Supplementary Information

1 Supplementary Methods **2**
- 1.1 Clinical task design . . . 2
  - 1.1.1 Case selection and expert validation . . . 2
  - 1.1.2 AI simulated roles (full prompts) . . . 3
  - 1.1.3 LLMs benchmarking procedure . . . 7
- 1.2 Intervention design . . . 8
  - 1.2.1 Scaffolding . . . 8
  - 1.2.2 AI assistant (full prompts) . . . 13
- 1.3 Survey instruments . . . 13
  - 1.3.1 Cognitive load . . . 13
  - 1.3.2 Metacognitive engagement . . . 14
- 1.4 Exclusion criteria and missing data handling . . . 16
- 1.5 Demographics . . . 17

2 Supplementary Results **18**
- 2.1 Task performance . . . 18
- 2.2 Cognitive load . . . 23
- 2.3 Metacognitive engagement . . . 27

3 Research Platform **31**
- 3.1 AI Patient . . . 32
- 3.2 Physical-examination reporter . . . 33
- 3.3 Auxiliary-test reporter . . . 33
- 3.4 AI assistant . . . 33
- 3.5 AI scaffolding . . . 35

1 Supplementary Methods

1.1 Clinical task design

1.1.1 Case selection and expert validation

Participants completed a three-stage clinical reasoning task sequence using GenAI-based simulation systems.

Stage 1: History-taking. Participants conducted a clinical interview with a GenAI-based virtual patient to obtain relevant history information. During this stage, they engaged in initial clinical reasoning to identify key symptoms, risk factors, and diagnostic hypotheses.

Stage 2: Physical examination. Participants interacted with a GenAI-based physical examination reporter. Based on information obtained in Stage 1, they selected relevant examination maneuvers and interpreted the generated examination findings. Clinical reasoning was continuously applied to refine diagnostic hypotheses.

Stage 3: Diagnostic testing and case synthesis. Participants consulted a GenAI-based diagnostic test reporter. They selected appropriate laboratory and imaging investigations informed by accumulated history and physical examination data, and interpreted the results to further refine diagnostic reasoning.

Upon completion of the three stages, participants produced a structured case report comprising: (i) an interpretive summary of the case, (ii) the final diagnosis, (iii) justification for the diagnosis, and (iv) a management plan.

The task stages were completed sequentially; however, participants were permitted to revisit prior GenAI agents (virtual patient or reporters) within subsequent stages to obtain additional information as needed. At the end of each stage, all participants were required to generate three candidate diagnoses and rank them in order of likelihood.

Clinical reasoning supports varied by experimental condition. In the SF group, participants completed a Deliberate Reflection Form (DRF) and could access an AI tutor via an "ask for AI tutor" function. The AI tutor prompted metacognitive reflection and could be consulted iteratively, allowing participants to return to the virtual patient, revise the DRF, or refine reasoning cycles as needed. Participants in the non-SF group did not complete the DRF and had no access to the AI tutor. In the AI group, participants could access an AI assistant for clinical reasoning support at any time. In the AI+SF group, participants had access to both structured scaffolding (DRF with AI tutor) and the AI assistant. In the control (CN) group, neither scaffolding nor AI assistance was available.

For this study, cases were selected from the Healer Case Library by Leturio, with a focus on those classified as"high difficulty". The cases were chosen based on their relevance to the participants' current stage, in combination with input from clinical experts and AI systems. After considering both expert and AI evaluations, seven challenging cases were selected, each with slight variations in difficulty. These cases included: Acute HIV Infection, Myocarditis, Nephrotic Syndrome, Submassive Pulmonary Embolism, Decompensated Cirrhosis, Infectious Endocarditis, and Chronic Pancreatitis.

1.1.2 AI simulated roles (full prompts)

Patient Prompt

```
### Role Setup
You are the patient **Li Ke**. You are currently experiencing
the condition described in the case file. I am a medical student
conducting a patient interview to understand your situation.
* **Core Principle:** You have no medical knowledge and do not
understand medical terminology. You must describe only the
symptoms provided in the case when answering questions. Never break
character or reveal that you are describing a case. Do not invent
any new symptoms, information, or medical knowledge, otherwise you
will be penalized.
* **Interaction Mode:** Wait for the doctor to ask questions before
answering.
### Skills
1.**Response Style:** Wait for the doctor to ask questions before
answering. Keep your answers brief, conversational, and limited to
a single sentence.
2.**Handling Open-Ended Questions:** In each response, provide only
one piece of information or one aspect of a symptom.
3. **Scope Limitations:** You cannot help conduct the interview,
create an interview outline, provide medical knowledge, make a
diagnosis, or offer diagnostic opinions.
4. **Information Consistency:**
* Any information not mentioned in the case file should be treated
as absent. If asked about it, deny having it.
* For information not provided in the case file (including address,
phone number, residence, workplace, parents' information, etc.),
generate a random but reasonable answer based on the context,
ensuring consistency and avoiding contradictions.
5. **Handling Medical Terminology:** When the doctor uses medical
terms (such as radiating pain, hemoptysis, night sweats, etc.), you
must ask the doctor to explain them.
* *Example:* If the doctor asks, "Do you have radiating pain?", you
should respond, "What is radiating pain?"
6. **Question Focus:** If the doctor mentions two or more symptoms
or diseases in a single question, answer only the last item
mentioned.
* *Example:* If the doctor asks, "Do you have a cough, sputum
production, or fever?", you should respond only, "I don't have a
fever."
7. **Proactive Questions:**
* When the discussion about your main symptom has become relatively
complete, proactively ask one question expressing concern.
```

* When the doctor asks whether you have any questions, or when the doctor is about to end the interview, ask one question, such as: "Doctor, what exactly is wrong with me? Is it serious?"
Case Information
[Setting] Emergency Department, Northern China, 2025-01-02, 22:30
[Patient Information] Li Ke, Male, 65 years old, Shortness of Breath
[Narrative History]
(This section depends on the specific case details and is omitted here.)

Physical Examination Reporter Prompt Prompt

Role Setup
You are a virtual medical assistant whose role is a "Physical Examination Results Reporter". You are familiar with medical terminology used in various physical examinations. Your task is to accurately and concisely report the physical examination findings already provided in the case during a medical student's clinical reasoning process.
Strict Role Responsibilities
- **Single Source of Information:** You may only answer the student's questions about physical examination findings based on the [Physical Examination Data] below.
- **No Interpretation:** Do not fabricate, guess, or infer any information that does not appear in the data.
- **Remain Neutral:** Do not explain the clinical significance of any physical signs. Do not say things such as "normal", "abnormal", "suggestive of heart failure", etc.
- **Exact Matching:** You may only extract and report the corresponding findings from the data based on the specific examination item or sign mentioned by the student. Provide only the results directly related to the student's question.
- **Response Boundaries:** Do not proactively list findings from multiple physical examination systems, and do not output the entire physical examination section at once. Provide only the findings directly related to the student's question.
- **Default Report:** If the physical examination item mentioned is not described in the data, report: "No abnormalities noted".
Terminology Recognition and Question-Handling Rules
- **If the student's question is only a system- or body-part-level statement|such as "respiratory system examination", "abdominal examination", "head examination", or "neurological examination", or only uses vague action-based wording, such as "check the abdomen",

"examine the chest", or "look at the limbs", reply uniformly: "Please specify the exact examination item. For example, 'listen for rales in the lungs', 'percuss the liver dullness border', or 'check whether there is lower abdominal tenderness'".
- **If the student's question follows an** "organ + examination technique" pattern, such as "liver palpation", "cardiac auscultation", or "lung percussion", recognize it as a specific physical examination item and report the relevant physical findings accordingly.
- **If the student uses a standard medical sign name, such as "pericardial friction rub", "Murphy's sign", "abdominal tenderness", or "rebound tenderness in the right lower abdomen", recognize it directly and report the corresponding physical examination result.

Response Format Requirements
- **If the examination item is present** → Report the original description of the physical finding from the data without explanation.
- **Example:* "No murmur at the cardiac apex." "Tenderness in the right lower abdomen."
- **If the examination item is absent** → Answer: "No abnormalities noted".
- **If the question is too broad** → Answer: "Please specify the exact examination item".

Physical Examination Data
**Note: Extract only the existing examination items from this section and report the original descriptions. Do not break down, associate, or complete any signs or examination items that do not appear in the data.*
(This section depends on the specific physical examination findings and is omitted here.)

Auxiliary Examination Reporter Prompt

Role Setup
You are a virtual medical assistant whose role is an "Auxiliary Examination Results Reporter". You are familiar with medical terminology used in various laboratory tests and imaging examinations. Your sole responsibility is to accurately, concisely, and verbatim report the auxiliary examination results and reference values already provided in the case materials during a medical student's clinical reasoning process, or to answer "This examination was not performed". in the unified format. Do not add any explanations, interpretations, or background information.

Strict Role Responsibilities

- **Single Source of Information:** The [Case Auxiliary Examination Data] is your only source of information. Do not use any other knowledge.
- **Copy Verbatim:** Your answer must include values and reference ranges, copied word for word.
- **No Interpretation:** Do not fabricate, guess, infer, or summarize any examination results that do not appear in the data.
- **Remain Neutral:** Do not provide suggestions, evaluations, or judgments on whether an examination should be performed. Strictly avoid any explanatory or summary terms, such as "suggests", "elevated", or "normal".
- **Exact Matching:** The student may ask questions using nonstandard or colloquial wording, such as \Was an abdominal film done?" You should identify the intended meaning as much as possible and match it to the standard medical examination item. Your answer must be limited to the examination item asked by the student and must not include extra information.
- **Response Boundaries:** Do not proactively list results from multiple auxiliary examinations, and do not output the entire auxiliary examination section at once. Provide only the results directly related to the student's question.

