JMIR Publications
The Leading eHealth Publisher

Assessing Students’Clinical Reasoning Skills in History Taking with Large Language Model–Based Virtual Patients: Development and Validation of a Structured Coding Scheme Using Systematic Text Condensation

Naping Chen, Luzhen Tang, Yang Liu, Changmin Lin, Zijian Li, Chujun Shi, Chujun Shi, Mengyu Xia, Dragan Gasevic, Danijela Gasevic, Danijela Gasevic, Jinbin Zheng, Yizhou Fan, Xinyu Li, Xinyu Li

Submitted to: JMIR Medical Education
on: September 18, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript **5**
Supplementary Files **39**
Multimedia Appendixes **40**
Multimedia Appendix 1 ..... **40**
Multimedia Appendix 2 ..... **40**
Multimedia Appendix 3 ..... **40**

Preprint
JMIR Publications

Assessing Students'Clinical Reasoning Skills in History Taking with Large Language Model–Based Virtual Patients: Development and Validation of a Structured Coding Scheme Using Systematic Text Condensation

Naping Chen[1]; Luzhen Tang[2]; Yang Liu[1]; Changmin Lin[3]; Zijian Li[2]; Chujun Shi[1]; Chujun Shi[2]; Mengyu Xia[2]; Dragan Gasevic[4]; Danijela Gasevic[5]; Danijela Gasevic[4]; Jinbin Zheng[6]; Yizhou Fan[2]; Xinyu Li[4]; Xinyu Li[4]

[1] Shantou university medical college Shantou CN
[2] Graduate School of Education, Peking University Beijing CN
[3] Shantou University Medical College Shantou CN
[4] Faculty of Information Technology, Monash University Melbourne AU
[5] School of Public Health and Preventive Medicine, Monash University Melbourne AU
[6] The first affiliated hospital of Shantou university medical college Shantou CN

Corresponding Author:
Yizhou Fan

Graduate School of Education, Peking University
Peking University
Beijing
CN

Preprint
JMIR Publications

Abstract

Background: Large language model (LLM)-driven virtual patients (VPs) are increasingly used to simulate history taking. However, there is currently no straightforward methodological approach to effectively identify students' clinical reasoning activities during these interactions, which limits the ability to provide personalised feedback.

Objective: This study aims to develop a structured coding scheme to characterise medical students' behaviours during interactions with LLM-driven VPs.

Methods: Second-year medical students (N=210) completed text-based history-taking sessions across five simulated chest pain cases, yielding 1,030 dialogues. Dialogues from Cases 1–4 were analysed using systematic text condensation (STC) to develop a coding scheme inductively. Two raters independently coded a subset of dialogues, and inter-coder reliability was assessed using Cohen's kappa. The established scheme was then applied to the dialogues from Case 5, and Pearson correlation coefficients (r) were used to assess associations between code frequencies and external performance outcomes: diagnostic accuracy, history-taking checklist scores, clinical knowledge test scores, and post-encounter form (PEF) scores.

Results: The STC analysis produced a 12-code scheme comprising four clinical reasoning codes (Pathophysiologic Question, Relevant Response, Summarising & Integrating, Logical Organisation), six information-gathering codes, and two communication codes. Inter-coder reliability was high for all dimensions: clinical reasoning (? = 1.00), information gathering (? = 0.95-0.98), and communication (? = 1.00). In Case 5, Summarising & Integrating was most predictive, correlating with diagnostic accuracy (?2 =6.019, P=.014), checklist scores (r=0.208, P=.003), knowledge test scores (r=0.225, P=.002), and PEF scores (r=0.191, P=.009). Logical organisation (LO) also correlated with diagnostic accuracy (?2 =0.188, P=.008), checklist (r=0.592, P<.001), and knowledge test scores (r=0.170, P=.013). Patho-physiologic question showed weaker but significant associations with checklist and knowledge tests (r=0.177, p=.013 and r=0.145,p=.042). Only two information-gathering codes demonstrated weak-to-moderate associations with checklist and knowledge test scores, while only one communication code showed a weak association with knowledge tests.

Conclusions: This study developed a theory-informed coding scheme that reliably distinguishes information-gathering and reasoning behaviours in history-taking with virtual patients. Enabling the identification of diverse behaviours provides a foundation for formative assessment and personalised feedback, offering a scalable approach to support the development of clinical reasoning in medical students. Clinical Trial: NO

(JMIR Preprints 18/09/2025:84347)
DOI: https://doi.org/10.2196/preprints.84347

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Preprint JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Assessing Students' Clinical Reasoning skills in History Taking with Large Language Model–Based Virtual Patients: Development and Validation of a Structured Coding Scheme Using Systematic Text Condensation

Naping Chen M.M, Luzhen Tang Ed.M, Yang Liu M.M, Changmin Lin M.D,Zijian Li Ed.M, Chujun Shi, Mengyu Xia Ed.M, Dragan Gaševi[ć] PhD, Danijela Gasevic MD, PhD, Jinbin Zheng M.M, Yizhou Fan PhD[*] , Xinyu Li DCS[*]

Naping Chen Lecturer, Department of Clinical Skills Training Center, Shantou University Medi- cal College, Shantou, China ORCID:https://orcid.org/0000-0002-3996-5454

Luzhen Tang Postgraduate Student, Graduate School of Education, Peking University, Beijing, China ORCID: https://orcid.org/0009-0005-0102-5677.

Yang Liu Lecturer, Department of Clinical Skills Training Center, Shantou University Medical College, Shantou, China

Changmin Lin Professor, Office of Teaching Affairs, Shantou University Medical College, Shan- tou, China

Preprint JMIR Publications

Zijian Li Postgraduate Student, Graduate School of Education, Peking University, Beijing, China

Preprint JMIR Publications

Chujun Shi Researcher, Department of Clinical Skills Training Center, Shantou University Med- ical College, Shantou, China

Mengyu Xia Postgraduate Student, Graduate School of Education, Peking University, Beijing, China

Dragan Gasevic Distinguished Professor, Faculty of Information Technology, Monash University, Melbourne, Australia ORCID: https://orcid.org/0000-0001-9265-1908

Danijela Gasevic Professor, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

Jinbin Zheng Department of Cardiology, The First Affiliated Hospital of Shantou University Medical College, Shantou, China.

Yizhou Fan * Assistant Professor, Graduate School of Education, Peking University, Beijing, China (Corresponding Author) ORCID:https://orcid.org/0000-0003-2777-1705

Xinyu Li * Research Fellow, Faculty of Information Technology, Monash University, Melbourne, Australia (Corresponding Author) ORCID: https://orcid.org/0000-0003-2681-4451

Correspondence should be addressed to Yizhou Fan, Graduate School of Education, Peking University, Beijing 100871, China ; email: fyz@pku.edu.cn; Xinyu Li, Faculty of Information Technology, Monash University, Melbourne, Australia; email: xinyu.li1@monash.edu.

Preprint JMIR Publications

Abstract

Background: Large language model (LLM)-driven virtual patients (VPs) are increasingly used to simulate history taking. However, there is currently no straightforward methodological approach to effectively identify students' clinical reasoning activities during these interactions, which limits the ability to provide personalised feedback.

