information

MDPI

Article

Using Large Language Models to Simulate History Taking: Implications for Symptom-Based Medical Education

Cheong Yoon Huh [1,2,†], Jongwon Lee [1,2,†], Gibaeg Kim [2], Yerin Jang [1,2], Hye-seung Ko [2,3], Min Jung Suh [2,3], Sumin Hwang [2,3], Ho Jin Son [2,4], Junha Song [2,4], Soo-Jeong Kim [5], Kwang Joon Kim [2,6], Sung Il Kim [2], Chang Oh Kim [6] and Yeo Gyeong Ko [2,7,*]

1 College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea; celinehuh@catholic.ac.kr (C.Y.H.); jah06190@catholic.ac.kr (J.L.); yerin7726@aitrics.com (Y.J.)
2 AITRICS, Inc., 218 Teheran-ro, Gangnam-gu, Seoul 06221, Republic of Korea; gb.kim@aitrics.com (G.K.); 2160006@ewhain.net (H.-s.K.); mngyy@aitrics.com (M.J.S.); sumed@ewhain.net (S.H.); hank221@khu.ac.kr (H.J.S.); jsong953@khu.ac.kr (J.S.); preppie@yuhs.ac (K.J.K.); ok2jinsung@aitrics.com (S.I.K.)
3 College of Medicine, Ewha Womans University, Seoul 07804, Republic of Korea
4 College of Medicine, Kyunghee University, Seoul 02447, Republic of Korea
5 Division of Hemato-Oncology, Department of Internal Medicine, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin 16995, Republic of Korea; alvin97@yuhs.ac
6 Division of Geriatrics, Department of Internal Medicine, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea; cokim@yuhs.ac
7 Department of Internal Medicine, Severance Hospital, Yonsei University College of Medicine, Seoul 03722, Republic of Korea
* Correspondence: choiko9120@yuhs.ac; Tel.: +82-2-569-5507
† These authors contributed equally to this work.

Abstract

Medical education often emphasizes theoretical knowledge, limiting students' opportunities to practice history taking, a structured interview that elicits relevant patient information before clinical decision making. Large language models (LLMs) offer novel solutions by generating simulated patient interviews. This study evaluated the educational potential of LLM-generated history-taking dialogues, focusing on clinical validity and diagnostic diversity. Chest pain was chosen as a representative case given its frequent presentation and importance for differential diagnosis. A fine-tuned Gemma-3-27B, specialized for medical interviews, was compared with GPT-4o-mini, a freely accessible LLM, in generating multi-branching history-taking dialogues, with Claude-3.5 Sonnet inferring diagnoses from these dialogues. The dialogues were assessed using a Chest Pain Checklist (CPC) and entropy-based metrics. Gemma-3-27B outperformed GPT-4o-mini, generating significantly more high-quality dialogues (90.7% vs. 76.5%). Gemma-3-27B produced diverse and focused diagnoses, whereas GPT-4o-mini generated broader but less specific patterns. For demographic information, such as age and sex, Gemma-3-27B showed significant shifts in dialogue patterns and diagnoses aligned with real-world epidemiological trends. These findings suggest that LLMs, particularly those fine-tuned for medical tasks, are promising educational tools for generating diverse, clinically valid interview scenarios that enhance clinical reasoning in history taking.

Keywords: large language models; history taking; medical interview; medical education; artificial intelligence in medicine

Academic Editors: Silvia Ceccacci, Catia Giaconi and Noemi Del Bianco

Received: 31 May 2025
Revised: 15 July 2025
Accepted: 28 July 2025
Published: 31 July 2025

Citation: Huh, C.Y.; Lee, J.; Kim, G.; Jang, Y.; Ko, H.-s.; Suh, M.J.; Hwang, S.; Son, H.J.; Song, J.; Kim, S.-J.; et al. Using Large Language Models to Simulate History Taking: Implications for Symptom-Based Medical Education. *Information* **2025**, *16*, 653. https://doi.org/10.3390/info16080653

1. Introduction

Medical education aims to help students effectively navigate real-world clinical environments. In line with this objective, educational strategies have shifted from faculty-led instruction to student-centered learning approaches [1]. However, several challenges remain to be resolved. Traditional medical curricula continue to rely on unidirectional teaching methods, such as disease-specific lectures and apprenticeship-based training [2,3]. Thus, students often rely primarily on passive observation of physicians when practicing history taking—the process of interviewing a patient to collect information about symptoms, medical history, and relevant personal context—and have limited opportunities to receive explicit feedback on their performance [4,5]. Additionally, students lack sufficient exposure to direct patient encounters, which further hinders their ability to integrate theoretical knowledge with real-world clinical practice, particularly in symptom-based clinical decision-making [6].

To address these challenges, recent studies have explored the use of large language models (LLMs) in medical education. With their advanced natural language processing capabilities, LLMs have been employed to support training in clinical history-taking processes. Holderried et al. and Li et al. proposed artificial intelligence (AI)-based educational frameworks in which LLMs act as virtual patients, responding to students' questions and providing feedback [7,8]. However, most previous studies have positioned LLMs primarily as passive responders or evaluators in student-led history taking, thereby overlooking their potential to function as autonomous questioners.

Therefore, this study investigates whether a fine-tuned LLM can generate clinically valid and educationally valuable history-taking dialogues. Specifically, we evaluate Gemma-3-27B's potential as a simulated interviewing partner for medical students practicing history-taking skills. While previous studies have explored LLMs as virtual patients, none have systematically evaluated a fine-tuned model's ability to generate structured interview dialogues or assessed their educational utility using objective clinical criteria. The ultimate aim is to support bi-directional, learner-centered education and facilitate the development of clinical reasoning skills through AI-generated clinical scenarios.

Importantly, this study was co-developed and reviewed by key stakeholders in medical education—including medical students, residents, and faculty members—to ensure relevance to practical teaching needs. We also compared its performance with GPT-4o-mini, a widely available general-purpose LLM. To our knowledge, this is among the first studies to illustrate how fine-tuned LLMs can contribute to medical education through stakeholder-informed design, while offering insights into their potential to improve accessibility, interactivity, and inclusivity in health professions training.

2. Materials and Methods

2.1. Dataset Construction and Assessment

To systematically evaluate history-taking performance of LLMs for educational purposes, this study focuses on an essential initial step: evaluating whether the model can generate evidence-based history-taking dialogues that encompass a broad range of clinically relevant conditions. We employed a three-model comparative framework (Figure 1), which comprises two LLMs as interviewing models that generate sequential clinical questions to collect patient history (fine-tuned Gemma-3-27B and GPT-4o-mini) and a third LLM that determines final diagnoses on the basis of interview transcripts (Claude-3.5 Sonnet).

