

Supplementary Information

Beyond Socratic Questioning: Designing GenAI-mediated Metacognitive Scaffolding for Clinical Reasoning in Medical Education

Context

Supplementary Note 1: Student-AI Tutor Dialogue Transcripts	2
1.1 Dialogue transcript between the medical student and the AI tutor under the Fixed Turn-Triggered Metacognitive Scaffolding condition in Cycle 1	2
1.2 Dialogue transcript between the medical student and the AI tutor under the Socratic Questioning-Based Metacognitive Scaffolding condition in Cycle 2	4
1.3 Dialogue transcript between the medical student and the AI tutor under the Adaptive On-Demand Metacognitive Scaffolding condition in Cycle 3	6
Supplementary Note 2: Interview guide for medical students in three cycles	11
2.1 Cycle 1: Fixed Turn-Triggered Metacognitive Scaffolding (FTMS)	11
1. Scaffolding Feedback: Recall the two automated prompt questions generated by the AI tutor during history taking	11
2. Interaction with the AI tutor	11
3. Scaffolding Expectations:	11
2.2 Socratic Dialogue-Based Metacognitive Scaffolding (SDMS)	12
1. Interaction with AI	12
2. Scaffolding use	12
3. AI tutor	12
4. Feedback on scaffolding refinement	12
2.3 Adaptive On-Demand Metacognitive Scaffolding (AODMS)	13
1. Users' subjective experience of the AI tutor	13
2. Influence of the AI tutor on clinical reasoning	13
3. Evaluation of the AI tutor's feedback quality	13
4. Relationship between case difficulty and AI tutor use	13
5. Suggestions for improving the AI tutor	13
6. Influence of the Deliberate Reflection Table on students' reasoning	13
7. Evaluation of the Deliberate Reflection Table design	13
8. Combined use of the Deliberate Reflection Table and the AI tutor	13

Supplementary Note 1: Student-AI Tutor Dialogue Transcripts

1.1 Dialogue transcript between the medical student and the AI tutor under the Fixed Turn-Triggered Metacognitive Scaffolding condition in Cycle 1

User Answer	AI tutor Questions
<i>AI tutor intervened at turn 20 of the student-virtual patient dialogue with the following questions.</i>	
	Now, please summarize the obtained information regarding the characteristics of the primary symptom.
The patient experienced a sudden onset of unexplained chest pain 2 hours ago. The pain is located in the mid-chest, presenting as severe, continuous tearing pain, accompanied by back and lumbar pain, diaphoresis, and pallor. The symptoms are not relieved by postural changes.	Could you try to summarize the key symptom characteristics reported by this patient?
Sudden, severe tearing pain in the mid-chest and back; continuous; accompanied by lumbar pain; no relief from postural changes.	An excellent summary. Next, based on these key symptom characteristics, what do you consider to be the possible diagnoses?
Aortic dissection, myocardial infarction.	Your considerations are on the right track. To further clarify the diagnosis, what additional medical history information do you think needs to be collected?
Accompanying symptoms, pertinent positive and negative symptoms, and past medical history.	Appropriate answer. Please proceed with a detailed inquiry and collection of the medical history. Now, please summarize the obtained information regarding the accompanying symptoms and relevant medical history.
<i>AI tutor intervenes at turn 40 of the student-virtual patient dialogue with the following questions.</i>	
	Please summarize these accompanying symptoms and relevant medical history information.
Accompanied by lumbar pain, diaphoresis, pallor, palpitations, and fatigue, concurrent with a history of hypertension.	A very comprehensive summary. Based on this information, how would you approach the preliminary and differential diagnoses for this patient?
Preliminary diagnosis: Aortic dissection.	You have formulated your preliminary diagnosis. Now, to further confirm the diagnosis, what specific medical history information do you need to inquire about?
Differential diagnoses could include acute	You have established a clear direction for the

myocardial infarction and aortic dissection.	differential diagnosis. Now, what additional medical history information do you require to help further differentiate between these potential diagnoses?
Aggravating or alleviating factors, accompanying symptoms, and the circumstances of onset.	Continue gathering relevant medical history information to help you further analyze and differentiate these possible diagnostic scenarios.