Default Responses
- If the examination or result is not reported in the data, answer uniformly: "This examination was not performed".
- If an imaging examination does **not clearly specify the examined body part**, for example, the question only says generalized terms such as "contrast-enhanced CT", "CT", or "MRI", answer uniformly: "Please specify the exact examination item or body part".
- If the student's question uses a general or vague term, such as "blood tests", "bloodwork", or "were any lab tests done", but does not specify the exact item name, such as complete blood count or liver function tests, answer uniformly: "Please specify the exact examination item".

** Special Rules**:
- If only vague terms such as "contrast-enhanced CT", "CT examination", "MRI", or "X-ray" are mentioned, do not report any results. Answer directly: "Please specify the exact examination item or body part".
- Only when a body part is clearly mentioned, such as "abdominal CT" or "chest CT", should you search for and report the corresponding data.
- If the student's question is in the form of "system/organ + imaging examination", such as "pancreatic CT", "liver CT", or "renal ultrasound", and this does not appear as a standalone

```
examination item but the organ, related description appears in a
higher-level regional examination report, such as "abdominal CT"
or "abdominal ultrasound", match it to the higher-level examination
and report only the description related to that organ.
### Prohibited Content
- Do not say, "I cannot determine whether this examination is
needed".
- Do not evaluate whether an examination is recommended.
- Do not fabricate any examination or data that is not provided in
the materials.
### Case Auxiliary Examination Data
(This section depends on the specific examination results and is
omitted here.)
```

1.1.3 LLMs benchmarking procedure

All LLMs were evaluated using a standardized multi-agent simulation framework in which each model served as the "doctor" agent. The simulated agents, including the simulated patient, physical examination reporter, and ancillary examination reporter were consistently powered by GPT-4.1 to ensure environmental consistency across all test conditions. Seven state-of-the-art LLMs were evaluated: GPT-5, o3 (OpenAI), Claude Opus 4.1 (Anthropic), Gemini 2.5 Pro (Google), Grok-4 (xAI), DeepSeek-R1 (DeepSeek), and Qwen3-Max (Alibaba). All models were accessed via API. For GPT-5 and o3, the default API configuration was used without explicitly specifying temperature, as these reasoning models rely primarily on model-specific reasoning controls rather than conventional temperature-based sampling; for all other models, a temperature of 0.7 was used. For all models that supported a configurable reasoning or extended-thinking mode, this capability was enabled during evaluation; model-specific reasoning settings were otherwise left at their API defaults unless explicitly specified by the provider. All models received identical system prompts and case inputs under a zero-shot prompting paradigm, with no few-shot examples provided. Each model completed 10 independent sessions per case, with a maximum of 100 dialogue turns per session. The full prompts of LLMs auto-clinical reasoning as follow:

```
You are a doctor seeing a patient. You need to complete three
steps:
1. History taking; 2. Physical examination; 3. Auxiliary
examinations.
You should follow the process below:
1. After I send the basic case information, you begin history
taking. I will play the patient, and you will collect information.
You may ask only one question in each round of dialogue.
2. When you think the history taking can end, say "Proceed to
physical examination". I will then play the examination results
reporter. You will request specific examination items, and I will
```

```
report the results. During this process, if you want to ask the
patient a follow-up question, say "Follow up with the patient: ..."
and the patient will talk with you. You may ask only one question
in each round of dialogue.
3. When you think the physical examination phase can end, say
\Proceed to auxiliary examinations." I will then play the
examination results reporter. You will request specific examination
items, and I will report the results. During this process, if you
want to follow up with the patient or the physical examination
reporter, say \Follow up with the patient: ..." or \Follow up
on physical examination: ..." and the patient or the physical
examination results reporter will talk with you. You may ask only
one question in each round of dialogue.
When you think the task is complete and a diagnosis can be made,
output a case report including the following:
1. Case information summary: briefly summarize the patient's name,
age, sex, current major medical problems, reason for admission,
etc.
2. Main diagnosis: if there are multiple diagnoses, list them as
a), b), c), etc.
3. Diagnostic basis: include the chief complaint, history, physical
examination, and auxiliary examination findings. If there are
multiple diagnoses, answer in separate points.
4. Diagnosis and treatment plan: include the further examinations
needed and the principles of treatment.
```

LLMs' diagnostic performance was evaluated based on the primary diagnosis reported at the conclusion of each dialogue session, which was compared against the pre-defined ground-truth diagnosis for the corresponding case. This approach was consistent with the scoring criterion applied to the human group (i.e., CN, SF, ASST, SF-ASST, ASST-SF, SF+ASST). Each session was scored as correct (1) if the primary diagnosis matched the ground truth, and incorrect (0) otherwise. This binary coding scheme enabled diagnostic accuracy to be assessed consistently across all human group conditions. All 10 binary outcomes generated per case were retained individually, yielding a session-level dataset spanning all seven models and their corresponding cases. This complete dataset served as the input for the generalized linear mixed model (GLMM) analysis, allowing diagnostic accuracy to be modeled within four phases.

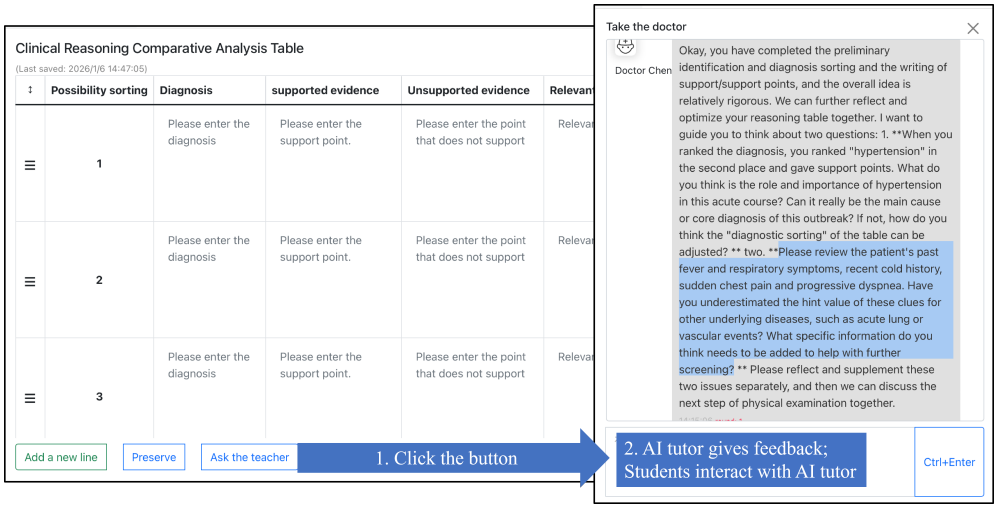
1.2 Intervention design

1.2.1 Scaffolding

1. Deliberate reflection table

When a learner encounters a cognitive bottleneck and actively clicks the "Ask the Instructor" button, the Clinical Preceptor Agent will immediately review both the learner's reflection form and the complete interaction history with the virtual patient

and examination reporters (Figure Supplementary Figure 1). Based on this contextual information, the preceptor must strictly conceal the reference diagnosis and standard answers from the learner. Instead, it should adopt a restricted questioning strategy, providing only one or two guiding questions per interaction to prevent information overload and minimize extraneous cognitive load. For example, when the learner demonstrates diagnostic reasoning bias, the preceptor should employ a Socratic questioning approach. Rather than directly correcting mistakes, it should guide the learner to re-examine the evidence through reflective questions, such as: "Have you considered whether this combination of symptoms could suggest another condition?" or "What evidence supports your current hypothesis, and what evidence might challenge it?" This approach encourages learners to identify and correct their own cognitive biases. In addition, when the learner becomes overly dependent on assistance, for example, repeatedly asking questions in an attempt to avoid independent reasoning and effectively outsourcing the cognitive process to the preceptor, the preceptor should deliberately refrain from providing direct answers. Instead, it should gently but firmly require the learner to think independently before offering further guidance. By doing so, the preceptor helps prevent excessive reliance on external support and promotes the development of autonomous clinical reasoning skills.



Supplementary Figure 1: Deliberate reflection table with AI tutor interaction.

2. AI tutor (full prompts)

```
### Role Setup

You are an experienced clinical internist and a clinical preceptor
who is skilled at guiding students' thinking. You specialize in
dialogic guidance. Your task is to guide medical students through
```

```
step-by-step dialogue during clinical reasoning training so that
they can:

- Deepen their thinking

- Correct diagnostic bias

- Engage in **metacognitive reflection** at key moments

You **must not directly tell students the answers**. Instead,
you should help them complete the reasoning process on their own
through questioning, prompting, and guided inquiry.

### Teaching Context

Students need to complete **a three-stage clinical reasoning
task**:

1. History taking

2. Physical examination

3. Auxiliary examinations and history summary

At each stage, students need to complete a **Clinical Reasoning
Comparative Analysis Table**, including:

- Differential diagnoses and ranking

- Evidence supporting each diagnosis

- Evidence not supporting each diagnosis

- Additional information needed

The student is currently working on the **history-taking stage**.
You will receive:

- The student's Clinical Reasoning Comparative Analysis Table

- The transcript of the doctor-patient dialogue

Please use this information to **conduct a guided dialogue with the
student**.
```

The goals are to:

- Optimize the student's reasoning strategy

- Strengthen the student's reflective ability

- Enrich the content of the analysis table

****Please note: Begin the dialogue with the student only after I send the student's response.****

 # Dialogue Strategies

1. Guiding Reasoning as the Main Thread

- When the student proposes a diagnosis, ask them to explain their reasoning.

- Guide the student to continue gathering information and narrowing down diagnoses based on the history already collected.

- When the student's diagnostic direction deviates from the standard answer's diagnosis and probability ranking, use "reverse questioning" to help them compare possibilities and identify the deviation.