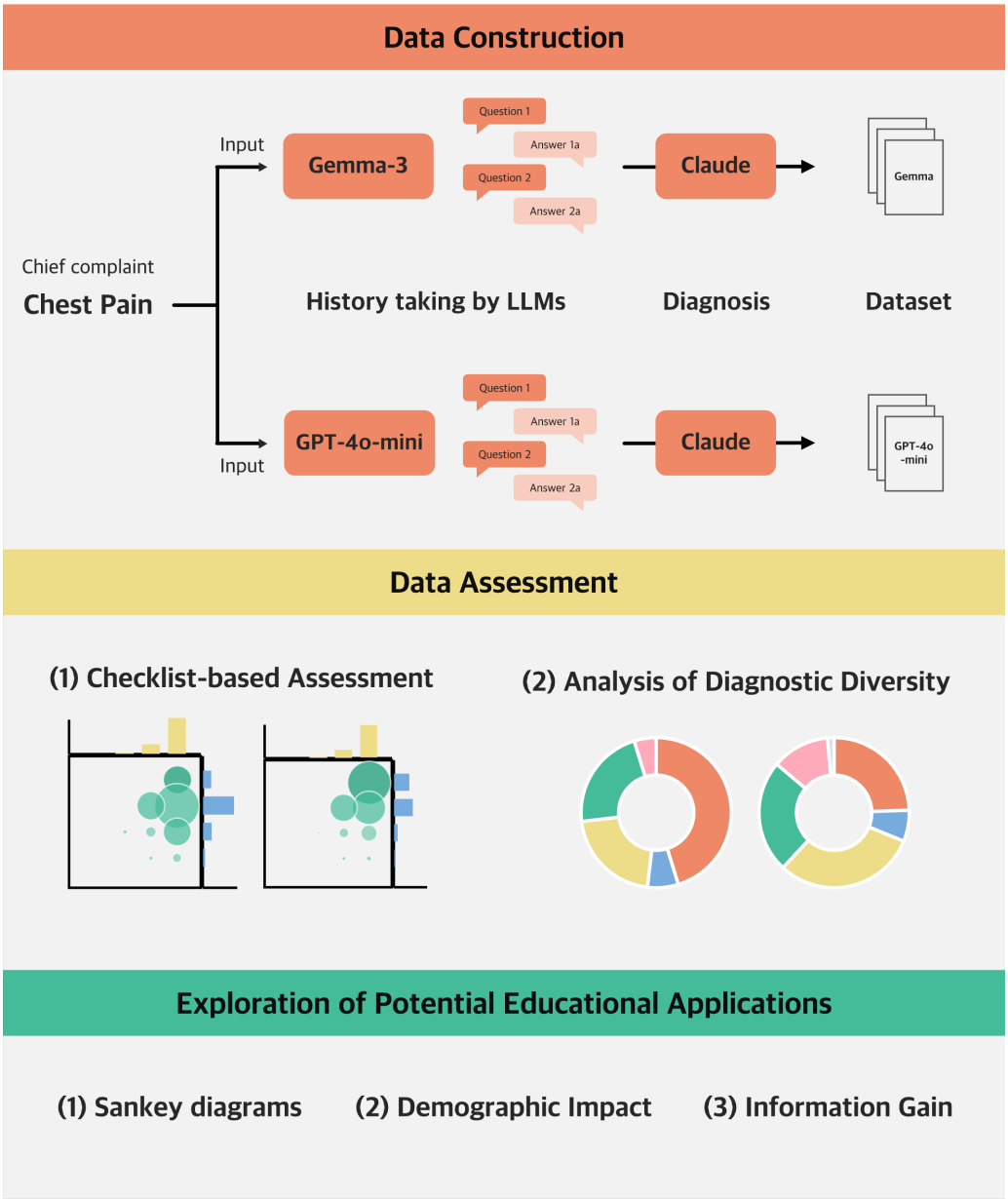


Figure 1. Overview of study design.

The two interviewing models were selected to represent contrasting approaches in contemporary LLM development and accessibility. Gemma-3-27B, fine-tuned on a synthetic dataset incorporating clinical protocols and medical reasoning, represents an LLM specifically optimized for history-taking tasks. Conversely, GPT-4o-mini was chosen as a general-purpose baseline LLM, given its free public availability and broad accessibility. This reflects a realistic option for students to use for academic purposes without technical or financial constraints, ensuring the practical relevance of our comparison.

A third model, Claude-3.5 Sonnet, was employed to infer diagnoses from the interviews. Assigning the diagnostic task to a third model, rather than to the interviewing models themselves, was essential to ensure a separation of history-taking performance from diagnostic performance and avoid potential confounding effects that could arise if a single LLM performed both tasks. This framework design enabled a controlled and focused comparison of the two models' interviewing performance, providing a standardized method for analyzing the diagnostic implications of the information collected by each

interviewing model. Claude-3.5 Sonnet was chosen as the diagnostic agent because of its established medical reasoning capabilities at the time of the study.

Both interviewing models received identical instructions to function as physicians and generated sequential clinical questions based on patient responses. Each model was provided with an identical initial scenario in which the patient presented with chest pain. Chest pain was selected as the chief complaint owing to its high clinical importance. It is a leading cause of emergency department visits [9], and covers a broad diagnostic spectrum, ranging from musculoskeletal disorders to life-threatening acute coronary syndromes. This breadth makes chest pain ideal for evaluating model performance across multiple medical domains while maintaining relevance to early clinical training. The models were instructed to begin with a standardized initial question on pain quality, which is a key element of chest pain assessment.

- Question: What does your chest pain feel like?
- Response options: Squeezing or tightening/sharp or stabbing/pressure-like and heavy/burning or aching/tearing or ripping.

To comprehensively evaluate the models' interviewing capabilities, we generated dialogue trees by exhaustively enumerating every possible conversational path that could occur when the patient sequentially selected each response option offered by the model. Each path began with one of the five preset answers to the initial question. For each answer, the models generated a relevant follow-up question with selectable patient response options. This branching expansion was iterated to a depth of five turns, creating a complete set of unique dialogue sequences for analysis. The total number of generated paths differed between the models due to inherent variations in their branching logic and the number of response options created, as detailed in the Results section.

To ensure data quality and meaningful evaluation, we implemented a strict filtering protocol. An entire dialogue path was excluded if any of its constituent patient responses was a non-informative option (e.g., "not sure"), as such responses disrupt the logical flow of the interview and hinder assessment of the models' interviewing ability. Filtering was conducted independently by medically trained researchers, with any disagreements resolved by consensus to ensure objective evaluation standards.

Each complete dialogue path was independently submitted to Claude-3.5 Sonnet for diagnostic evaluation. The agent determined the most likely diagnosis solely on the basis of the collected clinical information, ensuring standardized diagnostic criteria across all dialogues. The diagnoses were categorized into seven medical domains: (1) cardiovascular diseases, (2) respiratory diseases, (3) gastrointestinal diseases, (4) musculoskeletal disorders, (5) psychiatric disorders, (6) gynecological and breast diseases, and (7) miscellaneous. The initial classification followed International Classification of Diseases, 11th Revision standards, followed by refinement through consensus among medically trained researchers with reference to standard medical textbooks. The resulting dialogues and diagnoses were systematically evaluated for clinical relevance and educational utility, as described in the following sections.