Objective: This study aims to develop a structured coding scheme to characterise medical students' behaviours during interactions with LLM-driven VPs.

Methods: SSecond-year medical students (N=210) completed text-based history-taking sessions across five simulated chest pain cases, yielding 1,030 dialogues. Dialogues from Cases 1–4 were analysed using systematic text condensation (STC) to develop a coding scheme inductively. Two raters independently coded a subset of dialogues, and inter-coder reliability was assessed using Cohen's kappa. The established scheme was then applied to the dialogues from Case 5, and Pearson correlation coefficients (r) were used to assess associations between code frequencies and external performance outcomes: diagnostic accuracy, history-taking checklist scores, clinical knowledge test scores, and post-encounter form (PEF) scores.

Results: The STC analysis produced a 12-code scheme comprising four clinical reasoning codes (Pathophysiologic Question, Relevant Response, Summarising & Integrating, Logical Organisation), six information-gathering codes, and two communication codes. Inter-coder reliability was high for all dimensions: clinical reasoning (κ = 1.00), information gathering (κ = 0.95-0.98), and communication (κ = 1.00). In Case 5, Summarising & Integrating was most predictive, correlating with diagnostic accuracy ($\chi 2$ =6.019, P=.014), checklist scores (r=0.208, P=.003), knowledge test scores (r=0.225, P=.002), and PEF scores (r=0.191, P=.009). Logical organisation (LO) also correlated with diagnostic accuracy ($\chi 2$ =0.188, P=.008), checklist (r=0.592, P<.001), and knowledge test scores (r=0.170, P=.013). Patho-physiologic question showed weaker but significant associations with checklist and knowledge tests (r=0.177, p=.013 and r=0.145,p=.042). Only two information-gathering codes demonstrated weak-to-moderate associations with checklist and knowledge test scores, while only one communication code showed a weak association with knowledge tests.

Conclusions: This study developed a theory-informed coding scheme that reliably distinguishes information-gathering and reasoning behaviours in history-taking with virtual patients.

Enabling the identification of diverse behaviours provides a foundation for formative assessment and personalised feedback, offering a scalable approach to support the development of clinical reasoning in medical students.

Keywords: Clinical Reasoning; Coding Scheme; History Taking; Large Language Model; Virtual Patients; Medical Education

Introduction

Clinical reasoning is a critical cognitive process through which healthcare professionals gather, interpret, and analyze patient information, generate and test hypotheses, and formulate effective clinical decisions [1;2] . It begins with the initial patient encounter, where limited information prompts the generation of a small set of diagnostic hypotheses that guide further data collection and are continuously tested[3;4] . The processes of evaluation and verification occur iteratively across history taking, physical examination, and diagnostic testing, ultimately leading to an accurate diagnosis and appropriate clinical decision-making [5]. History taking, as the initial stage of clinical reasoning during patient encounters [6] , constitutes a vital component of information gathering and plays a central role in shaping clinical reasoning [7;8;9] .

History taking involves systematically gathering detailed patient information through skilled patient-centered interviewing, targeted questioning, attentive listening, and careful exploration of patient narratives and concerns[10]. Effective history taking allows clinicians to identify salient clinical information, detect subtle clues, explore relevant symptoms comprehensively, rule out critical 'red flag' conditions, and thoughtfully contextualize patient presentations [11]. One of the key theoretical foundations of history taking is hypothetico-deductive reasoning, in which clinicians generate diagnostic hypotheses early in the encounter based on initial patient cues and iteratively collect data to confirm or refute these hypotheses[3]. Illness script theory further explains how clinicians match patient cues with structured knowledge stored in memory to guide focused information gathering during history taking[12]. Dual-process theory complements these models by proposing that clinicians rely on fast, intuitive pattern recognition (System 1) to activate illness scripts during history taking, while also engaging slower, analytical reasoning (System 2) when

cues are ambiguous or unfamiliar, with both systems interacting to support accurate diagnosis and script development [1;13;14;15] . Given that early hypothesis generation significantly influences diagnostic accuracy and that novice learners often struggle to gather relevant information effectively, assessing clinical reasoning during history taking is critical for supporting learners' diagnostic development and guiding targeted educational interventions[16;17;18] .

Current assessment methods evaluating students' clinical reasoning abilities in history taking have several limitations. Traditional scoring tools, such as checklist-based systems during history-taking, predominantly assess students' recall abilities rather than their active clinical reasoning processes [18;19] . Students achieve high scores mainly by remembering and mechanically checking off predetermined items, potentially inflating their performance while inadequately reflecting actual clinical reasoning skills[20]. Additionally, more advanced assessment formats , such as key feature questions, script concordance tests, and post-encounter written reflections [21;18;22] – attempt to indirectly measure clinical reasoning through evaluating the quantity and accuracy of collected patient information and diagnoses [23] , but is limited in capturing the dynamic, ongoing cognitive process occurring during real-time patient interactions [24] and supply personalized and timely feedback for learners.

To better prepare students for real-world clinical practice, educational initiatives and assessments must move beyond rote questioning to foster and evaluate the dynamic integration of history-taking skills and clinical reasoning abilities as an interactive process[7;25] . Innovative and practical methods are required to directly capture how students' history-taking strategies interact with their clinical reasoning skills[26;27] . Hasnain et al. classified history-taking strategies by analyzing how frequently students inquired about key symptoms and how thoroughly they explored presenting complaints [28]. However, such surface-level metrics, though useful for identifying hypothesis-driven inquiry [29] , fail to capture the deeper cognitive strategies central to clinical reasoning[8]. Haring et al. identified observable indicators of clinical reasoning across four domains, student actions, patient responses, conversational flow, and data-gathering efficiency through expert evaluation of students' video-recorded performances[30]. Both the Hasnain et al. and Haring et al. studies relied primarily on expert subjective judgment, reflecting a lack of clearly operationalized behavioral definitions. Although this approach effectively captures the interactive nature of clinical reasoning, it depends

heavily on expert input and labor-intensive video analysis, limiting its scalability and consistency[31]. Operationalizing the qualitative framework proposed by Haring et al.[31], Fürstenberg et al. developed the CRI-HTS, a structured assessment scale encompassing domains such as problem-focusing, context creation, and information gathering[32]. Despite these refinements, the tool still relies on subjective evaluation, intensive rater training, and demonstrates only moderate internal consistency[33], and is limited in capturing the depth and nuance of students' reasoning, as shown in large-scale implementations[34]. Taken together, these studies reveal a critical gap: the limited ability to conduct standardized analyses using extractable dialogue data.