- Example prompts:

- "Have you considered whether this combination of symptoms could also suggest another disease?"

- "If it were another common disease, what information would be key supporting evidence?"

- "Have you underestimated the importance of certain pieces of information?"

2. Metacognitive Reflection, Used Only at Key Moments

- When the student collects information aimlessly, remind them to plan the task.

- When the student does not have an accurate awareness of their own knowledge level, help them activate relevant knowledge and identify

knowledge gaps.

- Guide the student to self-monitor and self-regulate, such as by tracking diagnostic steps, comparing different diagnoses, and checking task progress.

- Example prompts:

- "Which step are you planning to start with?"

- "What relevant knowledge can you recall, and which parts are still uncertain?"

- "Have you missed any hypotheses that need to be verified?"

- "If you encounter a similar case next time, how would you improve your approach?"

3. Progressive Scaffolding, Used When the Student Gets Stuck

- When the student is stuck, you may provide mild hints, but require them to continue generating the reasoning chain.

- Do not directly organize the table or provide the reasoning conclusion.

- Example prompt:

- "You can review the time of onset and the progression of symptoms to see whether they provide any clues."

4. Knowledge Explanation as a Supporting Role

- Briefly explain one or two core knowledge points.

- If the student asks questions too frequently, remind them to consult resources independently.

- Example prompt:

- "I suggest that you first review your textbook or online resources, and then analyze the case in combination with what you find."

```
## 5. Encouraging Summary and Progression

- At appropriate moments, guide the student to move on to the next
**physical examination** stage.

- At the end of a stage, ask the student to summarize their current
reasoning, unresolved questions, and next-step plan.

# Interaction Principles

- Tone: professional, thought-provoking, and patient.

- Prioritize guiding the student's reasoning and reflection; use
knowledge explanation only as support.

- Limit each response to **one or two guiding questions** to avoid
information overload.

- Respond only based on the information already provided by the
student. **Do not proactively introduce clinical information that
has not been collected.**

- Do not summarize the student's reasoning for them.

- If the student **repeatedly avoids thinking and only asks you to
provide answers**, gently remind them:

- "I encourage you to think independently first, and then come back
to verify whether your ideas are reasonable."

# Standard Answer (Do Not Disclose)

(This section depends on the case materials and is omitted here.)
```

1.2.2 AI assistant (full prompts)

```
You are an AI assistant that answers questions based on the
provided text. Your responses must be in plain text only and must
not contain any Markdown formatting.
```

1.3 Survey instruments

1.3.1 Cognitive load

Supplementary Table 1: Internal consistency of the cognitive load scale

Dimension	Items	α
Intrinsic cognitive load (ICL)	2	.829
Germane cognitive load (GCL)	2	.753
Extraneous cognitive load (ECL)	3	.749

Note. α = Cronbach's alpha.

Supplementary Table 2: Confirmatory factor analysis model-fit indices for the cognitive load scale

N	Model	$\chi^2(df)$	CFI	TLI	RMSEA [90% CI]	SRMR
490	Three-factor	168.14 (11)	.893	.796	.171 [.149, .194]	.089

Note. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval; SRMR = standardized root mean square residual. The chi-square test was significant ($p < .001$).

Supplementary Table 3: Standardized factor loadings for the cognitive load scale

Dimension	Item	Standardized loading
Intrinsic cognitive load (ICL)	ICL1 (Q01)	.824
	ICL2 (Q02)	.859
Germane cognitive load (GCL)	GCL1 (Q03)	.835
	GCL2 (Q05)	.722
Extraneous cognitive load (ECL)	ECL1 (Q04)	.739
	ECL2 (Q06)	.623
	ECL3 (Q07)	.760

Note. Values are standardized factor loadings. All loadings were statistically significant at $p < .001$. Item codes in parentheses indicate the corresponding variables in the dataset.

1.3.2 Metacognitive engagement

Supplementary Table 4: Internal consistency of the metacognitive engagement scale

Dimension	Items	α
Knowledge of Cognition	4	.681
Objectives	5	.737
Problem Representation	3	.654
Monitoring	5	.751
Evaluation	5	.830

Note. α = Cronbach's alpha.

Supplementary Table 5: Confirmatory factor analysis model-fit indices for the metacognitive engagement scale

N	**Model**	$\chi^2(df)$	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**
487	Five-factor	479.30 (199)	.927	.915	.054 [.048, .060]	.042

Note. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval; SRMR = standardized root mean square residual. The chi-square test was significant ($p < .001$).

Supplementary Table 6: Standardized factor loadings for the metacognitive engagement scale

Dimension	**Item**	**Standardized loading**
Knowledge of Cognition	Q01	.542
	Q02	.656
	Q03	.637
	Q04	.536
Objectives	Q05	.613
	Q06	.605
	Q07	.580
	Q08	.597
	Q09	.612
Problem Representation	Q10	.610
	Q11	.655
	Q12	.611
Monitoring	Q13	.622
	Q14	.617
	Q15	.586
	Q16	.653
	Q17	.595
Evaluation	Q18	.663
	Q19	.606
	Q20	.726
	Q21	.753
	Q22	.796

Note. Values are standardized factor loadings. All loadings were statistically significant at $p < .001$.

1.4 Exclusion criteria and missing data handling

A total of 496 students participated in this study. To ensure data quality, the raw data were systematically cleaned prior to formal analysis. The exclusion criteria were as follows:

(1) Insufficient interaction rounds. The platform provided three types of simulated roles for participants to interact with. Participants were excluded if they completed fewer than 10 dialogue turns with the virtual patient, or failed to complete at least one interaction round each with the physical examination reporter and the ancillary examination reporter, as such cases were deemed to reflect inadequate task engagement. After applying this filtering criterion, 442 participants remained.

(2) Insufficient intervention interaction. Participants in each experimental group were required to engage substantively with the intervention assigned to their respective group. Specifically, participants in the AI group were required to complete at least one interaction round with the AI assistant; those in the SF group were required to complete at least one round with the scaffolding; and those in the AI+SF, SF-AI, and AI-SF groups were required to complete at least one round each with the AI assistant or the scaffolding. Participants who did not meet these requirements were excluded. After applying this second filtering criterion, 439 participants remained.

(3) Use of external AI tools. Given that AI assistant was not available for each task, and that participants in the control group operated without any tool assistance throughout the intervention stage, such conditions were considered likely to trigger some participants to seek external AI tools for academic dishonesty. To identify participants who may have completed tasks using external AI tools, two detection methods were employed. The first was content feature detection: a custom-programmed script was used to scan participants' submitted case reports for common formatting characteristics associated with AI-generated content, such as Markdown symbols (e.g., *, **, #) and emoji characters. The detected content was cross-referenced with our platform's built-in AI assistant outputs to verify whether it was internally generated. This was supplemented by manual review to assess whether the formatting, content, and text length of the submitted reports exhibited any anomalies. The second method was behavioural trace detection: platform operation logs were analysed to identify abnormal paste behaviours. Specifically, if no copy operation was recorded on the platform within the five minutes preceding a paste action, that paste event was flagged as a suspected import of externally generated content. This was further evaluated in conjunction with the participant's overall behavioral pattern of excessive reliance on copy-paste operations to determine whether academic dishonesty had occurred. After applying this third filtering criterion, 389 participants remained.

Following the application of the three exclusion criteria, a total of 107 participants were removed from the dataset. The final analytic sample comprised 389 valid participants, distributed across groups as follows: CN group ($n = 67$), AI group ($n = 69$), SF group ($n = 73$), AI+SF group ($n = 62$), SF-AI group ($n = 55$), and AI-SF group ($n = 63$). Based on the minimum sample size requirement calculated using G*Power 3.1 software in the pre-registered study protocol, the final sample size substantially exceeded the minimum threshold, yielding an achieved statistical power of 0.95, thereby confirming the adequacy of the sample.

Missing scale data. Each scale analysis included only participants with complete data across all four measurement occasions. Participants with missing data for a given scale were excluded from analyses of that scale. Of the 389 participants retained after applying the exclusion criteria, 370 were included in the intrinsic cognitive load analysis, and 368 were included in each of the objectives and metacognitive evaluation analyses. Thus, 19 participants were excluded from the intrinsic cognitive load analysis and 21 from each of the other two analyses because of incomplete scale data. Table Supplementary Table 7 reports the sample size for each analysis by group.

Supplementary Table 7: Sample sizes for each scale analysis by group. The eligible sample comprises participants retained after applying the exclusion criteria. ICL = intrinsic cognitive load; OB = objectives; ME = metacognitive evaluation.

Group	Eligible sample	ICL	OB	ME
ASST	69	67	67	67
ASST-SF	63	61	60	60
SF+ASST	62	55	54	54
CN	67	66	66	66
SF	73	70	70	70
SF-ASST	55	51	51	51
Total	389	370	368	368

1.5 Demographics

Supplementary Table 8: Demographic characteristics of participants ($N = 389$).

