

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

Naping Chen

Shantou University Medical College

Xiaoyu Zhao

University Medical Center Groningen

Luzhen Tang

Graduate School of Education Peking University

Mengyu Xia

Graduate School of Education Peking University

Zijian Li

Graduate School of Education Peking University

Jiaqi Xu

Graduate School of Education Peking University

Xinyu Li

Faculty of Information Technology Monash University

Mingxue Xu

Graduate School of Education Peking University

Liping Lv

Shantou University Medical College

Oudong Xia

Shantou University Medical College

Changmin Lin

Shantou University Medical College

Ligia Maria Cayres Ribeiro

University Medical Center Groningen

Marco Antonio de Carvalho Filho

University Medical Center Groningen

Dragan Gašević

Faculty of Education & School of Computing and Data Science The University of Hong Kong

Danijela Gasevic

School of Public Health The University of Hong Kong

Yizhou Fan

fyz@pku.edu.cn

Graduate School of Education Peking University

Article

Keywords: generative artificial intelligence, metacognition, scaffolding, clinical reasoning, medical students

Posted Date: August 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10472890/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

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

Naping Chen^{1†}, Xiaoyu Zhao^{2,3†}, Luzhen Tang⁴, Mengyu Xia⁴,
Zijian Li⁴, Jiaqi Xu⁴, Xinyu Li⁶, Mingxue Xu⁴, Liping Lv¹,
Oudong Xia¹, Changmin Lin¹, Ligia Maria Cayres Ribeiro²,
Marco Antonio de Carvalho Filho², Dragan Gašević^{5,6},
Danijela Gasevic^{7,8}, Yizhou Fan^{4*}

¹Medical College, Shantou University, Jinping, Shantou, 515041,
Guangdong, China.

²Research Program LEARN (Lifelong Learning, Education and
Assessment Research Network), Wenckebach Instituut voor Onderwijs
en Opleiding, University Medical Center Groningen/ University of
Groningen, Antonius Deusinglaan 1, Groningen, 9713 AV, Groningen,
the Netherlands.

³Institute of Medical Education Research Rotterdam, Erasmus Medical
Center, Burgemeester Oudlaan 50, Rotterdam, 3062 PA, Rotterdam,
the Netherlands.

⁴Graduate School of Education, Peking University, Haidian, Beijing,
100871, Beijing, China.

⁵Faculty of Education & School of Computing and Data Science, The
University of Hong Kong, Pok Fu Lam Rd, Hong Kong, Hong Kong
SAR, China.

⁶Faculty of Information Technology, Monash University, 20 Exhibition
Walk, Clayton, 3800, Victoria, Australia.

⁷School of Public Health, The University of Hong Kong, 7 Sassoon Rd,
Sandy Bay, Hong Kong, Hong Kong SAR, China.

⁸School of Public Health and Preventive Medicine, Monash University,
553 St Kilda Road, Melbourne, 3004, Victoria, Australia.

*Corresponding author(s). E-mail(s): fyz@pku.edu.cn;

Contributing authors: s_chenmp@stu.edu.cn; x.zhao@umcg.nl;
tangluzhen@pku.edu.cn; Marine_xia@163.com; zijianli@stu.pku.edu.cn;
xjq@pku.edu.cn; xinyu.li1@monash.edu; mingxuexu25@stu.pku.edu.cn;
20lplv@alumni.stu.edu.cn; xiaodnf@163.com; cocolin@stu.edu.cn;
l.m.cayres.ribeiro@umcg.nl; m.a.de.carvalho.filho@umcg.nl;
dgasevic@hku.hk; daniijela@hku.hk;

[†]These authors contributed equally to this work.

Abstract

Clinical reasoning underpins clinical expertise, yet fostering its development remains challenging. Effective clinical reasoning requires students to monitor and control how their reasoning develops as a clinical problem unfolds. These processes are commonly described as metacognition. Existing approaches provide limited timely and adaptive support for students’ metacognition as reasoning evolves. Generative artificial intelligence (GenAI) creates opportunities for delivering such support, but how GenAI-mediated metacognitive scaffolding should be designed remains unclear. Using design-based research, we designed and iteratively refined such scaffolding within a simulated history-taking environment across three cycles. Fixed predefined questions functioned mainly as task reminders. Socratic questioning provided broader metacognitive support but was imbalanced and was sometimes experienced as endless and as knowledge testing. Adaptive on-demand scaffolding provided more balanced support and enabled students to select support aligned with their evolving reasoning needs. Effective GenAI-mediated metacognitive scaffolding requires context integration, balanced support, and preservation of student agency.

Keywords: generative artificial intelligence, metacognition, scaffolding, clinical reasoning, medical students

Introduction

Clinical reasoning (CR) is recognised as a core competency in medical education, underpinning safe and effective clinical practice [1, 2]. However, supporting the development of CR remains challenging due to the complexity of the reasoning process and the ill-structured nature of clinical problems [3–5]. Broadly defined, CR encompasses the cognitive processes through which healthcare providers “gather and synthesise information; generate hypotheses; and formulate a clinical impression, prognosis, diagnosis, treatment, care, and/or management plan” [2]. Given the complexity of CR, researchers have drawn on multiple theoretical traditions to explain different aspects of the reasoning process, including those emphasising knowledge organisation, cognitive processing, sociocultural influences, and metacognition [5]. Among these, metacognition has been recognised as an important component underpinning effective CR [6], as students must continuously generate, evaluate, and refine diagnostic hypotheses amid

uncertainty and incomplete information while monitoring the adequacy of their knowledge and reasoning processes [4, 6, 7]. Supporting metacognition during the reasoning process is therefore critical for CR development, as it enables students to identify knowledge gaps, evaluate their reasoning strategies, and adjust their understanding while engaging with clinical problems [6]. Without such timely support, students may have fewer opportunities to identify limitations in their reasoning and modify ineffective approaches. However, although certain instructional approaches, such as metacognitive checklists and deliberate reflection [8–10], have sought to promote metacognition in CR, few interventions have explicitly and systematically focused on supporting students’ metacognition throughout the reasoning process [6]. In addition, existing instructional approaches often rely on predefined scaffolds or prompts, limiting the opportunities for delivering timely, adaptive, and individualised metacognitive support as students’ reasoning evolves.