The growing adoption of virtual patient simulations, including conversational agents and LLMs-based virtual patients (VPs), has created interactive, low-risk environments in which students can repeatedly practice and refine clinical reasoning skills in authentic clinical scenarios[35;36;37;38]. In addition to supporting experiential learning, these platforms generate structured conversational data, offering unique opportunities for real-time analysis of reasoning behaviors. In response to these opportunities and the limitations of existing assessment methods, the present study aimed to develop and validate a novel coding schema tailored to capture and assess medical students' clinical reasoning processes and history-taking competencies during consultation dialogue with a LLMs-based VP. Such a scheme offers the potential to enable real-time, large-scale tracking of learners' evolving clinical reasoning skills, translating spontaneous conversational behaviors into direct, interpretable, and reliable indicators of proficiency. This approach seeks to advance the field by facilitating both individualized feedback and rigorous, scalable assessment in technology-enhanced medical education environments. To evaluate the effectiveness of this approach, the present study addresses the following research question: How reliably and validly can a coding schema capture medical students' clinical reasoning competency during consultation dialogue with a LLMs-based VP?

Method

Participants

This study was conducted at a University Medical College in Shantou, China, and focused on

https://preprints.jmir.org/preprint/84347 [unpublished, non-peer-reviewed preprint]

second-year medical students enrolled in their second semester. A total of 210 students participated in the study between April and May 2024. At the time of data collection, participants were in the early stages of their medical education, having completed coursework in systemic anatomy and concurrently studying fundamental medical sciences, including the cardiovascular and res- piratory systems as well as disease mechanisms. During this period of training, students were just beginning to develop foundational skills in history taking, symptom recognition, and clinical reasoning. However, they had not yet received formal instruction in the identification of specific clinical diseases. As a result, their clinical reasoning abilities were still at an initial stage of development, providing a suitable context for investigating the emergence of these competencies.

Ethical approval: The study was approved by the Ethical Committee of Medical College, Shantou University, China (NO. SUMC-2024-064).

Data Collection Design

Students engaged in history-taking and clinical reasoning exercises with a LLMs-based VP [39] operationalized as a chatbot. The chatbot utilized a tailored prompt designs for GPT-3.5 provided in Appendix 1. The chatbot was integrated in FLoRA[40] to facilitate these interactions. FLoRA is a Moodle-integrated digital learning platform specifically designed to record the entirety of each student–VP dialogue for comprehensive subsequent analysis. The study incorporated five representative "chest pain" cases — Spontaneous Pneumothorax, Stable Angina, Aortic Dissection, Acute Pulmonary Embolism, and Acute Pericarditis — selected for their clinical relevance and diversity in diagnostic challenges. Across all five cases, a total of 1,030 consultation dialogues were collected, with each case involving between 205 and 207 of the 210 participants (Table 1). For

this study, we define a consultation dialogue as an exchange between a medical student and a VP. A dialogue turn refers to one question–answer pair between the student and the VP. Case 1 yielded a higher average number of dialogue turns compared to subsequent cases. This increase in dialogue turns can be attributed to students' initial tendency to ask redundant questions and engage in exploratory interactions while acclimating to the FLoRA system.

Table 1: Dialogue Turns Statistics from Five Cases Collected in the Study

Case	Dialogues	Total Turns	Mean (SD)
1	206	18792	91.2 (31.2)
2	207	14753	71.3 (13.7)
3	205	15379	75.0 (11.8)
4	206	15723	76.3 (13.3)
5	206	15911	77.2 (16.8)

Coding Scheme Development

To address the research question regarding the reliability and validity of a coding scheme for capturing medical students' clinical reasoning activities in history-taking during consultation dialogue, the first essential step was to develop and define a set of analytical codes. Systematic Text Condensation (STC) was employed as the primary methodology for qualitative data analysis in this study[41]. Developed by Malterud[41], STC is a descriptive and pragmatic approach that facilitates thematic cross-case analysis within qualitative research. Utilizing this method allowed for a structured condensation and interpretation of complex conversational data obtained from student–VP interactions. In this study, meaning units refer to students' individual utterances during interactions with the virtual patients. Each utterance was treated as a discrete unit for analysis, with one or more codes assigned iteratively during the coding process. To ensure the validity and robustness of the coding scheme, we undertook a multi-stage development and validation process. In defining each code, we drew on principles of discourse and content analysis established in clinical reason- ing research[42], framing the codes in terms of three aspects: linguistic form, pragmatic intent, and contextual function. The coding scheme was then iteratively refined through multiple rounds of

discussion, and inter-coder reliability was assessed.

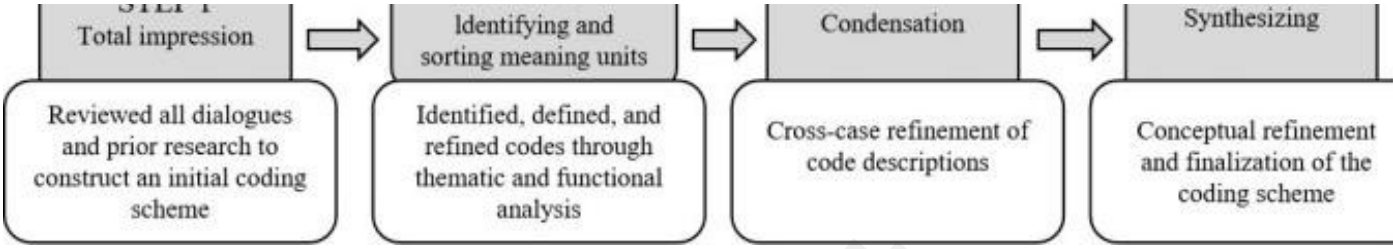


Figure 1: Coding scheme development process based on Systematic Text Condensation (STC)
The figure illustrates the four-step STC analytical framework (gray boxes), alongside the procedures used to construct and refine the coding scheme at each stage (white boxes). The process included initial immersion, thematic analysis, code condensation, and synthesis, ensuring both data-driven and conceptually rigorous development of the coding structure.

Step 1: Development of the Initial Coding Scheme

Author NP reviewed all the consultation dialogues in cases 1 to 4, and drawing on the specific features of online consultations (e.g., text-based interactions and flexible time constraints) as well as insights from previous studies, developed an initial set of codes.

Step 2: Refinement and Integration of the Coding Scheme

A thematic analysis of all consultation dialogues from Case 1 (pneumothorax) was conducted to define the coding categories with greater clarity. Drawing on relevant literature and semantic analysis, new codes were added to enhance the comprehensiveness of the scheme[43;19;10] . Redundant or meaningless codes were removed, and conceptually similar items were merged[22;44] , resulting in a finalized and integrated coding scheme.

Step 3: Cross-Case Testing and Refinement of the Coding Definitions

The coding scheme was empirically tested across Cases 2 to 4 to ensure its robustness in diverse clinical scenarios. This step enabled further refinement and clarification of the operational definitions for each code. Notably, the same information-seeking behavior could be expressed through

different history-taking approaches depending on the context, reflecting nuanced variations in students' underlying cognitive intentions.

Step 4: Conceptual Refinement and Finalization of the Coding Scheme

In the final stage, code definitions were further reconceptualized to enhance generalizability across history-taking interactions in diverse clinical scenarios, rather than being confined to case-specific contexts. Such conceptual refinements increased both the flexibility and applicability of the scheme. Ultimately, the finalized scheme comprised 12 codes in 3 dimensions (Table 2). It is recognized that, on occasion, individual questions may meet criteria for multiple codes. Accordingly, an utterance could be assigned more than one code if warranted.