1.2 Dialogue transcript between the medical student and the AI tutor under the Socratic Questioning-Based Metacognitive Scaffolding condition in Cycle 2

<i>Students documented reasoning in the Deliberate Reflection Table (DRT), then clicked the button "ask the tutor," and the entries in the DRT were sent to the AI tutor</i>	
User Questions	AI tutor Answers
The student would like to consult you. Here are the results of the student's Clinical Reasoning Comparative Analysis Form for this stage: Probability Ranking: 1 Diagnosis: Acute glomerulonephritis Supporting evidence: Oliguria, hypertension, edema, dyspnea Additional diagnostic evidence required: Urinalysis, complete blood count (CBC), costovertebral angle (CVA) tenderness, and pretibial pitting edema Probability Ranking: 2 Diagnosis: Liver cirrhosis Supporting evidence: Anorexia, oliguria, dyspnea Additional diagnostic evidence required: Abdominal ultrasound, CBC, and liver function tests (LFTs) Probability Ranking: 3 Diagnosis: Heart failure Supporting evidence: Dyspnea, edema, tachycardia Additional diagnostic evidence required: Electrocardiogram (ECG), echocardiogram	You selected acute glomerulonephritis as the primary diagnosis. Please analyze how the characteristics of the patient's disease course and symptoms align or conflict with the typical manifestations of acute glomerulonephritis.
The patient developed oliguria 3 days ago, and dyspnea and bilateral thigh edema 1 week ago, which may be caused by acute glomerulonephritis.	You mentioned acute glomerulonephritis as the primary diagnosis. Please analyze whether there are typical manifestations or symptoms in the patient's medical history that support acute glomerulonephritis.
Oliguria for 3 days, dyspnea, and bilateral thigh edema for 1 week.	These symptoms can indeed be associated with acute glomerulonephritis, but please evaluate whether the patient's past medical history and other symptoms consistently support this diagnosis.
The patient has a 20-year history of AIDS, hypertension, and osteoarthritis.	Considering the patient's medical history, please evaluate whether these past illnesses or medications might have an impact on the current symptoms and how they affect your diagnostic reasoning.

The patient has AIDS, leading to compromised immunity, which may result in nephritis. They also have hypertension, which could damage renal capillaries and cause nephritis.	Your reasoning has taken into account the potential impact of AIDS and hypertension on the kidneys. However, please consider whether the patient's dyspnea and weight gain can be more comprehensively explained by other systemic diseases.
Heart or liver?	Excellent, you are starting to consider possibilities in other systems. Please analyze whether the patient's symptoms of dyspnea, fatigue, and edema could possibly point to a cardiac function issue.
Yes, heart failure can present with dyspnea, fatigue, and edema.	3 Very good, you have identified the possibility of heart failure. Combining the patient's symptoms and medical history, please evaluate how likely acute decompensated heart failure is in this case.
Quite likely.	Very good. You have established acute decompensated heart failure as an important consideration. Please continue to analyze in depth: which specific symptoms and medical history information of the patient best support this diagnosis?
The patient developed persistent dyspnea a week ago, accompanied by bilateral leg edema, fatigue, and weakness, along with a history of hypertension.	You have well identified the symptoms and medical history associated with acute decompensated heart failure. Please maintain this line of reasoning, combine it with the patient's other information, and consider further diagnostic or therapeutic steps.