2.1.1. Checklist-Based Assessment of Medical Appropriateness

To evaluate whether the generated dialogues had sufficient medical appropriateness for potential use in medical education, a standardized Chest Pain Checklist (CPC) was developed. A checklist-based approach was selected to provide objective, reproducible assessment criteria grounded in established clinical practice, avoiding subjective expert review limitations noted in previous studies [10]. Researchers with medical expertise constructed the CPC on the basis of authoritative medical textbooks, e.g., Harrison's Principles of Internal Medicine, and established clinical guidelines [11–14]. The checklist

comprised core items that should be addressed when interviewing patients who present with chest pain.

Each question in the generated dialogue was systematically mapped to a corresponding CPC item using predefined criteria. If a question could not be mapped to any CPC item, it was categorized as "Others." This category was designed to capture instances of low-quality questions, including those that are clinically irrelevant, lacking informative value, or containing logical inconsistencies. Any discrepancies in item mapping were resolved by consensus among the researchers, including medical doctors.

The clinical relevance of each dialogue was evaluated using two metrics. Relevance was defined as the total number of questions in a dialogue that were successfully mapped to CPC items, excluding those labeled as "Others." This metric was intended to represent the medical relevance and overall quality of a dialogue. Since the first question in every dialogue was fixed and mapped to the "Quality" item, each dialogue had a minimum Relevance of 1. Variety was defined as the number of unique CPC items covered by the mapped questions in a dialogue. For example, if a dialogue consisted of 5 questions respectively mapped to CPC items "Quality," "Location," "Associated symptoms," "Associated symptoms," and "Others," the Relevance would be 4, and the Variety would be 3. The most ideal dialogue would achieve a maximum score of 5 points in both Relevance and Variety, with five questions covering five distinct core elements listed in the CPC (e.g., "Quality," "Location," "Associated symptoms," "Past medical history," and "Onset"), excluding the "Others."

2.1.2. Analysis of Diagnostic Diversity

Two complementary methods were used to evaluate the diagnostic diversity of each model: (1) frequency-based analysis of diagnostic domains, and (2) Shannon entropy-based quantification of diagnostic variability. Exposure to a diverse set of diseases can help medical students develop balanced diagnostic reasoning skills across multiple clinical domains. The distribution of the final diagnoses was analyzed at two hierarchical levels—individual diseases and diagnostic domains—based on the classification method described in Section 2.1. This dual-level approach enabled assessment of both specific diagnostic accuracy and broader categorical distribution patterns. Diagnostic diversity was quantified using Shannon entropy, a concept that captures the degree of uncertainty or variability in a given distribution [15]. Entropy values were calculated at both individual disease and domain levels and normalized to a value between 0 and 1 according to the number of diseases. Entropy values closer to 1 reflect a broader, more uniform distribution of diagnoses, whereas values near 0 indicate more concentrated and deterministic diagnostic outcomes.

2.2. Exploration of Potential Educational Applications

We conducted three complementary analyses to investigate the potential educational utility of the simulated dialogues: (1) visualization of the dialogue structure using a Sankey diagram; (2) analysis of age-specific and sex-specific dialogue patterns; and (3) evaluation of information gain (IG).

2.2.1. Visualization of Diagnostic Pathways

The Sankey diagram was selected as an effective visualization method that captures the branching nature and sequential flow of a dataset, while simultaneously reflecting the frequency of each branch [16]. To facilitate a more intuitive understanding of the question flow, we employed a Sankey diagram to visualize the branching structure of our dialogue dataset generated by the fine-tuned Gemma-3-27B. Visualization was performed using the Plotly library in Python (version 3.11.12).

2.2.2. Analysis of Age-Specific and Sex-Specific Dialogue Patterns

We investigated whether the interviewing model was capable of considering the demographic characteristics of a patient and adjusting the interview flow accordingly. Dialogue data were generated using the fine-tuned Gemma-3-27B and processed as previously described in Section 2.1, but with two key modifications: inclusion of patient demographic information as input to the model and reduction in the number of questions per dialogue to four.

Age and sex were the primary demographic variables of interest. Hsia et al. reported variations in the diagnosis of chest pain with age [17]. On the basis of their findings, we provided age information to the model as two distinct ranges: 18–44 years and 65 years or older. The intermediate range of 45–64 years was considered a transitional group that may present overlapping characteristics of younger and older populations and was excluded from data generation [18]. Shannon entropy was calculated to examine the effects of demographic variables on diagnostic variability.

2.2.3. Identifying High-Impact Questions for Diagnosis

To evaluate the educational value of the history-taking process, we analyzed how each question and its corresponding responses contributed to narrowing down the differential diagnoses. This analysis could provide students with insights into which clinical information is most relevant for evaluating a given chief complaint. Specifically, we used IG analysis, a measure derived from Shannon entropy, which quantifies the reduction in uncertainty when the data are divided into subsets. IG is a concept commonly used in medical diagnostics to assess how effectively a question reduces diagnostic ambiguity [15,19]. In our study, IG was calculated as the difference between the entropy before and after each question in the dialogue. For example, if a question significantly reduces the number of possible diagnoses, it results in a large decrease in uncertainty and thus a high IG value. The diagnostic uncertainty was measured at the disease category level to focus on clinically relevant differentiation.

2.3. Statistical Analysis

The Mann–Whitney U was performed to assess whether the mean Relevance and Variety of the two interviewing models were significantly different. Chi-square tests of independence were used to evaluate the differences in the distribution of CPC elements and final diagnostic domains between the two interviewing models, and to examine whether these distributions varied by patient age or sex. The universal first question of each dialogue was excluded from the assessment of the CPC item distribution to avoid bias. When the assumptions of the chi-square test were not met, the Fisher exact test was used. Statistical significance was defined as $p < 0.05$, and all statistical analyses were performed using the scipy.stats module in Python (version 3.11.12).

3. Results

The datasets generated with chest pain as the chief complaint using the fine-tuned Gemma-3-27B and GPT-4o-mini consisted of 4569 and 2481 dialogues, respectively. After excluding dialogues that included the selection of non-informative options (e.g., “not sure”), the numbers of remaining dialogues were 2020 and 2443, respectively. The history-taking dialogues from each model resulted in 111 and 119 unique diagnoses, respectively, which were categorized into seven domains. Figure 2 presents a sample dialogue from each model.