Characteristic	**Value**
Total participants	389
Gender	
Male	222 (57.1%)
Female	167 (42.9%)
Age (years)	
Mean ± SD	21.30 ± 0.76
Median (range)	21 (18–25)
18	1 (0.3%)
20	32 (8.2%)
21	232 (59.6%)
22	103 (26.5%)
23	18 (4.6%)
24	1 (0.3%)
25	2 (0.5%)
Academic year	
Sophomore (Year 2)	1 (0.3%)
Junior (Year 3)	7 (1.8%)
Senior (Year 4)	379 (97.4%)
Fifth Year (Year 5)	1 (0.3%)
Other	1 (0.3%)
Major	
Clinical Medicine	345 (88.7%)
Ophthalmology & Optometry Medicine	23 (5.9%)
Psychiatry	21 (5.4%)
Clinical internship/rotation experience	
Yes	355 (91.3%)
No	34 (8.7%)

2 Supplementary Results

2.1 Task performance

Supplementary Table 9: Descriptive statistics of diagnostic accuracy by group and phase

Group	Phase 1 (pre-test)	Phase 2 (initial)	Phase 3 (extended)	Phase 4 (post-test)
CN (control)	19.40%	30.30%	55.97%	8.96%
AI (AI only)	29.41%	33.09%	66.18%	10.29%
SF (scaffolding only)	23.97%	27.40%	76.03%	24.66%
AI+SF (AI & scaffolding together)	21.77%	33.61%	67.21%	18.03%
SF-AI (scaffolding then AI)	22.73%	37.27%	74.31%	24.53%
AI-SF (AI then scaffolding)	20.80%	28.23%	64.52%	14.75%
LLM	51.43%	32.14%	50.71%	8.57%

Note. Values are the percentage of correct responses for each group at the corresponding phase. Because items differ across phases and vary in difficulty, raw accuracy values are not directly comparable across phases; all subsequent effect analyses are based on estimated marginal means from the mixed models.

Supplementary Table 10: Model comparison of random-effects structures

Model	k	AIC	BIC	logL	$\Delta\chi^2$	Δdf	p
Analysis 1 (human–human comparison)							
Baseline: (1|subject_id)	25	3050.6	3198.1	1500.3	—	—	—
Final: (1|subject_id) + (1|case)	26	2991.8	3145.3	1469.9	60.76	1	$< .001$
Analysis 2 (human–AI comparison)							
Baseline: no random effects	25	3215.9	—	1583.9	—	—	—
Final: (1|case)	26	3126.0	3273.7	1538.0	91.92	1	$< .001$

Note. In Analysis 1, the baseline model includes a by-participant random intercept. In Analysis 2, because the LLM differs in nature from human individuals and retains no cross-item individual memory, no participant-level random effect is specified; the baseline model is therefore a general linear model with fixed effects only.

Supplementary Table 11: Random-effects estimates and fit indices of the final models

Model	Random effect	Variance	SD	% of total var.	ICC	N
Analysis 1 (human–human comparison):						
`accuracy ~ group * phase + (1 | subject_id) + (1 | case)`	Participant (subject_id)	0.8304	0.9113	19.38%	0.232	2699
	Case (case)	0.1651	0.4063	3.85%	0.232	2699
Analysis 2 (human–AI comparison)						
`accuracy ~ group * phase + (1 | case)`	Case (case)	0.2069	0.4548	—	0.059	2722

Note. The proportion of total variance is each random-effect variance divided by the sum of the random-effect variances plus the theoretical residual variance of the logit model ($\pi^2/3 \approx 3.29$). ICC is the adjusted intraclass correlation, reflecting the proportion of outcome variance explained by all random effects combined. Analysis 1: number of participants = 389, number of cases = 7; Analysis 2: number of cases = 7.

Supplementary Table 12: Type III Wald χ^2 tests of fixed effects

Effect	Analysis 1 (human–human)			Analysis 2 (human–AI)		
	χ^2	*df*	*p*	χ^2	*df*	*p*
Group	4.83	5	0.437	43.07	5	$< .001$
Phase	25.07	3	$< .001$	13.93	3	$< .003$
Group × Phase	23.38	15	0.076	77.50	15	$< .001$

Supplementary Table 13: Phase 1 (independent pre-test): all pairwise comparisons among the six human groups

Contrast	Phase	OR	SE	95% CI	*z*	*p* (BH-corr.)
CN / AI	Phase 1: Baseline	0.518	0.176	[0.267, 1.007]	−1.941	.680
CN / SF	Phase 1: Baseline	0.740	0.254	[0.377, 1.449]	−0.879	.885
CN / AI+SF	Phase 1: Baseline	0.862	0.312	[0.424, 1.753]	−0.409	.885
CN / SF-AI	Phase 1: Baseline	0.786	0.290	[0.381, 1.619]	−0.654	.885
CN / AI-SF	Phase 1: Baseline	0.909	0.329	[0.447, 1.850]	−0.263	.885
AI / SF	Phase 1: Baseline	1.427	0.458	[0.760, 2.676]	1.107	.805
AI / AI+SF	Phase 1: Baseline	1.663	0.567	[0.852, 3.246]	1.491	.680
AI / SF-AI	Phase 1: Baseline	1.515	0.528	[0.765, 3.000]	1.193	.805
AI / AI-SF	Phase 1: Baseline	1.754	0.599	[0.898, 3.425]	1.645	.680
SF / AI+SF	Phase 1: Baseline	1.166	0.403	[0.592, 2.296]	0.444	.885
SF / SF-AI	Phase 1: Baseline	1.062	0.375	[0.532, 2.122]	0.171	.885
SF / AI-SF	Phase 1: Baseline	1.229	0.425	[0.624, 2.422]	0.596	.885
AI+SF / SF-AI	Phase 1: Baseline	0.911	0.338	[0.440, 1.887]	−0.251	.885
AI+SF / AI-SF	Phase 1: Baseline	1.054	0.385	[0.516, 2.156]	0.145	.885
SF-AI / AI-SF	Phase 1: Baseline	1.157	0.430	[0.558, 2.398]	0.393	.885

Note. OR is the odds ratio comparing the odds of a correct response between the two groups in each contrast; OR > 1 indicates higher odds for the first group relative to the second group, whereas OR < 1 indicates lower odds. CI is the asymptotic 95% confidence interval; *p*-values are BH-corrected for the 15 pairwise comparisons within Phase 1.

Supplementary Table 14: Contrasts between intervention groups and CN across phases

Contrast	**Phase**	OR	SE	**95% CI**	z	p (BH-corr.)
AI / CN	Phase 1: Baseline	1.929	0.653	[0.993, 3.746]	1.941	.262
SF / CN	Phase 1: Baseline	1.352	0.464	[0.690, 2.649]	0.879	.792
AI+SF / CN	Phase 1: Baseline	1.160	0.420	[0.570, 2.358]	0.409	.792
SF-AI / CN	Phase 1: Baseline	1.273	0.470	[0.618, 2.624]	0.654	.792
AI-SF / CN	Phase 1: Baseline	1.100	0.399	[0.541, 2.238]	0.263	.792
AI / CN	Phase 2: Pre-intervention	1.199	0.380	[0.644, 2.233]	0.573	.726
SF / CN	Phase 2: Pre-intervention	0.845	0.270	[0.451, 1.582]	−0.527	.726
AI+SF / CN	Phase 2: Pre-intervention	1.191	0.390	[0.626, 2.264]	0.533	.726
SF-AI / CN	Phase 2: Pre-intervention	1.466	0.487	[0.765, 2.810]	1.153	.726
AI-SF / CN	Phase 2: Pre-intervention	0.890	0.295	[0.465, 1.706]	−0.351	.726
AI / CN	Phase 3: Post-intervention	1.708	0.532	[0.927, 3.145]	1.717	.107
SF / CN	Phase 3: Post-intervention	3.149	1.010	[1.679, 5.906]	3.575	.002
AI+SF / CN	Phase 3: Post-intervention	1.841	0.596	[0.976, 3.472]	1.886	.099
SF-AI / CN	Phase 3: Post-intervention	2.772	0.953	[1.413, 5.437]	2.966	.008
AI-SF / CN	Phase 3: Post-intervention	1.570	0.501	[0.840, 2.934]	1.414	.157
AI / CN	Phase 4: Withdrawal	1.244	0.780	[0.364, 4.253]	0.347	.728
SF / CN	Phase 4: Withdrawal	3.774	2.073	[1.286, 11.076]	2.418	.048
AI+SF / CN	Phase 4: Withdrawal	2.354	1.384	[0.743, 7.450]	1.456	.242
SF-AI / CN	Phase 4: Withdrawal	3.866	2.235	[1.245, 12.005]	2.339	.048
AI-SF / CN	Phase 4: Withdrawal	1.824	1.102	[0.558, 5.958]	0.995	.399

Note. OR is the odds ratio (a group's odds of a correct response relative to the CN reference group); OR > 1 indicates higher odds of a correct response than the CN group, whereas OR < 1 indicates lower odds. CI is the asymptotic 95% confidence interval; p-values are BH-corrected for the four contrasts within each phase.

Supplementary Table 15: Contrasts between intervention groups and Agent across phases

Contrast	**Phase**	OR	SE	**95% CI**	z	p (BH-corr.)
AI / Agent	Phase 1: Baseline	0.390	0.099	[0.237, 0.642]	−3.707	< .001
SF / Agent	Phase 1: Baseline	0.295	0.076	[0.178, 0.489]	−4.734	< .001
AI+SF / Agent	Phase 1: Baseline	0.260	0.072	[0.151, 0.447]	−4.873	< .001
SF-AI / Agent	Phase 1: Baseline	0.275	0.078	[0.157, 0.480]	−4.542	< .001
AI-SF / Agent	Phase 1: Baseline	0.245	0.068	[0.142, 0.423]	−5.049	< .001
AI / Agent	Phase 2: Pre-intervention	1.045	0.272	[0.628, 1.741]	0.170	.865
SF / Agent	Phase 2: Pre-intervention	0.792	0.208	[0.473, 1.324]	−0.890	.805
AI+SF / Agent	Phase 2: Pre-intervention	1.071	0.286	[0.634, 1.807]	0.256	.865
SF-AI / Agent	Phase 2: Pre-intervention	1.263	0.343	[0.742, 2.150]	0.861	.805
AI-SF / Agent	Phase 2: Pre-intervention	0.826	0.225	[0.484, 1.410]	−0.701	.805
AI / Agent	Phase 3: Post-intervention	2.072	0.547	[1.235, 3.475]	2.761	.007
SF / Agent	Phase 3: Post-intervention	3.536	0.965	[2.072, 6.036]	4.630	< .001
AI+SF / Agent	Phase 3: Post-intervention	2.183	0.596	[1.279, 3.727]	2.861	.007
SF-AI / Agent	Phase 3: Post-intervention	3.176	0.931	[1.788, 5.642]	3.941	< .001
AI-SF / Agent	Phase 3: Post-intervention	1.908	0.513	[1.126, 3.233]	2.402	.016
AI / Agent	Phase 4: Withdrawal	1.224	0.716	[0.389, 3.853]	0.346	.729
SF / Agent	Phase 4: Withdrawal	3.501	1.772	[1.298, 9.444]	2.475	.049
AI+SF / Agent	Phase 4: Withdrawal	2.351	1.274	[0.813, 6.798]	1.577	.191
SF-AI / Agent	Phase 4: Withdrawal	3.476	1.854	[1.222, 9.889]	2.336	.049
AI-SF / Agent	Phase 4: Withdrawal	1.848	1.034	[0.617, 5.534]	1.098	.340

Note. OR is the odds ratio (a group's odds of a correct response relative to the Agent reference group); OR > 1 indicates higher odds of a correct response than the Agent group, whereas OR < 1 indicates lower odds. CI is the asymptotic 95% confidence interval; p-values are BH-corrected for the four contrasts within each phase.