1.3 Dialogue transcript between the medical student and the AI tutor under the Adaptive On-Demand Metacognitive Scaffolding condition in Cycle 3

<i>Students documented reasoning in the Deliberate Reflection Table (DRT), then clicked the button "ask the tutor," and the entries in the DRT and student-AI patient dialogue in history-taking were sent to the AI tutor</i>	
User Questions	AI Tutor Answers
The student would like to consult you. Here are the results of the student's Clinical Reasoning Comparative Analysis Form: Probability Ranking: 1 Diagnosis: Chronic heart failure (early or compensated stage) Supporting evidence: Dyspnea after exertion, relieved after rest... Opposing evidence: The patient reports no history of heart disease and no paroxysmal nocturnal dyspnea... Probability Ranking: 2 Diagnosis: Hypertensive heart disease Supporting evidence: A history of hypertension for over 10 years, with unspecified blood pressure control; dyspnea after exertion... opposing evidence: The patient has no statements related to a history of heart disease... Probability Ranking: 3 Diagnosis: Early stage AIDS-related pulmonary infection Supporting evidence: A 20-year history of AIDS (immunocompromised, prone to opportunistic infections)... Opposing evidence: Currently, no common symptoms of pulmonary infection, such as fever, cough, or sputum... Probability Ranking: 4 Diagnosis: Early-stage obesity hypoventilation syndrome Supporting evidence: BMI 29.8 (obese); dyspnea after exertion, relieved after rest... Opposing evidence: No mention of relatively typical symptoms of obesity hypoventilation syndrome, such as snoring during sleep or morning headaches... Here is the dialogue between the student and the patient:	The content of your analysis form is quite comprehensive and has a certain degree of organization. Now I have a few questions. I hope you can think further, identify the strengths and weaknesses of your reasoning and information gathering, and see if there are any missing details worth adding. 1. You currently rank "chronic heart failure" first, but according to the medical history, the dyspnea has only been present for 1 week, and the symptoms are progressively worsening. Edema and fatigue have also progressed in a short period. Do you think this time progression fully matches the typical presentation of a patient with "chronic heart failure in the compensated stage"? In your classification, what is the difference between "chronic heart failure" and an "acute decompensated episode"? Will this affect your differential ranking? 2. You mentioned "no history of heart disease" as a basis for excluding heart-related diseases. But can you think of any common situations where patients "feel they have no history of heart disease," but in fact, heart problems may have been ignored or undiagnosed? How do you plan to further verify your hypothesis in this part? Please reflect on the above two questions, see if there are omissions or biases in your reasoning chain, and whether there are questions or analysis contents that need to be added to the interview.

Student: Hello, Mr. Li, I am your attending physician. May I ask what brought you to the hospital? AI Patient: Doctor, I have difficulty breathing, and I really can't stand it today. Student: How long has this been happening? AI Patient: About a week. Student: Don't worry, preliminarily, it doesn't look too serious. Next, I will arrange some tests for you, and we will have to rely on these tests to make a final judgment. AI Patient: Okay, thank you, doctor.	
<i>Students documented reasoning in the Deliberate Reflection Table (DRT), then clicked the button "ask the tutor," and the entries in the DRT and student-AI patient dialogue in history-taking were sent to the AI tutor</i>	
Continuing the dialogue, here are the results of the student's revised Clinical Reasoning Comparative Analysis Form: Probability Ranking: 1 Diagnosis: Acute heart failure decompensated episode Supporting evidence: The patient's dyspnea has only been present for 1 week and is progressively worsening... Opposing evidence: The patient reports no history of heart disease... Probability Ranking: 2 Diagnosis: Chronic heart failure (early or compensated stage) Supporting evidence: Dyspnea after exertion, relieved after rest... Opposing evidence: The patient reports no history of heart disease and no paroxysmal nocturnal dyspnea... Probability Ranking: 3 Diagnosis: Hypertensive heart disease Supporting evidence: A history of hypertension for over 10 years... Opposing evidence: ...No mention of relatively typical symptoms of hypertensive heart disease such as chest pain or edema. Probability Ranking: 4 Diagnosis: Early stage AIDS-related pulmonary infection Supporting evidence: A 20-year history of AIDS (immunocompromised, prone to opportunistic infections)...	The ranking and basis of your analysis form this time are more rigorous, showing obvious progress. Now I would like to ask you to think more deeply about two aspects: 1. Regarding "edema" and "decreased urine output": You used edema and decreased urine output as important supporting points for acute heart failure. Can you give examples of other diseases where patients might also have similar presentations? Do you think it is necessary in this case to gather more detailed information about the nature, distribution, and progression of the edema? If you were to explore these contents, how would you ask the questions? (Strategic reflection) 2. Thoughts on diagnostic ranking: You put "AIDS-related pulmonary infection" as a lower-ranked diagnosis, but the patient has a history of AIDS. Can you recall what key medical history points (or typical clinical manifestations) primarily targeting pulmonary infection you did not systematically follow up on? If it were indeed a pulmonary infection, how might these characteristic manifestations present? Please combine the above two questions for a self-check. Are there any "blind spots" in your current information-gathering strategy, and what parts of the analysis form can be supplemented or made clearer?