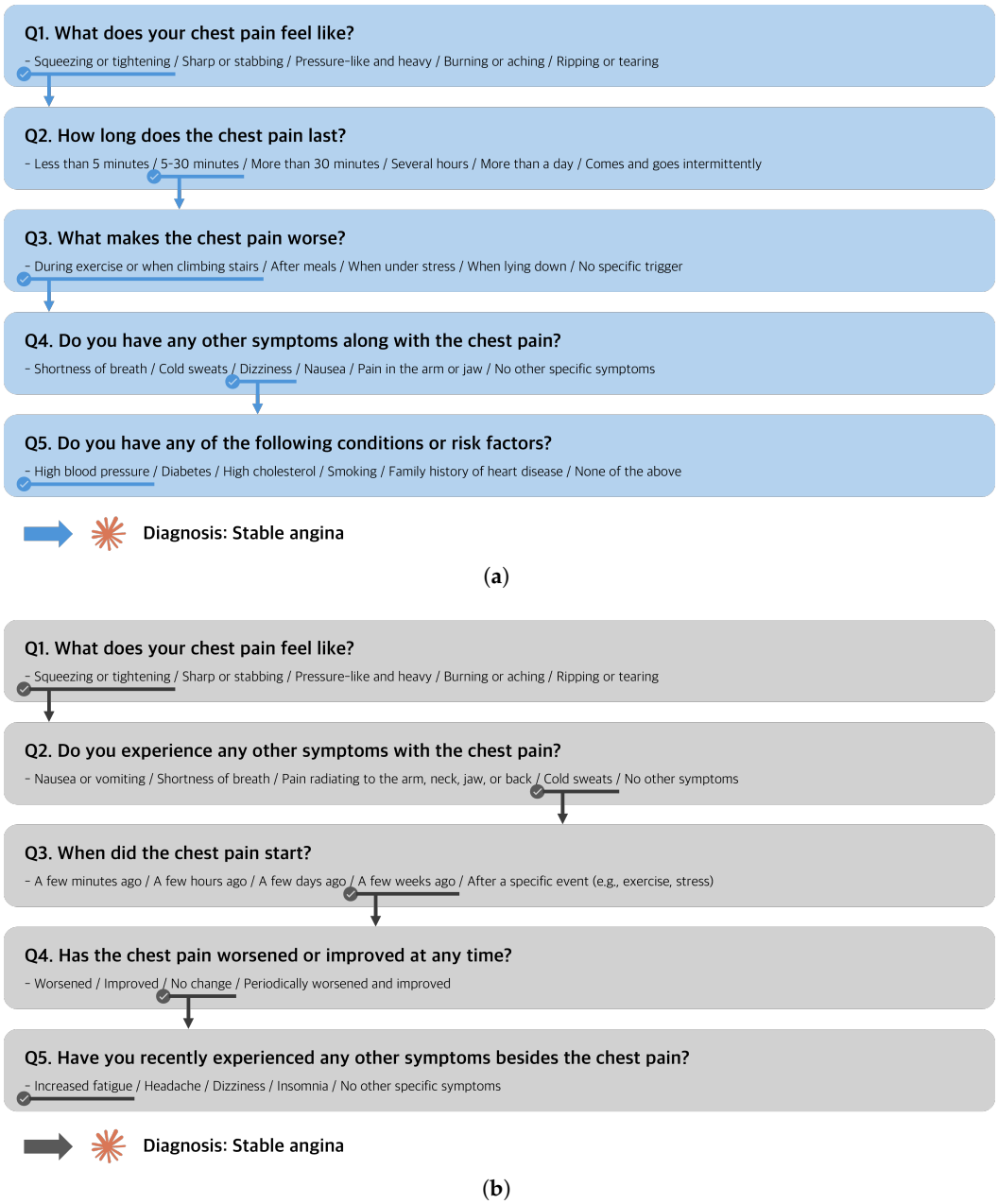


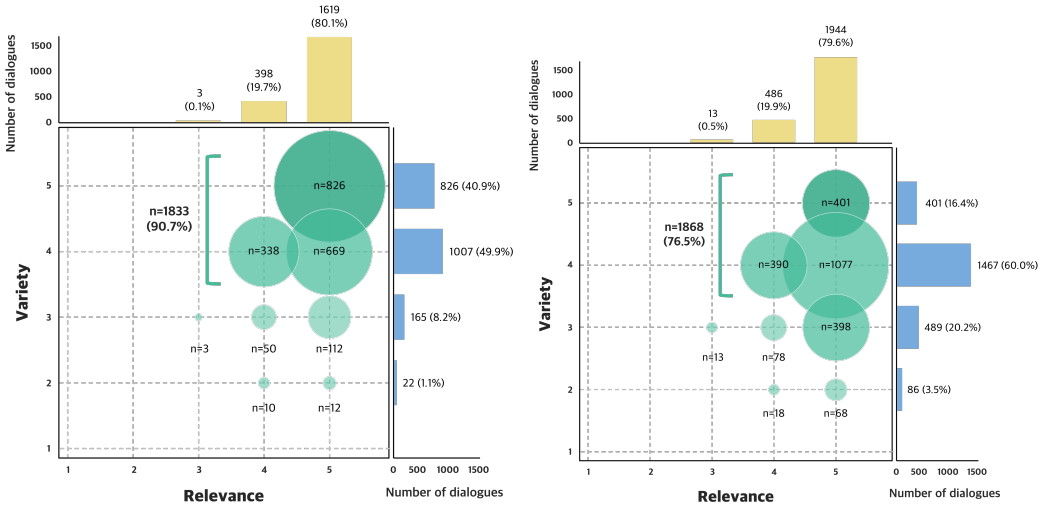
Figure 2. Sample five-turn dialogues for history taking generated by (**a**) the fine-tuned Gemma-3-27B and (**b**) GPT-4o-mini.

3.1. Evaluation of LLM-Generated History-Taking Dialogues

The CPC was constructed as a list of 10 items critical to the history of a patient with chest pain (Table 1). Using the CPC as a reference, the Relevance and Variety of each dialogue generated by the fine-tuned Gemma-3-27B and GPT-4o-mini were calculated (Figure 3). The average Relevances were 4.80 and 4.79 for the fine-tuned Gemma-3-27B and GPT-4o-mini, respectively. The average Variety of the fine-tuned Gemma-3-27B was 4.31, which was significantly higher than that of the GPT-4o-mini (3.89, $p < 0.05$). The distribution of the Relevance, Variety, and a combination of the two metrics differed significantly between the two datasets ($p < 0.05$).

Table 1. Chest Pain Checklist (CPC), consisting of 10 core history-taking items adapted from established clinical guidelines and medical textbooks [11–14].

No.	Item	Description
1	Quality	Character or nature of the chest pain (e.g., sharp, dull)
2	Location	Site of pain on the chest
3	Radiation	Site of radiated pain other than the chest (e.g., arm, jaw)
4	Onset	Timing of chest pain onset
5	Duration	How long a chest pain episode lasts
6	Aggravating or relieving factors	Factors that worsen or soothe the pain
7	Associated symptoms	Other symptoms present (e.g., shortness of breath, nausea)
8	Past medical history	Relevant pre-existing medical or surgical conditions
9	Patient's activity at onset	What the patient was doing at the time of chest pain onset (e.g., exertion, fall on the chest)
10	Severity	Intensity of the pain



(**a**) (**b**)

Figure 3. Relevance and Variety of history-taking dialogues generated by two large language models: (**a**) the fine-tuned Gemma-3-27B and (**b**) GPT-4o-mini.