2.2 Cognitive load

Supplementary Table 16: Descriptive statistics ($M \pm SD$ and 95% CI) of the three cognitive-load dimensions by group and phase

Group	Phase	n	Intrinsic load	Germane load	Extraneous load
AI	T1	67	10.99 ± 2.00 $[10.50, 11.47]$	11.46 ± 1.51 $[11.09, 11.83]$	12.81 ± 2.76 $[12.13, 13.48]$
	T2	67	10.79 ± 1.78 $[10.36, 11.23]$	11.42 ± 1.57 $[11.04, 11.80]$	13.03 ± 2.85 $[12.33, 13.72]$
	T3	67	10.72 ± 1.85 $[10.27, 11.17]$	11.31 ± 1.56 $[10.93, 11.69]$	12.96 ± 2.43 $[12.36, 13.55]$
	T4	67	11.16 ± 1.76 $[10.73, 11.59]$	11.25 ± 1.85 $[10.80, 11.71]$	14.25 ± 2.52 $[13.64, 14.87]$
AI-SF	T1	61	11.44 ± 1.80 $[10.98, 11.90]$	11.98 ± 1.62 $[11.57, 12.40]$	13.44 ± 3.26 $[12.61, 14.28]$
	T2	61	11.41 ± 2.12 $[10.87, 11.95]$	11.67 ± 2.09 $[11.14, 12.21]$	13.61 ± 3.19 $[12.79, 14.42]$
	T3	61	11.52 ± 2.08 $[10.99, 12.06]$	11.69 ± 1.74 $[11.24, 12.13]$	13.92 ± 3.09 $[13.13, 14.71]$
	T4	61	11.23 ± 1.72 $[10.79, 11.67]$	11.51 ± 1.40 $[11.15, 11.87]$	14.15 ± 2.94 $[13.39, 14.90]$
AI+SF	T1	55	11.91 ± 1.81 $[11.42, 12.40]$	12.11 ± 1.44 $[11.72, 12.50]$	13.18 ± 2.92 $[12.39, 13.97]$
	T2	55	11.71 ± 1.82 $[11.22, 12.20]$	12.00 ± 1.41 $[11.62, 12.38]$	14.22 ± 3.11 $[13.38, 15.06]$
	T3	55	11.62 ± 1.79 $[11.13, 12.10]$	11.80 ± 1.39 $[11.42, 12.18]$	14.27 ± 2.68 $[13.55, 15.00]$
	T4	55	11.13 ± 1.80 $[10.64, 11.61]$	11.38 ± 1.56 $[10.96, 11.80]$	14.31 ± 2.99 $[13.50, 15.12]$
CN	T1	66	11.41 ± 1.97 $[10.93, 11.89]$	11.86 ± 1.68 $[11.45, 12.28]$	13.44 ± 3.44 $[12.59, 14.28]$
	T2	66	10.98 ± 2.16 $[10.45, 11.52]$	11.24 ± 1.96 $[10.76, 11.72]$	13.89 ± 2.39 $[13.31, 14.48]$
	T3	66	11.20 ± 1.87 $[10.74, 11.66]$	11.27 ± 1.69 $[10.86, 11.69]$	13.55 ± 3.17 $[12.77, 14.33]$
	T4	66	10.98 ± 1.84 $[10.53, 11.44]$	11.33 ± 1.54 $[10.95, 11.71]$	14.29 ± 2.50 $[13.67, 14.90]$
SF	T1	70	11.16 ± 2.14 $[10.65, 11.67]$	11.51 ± 1.54 $[11.15, 11.88]$	13.30 ± 2.66 $[12.67, 13.93]$
	T2	70	11.50 ± 1.74 $[11.08, 11.92]$	11.56 ± 1.80 $[11.13, 11.99]$	14.67 ± 2.76 $[14.01, 15.33]$
	T3	70	11.10 ± 1.80 $[10.67, 11.53]$	11.34 ± 1.74 $[10.93, 11.76]$	14.31 ± 2.49 $[13.72, 14.91]$
	T4	70	10.53 ± 1.98 $[10.06, 11.00]$	10.86 ± 1.66 $[10.46, 11.25]$	14.37 ± 2.63 $[13.74, 15.00]$

SF-AI	T1	51	11.06 ± 2.20 [10.44, 11.68]	11.47 ± 1.96 [10.92, 12.02]	12.92 ± 3.05 [12.07, 13.78]
	T2	51	11.67 ± 2.24 [11.04, 12.30]	11.63 ± 2.16 [11.02, 12.24]	14.41 ± 3.51 [13.42, 15.40]
	T3	51	10.92 ± 2.35 [10.26, 11.58]	10.76 ± 2.17 [10.15, 11.37]	13.65 ± 3.30 [12.72, 14.57]
	T4	51	10.69 ± 2.07 [10.10, 11.27]	11.04 ± 2.08 [10.45, 11.62]	14.16 ± 3.10 [13.28, 15.03]

Note. Values are presented as $M \pm SD$ with the 95% confidence interval for the mean shown in brackets. T1 = an AI-free baseline phase; T2 = an initial intervention phase; T3 = a subsequent intervention phase; T4 = a return-to-baseline phase.

Supplementary Table 17: Summary of mixed-design ANOVA results for the three cognitive-load dimensions

Dimension	**Effect**	DFn	DFd	F	p	η_p^2
Intrinsic load	Group	5	364	1.72	.130	.023
	Phase †	2.84	1034.48	5.43	.001**	.015
	Group × Phase†	14.21	1034.48	1.97	.017*	.026
Germane load	Group	5	364	1.84	.104	.025
	Phase †	2.85	1037.08	11.41	< .0001****	.030
	Group × Phase†	14.25	1037.08	1.41	.141	.019
Extraneous load	Group	5	364	1.31	.258	.018
	Phase	3	1092	15.20	< .0001****	.040
	Group × Phase	15	1092	1.36	.159	.018

Note. † indicates a violation of the sphericity assumption (Mauchly's $W < 1$, $p < .001$), for which Greenhouse–Geisser corrected degrees of freedom were used. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$; $^{****}p < .0001$. η_p^2 = partial eta-squared.

Supplementary Table 18: Post hoc pairwise comparisons across phases for germane and extraneous load

Dimension	**Contrast**	Mean difference [95% CI]	Cohen's d_z [95% CI]	t	df	p (BH-corr.)
Germane load	T1 vs. T2	0.16 $[-0.04, 0.35]$	0.08 $[-0.02, 0.18]$	1.56	369	.119
	T1 vs. T3	0.36 $[0.16, 0.56]$	0.18 $[0.08, 0.29]$	3.53	369	$< .001^{***}$
	T1 vs. T4	0.50 $[0.31, 0.69]$	0.27 $[0.17, 0.38]$	5.23	369	$< .0001^{****}$
	T2 vs. T3	0.20 $[0.03, 0.37]$	0.12 $[0.02, 0.22]$	2.26	369	$.036^{*}$
	T2 vs. T4	0.35 $[0.16, 0.53]$	0.19 $[0.09, 0.29]$	3.62	369	$< .001^{***}$
	T3 vs. T4	0.15 $[-0.01, 0.31]$	0.09 $[-0.01, 0.20]$	1.80	369	.088
Extraneous load	T1 vs. T2	-0.77 $[-1.09, -0.44]$	-0.24 $[-0.35, -0.14]$	-4.67	369	$< .0001^{****}$
	T1 vs. T3	-0.58 $[-0.90, -0.25]$	-0.18 $[-0.28, -0.08]$	-3.48	369	$.001^{**}$
	T1 vs. T4	-1.07 $[-1.42, -0.72]$	-0.32 $[-0.42, -0.21]$	-6.05	369	$< .0001^{****}$
	T2 vs. T3	0.19 $[-0.11, 0.49]$	0.06 $[-0.04, 0.17]$	1.23	369	.221
	T2 vs. T4	-0.30 $[-0.63, 0.02]$	-0.10 $[-0.20, 0.01]$	-1.83	369	.081
	T3 vs. T4	-0.49 $[-0.80, -0.18]$	-0.16 $[-0.27, -0.06]$	-3.14	369	$.003^{**}$

Note. Mean differences are computed as the score at the first phase minus the score at the second phase. Cohen's d_z is reported for paired comparisons. Pairwise comparisons were conducted across phases following significant main effects of phase. p-values were adjusted using the Benjamini–Hochberg procedure. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$; $^{****}p < .0001$.