Recent advances in generative artificial intelligence (GenAI) have expanded opportunities for integrating AI into medical education, creating new opportunities for teaching, learning, assessment, and feedback [11–15]. A growing body of literature has explored the use of GenAI in medical education [12], including AI patient simulation [16–18], intelligent tutoring and Socratic questioning [19–22], clinical decision support [23], and assessment [24, 25]. These studies have demonstrated the feasibility of integrating GenAI and provided initial evidence of its potential educational value. However, they also suggest that the educational value of GenAI depends not only on its technological capabilities but also on how students engage with and evaluate AI-generated outputs. Medical students may struggle to critically engage with GenAI without appropriate guidance, indicating that the mere introduction of AI tools is insufficient [23]. Therefore, the transformative potential of GenAI may lie in purposeful instructional design that shapes how students interact with and learn through AI [13–15, 26]. Without purposeful instructional design, the use of GenAI may reduce active engagement of students, foster overreliance on GenAI [27], and ultimately contribute to never-skilling, referring to the failure of students to develop CR when relying on AI during the early formative years of clinical education [15, 28]. Given the complexity and uncertainty inherent in CR [3], the concerns about the use of GenAI are particularly salient when designing GenAI-mediated metacognitive scaffolding (temporary and structured assistance powered by GenAI that supports students’ metacognition). Therefore, research is needed to identify how GenAI-mediated metacognitive scaffolding can be designed to support students’ metacognition throughout the reasoning process.

The present study addressed the limited understanding of how GenAI-mediated metacognitive scaffolding should be designed and refined to support students’ metacognition during CR. A design-based research approach [29, 30] was used to design and iteratively refine the scaffolding within a history-taking activity and generate design principles for such scaffolding. The study aimed to support students’ metacognition during CR and generate design principles for GenAI-mediated metacognitive scaffolding. History-taking was selected as the learning context because it represents a particularly meaningful setting for investigating the metacognitive processes involved in CR, as students must continuously gather information, interpret emerging evidence,

and adjust their reasoning over time in history-taking [31]. To embed metacognitive scaffolding within this dynamic reasoning process, we developed a GenAI-mediated learning environment (Figure. 1). Within this environment, students engaged in history-taking and interacted with the GenAI-mediated metacognitive scaffolding during their ongoing CR processes. Specifically, students interacted with an **AI patient**, a GenAI-based virtual standardised patient designed to simulate patient encounters, and an **AI tutor**, a GenAI-based virtual tutor that provided metacognitive support (i.e., questions and feedback intended to elicit students' monitoring, evaluation, and planning activities). The GenAI-mediated metacognitive scaffolding was developed and refined through three iterative design cycles, each addressing emerging design challenges identified during the previous iteration. The following research questions guided this study: (1) How can GenAI-mediated metacognitive scaffolding be designed and refined to support the CR development of medical students? (2) What design principles can be derived through three iterative design cycles to inform the design of GenAI-mediated metacognitive scaffolding for the development of CR? To address these questions, each design cycle was examined through the following two sub-questions: a. For each design cycle, how does GenAI-mediated metacognitive scaffolding shape students' metacognitive processes during CR? b. For each design cycle, how do students perceive and experience the GenAI-mediated metacognitive scaffolding?

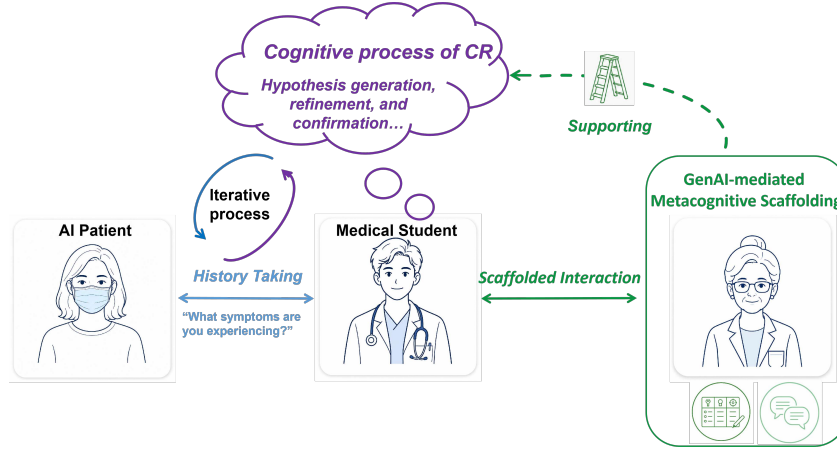


Fig. 1 Overview of the learning environment for supporting CR during history-taking.

Results

Across three iterative design cycles, the GenAI-mediated metacognitive scaffolding was progressively refined to address emerging challenges in supporting students' metacognition during CR. The design evolved from fixed, pre-defined scaffolding in Cycle 1, to Socratic questioning-based interaction in Cycle 2, and subsequently to adaptive scaffolding in Cycle 3. The following sections present the design and implementation of each scaffolding approach, students' metacognitive engagement and experiences, the

emerging design challenges and design implications identified across the three cycles. To understand how different scaffolding designs shaped students’ metacognitive processes during CR, we examined 1) the patterns of metacognitive support provided by the AI tutor and students’ subsequent engagement with the support, 2) students’ perceptions and experiences of the interventions.