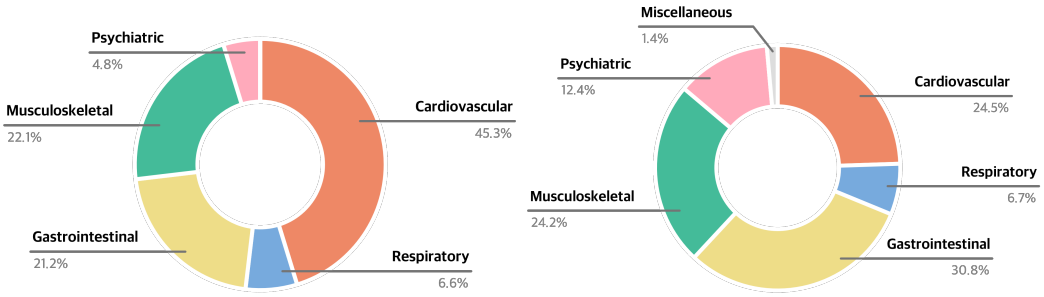
A total of 1833 dialogues (90.7%) generated by the fine-tuned Gemma-3-27B achieved a Relevance and Variety of 4 or higher. The most ideal dialogues, with the maximum Relevance and Variety of 5, were the most frequent, comprising 826 cases (40.9%). Dialogues with a Relevance of 5 and a Variety of 4 were the second most common, accounting for 669 cases (33.1%). Dialogues with a Relevance and Variety of 4 ranked third, with 338 cases (16.7%).

In the GPT-4o-mini dataset, 1868 dialogues (76.5%) achieved a Relevance and Variety ≥ 4, representing a significantly smaller proportion than that of the fine-tuned

Gemma-3-27B ($p < 0.05$). Among these, a significantly smaller proportion of 401 dialogues (16.4%) were assigned the maximum Relevance and Variety of 5 ($p < 0.05$). Dialogues with a Relevance of 4 and a Variety of 5 comprised 1077 cases (44.1%), and those with a Relevance and Variety of 4 accounted for 390 cases (16.9%).

Among the diagnoses following the fine-tuned Gemma-3-27B dialogue, the five most frequently identified conditions were costochondritis (20.0%), gastroesophageal reflux disease (17.2%), stable angina (8.1%), angina pectoris (6.9%), and aortic dissection (6.4%). Conversely, the GPT-4o-mini showed a narrower diagnostic distribution with gastroesophageal reflux disease appearing most frequently (29.3%), followed by costochondritis (22.3%), anxiety disorder (8.6%), angina pectoris (6.6%), and aortic dissection (2.8%).

The distribution of diagnoses across disease domains is illustrated in Figure 4. The most frequently predicted category based on the fine-tuned Gemma-3-27B dialogues was cardiovascular diseases (45.2%), followed by musculoskeletal disorders (22.1%), gastrointestinal disorders (21.2%), respiratory diseases (6.6%), and psychiatric disorders (4.8%). In contrast, dialogues from the GPT-4o-mini were more frequently associated with gastrointestinal diagnoses (31.0%), while showing lower proportions for the cardiovascular (24.5%), musculoskeletal (24.2%), psychiatric (12.4%), and respiratory (6.7%) categories. The overall distribution of the diagnostic categories differed significantly between the two datasets ($p < 0.05$).



(**a**) (**b**)

Figure 4. The distribution of disease categories of (**a**) the fine-tuned Gemma-3-27B and (**b**) GPT-4o-mini.

The normalized Shannon entropy values were 0.658 and 0.564 at the disease level for the fine-tuned Gemma-3-27B and GPT-4o-mini, respectively, indicating that the diagnoses derived from the fine-tuned Gemma-3-27B dialogues were distributed across a broader set of conditions. Conversely, the entropy value at the domain level was higher for the GPT-4o-mini (0.979) than for the fine-tuned Gemma-3-27B (0.837), indicating that the predictions of the former model were more diffusely spread across the major diagnostic domains.

3.2. Exploration of Potential Educational Applications

3.2.1. Sankey Diagram

A Sankey diagram was employed to visualize the flow of the history-taking dialogues generated by the fine-tuned Gemma-3-27B (Figure 5). Following "Quality," which was used as the fixed first question, the model chose to inquire about "Duration," "Onset," "Location," and "Aggravating or relieving factors" as the second questions. "Past medical history" appeared only in the last two columns, while "Severity" appeared only in the very last column. The diagram is provided as an HTML file, which enables users to obtain additional information regarding each link and node.

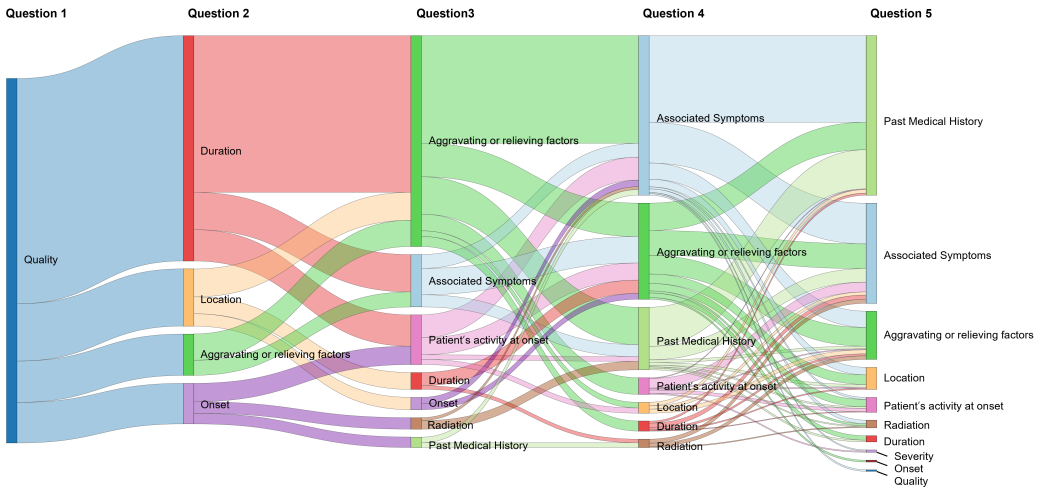


Figure 5. A Sankey diagram visualizing the flow of dialogues generated by the fine-tuned Gemma-3-27B.

3.2.2. Age-Specific and Sex-Specific Dialogues

When provided with age or sex information as a patient profile, the fine-tuned Gemma-3-27B generated 985, 769, 835, and 795 dialogues for younger adults (18–44 years), older adults (≥65 years), males, and females, respectively. After data cleaning, the numbers of dialogues were reduced to 567, 422, 452, and 475, respectively. The CPC item distribution differed significantly by age and sex (both, $p < 0.05$).

In the age-based analysis (Figure 6a), musculoskeletal disorders (38.1%), gastrointestinal disorders (29.3%), and psychiatric disorders (9.0%) were more frequently diagnosed in dialogues with younger age inputs, whereas cardiovascular diseases were uncommon (18.5%). However, cardiovascular diseases accounted for most diagnoses (57.8%) in dialogues with older age inputs, whereas musculoskeletal (14.2%) and psychiatric disorders (0.9%) were less prevalent. The Fisher exact test confirmed that these age-related differences in the diagnostic category distribution were statistically significant ($p < 0.05$).