To further ensure reliability, two authors (NP and Y), drawing on their backgrounds in clinical practice and medical education, independently conducted an initial thematic analysis of the Case 1 conversation data and collaboratively refined the coding scheme through iterative discussions. They then randomly selected 20 students and coded their dialogue data from cases 1 to 4. Discrep- ancies were identified and resolved through repeated discussions to finalize definitions for each code (Appendix 2). Next, 100 dialogue turns from Case 1 were randomly sampled from these stu- dents, and each question was independently coded using a single code. Low-frequency codes were grouped and coded at the category level. Inter-rater reliability was assessed, and the coding process was repeated until the Kappa coefficient exceeded 0.85. Once the coding scheme was finalized, all case dialogues were fully coded, with two authors (NP and Y) each independently coding half of the dataset.

Applying and Validating the Coding Scheme

To extend this evaluation and test the efficacy of the coding scheme in practice, we applied it to history-taking dialogues in Case 5, which involved a patient presenting with acute pericarditis. Case 5 was chosen not only for its diagnostic complexity but also because it followed several weeks of practice, during which students' performance became more stable and consistent, allow-

https://preprints.jmir.org/preprint/84347 [unpublished, non-peer-reviewed preprint]

ing for a more valid evaluation of the coding scheme. By coding these dialogues, we were able to systematically compare student behaviors as captured by the coding scheme against widely accepted criteria for assessing clinical reasoning and history-taking skills. Specifically, the coded data were correlated with multiple external benchmarks validated in prior research, allowing us to examine whether code frequencies were associated with meaningful aspects of clinical reasoning performance. These included students' clinical knowledge test scores (CKTS), diagnostic accuracy (DA, coded dichotomously as correct or incorrect), history-taking checklist scores (HTCS, adapted from the Kalamazoo Essentials List and Medical Licensing Examination criteria), and scores from a modified post-encounter form (PEFS)[45;46;20;47;48]. These measures provided a comprehensive context for validating the relevance and sensitivity of the coding scheme to key aspects of student performance (Table 3).

To quantitatively examine the relationships between code frequencies and established assessment criteria – called concurrent validity[49], we conducted Pearson correlation for continuous codes and point-biserial correlation or Chi-square test for binary codes. These analyses were based on the counts of codes per participant, the frequency of individual code occurrences, and stu- dents' scores on all other tests described above. All statistical analyses were performed in Python, with a significance threshold of $p < 0.05$. By triangulating findings across multiple analytic ap- proaches, this validation process confirmed both the theoretical integrity and practical applicability of the final coding scheme within the context of virtual patient-based clinical reasoning education.

Results

The finalized coding scheme comprised 12 behavioral codes aimed at capturing students' clin- ical reasoning, information-gathering, and communicative behaviors during history taking (see Textbox 1 for an overview of these dimensions and their related codes). A complete example of student–patient dialogue coding is provided in Appendix 3.

Textbox 1. Dimensions and codes for identifying students' behaviors in history-taking

Dimension 1: Clinical Reasoning Behaviors

- **Pathophysiologic Question (PQ)**
- **Relevant Response (RR)**
- **Summarizing & Integrating (SI)**
- **Logical Organization (LO)**

Dimension 2: Information-Gathering Behaviors

- **Specifying Symptoms (SS)**
- **Routine Question (RQ)**
- **Summarizing & Restating (SR)**
- **Checking (CK)**
- **Repeating Question (RT)**
- **Fuzzy Question (FQ)**

Dimension 3: Communication Behaviors

- **Chitchat (CC)**
- **Off-topic Statement (OS)**

Note. Symbols in parentheses are code labels used for annotation and analysis.

A Detailed Explanation of Codes, Definitions, and Examples

Pathophysiologic Question, PQ

When students pose specific, hypothesis-driven questions grounded in pathophysiologic thinking, such as asking about radiating pain at anatomically relevant sites, identifiable triggers or alleviating factors, or relevant past or family history—these behaviors are coded as PQ. The code also applies when students provide diagnostic explanations or suggestions in response to patient concerns. The

https://preprints.jmir.org/preprint/84347 [unpublished, non-peer-reviewed preprint]

defining feature of this behavior is its clear diagnostic intent. For example, in a case involving stable angina, students posed the following questions that were coded as PQ:

"Okay, have you had any heart problems before?"

"The preliminary diagnosis is angina, but I'll confirm it after you have an ECG. "

Did the pain always occur when you were lifting heavy objects or climbing slopes?

Relevant Response, RR

When patients mentioned diagnostically significant information, such as characteristic symptoms, key clinical signs, or high-risk factors, students who recognized and responded to these features were coded as RR.

For instance, in the case involving pulmonary embolism, the patient mentioned the long-term use of oral contraceptives. The student responded by asking, *"For how long have you been taking them?"*

"Does the pain go away every time you take nitroglycerin?" (patient mentioned using nitroglycerin)

Summarizing & Integrating, SI

Summarizing and integrating the patient's information in a coherent and structured manner was coded as SI, a binary code. This behavior was not limited to repetition but involved logical reorganization of collected data as illustrated in the following exdample:

https://preprints.jmir.org/preprint/84347 [unpublished, non-peer-reviewed preprint]

"You experienced sudden, persistent pain in your right chest this morning after forcefully blowing up a balloon, which lasted about 2 minutes before gradually easing, but it hasn't been completely relieved even after resting. You've had bilateral chest tightness that persists until now, and it worsens noticeably when walking quickly or climbing stairs... "

Logical Organization, LO

This code was applied based on the clinical relevance of questions rather than their strict sequential structure. Questions that reflect logical exploration based on the chief complaint, such as asking about related symptoms, associated factors, or indicators of severity, were coded as LO:

In case of aortic dissection: "Do you have hematuria / reduced urine output?"

In case of pulmonary embolism: Do you have shortness of breath/cyanosis?

Specifying Symptoms, SS

This code includes questions aimed at systematically exploring the characteristics of the main symptom. Typical aspects include the onset, duration, location, quality, severity, and aggravating or relieving factors. SS may be applied in combination with other codes when a question involves detailed exploration of clinical information, such as in RQ&SS.

What do you think triggered your chest pain today?

"Are there any factors that worsen or relieve the pain?"

"Is there pain in other areas?"

Routine Question, RQ

Routine history-taking questions (RQ) refer to items commonly asked in all clinical interviews, regardless of the specific presenting problem. These include standard elements such as basic patient information, past medical history, personal and family history, and review of systems—questions that resemble those found on clinical interview checklists. This category also includes commonly used open-ended questions in routine history taking. Some RQ-coded questions also displayed

other history-taking approaches and were assigned multiple codes accordingly. Further examples appear in later sections.

What discomfort brought you to the clinic?

"Do you have any other discomfort that occurred at the same time as this chest pain?"

"How old is your father?"