Specifically, Table 1 presents interaction characteristics and the extent of metacognitive support provided by AI tutor across the three design cycles. These indicators describe how different scaffolding designs operationalised metacognitive support and the extent to which students engaged with such support during CR. Figure 2 illustrates the correspondence between different types of metacognitive support provided by the AI tutor and students’ subsequent metacognitive engagement across the three cycles. Metacognitive support and engagement were coded across six dimensions of metacognition: information gathering monitoring (M1), knowledge limitation monitoring (M2), evidence linkage evaluation (E1), differential diagnosis evaluation (E2), further history-taking planning (P1), and physical examination and test planning (P2). Detailed definitions and examples are provided in Table 3 in the Methods section. The percentages in Figure 2 indicate the proportion of total AI tutor turns that contained a specific type of metacognitive support. For example, an M1 rate of 1.89% in Cycle 1 indicates that 4 out of the 212 total tutor turns contained prompts targeting information gathering monitoring. Furthermore, the values in parentheses (n+/n-) denote the breakdown of students’ responses to these specific supports. Specifically, n+ represents instances of successful metacognitive engagement (i.e., the student provided a substantive metacognitive response), while n- represents non-engagement (i.e., the student ignored the prompt or provided a superficial reply lacking metacognitive effort). Together, these indicators enabled comparison of how different scaffolding designs shaped the provision of metacognitive support and students’ engagement with that support across the three cycles.

Table 1 Descriptive characteristics of coded student–AI tutor interactions across three design cycles.

Measure	Cycle 1 (FTMS)	Cycle 2 (SQMS)	Cycle 3 (AODMS)
<i>Analytical sample</i>			
Students, n	16	14	15
Student–case sessions, n	64	82	77
<i>Interaction characteristics</i>			
Dialogue exchanges, n	212	446	162
Interaction episodes, n	64	82	77
Episodes per student–case session, median [IQR]	4 [4–4]	5.5 [4.25–6.75]	5 [4–8]
Exchanges per episode, mean (SD)	3.31 (1.44)	5.46 (4.53)	2.18 (1.94)
Exchanges per episode, median [IQR] (range)	3 [2–4.25] (1–6)	4 [2–8.5] (1–21)	1 [1–3] (1–7)
Single-exchange episodes, n/N (%)	7/64 (10.9%)	17/82 (20.7%)	50/77 (64.9%)
<i>Metacognitive support</i>			
Support elements, n	41	563	518
Tutor turns with ≥ 1 support element, n/N (%)	41/212 (19.3%)	436/446 (97.8%)	161/162 (99.4%)
Support elements per tutor turn, mean (SD)	0.20 (0.41)	1.29 (0.61)	3.27 (1.19)
Support elements per tutor turn, median [IQR] (range)	0 [0–0] (0–2)	1 [1–2] (0–4)	3 [2.25–4] (0–6)

Note. Values describe the coded-log sample. A student–case session is one student’s interaction with one clinical case. A dialogue exchange comprises one student turn and one AI tutor turn; an interaction episode comprises one or more exchanges. Multiple support elements could be coded in one tutor turn. Percentages use interaction episodes or tutor turns, as specified by the row denominator. The cycles were not compared inferentially. FTMS, Fixed Turn-Triggered Metacognitive Scaffolding; SQMS, Socratic Questioning-Based Metacognitive Scaffolding; AODMS, Adaptive On-Demand Metacognitive Scaffolding; IQR, interquartile range; SD, standard deviation.

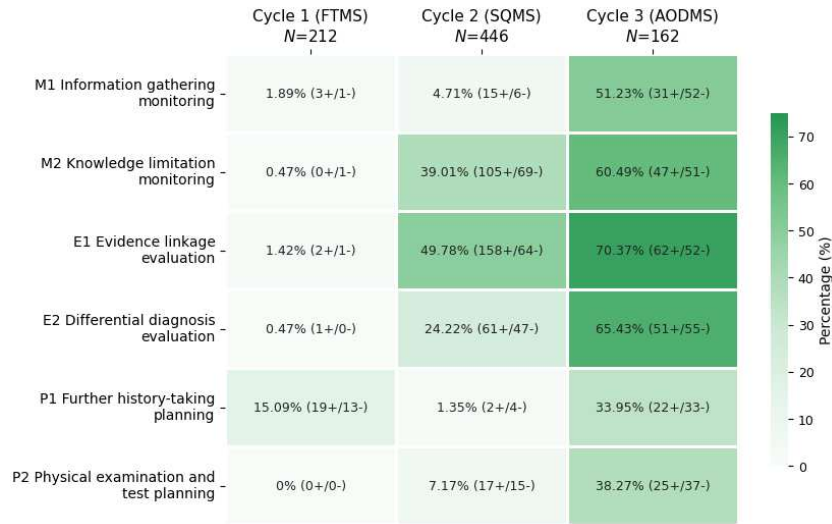


Fig. 2 Frequencies and distributions of metacognitive support provided by the AI tutor and students' corresponding metacognitive engagement across three design cycles.

Note. The figure presents the distribution of metacognitive support provided by the AI tutor across the six metacognitive dimensions (M1, M2, E1, E2, P1, and P2) and the corresponding student engagement across the three design cycles. Detailed definitions and examples of these six dimensions are provided in Table 3 in the Methods section. The capital N in the column headers represents the total number of AI tutor turns for each cycle. Percentages indicate the proportion of total AI tutor turns that contained a specific type of metacognitive support. Values in parentheses denote the breakdown of successful (n+) versus unsuccessful (n-) student metacognitive engagement in response to these supports. Darker shading reflects higher proportions of support instances.

Cycle 1: exploring the instructional potential of Fixed Turn–Triggered Metacognitive Scaffolding (FTMS)

Design approach: turn-triggered predefined reflective questions

In Cycle 1, we introduced FTMS as an initial design to explore whether and how strategically placed support could facilitate students’ metacognition during the CR process. Given the importance of timely reflection during CR, FTMS was designed based on the assumption that predefined reflective questions at strategic points could trigger students’ metacognitive processes without disrupting the history-taking flow. FTMS was implemented through AI tutor–initiated interventions at two predefined history-taking points (turns 20 and 40). At these points, an automated dialogue box was seamlessly integrated into the interface, presenting students with a series of structured reflective questions without forcing an interruption to their consultation. These predefined questions guided students to synthesise collected information, consider possible diagnoses, and identify additional information needed for further reasoning (sample questions are provided in Appendix A).

Student experience: low-density metacognitive support functioned as task-oriented reminders

As shown in Table 1, FTMS provided the lowest density of metacognitive support, with an average of 0.20 (SD = 0.41) metacognitive support elements per AI tutor turn. The metacognitive support provided was not only infrequent but also singularly focused. Most coded elements were restricted to P1, appearing in only 15.09% of AI tutor turns, even though roughly 60% of these support instances effectively elicited students’ metacognitive engagement (19+/13-) (Figure 2).