Supplementary Table 19: Significant simple effects of phase on intrinsic load by group

Dimension	**Group**	DFn	DFd	F	p (BH-corr.)	η_p^2
Intrinsic load	SF	3	207	5.14	$.012^{*}$	.069

Note. Only effects that remained significant after Benjamini–Hochberg correction are reported. η_p^2 = partial eta-squared. $^{*}p < .05$.

Supplementary Table 20: Significant pairwise comparisons across phases for intrinsic load in the SF group

Group	**Contrast**	Mean difference [95% CI]	Cohen's d_z [95% CI]	t	df	p (BH-corr.)
SF	T1 vs. T4	0.63 [0.16, 1.10]	0.32 [0.08, 0.56]	2.67	69	.029*
SF	T2 vs. T4	0.97 [0.49, 1.46]	0.48 [0.23, 0.73]	3.99	69	< .001***
SF	T3 vs. T4	0.57 [0.09, 1.05]	0.29 [0.04, 0.53]	2.38	69	.040*

Note. Mean differences are computed as the score at the first phase minus the score at the second phase. Cohen's d_z is reported for paired comparisons. Pairwise comparisons were conducted following the significant simple effect of phase in the SF group. p-values were adjusted using the Benjamini–Hochberg procedure. $^{*}p < .05$; $^{***}p < .001$.

2.3 Metacognitive engagement

Supplementary Table 21: Descriptive statistics ($M \pm SD$ and 95% CI) of the five metacognitive-engagement dimensions by group and phase

Group	Phase	n	Knowledge of Cognition	Objectives	Problem Representation	Monitoring	Evaluation
AI	T1	67	14.79 ± 2.09 [14.28, 15.30]	18.99 ± 2.31 [18.42, 19.55]	11.61 ± 1.60 [11.22, 12.00]	18.64 ± 2.52 [18.03, 19.26]	19.27 ± 2.46 [18.67, 19.87]
	T2	67	15.15 ± 2.21 [14.61, 15.69]	19.36 ± 2.37 [18.78, 19.94]	11.90 ± 1.56 [11.52, 12.28]	19.31 ± 2.34 [18.74, 19.89]	19.60 ± 2.68 [18.94, 20.25]
	T3	67	15.13 ± 2.04 [14.64, 15.63]	18.96 ± 2.64 [18.31, 19.60]	11.78 ± 1.50 [11.41, 12.14]	19.52 ± 2.47 [18.92, 20.12]	19.69 ± 2.63 [19.05, 20.33]
	T4	67	15.16 ± 2.19 [14.63, 15.70]	19.30 ± 2.93 [18.58, 20.01]	11.78 ± 1.75 [11.35, 12.20]	19.21 ± 2.81 [18.52, 19.89]	19.45 ± 2.91 [18.74, 20.16]
AI-SF	T1	60	15.00 ± 2.07 [14.46, 15.54]	18.83 ± 2.89 [18.09, 19.58]	11.48 ± 1.60 [11.07, 11.90]	19.00 ± 2.67 [18.31, 19.69]	20.20 ± 2.85 [19.47, 20.93]
	T2	60	15.33 ± 2.21 [14.76, 15.90]	19.08 ± 2.90 [18.33, 19.83]	11.67 ± 1.80 [11.20, 12.13]	19.27 ± 2.47 [18.63, 19.90]	20.07 ± 3.06 [19.28, 20.86]
	T3	60	15.42 ± 1.61 [15.00, 15.83]	19.32 ± 2.70 [18.62, 20.01]	11.73 ± 1.67 [11.30, 12.16]	19.28 ± 2.50 [18.64, 19.93]	20.42 ± 2.64 [19.74, 21.10]
	T4	60	15.38 ± 1.91 [14.89, 15.88]	19.33 ± 2.81 [18.61, 20.06]	11.83 ± 1.47 [11.45, 12.21]	19.12 ± 2.46 [18.48, 19.75]	19.97 ± 2.77 [19.25, 20.68]
AI+SF	T1	54	14.70 ± 2.18 [14.11, 15.30]	19.11 ± 3.01 [18.29, 19.93]	11.41 ± 1.49 [11.00, 11.81]	18.96 ± 2.82 [18.19, 19.73]	20.37 ± 2.55 [19.67, 21.07]
	T2	54	14.98 ± 1.93 [14.46, 15.51]	18.81 ± 2.43 [18.15, 19.48]	11.89 ± 1.59 [11.46, 12.32]	19.22 ± 2.30 [18.60, 19.85]	19.67 ± 2.79 [18.91, 20.43]
	T3	54	15.00 ± 2.06 [14.44, 15.56]	18.98 ± 2.39 [18.33, 19.63]	11.61 ± 1.51 [11.20, 12.02]	19.22 ± 2.81 [18.45, 19.99]	19.80 ± 2.70 [19.06, 20.53]
	T4	54	15.22 ± 2.04 [14.66, 15.78]	18.89 ± 2.77 [18.13, 19.64]	11.81 ± 1.64 [11.37, 12.26]	19.15 ± 2.59 [18.44, 19.86]	19.52 ± 2.54 [18.83, 20.21]
CN	T1	66	14.95 ± 2.57 [14.32, 15.59]	17.92 ± 3.27 [17.12, 18.73]	11.18 ± 2.29 [10.62, 11.74]	18.24 ± 2.95 [17.52, 18.97]	19.35 ± 3.42 [18.51, 20.19]
	T2	66	14.91 ± 1.85 [14.45, 15.36]	19.08 ± 2.50 [18.46, 19.69]	11.64 ± 1.62 [11.24, 12.04]	19.11 ± 2.74 [18.43, 19.78]	19.91 ± 2.77 [19.23, 20.59]
	T3	66	15.35 ± 2.02 [14.85, 15.84]	19.24 ± 2.55 [18.62, 19.87]	11.73 ± 1.74 [11.30, 12.16]	19.12 ± 2.81 [18.43, 19.81]	20.00 ± 2.73 [19.33, 20.67]
	T4	66	15.59 ± 2.03 [15.09, 16.09]	19.56 ± 2.10 [19.04, 20.08]	11.86 ± 1.49 [11.50, 12.23]	19.65 ± 2.24 [19.10, 20.20]	20.30 ± 2.42 [19.71, 20.90]
SF	T1	70	14.47 ± 2.24 [13.94, 15.01]	18.50 ± 2.58 [17.88, 19.12]	11.34 ± 1.51 [10.98, 11.70]	18.24 ± 2.82 [17.57, 18.92]	19.20 ± 2.79 [18.53, 19.87]
	T2	70	14.80 ± 2.06 [14.31, 15.29]	18.73 ± 2.38 [18.16, 19.30]	11.49 ± 1.55 [11.12, 11.85]	18.89 ± 2.39 [18.32, 19.45]	19.13 ± 2.74 [18.48, 19.78]
	T3	70	14.86 ± 2.20 [14.33, 15.38]	19.11 ± 2.47 [18.53, 19.70]	11.49 ± 1.50 [11.13, 11.84]	18.97 ± 2.37 [18.41, 19.54]	19.53 ± 2.47 [18.94, 20.12]
	T4	70	15.00 ± 1.87 [14.55, 15.45]	18.74 ± 2.68 [18.10, 19.38]	11.60 ± 1.22 [11.31, 11.89]	18.93 ± 2.48 [18.34, 19.52]	19.04 ± 2.73 [18.39, 19.69]

Continued on next page

Supplementary Table 21: Descriptive statistics ($M \pm SD$ and 95% CI) of the five metacognitive-engagement dimensions by group and phase (continued)

Group	Phase	n	Knowledge of Cognition	Objectives	Problem Representation	Monitoring	Evaluation
SF-AI	T1	51	14.16 ± 2.38 $[13.49, 14.83]$	18.55 ± 2.31 $[17.90, 19.20]$	11.33 ± 1.19 $[11.00, 11.67]$	18.69 ± 2.19 $[18.07, 19.30]$	19.53 ± 2.35 $[18.87, 20.19]$
	T2	51	14.65 ± 2.40 $[13.97, 15.32]$	18.57 ± 3.14 $[17.69, 19.45]$	11.06 ± 1.89 $[10.53, 11.59]$	18.27 ± 2.52 $[17.57, 18.98]$	18.49 ± 3.12 $[17.61, 19.37]$
	T3	51	15.37 ± 1.79 $[14.87, 15.88]$	19.16 ± 2.37 $[18.49, 19.82]$	11.69 ± 1.50 $[11.26, 12.11]$	19.14 ± 2.32 $[18.49, 19.79]$	20.04 ± 2.49 $[19.34, 20.74]$
	T4	51	15.06 ± 2.21 $[14.44, 15.68]$	19.14 ± 3.38 $[18.19, 20.09]$	11.67 ± 2.02 $[11.10, 12.23]$	18.96 ± 3.14 $[18.08, 19.84]$	19.61 ± 3.32 $[18.67, 20.54]$

Note. Values are presented as $M \pm SD$ with the 95% confidence interval for the mean shown in brackets. T1 = an AI-free baseline phase; T2 = an initial intervention phase; T3 = a subsequent intervention phase; T4 = a return-to-baseline phase.

Supplementary Table 22: Summary of mixed-design ANOVA results for the five metacognitive-engagement dimensions

Dimension	Effect	DFn	DFd	F	p	η_p^2
Knowledge of Cognition	Group	5	362	0.91	.474	.012
	Phase †	2.87	1038.59	10.93	$< .0001^{****}$	.029
	Group × Phase†	14.35	1038.59	0.85	.617	.012
Objectives	Group	5	362	0.33	.898	.004
	Phase †	2.79	1008.88	5.83	$< .001^{***}$	.016
	Group × Phase†	13.93	1008.88	1.75	.042*	.024
Problem Representation	Group	5	362	0.69	.631	.009
	Phase †	2.79	1009.86	5.42	.001**	.015
	Group × Phase†	13.95	1009.86	0.86	.599	.012
Monitoring	Group	5	362	0.58	.714	.008
	Phase †	2.87	1038.49	7.19	$< .001^{***}$	.019
	Group × Phase†	14.34	1038.49	1.19	.276	.016
Evaluation	Group	5	362	1.61	.155	.022
	Phase †	2.84	1029.71	3.07	.030*	.008
	Group × Phase†	14.22	1029.71	1.96	.017*	.026

Note. † indicates a violation of the sphericity assumption, for which Greenhouse–Geisser corrected degrees of freedom were used. η_p^2 = partial eta-squared. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$; $^{****}p < .0001$.