Opposing evidence: Currently, no common symptoms of pulmonary infection, such as fever, cough, or sputum... Probability Ranking: 5 Diagnosis: Early-stage obesity hypoventilation syndrome Supporting evidence: BMI 29.8 (obese); dyspnea after exertion... Opposing evidence: No mention of relatively typical symptoms of obesity hypoventilation syndrome, such as snoring during sleep or morning headaches... Here is the dialogue between the student and the patient: Student: Hello, Mr. Li, I am your attending physician. May I ask what brought you to the hospital? AI Patient: Doctor, I have difficulty breathing, and I really can't stand it today. Student: How long has this been happening? AI Patient: About a week. ... Student: Don't worry, preliminarily, it doesn't look too serious. Next, I will arrange some tests for you, and we will have to rely on these tests to make a final judgment. AI Patient: Okay, thank you, doctor.	
Pulmonary embolism, pneumothorax, and severe anemia can also have similar presentations.	Your thought is very good, and you can already actively associate other diseases that cause similar edema and dyspnea. So, next, I would like to ask you to: 1. Metacognitive reflection: How did you decide which differential diagnoses to prioritize and which to temporarily set aside during the case reasoning process? When facing "common symptoms" of complex etiologies, do you have a systematic method for induction and elimination? Do you feel any part of your thinking is still lacking? 2. Regarding pulmonary embolism, pneumothorax, and severe anemia, you mentioned: - Are there any clues in the medical history you gathered that can rule them out? Are there any aspects of the inquiry that could have been more specific to further confirm or rule out these diseases? - If you need to rule out these emergencies in the next step, what signs will you focus on during the physical examination phase? Please answer sequentially and consider: How do these