Interview data indicated that fixed reflective questions helped some students pause during history-taking, review collected information, and form preliminary diagnostic hypotheses (*“slow down, so fewer things would be missed”* (V1S10); *“what I had asked, what diagnoses I might have...”*, V1S08). Nevertheless, multiple students reported that the scaffolding functioned mainly as task-oriented reminders (*“strengthened reminders, but I didn’t pay much attention”*, V1S12) rather than supporting sustained metacognitive regulation (*“I feel like I hadn’t missed anything”*, V1S01; *“I had asked everything, so I ignored it”*, V1S04). Consequently, the scaffolding rarely promoted continuous metacognitive engagement and appeared insufficient to disrupt cognitive dispositions to respond, such as premature closure and search satisficing (*“I didn’t feel the AI tutor made much difference”*, V1S03). Students recognised that fixed questions could encourage summarization and reflection, but acceptance was limited by generic metacognitive triggers, poor timing, and lack of follow-up. Some students described the questions as (*“just another way of asking me to summarise”*, V1S05), and others felt that it did not provide any feedback on whether their reasoning was appropriate (*“He didn’t say whether I was right or wrong”*, V1S14).

Design implications: scaffold rigidity in predefined reflective questions and the need for dynamic dialogic scaffolding

FTMS provided an initial approach to embedding GenAI-mediated metacognitive scaffolding within the history-taking activity. The structured reflective questions successfully created brief reflective pauses, encouraging students to review collected information and generate preliminary diagnostic hypotheses. However, these effects were limited by three key challenges: restricted metacognitive coverage, with reflective questions primarily eliciting planning-related reflection; limited alignment with students' evolving reasoning processes; and the absence of adaptive support and feedback. Consequently, students often perceived FTMS as a task reminder rather than as ongoing support for the metacognitive process. These findings suggested that effective metacognitive scaffolding requires more than triggering reflection at predefined moments; it requires interactive support that is responsive to students' evolving reasoning processes.

Cycle 2: enhancing metacognitive support through Socratic Questioning-Based Metacognitive Scaffolding (SQMS)

Design approach: integrating structured reflection and iterative Socratic questioning

Building on the limitations identified in Cycle 1, Cycle 2 introduced SQMS, an interactive scaffolding design integrating a Deliberate Reflection Table (DRT) [9], a structured tool that guided students to articulate and evaluate their diagnostic hypotheses, with an AI tutor employing Socratic questioning to scaffold students' metacognition. Specifically, to address the limited depth of reflection and lack of context-sensitive support observed in Cycle 1, the DRT was introduced to externalise students' evolving diagnostic hypotheses and provide a structured basis for further dialogue between the student and the AI tutor. Within the DRT, students documented their prioritized differential diagnoses, alongside supporting and opposing evidence, and identified the additional clinical information required to evaluate each hypothesis. Then students could choose to interact with the AI tutor by clicking an "Ask the Tutor" button to submit their DRT and ask for further reasoning support from the AI tutor. The AI tutor subsequently engaged students in Socratic dialogue, posing iterative questions to prompt students to reflect on their diagnostic reasoning. This enabled an ongoing feedback loop: participants could revise their DRTs and continue the dialogue iteratively, allowing them to progressively refine their diagnostic hypotheses throughout the task.

Student experience: high-density, imbalanced metacognitive support deepened diagnostic reflection yet risked interaction overload

As shown in Table 1, SQMS substantially increased student-AI tutor interaction, lengthening the interactions to an average of 5.46 dialogue exchanges per student-initiated episode. SQMS also substantially increased the density of metacognitive

support, with an average of 1.29 (SD = 0.61) metacognitive support elements per AI tutor turn. Compared to the FTMS approach, SQMS provided a broader range of metacognitive support and elicited more frequent metacognitive activities among students, particularly M2 (39.01%, 105+/69-), E1 (49.78%, 158+/64-), and E2 (24.22%, 61+/47-) (Figure 2). However, the metacognitive support provided by SQMS was heavily imbalanced. As illustrated in Figure 2, while SQMS successfully prompted students to evaluate existing evidence and knowledge limitations, it fell short in promoting forward-looking clinical actions. Metacognitive support related to data gathering and planning (M1, P1, P2) remained disproportionately limited.

Interview data further indicated how the integration of the DRT and Socratic questioning supported students' metacognition. Most students valued the DRT as a cognitive mapping tool that helped organize clinical information and synthesize collected findings, while reducing the likelihood of overlooking relevant information (*"...like a mind map... a form of information structuring," V2S13*). Students also reported that the AI tutor generated contextually relevant inquiries by cross-referencing their DRT inputs (*"It combines information points in the table ... to generate more targeted questions," V2S09*). Furthermore, students reported that the interactions with the AI tutor encouraged them to reconsider knowledge gaps and critically evaluate evidence in relation to their hypotheses (*"Why didn't he ask me about... instead he asked about...? This case clearly doesn't fit... I thought about it later, and maybe it is..."*, V2S05). However, the interview data also suggested limitations of relying primarily on Socratic questioning for metacognitive scaffolding. Students frequently described the AI tutor's inquiries as *"endless"* (V2S12), expressing frustration over the lack of clear stopping points or confirmatory feedback (*"I've already provided all the information I could answer; shouldn't he just stop? Don't ask any more," V2S13*). Additionally, students perceived that the AI tutor occasionally drifted from scaffolding CR to conventional knowledge testing, particularly when dialogue focused on pathophysiological mechanisms or clinical manifestations (*"...asking about the pathological mechanism... testing me, rather than indicating that there was a problem with the diagnosis," V2S06*).