In the sex-specific setting (Figure 6b), cardiovascular diseases were more frequently diagnosed in males (37.6%), whereas musculoskeletal disorders were predominantly diagnosed in females (39.6%). Gastrointestinal disorders occurred at similar rates in both sexes (males, 29.0%; females, 31.6%). Additionally, 9 female dialogues (1.9%) were associated with breast-related diagnoses, including fibrocystic breast disease, fibroadenoma, and acute mastitis. These differences between the sexes were statistically significant ($p < 0.05$).

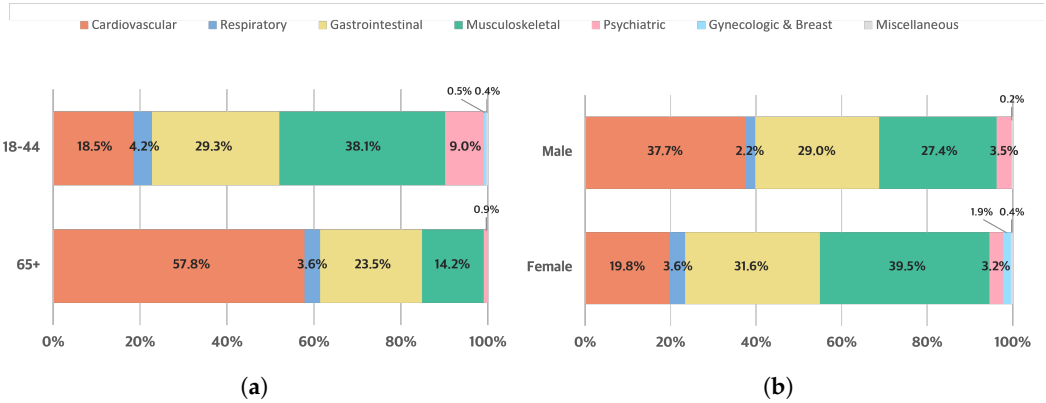


Figure 6. The proportion of disease categories with (**a**) the age information and (**b**) the sex information.

Regarding entropy analysis, disease-level entropy was lower in the 18–44-year-old dialogues (0.565) than in the 65 years and older dialogues (0.758), indicating a narrower spread of specific diagnoses in the younger dialogues. Conversely, at the disease category level, entropy was higher in younger dialogues (0.739) than in older dialogues (0.682). Among the dialogues with sex information, the male dialogues exhibited higher entropy at the disease level (0.664) and the category level (0.727) than their female counterparts (0.567 and 0.673, respectively), implying greater diagnostic dispersion in male dialogues.

3.2.3. IG Analysis

The IG of each CPC category associated with each question varied across the different stages of the dialogues generated by the fine-tuned Gemma-3-27B (Figure 7). Later-turn questions exhibited a higher IG than earlier turns. Among the second-turn questions, those related to duration showed the highest IG, followed by location, onset, and aggravating or relieving factors. From the third to the fifth turns, questions concerning aggravating or relieving factors and associated symptoms consistently showed high IG values.

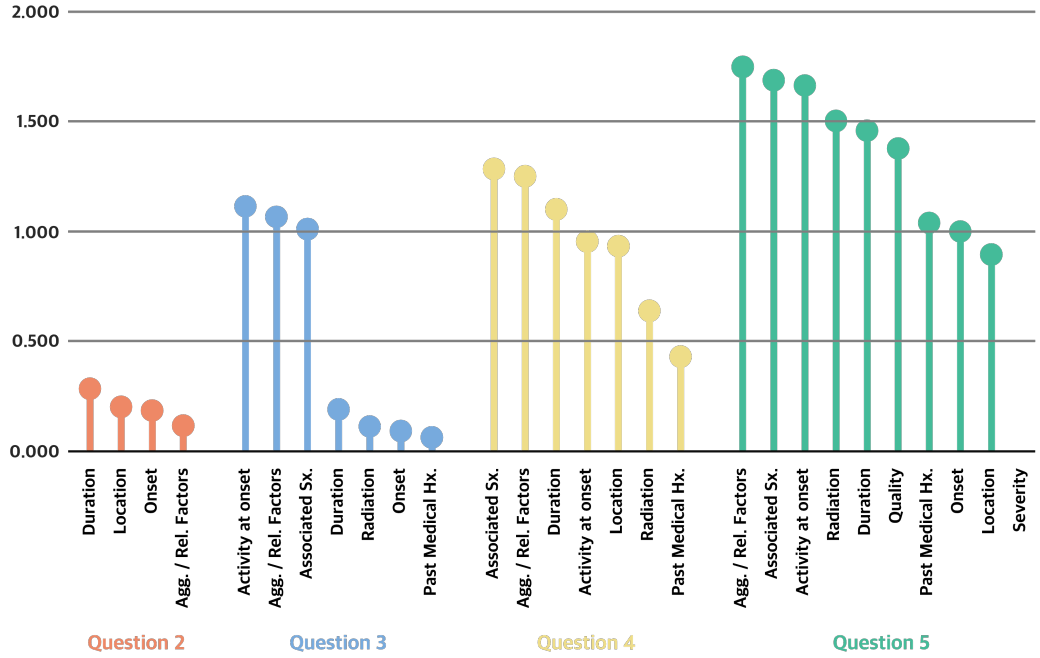


Figure 7. Information gain of each question category across different turns in the history-taking dialogues. Abbreviations: Agg./Rel., Aggravating or Relieving; Sx., Symptoms; Hx., History.

4. Discussion

This study explored the potential of LLMs as a novel tool in medical education, particularly in bridging the persistent gap between theoretical knowledge and its application in medical interviews. While recent efforts have explored the use of LLMs in medical education, particularly for training history-taking skills, most have focused on implementing LLMs as simulated patients that respond passively to students' questions [7,8,20,21]. A few studies have employed LLMs in the physician role, generating interview questions or full dialogues, but these were primarily designed to evaluate diagnostic performance rather than their educational value [22–24]. Moreover, prior studies assessing the appropriateness of LLM-generated dialogues have often relied on subjective expert reviews, raising concerns regarding bias and reproducibility [10].

In contrast, our study proposes an objective and standardized evaluation framework grounded in established medical guidelines to assess not only diagnostic outcomes but also the structure and content of dialogues. This allows for a more reliable exploration of how

LLM-generated interviews can be used as pedagogically meaningful tools in preclinical training. Moreover, this study involved a diverse group of researchers representing all levels of medical education from medical students to residents and faculty members, thereby encompassing the full continuum of medical education. Our approach comprises two main components: validating the performance of LLMs in history-taking tasks and proposing methods to leverage them as informative educational tools.

