **2023**, *13*, 16492. [CrossRef]
13. Garabet, R.; Mackey, B.P.; James Cross, M.; Weingarten, M. ChatGPT-4 performance on USMLE Step 1 questions and its implications for medical education: A comparative study across systems and disciplines. *Med. Sci. Educ.* **2024**, *34*, 145–152. [CrossRef]
14. Giray, L. Prompt engineering with ChatGPT: A guide for academic writers. *Ann. Biomed. Eng.* **2023**, *51*, 2629–2633. [CrossRef] [PubMed]
15. Ekin, S. Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices. *TechRxiv* **2023**. [CrossRef]
16. Jang, D.; Yun, T.-R.; Lee, C.-Y.; Kwon, Y.-K.; Kim, C.-E. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. *PLOS Digit. Health* **2023**, *2*, e0000416. [CrossRef]
17. Development, National Institute for Korean Medicine. Clinical Performance Examination Module. Available online: https://nikom.or.kr/nckm/html.do?menu_idx=225 (accessed on 12 May 2024).
18. Cho, J.-H.; Yoo, S.-R.; Cho, J.-K.; Son, C.-G. Analytic Study for Syndrome-differentiation and Sasang-constitution in 72 Adults with Chronic Fatigue. *Korean J. Orient.Int. Med.* **2007**, *28*, 791–796.
19. Kim, J.H.; Ku, B.C.; Kim, J.E.; Kim, Y.S.; Kim, K.H. Study on Reliability and Validity of the 'Qi Blood Yin Yang Deficiency Questionnaire'. *Korean J. Orient. Physiol. Pathol.* **2014**, *28*, 346–354. [CrossRef]
20. Cook, D.A. Creating virtual patients using large language models: Scalable, global, and low cost. *Med. Teach.* **2025**, *47*, 40–42. [CrossRef]
21. Richey, R.C.; Klein, J.D. *Design and Development Research: Methods, Strategies, and Issues*; Routledge: England, UK, 2014.
22. Chang, C.-C.; Yan, C.-F.; Tseng, J.-S. Perceived convenience in an extended technology acceptance model: Mobile technology and English learning for college students. *Australas. J. Educ. Technol.* **2012**, *28*, 809–826. [CrossRef]
23. 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]
24. Kim, J.; Lee, H.-Y.; Kim, J.-H.; Kim, C.-E.; 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]

25. Lee, W.-Y.; Han, S.Y.; Lee, S. Exploring the feasibility of developing an education tool for pattern identification using a large language model: Focusing on the case of a simulated patient with fatigue symptom and dual deficiency of the heart-spleen pattern. *Herb. Formula Sci.* **2024**, *32*, 1–9.
26. Wang, L.; Chen, R.; Li, L. Knowledge-guided prompt learning for few-shot text classification. *Electronics* **2023**, *12*, 1486. [CrossRef]
27. Alsafari, B.; Atwell, E.; Walker, A.; Callaghan, M. Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants. *Nat. Lang. Process. J.* **2024**, *8*, 100101. [CrossRef]
28. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
29. Cui, C.; Zhou, Y.; Yang, X.; Wu, S.; Zhang, L.; Zou, J.; Yao, H. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv* **2023**, arXiv:2311.03287.
30. Di Plinio, S. Panta Rh-AI: Assessing multifaceted AI threats on human agency and identity. *Soc. Sci. Humanit. Open* **2025**, *11*, 101434. [CrossRef]
31. Oyekunle, D.O.T.; Boohene, D.; Preston, D. Ethical Considerations in AI-Powered Work Environments: A Literature Review and Theoretical Framework for Ensuring Human Dignity and Fairness. Ethical Considerations in AI-Powered Work Environments: A Literature Review and Theoretical Framework for Ensuring Human Dignity and Fairness. *Int. J. Sci. Res. Manag. (IJSRM)* **2024**, *12*, 6166–6178.

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.