Design implications: interactional overload in Socratic questioning and the need for learner-driven calibration

Research on SQMS demonstrated the value of combining DRT with Socratic questioning to support metacognitive engagement. The DRT helped students organise clinical information, integrate findings, and make their reasoning process visible. Drawing on the information externalised in the DRT, the AI tutor could pose targeted questions that encouraged students to identify knowledge gaps and critically evaluate evidence in relation to their hypotheses. However, the findings also suggested that using GenAI-mediated Socratic questioning alone provided limited support for effective metacognitive scaffolding, as it could lead to uneven metacognitive engagement, prolonged interactions, and students' perceptions of being tested rather than supported in the reasoning process. These insights highlighted the need for a scaffolding approach that enabled students to calibrate and customise metacognitive support

according to their evolving reasoning needs, moving beyond merely responding to Socratic questions.

Cycle 3: advancing toward Adaptive On-Demand Metacognitive Scaffolding (AODMS)

Design approach: context-sensitive support and metacognitive menu integration

Drawing on lessons from SQMS, Cycle 3 introduced AODMS. This cycle incorporated three design refinements to realize a fully learner-driven and context-sensitive scaffolding environment. First, the redesigned DRT provided students with greater flexibility to update their evolving diagnostic hypotheses by adding, deleting, and reordering differential diagnosis entries. Second, whenever students requested support from the AI tutor, the complete student–AI patient dialogue record and the current DRT entries were supplied to the AI tutor, providing a real-time, comprehensive map of the student’s reasoning trajectory. Third, leveraging this integrated context, the AI tutor delivered adaptive, on-demand scaffolding through a metacognitive menu, a structured set of questions targeting distinct dimensions of metacognition (see Appendix A). Crucially, students exercised full control over this support: they could actively select and respond to the specific questions that best matched their immediate reasoning needs, or merely review the AI tutor’s questions before independently advancing the clinical task.

Student experience: balanced, high-density and on-demand metacognitive support fostered safe exploration

AODMS generated fewer total dialogue exchanges than SQMS (N=162 vs. N=446) and reached the highest metacognitive support density across the three design cycles, with an average of 3.27 (SD = 1.19) metacognitive support elements per AI tutor turn (Table 1). Compared with previous cycles, AODMS showed a more balanced distribution of metacognitive support across dimensions. As shown in Figure 2, planning-related support elements accounted for 33.95% (P1) and 38.27% (P2), whereas evaluation-related support elements accounted for 70.37% (E1) and 65.43% (E2). AODMS also showed the highest proportion of single-exchange episodes (64.9%).

Interview data provided further explanations for the quantitative patterns observed in Cycle 3, particularly the high rate of single exchanges coupled with the dense volume of questions per AI tutor turn. These patterns may not indicate a lack of engagement; rather, they reflect students’ perception that sufficient support could be obtained within a limited number of interactions. One student explained, (*“after submitting... he would help me differentiate... so I could make a clearer diagnosis... I didn’t talk with him much”, V3S01*), illustrating that some students may obtain meaningful support without engaging in prolonged dialogue. Furthermore, although the high density of metacognitive support occasionally led to broad and lengthy questions, students valued the detailed guidance as a worthwhile investment to broaden their diagnostic space (*“too much information at once... it was very useful”, V3S15*; *“could better expand my divergent thinking”, V3S12*; *“guided thinking in more diverse directions”, V3S06*).

Even when continuing the dialogue, students no longer perceived the questioning as burdensome (*“it would keep asking, it was not necessarily redundant”, V3S04*).

In summary, students valued three key aspects of AODMS: the redesigned DRT, more context-sensitive guidance, and a psychologically supportive learning environment. Regarding the redesigned DRT, students particularly appreciated its flexibility in managing diagnostic hypotheses (*“Very good, especially the drag-and-sort function,” V3S02*). Students also perceived that the AI tutor effectively integrated the structured DRT information with their ongoing history-taking interactions, providing context-sensitive guidance particularly for evidence weighting and diagnostic prioritization (*“The tutor combined the patient’s situation with my previous interactions to guide me”, V3S06*). Furthermore, students perceived that the AI tutor fostered a highly supportive psychological environment for CR (*“more patient, more approachable, and less stressful”, V3S12*). The environment created a safer space for diagnostic exploration and trial and error.

Despite these benefits, students also identified remaining challenges. While students acknowledged the value of the expansive guidance, they expressed a clear need for explicit, directive closure once a reasoning pathway had been sufficiently explored. One student emphasised this need, stating, (*“When it follows up multiple times, it really should provide a more definitive answer”, V3S02*). Additionally, navigating these highly detailed interactions resulted in a higher overall time cost (*“I don’t have much time to keep spending on this”, V3S04*).

Design implications: derived design principles for GenAI-mediated metacognitive scaffolding

Although AODMS achieved a more adaptive and learner-responsive form of metacognitive scaffolding, the remaining challenges identified by students highlighted an ongoing tension between providing sufficient exploratory support and maintaining interaction efficiency. Building on the insights and iterative refinements across three design cycles, we derived five core design principles (Table 2). These principles emphasising the critical importance of strategically timing scaffolding, externalising students’ reasoning, adapting support to evolving reasoning needs, calibrating guidance intensity, and preserving student agency. These principles suggest that the effectiveness of GenAI-mediated metacognitive scaffolding lies in designing flexible interaction opportunities that enable students to selectively engage in metacognitive activation aligned with their evolving reasoning needs.

Discussion

In this design-based research, we developed and refined a GenAI-mediated metacognitive scaffolding design to support medical students’ CR development during history-taking. Through three iterative design iterations, we found that Socratic questioning can effectively promote students’ metacognitive engagement. Specifically, Socratic questioning successfully prompted students to articulate their reasoning, critically evaluate the rationale and evidence underlying their judgments, and consider limitations and alternative explanations. This is consistent with previous research that

Table 2 Design principles for AI-supported metacognitive scaffolding

Design element	Design principle
Intervention mechanism	Enable student-initiated support at key reasoning points to promote pausing, monitoring, and adjustment.
Reasoning representation	Externalize reasoning through visual, comparable, and revisable representations.
Questioning strategy	Use targeted questions to identify reasoning gaps and elicitation awareness, reflection, and revision.
Feedback strategy	Provide low-burden, focused, and progressive feedback without replacing students’ reasoning.
Context integration	Integrate students’ reasoning products with authentic interaction processes to support task monitoring.