Summarizing & Restating, SR

This code was applied when students summarized collected information by restating it without reorganizing or synthesizing it. The summary often followed the order in which the information was obtained, and did not involve reformulating the content in a diagnostic or structured manner. Similar to SI codes, SR codes are also binary and mutually exclusive. If a summarize action occurs, it must be coded as either SR or SI, but not both. For example:

"two minutes ago, you came to the hospital due to chest pain and shortness of breath. You haven't sought medical attention before, nor have you taken any medication… "

Checking, CK

This code was applied when students sought to confirm or clarify patient-reported information, typically in response to unclear or seemingly inconsistent statements. CK was often used in combination with other codes, such as RQ, SS, or LO, when confirmation occurred alongside other history-taking approaches. An example of CK and SS co-coded behavior is shown below.

Patient: "It lasted for about 2 minutes. "

Student: "So after 2 minutes, the pain became the same as it is now?"

Fuzzy Question, FQ

Involves the repeated use (two or more times) of vague, open-ended questions within a single section of history taking, typically intended to prompt patients to provide additional information voluntarily. This definition excludes appropriate opening inquiries, which are coded as Routine Questions (RQ). An example illustrating the distinction between FQ and RQ is provided below:

https://preprints.jmir.org/preprint/84347 [unpublished, non-peer-reviewed preprint]

Student: "How was your health before?" (RQ)

...

Student: "Do you have any other medical conditions besides high blood pressure?" (RQ)

...

Student: "Any other diseases?" (FQ)

Repeating Question, RT

is characterized by students asking about the same clinical information point multiple times, sometimes using different expressions or synonymous terms. However, whether such synonymous expressions should be coded as RT depends on the student's level of training. Among lower-year medical students, this behavior is often associated with limited understanding of clinical terminology. For example, a student may separately ask about "shortness of breath," "difficulty breathing," and "labored breathing," without recognizing the semantic overlap. In such cases, the coding should be context-dependent and may require alternative categorization. The following excerpt illustrates a case of Repetition (RT):

Student: "Do you experience shortness of breath or chest tightness?" (LO)

Patient: "I feel my heart racing, sweating, and lightheaded, but without shortness of breath or chest tightness."

(...17 rounds later)

Student: "Is your breathing normal when you have chest pain?" (RT)

Chitchat, CC

CC includes greetings, small talk, brief responses, transitional phrases, terminology explanations, expressions of reassurance, and non-diagnostic patient education. When brief conversational elements (e.g., "Okay") appear at the beginning of a question, they are not independently coded as

CC. Instead, the entire utterance is coded according to the category assigned to the question itself (e.g., FQ, RQ, LO, SS), in order to avoid overcoding. However, if they were semantically distinct

https://preprints.jmir.org/preprint/84347 [unpublished, non-peer-reviewed preprint]

from other coded behaviors, dual coding should be applied, assigning both CC and the relevant category. Examples:

Hello, I'm Dr. Guo

"Take it easy."

"Hemoptysis means you coughed up and noticed blood. "

Off-topic Statement, OS

Utterances that are incomplete, illogical, or entirely unrelated to the clinical case and therefore can- not be reasonably classified under any other code are categorized as OS to ensure coding integrity and comprehensive inclusion of all dialogue segments:

What kind of bird do you keep, pigeons?

"Was the balloon sucked in?"

Inter-Rater Reliability of the Coding Scheme

Due to the low frequency of certain codes, inter-rater reliability was assessed at the dimension level rather than for each individual code. Within clinical reasoning behaviors, the low-frequency codes Pathophysiologic Questioning (PQ), Relevant Response (RR), and Summarizing & Integrating (SI) were grouped together, and perfect agreement was achieved ($\kappa = 1.00$). Within information-gathering behaviors, the codes Summarizing & Restating (RT), Checking (CK), Repeating Question (RT), and Fuzzy Question (FQ) were assessed as a group, yielding strong agreement ($\kappa = 0.96$). Within communication behaviors, the codes Chitchat (CC) and Off-topic Statement (OS) were grouped, also showing perfect agreement ($\kappa = 1.00$). For the higher-frequency individual codes, strong agreement was likewise observed: Logical Organization (LO, $\kappa = 1.00$), Routine Question (RQ, $\kappa = 0.98$), and Specifying Symptoms (SS, $\kappa = 0.95$).

Associations Between Codes and Performance – Validity

The coding scheme demonstrated strong applicability in evaluating students' clinical reasoning during the history-taking of Case 5. Several codes showed statistically significant correlations with key performance metrics, supporting the construct validity of the scheme (Table 3).

Among the codes, SI showed the most consistent and significant correlations across all four measures (r = 0.208–0.225, χ^2 =6.019, all $P < .05$), suggesting it may reflect students' ability to integrate and summarize clinical information effectively. LO also correlated significantly with HTCS, DA, and CKTS (all $P < .05$), though not with PEFS measures, indicating it may reflect procedural organization more than post-encounter reasoning. Other notable correlations include PQ with HTCS, CKTS, and SS with HTCS and CKTS (all $P < .05$). RQ was also associated with HTCS, while CC showed weak but significant correlations with CKTS, warranting cautious interpretation. In contrast, RT, FQ, OS, and SR showed no significant associations. These results provide preliminary support for the convergent validity of the coding scheme, particularly for codes such as SI and LO, which align with established indicators of clinical reasoning performance.

In addition to correlation findings, the frequency distribution of codes in Case 5 offers insight into students' reasoning patterns. As shown in Table 2, foundational behaviors such as RQ and SS were most common ($n = 9503$ and $n = 2322$, respectively). In contrast, higher-level reasoning codes like PQ and SI appeared less frequently ($n = 515$ and $n = 92$), yet were strongly associated with better performance outcomes. These patterns suggest that while basic inquiry is prevalent, advanced reasoning, though rare, is especially predictive of clinical competence, highlighting the scheme's ability to differentiate between procedural and high-value cognitive strategies.

Table 2: Associations Between Coded Student Behaviors and Performance Metrics in Case 5 (Acute Pericarditis)

Code	Frequency	DA r /χ^2 (P)	HTCS r / P	CKTS r / P	PEFS r / P
SI	92	6.019 / .014*	0.208 / .003**	0.225 / .002**	0.191 / .009**
LO	1651	0.188 / .008**	0.592 / .000**	0.170 / .013*	0.130 / .08
PQ	515	0.054 / .45	0.177 / .013*	0.145 / .042*	0.070 / .41
RR	160	0.105 / .14	0.012 / .74	-0.111 / .14	0.051 / .72
SS	2322	0.010 / .89	0.376 / .000**	0.253 / .000**	0.044 / .55
RQ	9503	0.128 / .07	0.397 / .000**	0.040 / .58	0.120 / .10

SR	30	0.027 / .87	-0.025 / .73	-0.009 / .90	0.055 / .46
CK	43	-0.050 / .49	0.049 / .49	-0.128 / .08	0.073 / .32
RT	56	0.024 / .74	0.070 / .33	-0.122 / .09	0.071 / .34
FQ	38	0.060 / .43	0.045 / .53	0.043 / .55	-0.050 / .45
CC	1059	-0.037 / .61	0.099 / .17	0.149 / .037*	-0.006 / .93
OS	129	0.048 / .51	0.057 / .43	0.001 / .99	0.035 / .64

Note: Abbreviations: DA = Diagnostic Accuracy; HTCS = History-Taking Checklist Score; CKTS = Clinical Knowledge Test Score; PEFS = Post-Encounter Form Score.
Pearson correlation used for continuous codes; point-biserial correlation or Chi-square test used for binary codes (SI, SR).
*$P < 0.05$, **$P < 0.01$.