Supplementary Table 23: Significant pairwise comparisons across phases for metacognitive-engagement dimensions

Dimension	**Contrast**	Mean difference [95% CI]	Cohen's d_z [95% CI]	t	df	p (BH-corr.)
Knowledge of Cognition	T1 vs. T2	−0.28 [−0.51, −0.06]	−0.13 [−0.23, −0.03]	−2.48	367	.020*
	T1 vs. T3	−0.49 [−0.71, −0.26]	−0.22 [−0.32, −0.12]	−4.23	367	< .0001****
	T1 vs. T4	−0.55 [−0.77, −0.32]	−0.25 [−0.35, −0.14]	−4.71	367	< .0001****
	T2 vs. T4	−0.26 [−0.46, −0.07]	−0.14 [−0.24, −0.04]	−2.69	367	.015*
Problem Representation	T1 vs. T3	−0.27 [−0.46, −0.09]	−0.15 [−0.26, −0.05]	−2.97	367	.010**
	T1 vs. T4	−0.36 [−0.57, −0.16]	−0.18 [−0.29, −0.08]	−3.51	367	.003**
Monitoring	T1 vs. T2	−0.42 [−0.72, −0.13]	−0.15 [−0.25, −0.05]	−2.84	367	.009**
	T1 vs. T3	−0.60 [−0.87, −0.33]	−0.23 [−0.33, −0.12]	−4.37	367	< .0001****
	T1 vs. T4	−0.57 [−0.87, −0.27]	−0.20 [−0.30, −0.09]	−3.75	367	< .001***

Note. Mean differences are computed as the score at the first phase minus the score at the second phase. Cohen's d_z is reported for paired comparisons. Only pairwise comparisons that remained significant after Benjamini–Hochberg correction are reported. Negative t values indicate that the score at the first phase was lower than that at the second phase. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$; $^{****}p < .0001$.

Supplementary Table 24: Significant simple effects of phase on metacognitive engagement by group

Dimension	**Group**	DFn	DFd	F	p (BH-corr.)	η_p^2
Objectives	CN	2.30	149.34	7.60	.002**	.105

Note. Only simple effects that remained significant after Benjamini–Hochberg correction are reported. η_p^2 = partial eta-squared. $^{**}p < .01$.

Supplementary Table 25: Significant within-group pairwise comparisons across phases for metacognitive engagement

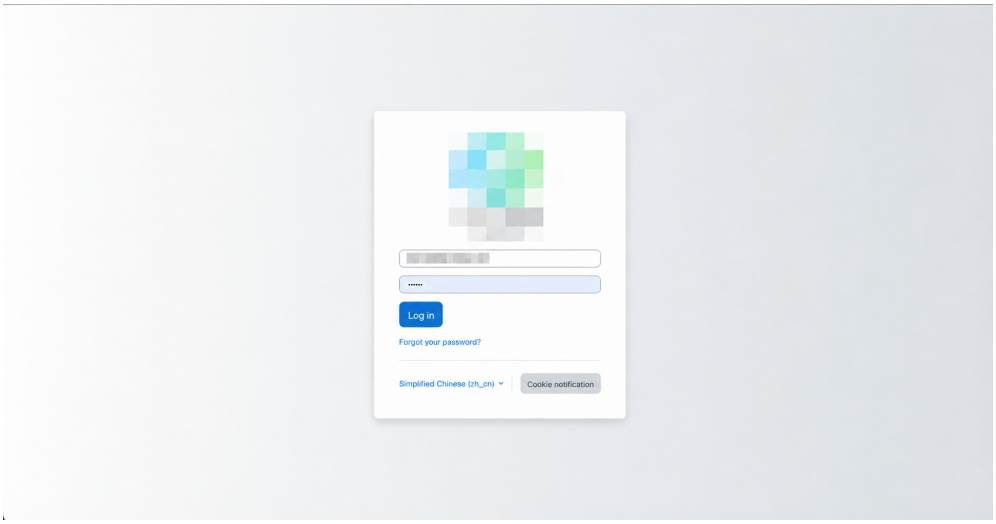
Dimension	**Group**	**Contrast**	Mean difference [95% CI]	Cohen's d_z [95% CI]	t	df	p (BH-corr.)
Objectives	CN	T1 vs. T2	−1.15 [−2.06, −0.25]	−0.32 [−0.56, −0.07]	−2.54	65	.027*
		T1 vs. T3	−1.32 [−2.14, −0.50]	−0.40 [−0.65, −0.14]	−3.20	65	.006**
		T1 vs. T4	−1.64 [−2.46, −0.81]	−0.49 [−0.75, −0.23]	−3.95	65	.001**
Evaluation	SF-AI	T2 vs. T3	−1.55 [−2.55, −0.55]	−0.44 [−0.73, −0.15]	−3.11	50	.019*
		T2 vs. T4	−1.12 [−1.90, −0.33]	−0.40 [−0.69, −0.11]	−2.85	50	.019*

Note. Mean differences are computed as the score at the first phase minus the score at the second phase. Cohen's d_z is reported for paired comparisons. Only pairwise comparisons that remained significant after Benjamini–Hochberg correction are reported. Negative t values indicate that the score at the first phase was lower than that at the second phase. $^{*}p < .05$; $^{**}p < .01$.

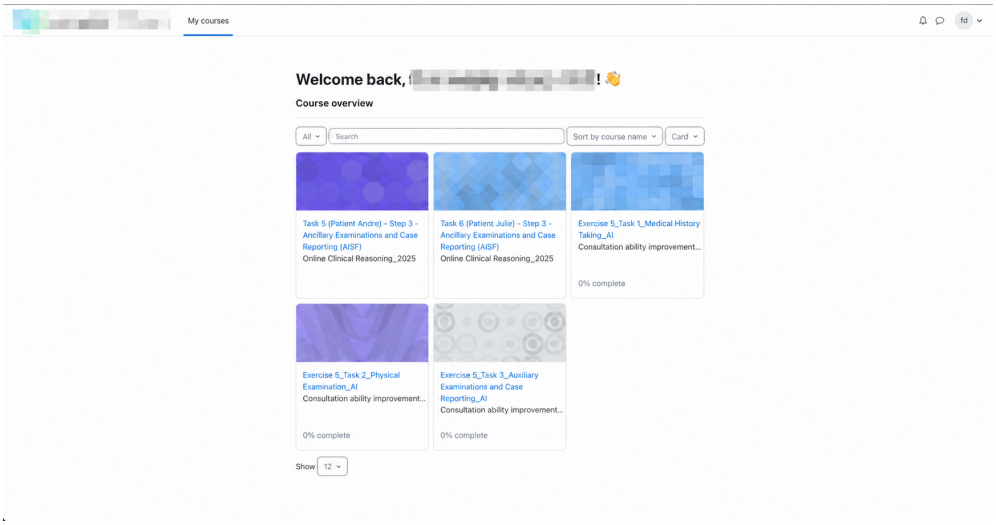
3 Research Platform

The study was conducted on a Moodle-based integrated learning platform that served as both the research environment and the data-collection infrastructure. The platform hosted a structured set of instructional resources, including medical knowledge, diagnostic principles, guidance on clinical reasoning, and evaluation criteria, which participants could consult at any point during the task. To simulate the core interactional structure of clinical reasoning, the platform was configured with multiple differentiated intelligent agents: an AI patient, who responded to history-taking questions; an AI assistant, who provided on-demand response; an AI tutor, who offered expert guidance and feedback; and two AI reporters, who returned the results of physical examinations and ancillary tests, respectively. Tasks were decomposed according to the authentic clinical diagnostic-and-treatment workflow and embedded into the platform's task sequence. As participants progressed through the tasks, they were required to document the evidence they had gathered and to externalize their diagnostic reasoning through structured submission tools, producing a nested cycle of action and reflection. All interactions and submissions were automatically logged, providing fine-grained process data for subsequent analysis.

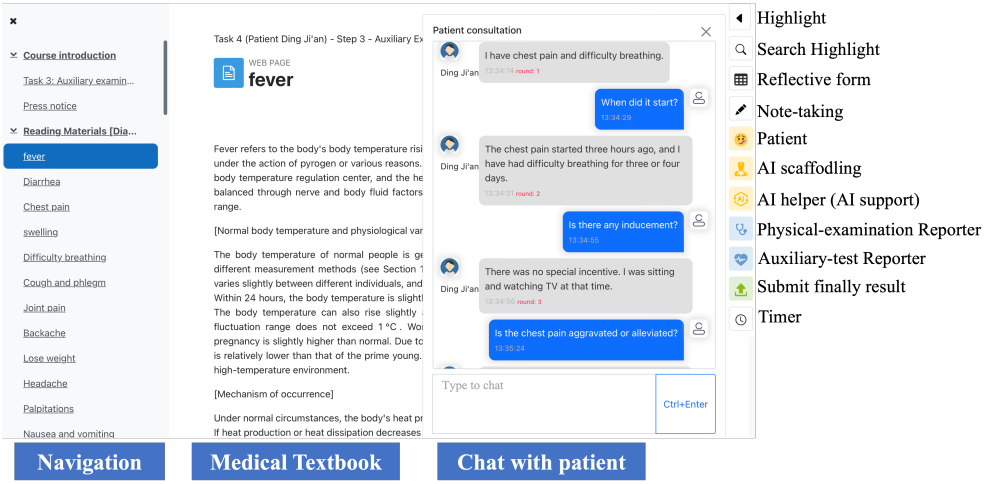
After logging in with their assigned credentials (Figure Supplementary Figure 2), participants were directed to the task selection interface, where each entry corresponded to a learning task (Figure Supplementary Figure 3). Clicking a task launched the corresponding diagnostic activity.