Socratic questioning can make students’ reasoning processes more visible and facilitate critical examination of diagnostic judgments [26, 32]. Given its dialogic nature, Socratic questioning represents a promising approach for integration with GenAI, and emerging evidence supports its potential for facilitating students’ reflection and reasoning in GenAI-supported environments [33, 34].

However, our findings further suggest that Socratic questioning alone is insufficient to constitute effective metacognitive scaffolding in CR; without appropriate alignment with students’ immediate needs, open-ended questioning may become prolonged, cognitively demanding, and perceived as knowledge testing. In our study, an overemphasis on knowledge explanations appeared to narrow metacognitive engagement to identifying knowledge gaps, causing some students to feel tested rather than supported [35–37]. Beyond the content of the questions, the way questions were delivered also shaped students’ experiences. When questions from the AI tutor were perceived as overly challenging, repetitive, or insufficiently connected to the clinical case, they risked resembling the “pimping” trap [38] described in medical education, which may undermine students’ confidence and inhibit critical reflection [39]. Although GenAI-mediated Socratic questioning may provide a psychologically safer environment through private, one-to-one interaction [40, 41], it may lack the situated pedagogical judgment required to adapt questioning according to students’ evolving reasoning and interactional cues [42].

The limitations of Socratic questioning in our study may reflect the inherent complexity of CR, where students must continuously navigate uncertainty, incomplete information, evolving evidence, and multiple plausible diagnostic explanations [3]. Unlike well-structured tasks with clearly defined solution pathways, CR represents an ill-structured problem domain in which students must construct, evaluate, and revise multiple possible solutions rather than follow predetermined procedures [43]. These characteristics indicate that effective metacognitive scaffolding in CR requires going beyond Socratic questioning alone to include adaptive orchestration of support aligned with students’ evolving reasoning processes. In our study, AODMS achieved

such adaptive orchestration of support through two key design features. First, AODMS enabled the AI tutor to provide adaptive metacognitive support aligned with students’ evolving reasoning needs by integrating two sources of learner information: externalised representations of students’ diagnostic reasoning captured through the DRT and authentic student–AI patient dialogue. Recent research has highlighted the importance of tailoring digital scaffolding to learner characteristics, behaviours, and learning needs to provide more targeted support [44]. Extending this perspective, our study further demonstrated a design mechanism through which structured external representations combined with authentic dialogue data can make learners’ evolving reasoning processes available to GenAI in complex CR tasks, thereby enabling adaptive metacognitive support. Second, AODMS provided student-controlled access to multiple forms of metacognitive support, allowing students to decide when and what support to seek based on their perceived needs, addressing the limitations of relying solely on Socratic questioning, including fixed interaction trajectories, limited learner control, and potential interaction overload. This design aligns with recent perspectives on personalized and hybrid human–AI scaffolding, which emphasise adapting support while preserving students’ role as active decision-makers [45]. These findings align with Jonassen’s argument that ill-structured problem solving can be better supported through constructivist and situated approaches that enable students to construct and evaluate multiple possible solutions [43]. Although this study was conducted in the context of CR, design insights about Socratic questioning may have broader relevance for other ill-structured problem domains. Further research is needed to examine how these findings can be adapted across different educational contexts.

This study advances research on metacognition in CR by illustrating how metacognitive processes can be deliberately scaffolded within the reasoning process through GenAI-mediated metacognitive scaffolding. Although metacognition has long been recognised as important for supporting CR [46] and previous research has examined instructional strategies that can foster students’ metacognition [47, 48], those strategies have often been positioned before or after the reasoning process itself. Our findings showed that metacognition can be understood as a process that can be deliberately scaffolded within the reasoning process. Furthermore, while prior CR research has often focused on outcome measures such as diagnostic accuracy when evaluating instructional interventions [47], more research is needed to understand the underlying processes (e.g., metacognitive process) through which students engage in CR [4, 49]. Our study provides complementary process-level insights: targeted, context-sensitive scaffolds may help students identify overlooked symptoms or alternative explanations, thereby creating opportunities to reflect on cognitive dispositions to respond, such as search satisficing, insufficient inquiry, confirmation bias, and premature closure [49, 50]. Although further controlled studies are required to determine whether these process-level benefits ultimately translate into improvements in diagnostic performance, our study provides initial evidence that carefully designed GenAI-mediated metacognitive scaffolding can meaningfully support students’ metacognition during CR.

Our findings further suggest that one of the transformative potentials of GenAI in medical education lies in its capacity to foster student agency through adaptive

and context-sensitive support [51–53]. Student agency is increasingly recognised as important in the GenAI era because the abundance, accessibility, and adaptability of GenAI support require students to actively determine when and how to engage with GenAI while navigating complex and uncertain learning environments [54–56]. Without such agency, students may increasingly delegate cognitive activities to GenAI, potentially compromising the development of core competencies and contributing to concerns about GenAI overreliance [27] and deskilling [28]. In our study, student agency was supported by the combination of externalised reasoning representations and student-controlled access to metacognitive support, reflecting broader conceptualisations of student agency, which view agency not merely as an individual capacity but as emerging through interactions among students, feedback processes, and learning environments [56]. The relevance of these design features extends to human–AI interaction in digital medicine, where the effects of AI assistance vary across clinicians and can be influenced by workflow and explanation format [57–59]. While the present study does not establish that these features reduce overreliance or strengthen student agency, it identifies externalised reasoning representations and student-controlled access to support as testable design principles for future research on student agency and appropriate reliance during GenAI-supported CR.