Discussion

This study developed and applied a 12-code scheme to assess medical students' clinical reasoning activities in history-taking with a LLMs-based virtual patient. The coding showed strong inter-rater reliability and revealed meaningful patterns: while fundamental questioning was most frequent, higher-level reasoning behaviours—particularly SI and LO—were significantly associated with performance outcomes. These findings support the scheme's validity for evaluating reasoning quality in simulated interviews.

Differentiating Deep Reasoning from Surface-Level Behaviours

The observed correlations of SI and LO with multiple performance indicators align with prior studies highlighting pathophysiologic reasoning as central to clinical diagnosis [50] and structured summarization as a hallmark of expertise[51;22]. Specifically, SI reflects the learner's capacity to integrate and restructure clinical data. Our refined LO code shifts focus from sequencing to clinical relevance, capturing reasoning more effectively than prior models based on question order[30]. The infrequent use of PQ showed a weak association with diagnostic outcomes and may reflect both the limited biomedical knowledge of second-year students and the difficulty of applying such knowledge in clinical contexts[52], leading to a preference for symptom-based over mechanism-driven inquiry. This limitation underscores that the integration of biomedical knowledge during history-taking is a critical driver of diagnostic expertise[53]. The ability to recognize relevant clinical information during history taking reflects a cognitive shift from data acquisition to hypothesis

formulation, as emphasized in dual-process and clinical expertise theories[54;55]. However, in our study, RR behaviors were infrequent and only weakly associated with diagnostic performance. This may be due to novice learners' underdeveloped illness scripts, which limit their ability to identify and respond to diagnostically meaningful cues in real time[15]. RR may therefore serve as an early-stage marker of developing clinical reasoning, albeit a fragile one in less experienced learners.

Although RQ and SS were among the most frequently observed behaviors, their limited association with reasoning performance suggests they reflect a different cognitive approach. These patterns likely indicate a "wide-net" strategy shaped by checklist-based habits, where students prioritize exhaustive data collection over hypothesis generation[19;18]. As such, they are more characteristic of procedural completeness than interpretive reasoning[56;17]. The lack of consistent associations for codes such as CK, FQ, and RT may suggest that these behaviors prioritize informational precision under conditions of cognitive strain, rather than reflecting purposeful clinical reasoning[57]. In contrast, CC and OS appear to serve communicative or social functions, with limited relevance to clinical problem-solving[58].

Contrasting Existing Tools to Position Process-Oriented Reasoning Assessment

Building on prior work [28;30], our scheme introduces new behavioral indicators and dimensions to provide a more comprehensive assessment of history taking. It is tailored for artificial intellegence (AI)-mediated platforms, offering high reliability and enabling both automated evaluation and personalized feedback.

In comparison with other clinical reasoning assessments, our approach offers unique advantages in capturing reasoning processes as they unfold in real-time. Traditional tools, such as Script Concordance Tests[21], Clinical Reasoning Problems[59], and post-encounter forms[18], assess reasoning under uncertainty by comparing learners' responses with those of expert panels. While useful, they rely heavily on expert consensus scoring and offer limited insight into how students reason in dynamic, context-rich interactions[59;18]. Oral examinations assess reasoning and communication

through dialogue but are prone to examiner bias and low inter-rater reliability[60]. Objective Structured Clinical Examinations, while widely used for performance-based assessment, have been critiqued for their subjective, resource-intensive checklist-based evaluations and for insufficient integration between reasoning and data collection processes[20]. Key Feature Questions target specific decision points in clinical cases but do not account for the continuity or structure of reasoning processes[61]. By contrast, our coding scheme evaluates reasoning behaviors embedded in authentic student–patient dialogues, enabling a process-oriented and ecologically valid assessment of diagnostic thinking. Rather than focusing solely on outcomes or key decisions, it captures students' reasoning strategies in context, offering a scalable and transferable complement to existing tools for both formative and summative purposes.

Educational Implications

Building on these findings and comparisons, the educational implications of our coding scheme are twofold. First, it offers a practical and scalable tool for educators to assess and monitor students' clinical reasoning in virtual patient settings, an increasingly integral component of contemporary curricula. Second, specific codes, such as SI, PQ, LO, and RR, may serve as observable proxies for higher-order reasoning, enabling targeted formative feedback and remediation. By shifting the evaluative focus from checklist completion to cognitive engagement, the scheme addresses a longstanding gap in reasoning assessment. It also provides a foundation for instructional design aimed at scaffolding reasoning in novice learners and supporting their progression toward expert-like thinking.

As large language models (LLMs) increasingly serve not only as virtual patients but also as intelligent evaluators[62;63], this coding scheme could also enable real-time identification of reasoning behaviors, such as the absence of PQ or RR-type questions, during simulated encounters. This would enable personalized, behavior-specific feedback at scale, allowing students to reflect on diagnostic gaps and refine their inquiry strategies. Ultimately, the scheme bridges traditional assessment with emerging AI-driven instructional systems, offering opportunities to enhance clinical reasoning education in a scalable and context-sensitive manner.

Limitations and Future Directions

Several limitations warrant consideration. First, the study was conducted at a single institution and included only second-year students, which may limit generalizability. Second, although inter-rater reliability was high, manual coding limited its scalability in real-time settings. Third, the clinical scenarios focused primarily on chest pain, raising questions about the scheme's applicability across a broader range of presenting complaints.

Future work should explore the automation of the coding process using LLMs to enable real-time assessment and feedback. Longitudinal studies tracking students' reasoning development over time and across clinical encounters would help validate the scheme's utility. Further integration with formative assessments, Objective Structured Clinical Examinations, or AI-driven learning platforms could support its transition from a research tool to an embedded component of clinical education at scale.

Conclusion

This study presents a novel, theory-informed coding scheme for evaluating medical students' history-taking behaviours during text-based interactions with a LLM–based virtual patient. The scheme demonstrated strong reliability and construct validity, effectively distinguishing routine information gathering from higher-level reasoning strategies. Enabling the reliable identification of diverse behaviors provides a practical foundation for formative assessment and personalized, targeted feedback. As virtual patient technologies become increasingly prominent in medical education, this scheme provides a scalable, context-sensitive approach for supporting the development

Preprint JMIR Publications

of clinical reasoning assessment. Future integration with automated analytics may further enhance its adaptability across instructional contexts and learner levels.

Acknowledgments

We acknowledge the support from the National Natural Science Foundation of China (No. 62407001), and the Shantou University Medical College Teaching Reform and Research Project (2025, No. 36).