4.1. Can the Fine-Tuned Gemma-3-27B Perform Effective History Taking?

We evaluated the history-taking performance of a fine-tuned Gemma-3-27B and compared its performance with that of the GPT-4o-mini. The CPC was constructed as a standard list of essential elements that should be asked to a patient reporting chest pain. On the basis of the CPC, we assessed the quality of the interview dialogues generated using the two models. The fine-tuned Gemma-3-27B outperformed the GPT-4o-mini, producing a significantly higher proportion of high-quality dialogues and demonstrating significantly greater Variety. However, the GPT-4o-mini often fixated on a certain CPC element and repeatedly asked about it, even when the information was no longer diagnostically useful. For instance, after identifying "Dizziness or syncope" as an associated symptom, the model would focus excessively on details of this symptom, while neglecting the patient's chief complaint of chest pain, which is likely more critical for diagnosis. This tendency likely contributed to the significantly lower Variety and reduced proportion of ideal dialogues in the GPT-4o-mini.

Diversity and entropy analyses further support this distinction, showing opposite patterns depending on the level of aggregation. The fine-tuned Gemma-3-27B exhibited higher entropy at the individual disease level but lower entropy at the category level, indicating that the model was associated with various specific diagnoses within a focused range of categories. This pattern suggests that the fine-tuned Gemma-3-27B was effective in extracting detailed differential cues within a particular medical domain, whereas the GPT-4o-mini tended to pursue a broader but less targeted line of inquiry.

Taken together, these findings support the validity of employing LLMs to generate exemplary history-taking dialogues and underscore the greater potential of the fine-tuned Gemma-3-27B over the GPT-4o-mini for this purpose, given its superior performance in the clinical relevance and breadth of questions as well as its focused approach toward appropriate diagnoses. These results indicate that history-taking agents based on well-tuned LLMs can serve as reliable tools for medical learners.

4.2. How Can LLM-Generated Dialogues Support Medical Students in Learning History Taking?

We investigated strategies for applying LLM-generated history-taking dialogues in medical education. A Sankey diagram was used to visualize the dialogue dataset generated by the fine-tuned Gemma-3-27B, illustrating the sequential flow and frequency of each CPC item addressed in the dialogues. This visualization enabled learners to identify which CPC items were prioritized in the earlier stages of the interview and which were addressed later, thereby informing them of the relative priority of the questions. Thus, the Sankey diagram may serve as an educational tool to guide medical students regarding the appropriate sequence of questions to ask when encountering patients with chest pain.

Incorporating demographic features into clinical reasoning is a key challenge for medical learners. We demonstrated that a fine-tuned LLM successfully exhibited distinct history-taking patterns when provided with patient age or sex information, resulting in significant shifts in the final diagnoses. These changes are consistent with known epidemiological trends: cardiovascular conditions are more commonly diagnosed through dialogue with older patients, whereas musculoskeletal and psychiatric disorders prevail in

dialogue with younger patients [17]. Similarly, the predominance of cardiovascular diseases in male dialogues and musculoskeletal disorders in female dialogues further support this pattern [25]. In addition, the higher entropy at the disease level but lower entropy at the category level in dialogues for older adults suggests that diagnoses were concentrated within certain domains, despite including a variety of specific conditions. These findings suggest that LLMs may help medical students better understand how demographic context influences clinical reasoning.

Students may also benefit from learning how effective a question is in medical interviews. IG analysis evaluated the effectiveness of each question category in narrowing the potential diagnostic domains. By quantifying the IG values, this analysis helps students recognize which questions have the greatest impact and develop more effective questioning strategies. The observed tendency of later-turn questions to exhibit higher IG values than earlier turns emphasizes the cumulative nature of clinical reasoning, where successive questions build on prior responses to become increasingly focused and diagnostically decisive. Because IG values varied across the stages of questioning, comparisons were conducted within each stage to accurately reflect their specific contributions. If a more comprehensive understanding of high-impact questions throughout the history-taking process is desired, more advanced methods, e.g., calculating entropy based on stage-specific probability distributions, can be employed.

These approaches may be combined to enhance students' self-directed learning of history-taking skills. For example, an LLM-powered online education platform can be designed to provide a simulated history-taking dialogue based on the chief complaints and patient demographics selected by students. Such platforms may also visualize the overview structure of dialogues using Sankey diagrams and highlight high-impact questions to help students develop structured and clinically relevant interview strategies. This will support more effective training by bridging the gap between theoretical knowledge and its application in real-world patient encounters.

4.3. Limitations

While this study highlights the potential of LLM-powered educational tools, certain aspects require further exploration. First, the fine-tuned Gemma-3-27B, which demonstrated superior performance in generating history-taking dialogues, was trained using Korean data. Future studies could examine how its performance differs across diverse linguistic and cultural contexts and facilitate adaptation of the model for broader, global applications.

Second, this study did not directly evaluate the educational effect of LLM-generated history-taking dialogues. Rather, it focused on exploring the feasibility of employing LLMs to generate valid and diverse patient interviews, as a foundational step toward educational application. Given the limited prior research in this area, demonstrating this feasibility was considered a necessary first step. To address this limitation, a follow-up study involving medical students is currently underway to empirically assess the impact of LLM history-taking agents on learning outcomes.

Third, our evaluation relied on a single LLM, Claude-3.5 Sonnet, to infer diagnosis from interview transcripts. While this design enabled a controlled comparison of the interviewing models' performance, it may have introduced model-specific biases and limited generalizability. Future research could address this limitation by comparing diagnostic outputs from multiple LLMs or involving human experts to validate LLM-generated diagnoses.

5. Conclusions

This study demonstrates the validity of LLMs, particularly the fine-tuned Gemma-3-27B specifically optimized for medical interviewing tasks, as educational resources capable

of generating exemplary history-taking dialogues. By providing diverse yet structured clinical scenarios that align with epidemiological patterns, properly evaluated and fine-tuned LLMs may enhance the understanding of history taking and support students in considering broader differential diagnoses. We also proposed practical and scalable strategies for incorporating these tools into medical education. As one of the earliest studies to apply AI-generated patient-interview dialogues in preclinical education, this approach may complement traditional methods and support students in bridging the gap between theoretical knowledge and real-world clinical practice.

Author Contributions: Conceptualization, C.Y.H., J.L., and Y.G.K.; methodology, C.Y.H. and J.L.; software, G.K.; validation, C.Y.H., J.L., G.K., and Y.G.K.; formal analysis, C.Y.H. and J.L.; investigation, C.Y.H., J.L., and Y.J.; resources, G.K.; data curation, C.Y.H., G.K., Y.J., H.-s.K., M.J.S., S.H., H.J.S., and J.S.; writing—original draft preparation, C.Y.H., J.L., G.K., Y.J., H.-s.K., M.J.S., S.H., H.J.S., and J.S.; writing—review and editing, C.Y.H., J.L., G.K., S.I.K., K.J.K., and Y.G.K.; visualization, C.Y.H., J.L., S.H., and H.J.S.; supervision, S.-J.K., K.J.K., S.I.K., C.O.K., and Y.G.K.; project administration, K.J.K. and Y.G.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Authors Cheong Yoon Huh, Jongwon Lee, Gibaeg Kim, Yerin Jang, Hye-seung Ko, Min Jung Suh, Sumin Hwang, Hojin Son, Junha Song, Kwang Joon Kim, Sung Il Kim and Yeo Gyeong Ko were employed by the AITRICS, Inc. 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.