This study further underscores the importance of pedagogical design for maximising the transformative potential of GenAI in medical education. While current research on integrating GenAI into medical education has successfully enhanced system reliability and usability by improving output accuracy and mitigating hallucinations [16, 23, 24], these advances primarily strengthen GenAI’s technological capabilities. With these advances, GenAI systems can more reliably leverage their vast knowledge repositories, information retrieval capabilities, and linguistic fluency to serve as highly effective question-answering resources, as exemplified by AI clinical clerkship mentors and teaching assistants [26, 60, 61]. However, the educational value of these technology-oriented approaches is also shaped by how their affordances are leveraged through pedagogical designs [62]. Our study provides insights into how this pedagogical challenge can be approached through iterative design-based research focus on instructional design, illustrating the evolution of GenAI’s pedagogical role from a passive task reminder toward an adaptive metacognitive scaffold. Across the design iterations, students’ engagement with GenAI evolved from responding to predefined prompts to actively leveraging AI-mediated support to reflect on and regulate their own reasoning processes. These findings reinforce the broader view that the educational potential of technology emerges through the interaction between technological affordances and instructional design, which together shape how students engage with and learn through technology [63].

For medical education, this study suggests that GenAI should be deliberately designed to elicit students’ reasoning; for example, by prompting them to articulate their assumptions, justify clinical decisions, consider alternative explanations, and identify gaps in their knowledge, rather than simply providing answers. Such scaffolding could offer timely and potentially scalable metacognitive support during independent learning while complementing, rather than replacing, educator guidance. More broadly, the implication for digital medicine lies in preparing learners

for AI-enabled clinical practice by developing their capacity to use AI while remaining actively engaged in and accountable for their own CR [64]. Evidence that access to a capable LLM did not necessarily improve physicians’ diagnostic reasoning, and that misleading AI-generated explanations reduced novice students’ diagnostic accuracy and disrupted confidence–accuracy calibration, underscores the need to evaluate human reasoning alongside system performance [65, 66]. Future research should therefore move beyond evaluating GenAI capabilities alone to investigate how instructional designs can harness these capabilities to support meaningful learning processes. In particular, research should determine whether metacognitive gains are sustained as scaffolding is gradually withdrawn and whether they transfer to independent CR, diagnostic performance, and patient care.

This study has two main limitations. First, a defining characteristic of design-based research is that it involves iterative cycles of design, implementation, analysis, and refinement conducted in authentic and evolving educational contexts rather than under controlled and identical conditions [29, 30]. Therefore, the three design cycles in this study were conducted with evolving student populations and implementation contexts, with differences across cycles in students’ year levels, prior clinical exposure, case characteristics, participation modes and the ChatGPT models used to support the AI patient and AI tutor functions. Although Cycle 1 and Cycle 3 involved students from the same institutional cohort at different stages of training, individual-level changes over time were not examined. Therefore, findings across cycles should be interpreted as evidence of iterative refinement and the development of design knowledge, rather than as controlled comparisons of scaffold effectiveness. Second, although multiple data sources were used, students’ metacognitive processes were examined indirectly through observable interactions (e.g., interaction logs) and students’ reported experiences. Future research could incorporate complementary methods, such as think-aloud protocols or physiological/learning analytics measures, to further elucidate students’ metacognitive processes during CR.

Conclusion

Using a design-based research approach, this study explored the iterative design and refinement of GenAI-mediated metacognitive scaffolding to support medical students’ CR development. Our findings revealed that embedding Socratic questioning within GenAI does not automatically constitute effective metacognitive scaffolding; rather, effective scaffolding requires instructional design that balances adaptive support with the preservation of student agency. This study provides a pathway for how GenAI can transform medical education by moving beyond information provision toward adaptive, student-centred scaffolding that supports students in navigating complexity in medicine while mitigating the risks of overreliance and deskilling.

Methods

Study Design

This study employed a design-based research approach to iteratively develop, implement, and refine a GenAI-mediated metacognitive scaffolding during history-taking, aimed at fostering CR by supporting students' metacognition development. Design-based research is particularly suitable for investigating complex educational interventions in authentic settings, as it allows the integration of design, implementation, analysis, and redesign across multiple iterative cycles [29, 30]. In this study, the design-based research approach served both a practical purpose of improving the design of the intervention and an analytical purpose to derive design principles that can inform future educational practices.

Learning environment

The intervention was implemented through a web-based learning platform [45, 67], which provides medical materials and is specifically designed to support history-taking and metacognition. This platform incorporated two roles: an *AI patient*, which responded to students' inquiries by providing case-specific symptoms, and an *AI tutor*, which delivered metacognitive support through questioning, aiming to support students' reasoning processes (Figure 3). The platform was powered by the latest available OpenAI ChatGPT models at each design cycle (Cycle 1: ChatGPT-3.5 for the AI patient and ChatGPT-4 for the AI tutor; Cycle 2: ChatGPT-4o for both roles; Cycle 3: ChatGPT-4 for the AI patient and ChatGPT-4o for the AI tutor). Students engaged in an interactive dialogue with the AI patient, during which they actively elicited relevant information, interpreted emerging findings, and formulated initial diagnostic considerations. Through this interaction, the activity simulated an authentic clinical encounter in which information gathering and reasoning unfolded simultaneously. To support metacognitive engagement, scaffoldings were embedded within the interaction, evolving from solely relying on the AI tutor's questions in Cycle 1, to incorporating a structured table alongside the tutor in Cycles 2 and 3. The triggering mechanism evolved across the design cycles: while scaffolding was automatically initiated at fixed turns of student-AI patient conversation in Cycle 1 (Figure 4), Cycles 2 and 3 employed a learner-initiated approach, activating the AI tutor only after students completed the table and explicitly clicked "Ask the tutor" (Figure 5).