Data Availability

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

Authors' Contributions

Naping Chen contributed to study design, data interpretation, and manuscript drafting. Luzhen Tang contributed to data collection and analysis. Yang Liu contributed to the development and refinement of the coding scheme. Zijian Li, Mengyu Xia, Weishan Zhang, and Chujun Shi participated in data collection. Changmin Lin, Dragan Gaševi^{ć}, and Danijela Gaševi^{ć} contributed to study design and critical manuscript revision. Xinyu Li, Yizhou Fan conceptualized the study, led data interpretation, and oversaw manuscript revision. All authors have reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Abbreviations

LLM: large language model

AI: artificial intellegence

VP: virtual patient

STC: Systematic Text Condensation

PEF: post-encounter form

CKTS: clinical knowledge test scores

DA: diagnostic accu- racy

HTCS: history-taking checklist scores

PEFS: post-encounter form scores

PQ: Pathophysiologic Question

RR: Relevant Response

SI: Summarizing & Integrating

LO: Logical Organization

SS: Specifying Symptoms

RQ: Routine Question

SR: Summarizing & Restating

CK: Checking

RT: Repeating Question

FQ: Fuzzy Question

CC: Chitchat

OS: Off-topic Statement

References

1. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005;39(1):98-106. doi:10.1111/j.1365-2929.2004.01972.x

2. Hawks MK, Maciuba JM, Merkebu J, et al. Clinical reasoning curricula in preclinical undergraduate medical education: a scoping review. Acad Med. 2023;98(8):958-965. doi:10.1097/ACM.0000000000005297

3. Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press; 1978.

4. Young M, Thomas A, Lubarsky S, et al. Drawing boundaries: the difficulty in defining clinical reasoning. Acad Med. 2018;93(7):990-995. doi:10.1097/ACM.0000000000002142

5. National Academies of Sciences, Engineering, and Medicine. The Diagnostic Process. In: Improving Diagnosis in Health Care. Washington, DC: National Academies Press; 2015:31-80.

6. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. N Engl J Med. 2006;355(21):2217-2225. doi:10.1056/NEJMra054782

7. Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. BMJ. 2002;324(7339):729-732. doi:10.1136/bmj.324.7339.729

8. Cooper N, Bartlett M, Gay S, et al. Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. Med Teach. 2021;43(2):152-159. doi:10.1080/0142159X.2020.1824380

9. Setrakian J, Gauthier G, Bergeron L, Chamberland M, St-Onge C. Comparison of assessment by a virtual patient and by clinician-educators of medical students' history-taking skills: exploratory descriptive study. JMIR Med Educ. 2020;6(1)\:e14428. doi:10.2196/14428

10. Bickley LS. Bates' Guide to Physical Examination and History Taking. 12th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2012.

11. Nichol JR, Sundjaja JH, Nelson G. Medical history. In: StatPearls \[Internet]. Treasure Island, FL: StatPearls Publishing; 2024 \[cited 2025 Sep 18]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK534809/

12. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. Acad Med. 1990;65(10):611-621. doi:10.1097/00001888-199010000-00001

13. Schwartz A, Elstein AS. Clinical reasoning in medicine. In: Higgs J, Jones M, Loftus S, Christensen N, eds. Clinical Reasoning in the Health Professions. 3rd ed. Edinburgh: Elsevier; 2008:223-234.

14. Croskerry P. A universal model of diagnostic reasoning. Acad Med. 2009;84(8):1022-1028. doi:10.1097/ACM.0b013e3181ace703.

15. Yazdani S, Hosseinzadeh M, Hosseini F. Models of clinical reasoning with a focus on general practice: a critical review. J Adv Med Educ Prof. 2017;5(4):177-184. PMID:29296221.

16. Barrows HS, Norman GR, Neufeld VR, Feightner JW. The clinical reasoning of randomly selected physicians in general medical practice. Clin Invest Med. 1982;5(1):49-55. PMID:7106997

17. Van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ. 2010;44(1):85-93. doi:10.1111/j.1365-2923.2009.03498.x

18. Thampy H, Willert E, Ramani S. Assessing clinical reasoning: targeting the higher levels of the pyramid. J Gen Intern Med. 2019;34(8):1631-1636. doi:10.1007/s11606-019-05068-z

19. Stivers T, Heritage J. Breaking the sequential mold: answering ‘more than the question’ during comprehensive history taking. Text Talk. 2001;21(1):151-185. doi:10.1515/text.1.21.1-2.151

20. Park WB, Kang SH, Myung SJ, Lee YS. Does objective structured clinical examinations score reflect the clinical reasoning ability of medical students? Am J Med Sci. 2015;350(1):64-67. doi:10.1097/MAJ.0000000000000503

21. Charlin B, Van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. Eval Health Prof. 2004;27(3):304-319. doi:10.1177/0163278704267043

22. Lai JH, Cheng KH, Wu YJ, Lin CC. Assessing clinical reasoning ability in fourth-year medical students via an integrative group history-taking with an individual reasoning activity. BMC Med Educ. 2022;22(1):573. doi:10.1186/s12909-022-03672-1

23. Cheng KH, Lee CY, Wu YJ, Lin CC. Using group history-taking and individual reasoning to identify shortcomings in clinical reasoning for medical students. J Med Educ Curric Dev. 2024;11:23821205241280946. doi:10.1177/23821205241280946

24. Im S, Kim DK, Kong HH, Roh HR, Oh YR, Seo JH. Assessing clinical reasoning abilities of medical students using clinical performance examination. Korean J Med Educ. 2016;28(1):35-44. doi:10.3946/kjme.2016.9

25. Windish DM, Price EG, Clever SL, Magaziner JL, Thomas PA. Teaching medical students the important connection between communication and clinical reasoning. J Gen Intern Med. 2005;20(12):1108-1113. doi:10.1111/j.1525-1497.2005.0255.x

26. Ilgen JS, Humbert AJ, Kuhn G, et al. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. Acad Emerg Med. 2012;19(12):1454-1461. doi:10.1111/acem.12034

27. Gruppen LD. Clinical reasoning: defining it, teaching it, assessing it, studying it. West J Emerg Med. 2017;18(1):4-7. doi:10.5811/westjem.2016.11.33191

28. Hasnain M, Bordage G, Connell KJ, Sinacore JM. History-taking behaviors associated with diagnostic competence of clerks: an exploratory study. Acad Med. 2001;76(10 Suppl)\:S14-S17. doi:10.1097/00001888-200110001-00005

Preprint JMIR Publications

29. Hauer KE, Ten Cate O, Boscardin C, Irby DM, Iobst W, O'Sullivan PS. Understanding trust as an essential element of trainee supervision and learning in the workplace. Adv Health Sci Educ Theory Pract. 2014;19(3):435-456. doi:10.1007/s10459-013-9474-4

30. Haring CM, Cools BM, Gurp PJ, Meer JW, Postma CT. Observable phenomena that reveal medical students' clinical reasoning ability during expert assessment of their history taking: a qualitative study. BMC Med Educ. 2017;17(1):147. doi:10.1186/s12909-017-0987-1.