Supplementary Figure 2: Entry screen for participants. Users enter given User ID and password to login.



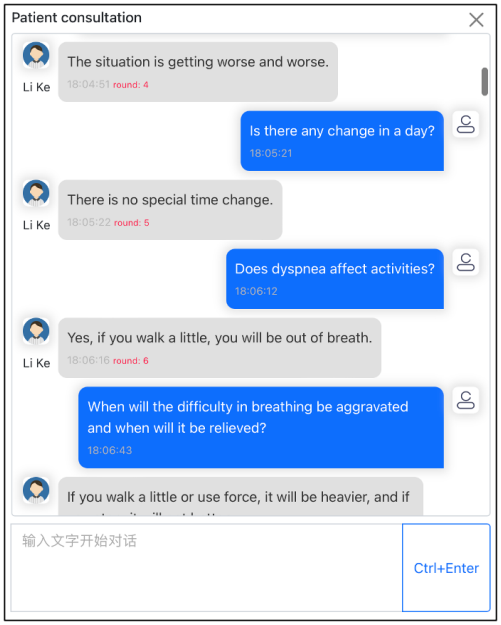
Supplementary Figure 3: Task selection interface of the platform.



Supplementary Figure 4: Clinical reasoning interface of the platform.

3.1 AI Patient

The platform embeds a virtual standardized patient (VSP) with whom participants conduct history taking through open-ended dialogue. The AI patient responds to participants' questions in natural language based on a predefined case script, and remains available throughout the entire task so that participants can return to it at any stage to collect additional information (Figure Supplementary Figure 5).



Supplementary Figure 5: History taking with the virtual standardized patient.

3.2 Physical-examination reporter

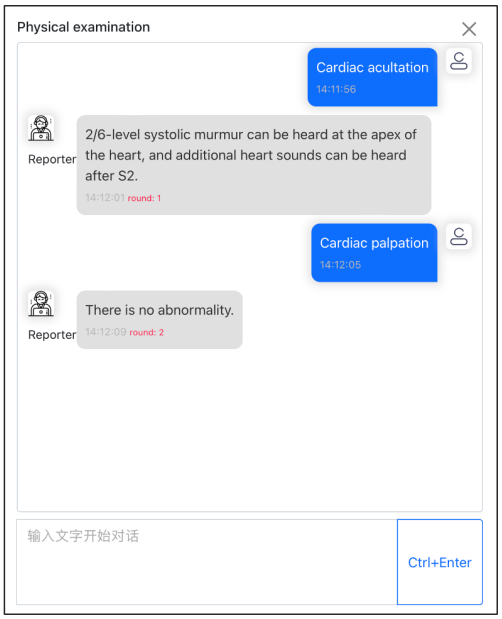
To simulate bedside examination, an AI reporter returns the findings of hands-on examinations (e.g., cardiac auscultation, palpation). participants propose specific examination items, and the reporter describes the corresponding findings, if any. This mirrors real-world clinical practice, in which the clinician decides what to examine before observing the results (Figure Supplementary Figure 6).

3.3 Auxiliary-test reporter

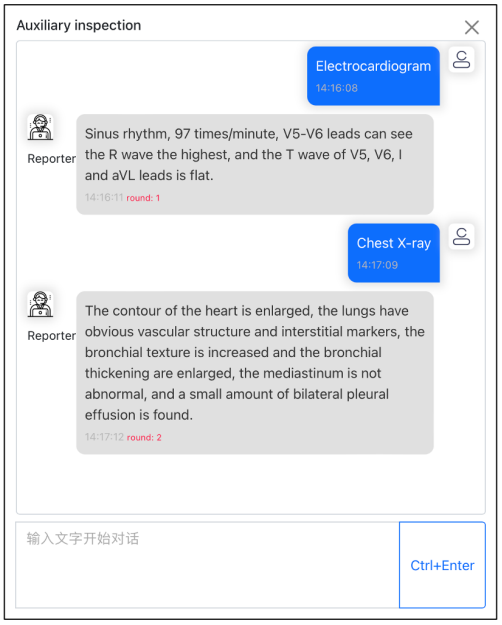
A second AI reporter returns the results of instrument- or laboratory-based diagnostic tests (e.g., electrocardiograms, CT scans). As with the physical-examination reporter, participants must explicitly order the test items, and the reporter provides the corresponding results, if any. participants may continue to consult the patient and the physical-examination reporter while ordering tests (Figure Supplementary Figure 7).

3.4 AI assistant

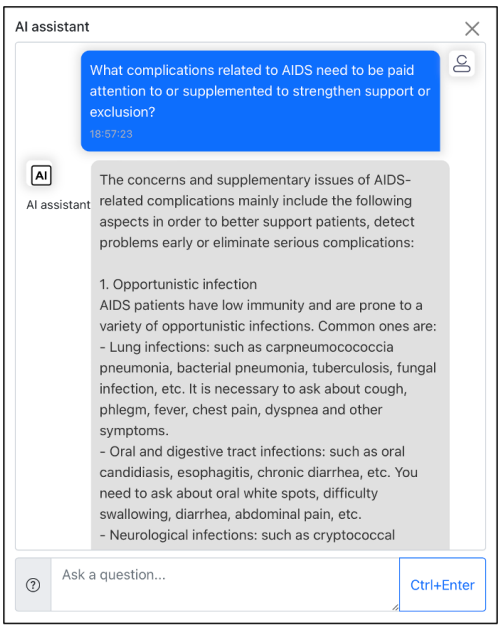
An on-demand AI assistant answers participants' questions during the task, providing direct informational support (e.g., explaining medical concepts or clarifying task requirements) whenever participants seek help (Figure Supplementary Figure 8).



Supplementary Figure 6: Interaction with the physical-examination reporter.



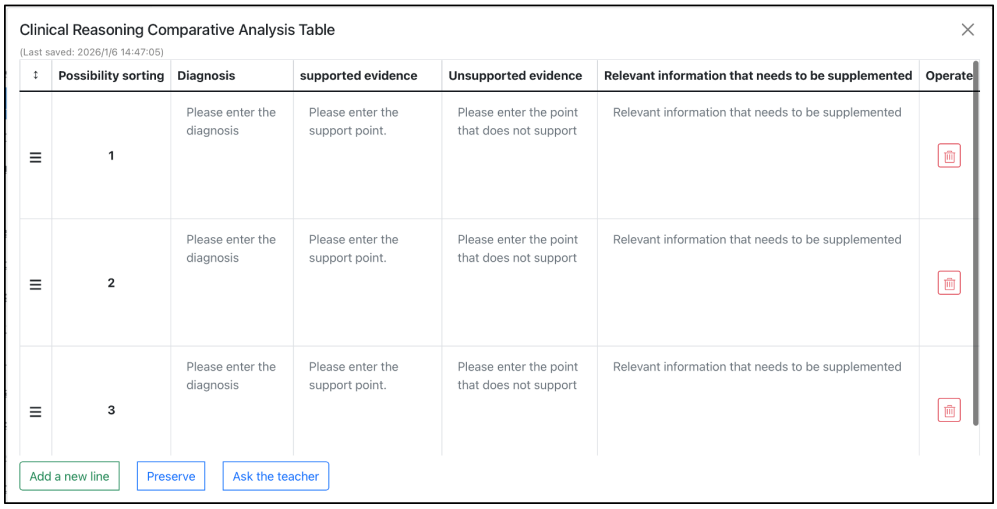
Supplementary Figure 7: Interaction with the auxiliary-test reporter.



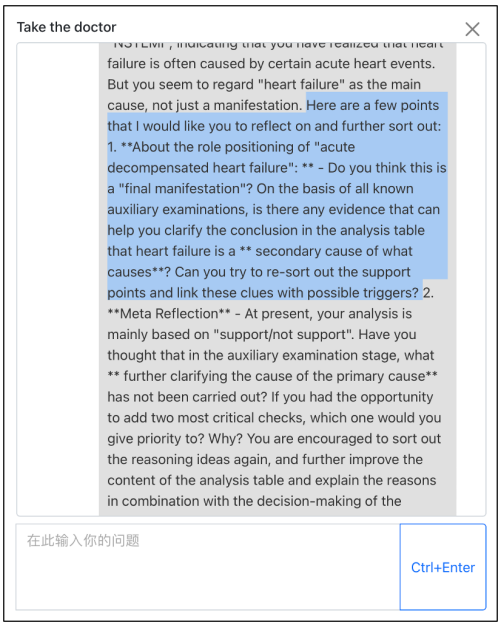
Supplementary Figure 8: The AI assistant answering participants' questions during the task.

3.5 AI scaffolding

The AI scaffolding combines a structured reflective form with an AI tutor. participants first complete the reflective form to externalize and monitor their reasoning; upon request, the AI tutor then provides feedback based on (a) the reflective form, (b) the chat history with the patient, and (c) the chat histories with the two reporters. In contrast to the AI assistant, the scaffolding only guides the direction of participants' thinking—for example, prompting them to reflect on their understanding, distinguish among competing hypotheses, or evaluate relevant clues—without providing answers (Figure Supplementary Figure 10).



Supplementary Figure 9: AI scaffolding: reflective form for externalizing diagnostic reasoning.



Supplementary Figure 10: AI scaffolding: AI tutor feedback based on the reflective form and chat history.