	reflections help improve your differential diagnosis analysis form and avoid diagnostic bias?
1. Differential priority and elimination methods: Prioritize diagnoses with close symptom association, urgent conditions, and underlying diseases (such as acute heart failure), and set aside diseases with insufficient evidence or low risk. Use "medical history collection - multi-system analysis - elimination method - evidence evaluation" to check for common symptoms. Lacking consideration for rare diseases and dynamic symptom changes. 2. Exclusion of pulmonary embolism, etc.: The medical history showed no chest pain, hemoptysis (pulmonary embolism), sudden chest pain (pneumothorax), or pallor and dizziness (anemia), but the inquiry did not detail high-risk factors. Physical examination focus: check for lower limb swelling and jugular vein distension for pulmonary embolism; check chest symmetry and breath sounds for pneumothorax; check for a pale complexion and nail beds for anemia. Reflection value: Improve the analysis form, supplement missing information, reduce diagnostic bias, and improve differentiation accuracy.	You are already able to use a relatively structured approach to organize diagnostic priorities, and you are aware that consideration for rare diseases and dynamic symptom changes needs strengthening. This is a very good metacognitive awareness. Next, I have two guiding suggestions to help you further refine your reasoning: 1. Information supplementation: You mentioned that you did not detail high-risk factors, such as the source of emboli for pulmonary embolism, a history of venous thrombosis, prolonged bed rest, tumors, or hormone use. Given the background of this case, what other specific high-risk factors do you think need to be gathered? If you were asked to ask follow-up questions, how would you organize the questions? 2. Layered differential thinking: You just focused on the physical examination plan. If you find signs inconsistent with your preliminary deduction during the physical examination phase (such as signs of pneumothorax or obvious signs of anemia), how would you adjust your possible ranking of the etiologies? What kind of "self-monitoring and correction" strategy do you have? Please answer each question step by step. This helps cultivate your self-regulation ability in diagnostic reasoning.
1. Collection and follow-up on high-risk factors: Need to collect: recent infection history, hypertension control status, AIDS CD4 count, and history of lower extremity deep vein thrombosis. Follow-up questions: "Have you had a cold or fever in the past week? What is your usual blood pressure? Have you checked your AIDS CD4 cells recently? Have your legs ever been swollen or painful?" 2. Layered differential adjustment and correction: If signs of pneumothorax are found, prioritize pneumothorax; if signs of anemia are seen, elevate the priority of severe anemia. Correction strategy: Use physical signs as anchors, supplement with targeted examinations (such as a chest X-ray and a complete blood count), re-evaluate the correlation between medical history and signs, and exclude contradictory	Your follow-up questioning and correction ideas are very clear, reflecting the ability to dynamically adjust and monitor metacognition. To make your clinical reasoning more systematic and self-consistent, consider the following two aspects: 1. Your collection of high-risk factors is mostly based on etiology inference. What potential omissions on the "timeline related to symptom evolution" do you think still exist in your current inquiry? For example, are the symptoms continuous, recurrent, or fluctuating with relief and aggravation? Could you give a specific example of how you would ask a follow-up question? 2. When you plan to rely on physical examination to confirm hypotheses, how do you avoid being swayed by preconceived impressions? Are there any specific self-reminder or reflection methods to ensure that you

diagnoses.	can objectively evaluate new physical findings? Please briefly share your thoughts. After completing these two aspects, you can summarize the highlights and areas for improvement in your current reasoning, and consider where you will place the focus of your examination in the next step of the physical examination.
Okay, thank you for your guidance, teacher.	Very good. It's great to see you constantly reflecting and actively improving your reasoning process. You've made significant progress. If you are willing, you can continue to summarize: What are the highlights and shortcomings of your current medical history collection and reasoning? For the next phase of the physical examination, which systems and signs do you expect to focus on? If you are ready, we can move into the reasoning analysis for the physical examination phase. You can first state your plan, or directly start the simulated collection and fill out the analysis form. If you encounter any uncertainties, feel free to raise your thoughts at any time, and we can discuss how to optimize them together.

Supplementary Note 2: Interview guide for medical students in three cycles

2.1 Cycle 1: Fixed Turn-Triggered Metacognitive Scaffolding (FTMS)

1. **Scaffolding Feedback:** Recall the two automated prompt questions generated by the AI tutor during history taking

- ① What was your perspective on these questions, and to what extent did you find them beneficial?
- ② How did you formulate your responses?
- ③ What specific cognitive processes did these prompts trigger, and how did they impact your final diagnostic decision-making?

2. **Interaction with the AI tutor**

- ① How did you engage in dialogue with the AI tutor?
- ② How would you evaluate the quality and utility of these interactions?
- ③ In what ways did this dialogue influence your clinical reasoning process?
- ④ Specifically, did it have an observable impact, and by what mechanisms?
- ⑤ How did it alter your subsequent history-taking trajectory?
- ⑥ Furthermore, how did the formulation of potential diagnoses inform your subsequent data gathering?