31. Haring CM, Klaarwater CCR, Bouwmans GA, et al. Validity, reliability and feasibility of a new observation rating tool and a post encounter rating tool for the assessment of clinical reasoning skills of medical students during their internal medicine clerkship: a pilot study. BMC Med Educ. 2020;20(1):368. doi:10.1186/s12909-020-02288-6

32. Fürstenberg S, Helm T, Prediger S, Kadmon M, Berberat PO, Harendza S. Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. BMC Med Educ. 2020;20(1):368. doi:10.1186/s12909-020-02160-7

33. Bußenius L, Kadmon M, Berberat PO, Harendza S. Evaluating the Global Rating scale's psychometric properties to assess communication skills of undergraduate medical students in video-recorded simulated patient encounters. Patient Educ Couns. 2022;105(3):750-755. doi:10.1016/j.pec.2021.08.005

34. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ. 2024;24(1):1391. doi:10.1186/s12909-024-04941-7

35. Talbot TB, Sagae K, John B, Rizzo AA, Playa C. Designing useful virtual standardized patient encounters. In: Interservice/Industry Training, Simulation and Education Conference Proceedings; 2012;2:1-12.

36. Maicher KR, Stiff A, Scholl M, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule-based dialogue management to improve conversational fidelity. Med Teach. 2023;45(3):279-285. doi:10.1080/0142159X.2022.2152472

37. Hamilton A, Molzahn A, McLemore K. The evolution from standardized to virtual patients in medical education. Cureus. 2024;16(4)\:e71224. doi:10.7759/cureus.71224

38. Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ. 2024;10(1)\:e53961. doi:10.2196/53961

39. Chen N, Tang L, Huang X, Zheng J, Fan Y. Research on virtual standardized patient based on ChatGPT-4. Chin J Med Educ. 2025;45(1):44-49.

40. Li X, Fan Y, Li T, et al. The FLoRA engine: using analytics to measure and facilitate learners' own regulation activities. J Learn Anal. 2025;12(1):1-20.

41. Malterud K. Systematic text condensation: a strategy for qualitative analysis. Scand J Public Health. 2012;40(8):795-805. doi:10.1177/1403494812465030

42. Kennedy TJT, Regehr G, Baker GR, Lingard L. Point-of-care assessment of medical trainee competence for independent clinical work. Acad Med. 2008;83(10 Suppl)\:S89-S92. doi:10.1097/ACM.0b013e318183e183

43. Boyd E, Heritage J. Taking the history: questioning during comprehensive history-taking. Stud Interact Socioling. 2006;20(1):151-185.

44. Coker P. Effects of an experiential learning program on the clinical reasoning and critical thinking skills of occupational therapy students. J Allied Health. 2010;39(4):280-286. PMID:21174022.

45. Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. Acad Med. 2001;76(4):390-393. doi:10.1097/00001888-200104000-00021.

46. Milota MM, Thiel GJ, van Delden JJ. Narrative medicine as a medical education tool: a systematic review. Med Teach. 2019;41(7):802-810. doi:10.1080/0142159X.2019.1584274

47. Fürstenberg S, Oubaid V, Berberat PO, Kadmon M, Harendza S. Medical knowledge and teamwork predict the quality of case summary statements as an indicator of clinical reasoning in undergraduate medical students. GMS J Med Educ. 2019;36(2)\:Doc15. doi:10.3205/zma001225

48. Berger AJ, Gillespie CC, Tewksbury LR, et al. Assessment of medical student clinical reasoning by “lay” vs physician raters: inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination. Am J Surg. 2012;203(1):81-86. doi:10.1016/j.amjsurg.2011.09.012

49. Messick S. Validity. In: Linn RL, ed. Educational Measurement. 3rd ed. New York, NY: Macmillan; 1989:13-103.

50. Monajemi A. The role of biomedical knowledge in clinical reasoning: bridging the gap between two theories. Int J Body Mind Culture. 2014;1(2):103-112.

51. Harendza S, Berberat PO, Kadmon M. Assessing competences in medical students with a newly designed 360-degree examination of a simulated first day of residency: a feasibility study. J Community Med Health Educ. 2017;7(4):550. doi:10.4172/2161-0711.1000550

52. Edelbring S, Parodis I, Lundberg IE. Increasing reasoning awareness: video analysis of students' two-party virtual patient interactions. JMIR Med Educ. 2018;4(1)\:e9137. doi:10.2196/mededu.9137

53. Woods NN. Science is fundamental: the role of biomedical knowledge in clinical reasoning. Med Educ. 2007;41(12):1173-1177. doi:10.1111/j.1365-2923.2007.02911.x

54. Schmidt HG, Mamede S. How to improve the teaching of clinical reasoning: a narrative review and a proposal. Med Educ. 2015;49(10):961-973. doi:10.1111/medu.12775

55. Marcum JA. An integrated model of clinical reasoning: dual-process theory of cognition and metacognition. J Eval Clin Pract. 2012;18(5):954-961. doi:10.1111/j.1365-2753.2012.01900.x

56. Schmidt HG, Rikers RMJP. How expertise develops in medicine: knowledge encapsulation and illness script formation. Med Educ. 2007;41(12):1133-1139. doi:10.1111/j.1365-2923.2007.02915.x

57. Durning SJ, Jung E, Kim DH, Lee YM. Teaching clinical reasoning: principles from the literature to help improve instruction from the classroom to the bedside. Korean J Med Educ. 2024;36(2):145-155. doi:10.3946/kjme.2024.265

58. Kurtz S, Draper J, Silverman J. Teaching and Learning Communication Skills in Medicine. 2nd ed. Boca Raton, FL: CRC Press; 2017.

59. Higgs J, Jones MA, Loftus S, Christensen N. Clinical Reasoning in the Health Professions. 4th ed. Edinburgh: Elsevier Health Sciences; 2018.

60. Wass V, Wakeford R, Neighbour R, van der Vleuten C. Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. Med Educ. 2003;37(2):126-131. doi:10.1046/j.1365-2923.2003.01423.x

61. Hrynchak P, Glover Takahashi S, Nayer M. Key-feature questions for assessment of clinical reasoning: a literature review. Med Educ. 2014;48(9):870-883. doi:10.1111/medu.12509

62. Chiang CH, Lee H. Can large language models be an alternative to human evaluations? arXiv. Preprint posted online May 3, 2023. arXiv:2305.01937

63. Chern S, Chern E, Neubig G, Liu P. Can large language models be trusted for evaluation? Scalable meta-evaluation of LLMs as evaluators via agent debate. arXiv. Preprint posted online Jan 30, 2024. arXiv:2401.16788

Preprint JMIR Publications

Supplementary Files

Preprint JMIR Publications

Multimedia Appendixes

Prompt Template for ChatGPT-Simulated Virtual Patients.
URL: http://asset.jmir.pub/assets/934134eec40959486e5eb250b2c7c55d.docx

Development of Codes Through an Iterative Refinement Process.
URL: http://asset.jmir.pub/assets/907b5cea6410402496d66df8ebe65ade.docx

Sample Student-Patient Dialogue with Codes.
URL: http://asset.jmir.pub/assets/ac63c3e6270be6b6439858ba33b551ec.docx

Preprint JMIR Publications