Abbreviations

The following abbreviations are used in this manuscript:

LLMs	Large language models
AI	Artificial intelligence
CPC	Chest Pain Checklist
IG	Information gain

References

1. Peterson, P.; Baker, E.; McGaw, B. *International Encyclopedia of Education*; Elsevier: Amsterdam, The Netherlands, 2009.
2. Bösner, S.; Pickert, J.; Stibane, T. Teaching differential diagnosis in primary care using an inverted classroom approach: Student satisfaction and gain in skills and knowledge. *BMC Med. Educ.* **2015**, *15*, 63. [CrossRef] [PubMed]
3. Kiesewetter, J.; Ebersbach, R.; Tsalas, N.; Holzer, M.; Schmidmaier, R.; Fischer, M.R. Knowledge is not enough to solve the problems–The role of diagnostic knowledge in clinical reasoning activities. *BMC Med. Educ.* **2016**, *16*, 303. [CrossRef] [PubMed]
4. Faustinella, F.; Jacobs, R.J. The decline of clinical skills: A challenge for medical schools. *Int. J. Med. Educ.* **2018**, *9*, 195–197. [CrossRef] [PubMed]
5. Schopper, H.; Rosenbaum, M.; Axelson, R. 'I wish someone watched me interview:' Medical student insight into observation and feedback as a method for teaching communication skills during the clinical years. *BMC Med. Educ.* **2016**, *16*, 286. [CrossRef] [PubMed]
6. Alrasheedi, A.A. Deficits in history taking skills among final year medical students in a family medicine course: A study from KSA. *J. Taibah Univ. Med. Sci.* **2018**, *13*, 415–421. [CrossRef] [PubMed]
7. Li, Y.; Zeng, C.; Zhong, J.; Zhang, R.; Zhang, M.; Zou, L. Leveraging large language model as simulated patients for clinical education. *arXiv* **2024**, arXiv:2404.13066. [CrossRef]

8. Holderried, F.; Stegemann-Philipps, C.; Herrmann-Werner, A.; Festl-Wietek, T.; Holderried, M.; Eickhoff, C.; Mahling, M. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Med. Educ.* **2024**, *10*, e59213. [CrossRef] [PubMed]
9. Cairns, C.; Kang, K. National Hospital Ambulatory Medical Care Survey: 2020 Emergency Department Summary Tables. 2022. Available online: https://stacks.cdc.gov/view/cdc/121911 (accessed on 15 July 2025).
10. Johri, S.; Jeong, J.; Tran, B.A.; Schlessinger, D.I.; Wongvibulsin, S.; Barnes, L.A.; Zhou, H.Y.; Cai, Z.R.; Van Allen, E.M.; Kim, D.; et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **2025**, *31*, 77–86. [CrossRef] [PubMed]
11. Walls, R.; Hockberger, R.; Gausche-Hill, M.; Erickson, T.; Wilcox, S. *Rosen's Emergency Medicine: Concepts and Clinical Practice*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 202–210.
12. Loscalzo, J.; Fauci, A.; Kasper, D.; Hauser, S.; Longo, D.; Jameson, J.L. *Harrison's Principles of Internal Medicine*, 21st ed.; McGraw-Hill Education: New York, NY, USA, 2022.
13. Henderson, M.C.; Tierney, L.M., Jr.; Smetana, G.W. *The Patient History: An Evidence-Based Approach to Differential Diagnosis*, 2nd ed.; The McGraw-Hill Companies: New York, NY, USA, 2012; pp. 261–272.
14. Gulati, M.; Levy, P.D.; Mukherjee, D.; Amsterdam, E.; Bhatt, D.L.; Birtcher, K.K.; Blankstein, R.; Boyd, J.; Bullock-Palmer, R.P. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of chest pain: A report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **2021**, *78*, e187–e285. [CrossRef] [PubMed]
15. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
16. Otto, E.; Culakova, E.; Meng, S.; Zhang, Z.; Xu, H.; Mohile, S.; Flannery, M.A. Overview of Sankey flow diagrams: Focusing on symptom trajectories in older adults with advanced cancer. *J. Geriatr. Oncol.* **2022**, *13*, 742–746. [CrossRef] [PubMed]
17. Hsia, R.Y.; Hale, Z.; Tabas, J.A. A national study of the prevalence of life-threatening diagnoses in patients with chest pain. *JAMA Intern. Med.* **2016**, *176*, 1029–1032. [CrossRef] [PubMed]
18. Kim, H.S.; Han, H.S.; Kim, W.; Kim, C.; Jang, J.Y.; Kwon, W.; Heo, J.S.; Shin, S.H.; Hwang, H.K.; Park, J.S. Clinical implications of young-onset pancreatic cancer patients after curative resection in Korea: A Korea Tumor Registry System Biliary Pancreas database analysis. *HPB* **2023**, *25*, 146–154. [CrossRef]
19. Bertolini, S.; Maoli, A.; Rauch, G.; Giacomini, M. Entropy-driven decision tree building for decision support in gastroenterology. In *Data and Knowledge for Medical Decision Support*; IOS Press: Amsterdam, The Netherlands, 2013; pp. 93–97.
20. Yamamoto, A.; Koda, M.; Ogawa, H.; Miyoshi, T.; Maeda, Y.; Otsuka, F.; Ino, H. Enhancing Medical Interview Skills Through AI-Simulated Patient Interactions: Nonrandomized Controlled Trial. *JMIR Med. Educ.* **2024**, *10*, e58753. [CrossRef] [PubMed]
21. Yi, Y.; Kim, K.J. The feasibility of using generative artificial intelligence for history taking in virtual patients. *BMC Res. Notes* **2025**, *18*, 80. [CrossRef] [PubMed]
22. Sun, Z.; Luo, C.; Liu, Z.; Huang, Z. Conversational disease diagnosis via external planner-controlled large language models. *arXiv* **2024**, arXiv:2404.04292. [CrossRef]
23. Du, Z.; Zheng, L.; Hu, R.; Xu, Y.; Li, X.; Sun, Y.; Chen, W.; Wu, J.; Cai, H.; Ying, H. LLMs Can Simulate Standardized Patients via Agent Coevolution. *arXiv* **2024**, arXiv:2412.11716. [CrossRef]
24. Tu, T.; Schaekermann, M.; Palepu, A.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Cheng, Y.; et al. Towards conversational diagnostic artificial intelligence. *Nature* **2025**, *642*, 442–450. [CrossRef] [PubMed]
25. Bösner, S.; Haasenritter, J.; Hani, M.A.; Keller, H.; Sönnichsen, A.C.; Karatolios, K.; Schaefer, J.R.; Baum, E.; Donner-Banzhoff, N. Gender differences in presentation and diagnosis of chest pain in primary care. *BMC Fam. Pract.* **2009**, *10*, 79. [CrossRef] [PubMed]