3. **Scaffolding Expectations:**

- ① Following the scaffolding intervention, what longitudinal effects have you observed on your clinical reasoning capabilities?
- ② What was your specific cognitive approach to clinical reasoning during the history-taking in the future? (For instance, did you exhibit a habitual tendency to synthesize the characteristics of the chief complaint alongside the history of associated symptoms?)
- ③ What improvements would you suggest for these prompt questions?
- ④ In the context of a consultation, if the objective is to stimulate clinical reasoning, facilitate the analysis of collected data, guide the subsequent history-taking trajectory, and support the formulation of a preliminary diagnosis. What specific types of questioning would provide the most optimal guidance? Please provide illustrative examples.

2.2 Socratic Dialogue-Based Metacognitive Scaffolding (SDMS)

1. Interaction with AI

- ① How did the feedback or questions from the AI tutor influence your thinking at that moment?
- ② How did you interpret the AI's suggestion?
- ③ Were you satisfied with the AI's response? Why or why not?

2. Scaffolding use

- ① Why did you choose to use the AI-supported scaffolding, including the Deliberate Reflection Table and the interaction with the AI tutor?
- ② In what ways did it help you?
- ③ Did you encounter any operational difficulties or points of confusion when using the scaffolding?

3. AI tutor

- ① How did you perceive the responses provided by the AI tutor?
- ② What were your views on the AI tutor's responses?
- ③ How useful did you find its functions? How did you respond when the AI tutor restated or summarised your reasoning?
- ④ Did this prompt you to think more deeply, such as by asking the AI further questions or consulting the reading materials?

4. Feedback on scaffolding refinement

- ⑤ Did you notice any of these changes during this round of use?
- ⑥ How did you evaluate these modifications?

2.3 Adaptive On-Demand Metacognitive Scaffolding (AODMS)

1. Users' subjective experience of the AI tutor

- ① Could you describe your overall experience of using the AI tutor?
- ② Probe if needed: What was the dialogue with the AI tutor like? Was the interaction smooth?

Did you encounter any obstacles or misunderstandings? How did the AI tutor differ from a real clinical supervisor? In which parts of the interaction did the process feel smooth, obstructed, or misaligned with your understanding?

2. Influence of the AI tutor on clinical reasoning

- ① How did the AI tutor participate in your clinical reasoning process? Did it influence your clinical reasoning?
- ② Probe if needed: To what extent did it help you reason through and analyze the case? Which aspects were useful, and which were less useful? How was the process of history-taking and case analysis different when using the AI tutor compared with not using it? Please provide examples where possible.

3. Evaluation of the AI tutor's feedback quality

- ① How would you evaluate the feedback provided by the AI tutor? Which types of feedback were valuable, and which were less applicable?
- ② Probe if needed: What would the ideal content and format of feedback look like to you?

4. Relationship between case difficulty and AI tutor use

- ① How would you evaluate the difficulty of the practice cases? Which cases were relatively easy, and which were more difficult?
- ② Compared with easier cases, was your interaction with the AI tutor different when working on more difficult cases? In which types of cases did the tutor provide greater support?

5. Suggestions for improving the AI tutor

- ① If you could improve the AI tutor, what changes would you make?
- ② If only one aspect could be changed, what would be the most important improvement?

6. Influence of the Deliberate Reflection Table on students' reasoning

- ① How did you experience using the Deliberate Reflection Table?
- ② How did it influence your diagnostic thinking?

7. Evaluation of the Deliberate Reflection Table design

- ① How would you evaluate the design of the table? Which components were useful, and which were redundant or difficult to complete?
- ② Probe if needed: How did fields such as "likelihood ranking," "diagnosis," "supporting evidence," and "opposing evidence" help you analyze and reason through the case? How did you perceive table functions such as adding, deleting, and drag-and-drop ranking?

8. Combined use of the Deliberate Reflection Table and the AI tutor

- ① How did you experience using the table together with the AI tutor?
- ② Probe if needed: How do you think the table and AI tutor should be used together? How did you actually use them? What role did each play in your reasoning process? Which one did you usually use first, and how did you proceed afterward?