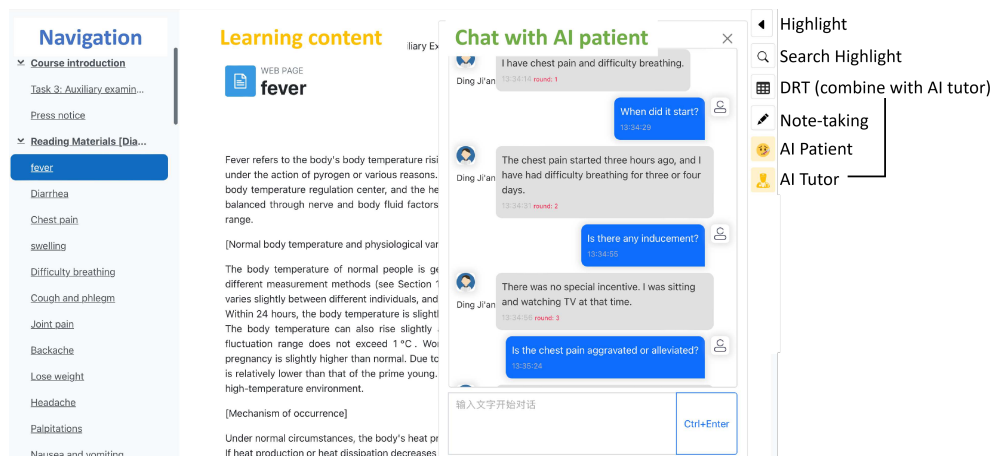


Fig. 3 Learning platform and virtual roles

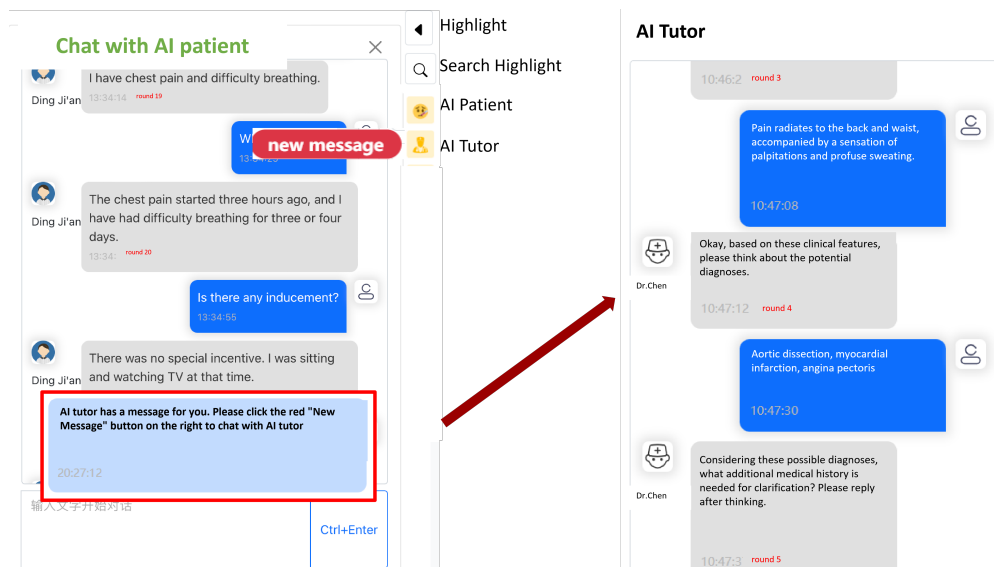


Fig. 4 AI tutor-initiated metacognitive scaffolding (FTMS) in Cycle 1

Note. The interface demonstrates the fixed turn-triggered mechanism. The AI tutor automatically inserts a pop-up message into the dialogue at predefined turns, presenting students with structured reflective questions.

Participants and procedure

Cycle 1: exploring the instructional potential of FTMS

In Cycle 1, to evaluate the feasibility of integrating a GenAI-mediated metacognitive scaffolding into history-taking practice, the study implemented FTMS. Guided by clinical teaching experience, the AI tutor was programmed to intervene at predefined conversational milestones: turn 20 (approximating the completion of chief complaint and major symptom inquiry) and turn 40 (approximating the completion of the history of present illness). At each milestone, three sequential questions were progressively introduced, prompting students to summarise clinical findings, formulate diagnostic hypotheses, and identify directions for further information gathering. This structured guidance aimed to cultivate deliberate CR and discourage superficial, unreflective questioning, thereby enabling students to execute a more comprehensive and efficient history-taking process.

Cycle 1 involved 105 second-year undergraduate medical students (from the 2022 enrollment cohort) enrolled in the "History-Taking and Symptomatology" course at a medical college in China (hereafter referred to as the primary institution), from April to May 2024. Having completed foundational medical courses, these students possessed the requisite theoretical knowledge but were at an early stage of CR development. Participants practised with two cases adapted from authentic clinical scenarios: an aortic dissection case and a pulmonary embolism case. Both chest pain cases were meticulously developed and reviewed by two clinical educators with extensive medical and teaching backgrounds to ensure clinical plausibility, educational relevance, and suitability for novice students.

To inform subsequent design refinement, a multi-method data collection approach was adopted. As a key qualitative component of this approach, semi-structured interviews were conducted with a randomly selected subsample of 16 students drawn from the original cohort of 105 participants. These interviews were administered after the students completed all CR exercises, with each session lasting approximately 30 minutes. The primary objective was to capture detailed insights into students' real-time metacognitive engagement with the FTMS, specifically exploring how they perceived and reacted to the AI tutor's support.

Cycle 2: enhancing metacognitive support through SQMS

Informed by the preliminary evaluation of Cycle 1, which highlighted the need for more adaptive support to stimulate deeper cognitive engagement, Cycle 2 shifted from a fixed turn-intervention to a learner-initiated mechanism. Furthermore, the scaffolding paradigm transitioned from template-based prompts to Socratic questioning, an approach highly suitable for healthcare education that prompts students to articulate their reasoning, thereby eliciting self-explanation processes to facilitate deeper conceptual understanding and knowledge restructuring [68, 69]. Specifically, the AI tutor was engineered to generate adaptive Socratic questions based on students' externalised reasoning, deliberately targeting the analysis, synthesis, and evaluation levels of Bloom's taxonomy [70].

To validate the findings of Cycle 1 and evaluate the generalizability of the GenAI-mediated metacognitive scaffolding across diverse educational environments, Cycle 2 employed a purposeful cross-institutional and cross-grade (spanning the 2020–2023 enrollment cohorts) sampling strategy. Between July and August 2025 (during the summer recess), this approach yielded a focused cohort of 14 undergraduate medical students from three institutions in China. Approximately one-third of these participants were drawn from the primary institution. To capture a wide spectrum of cognitive workflows, this cohort purposefully included five junior students (one Year 2, four Year 3) who had generally completed foundational biomedical coursework, alongside nine senior students (seven Year 4, two Year 5) possessing more advanced training and early clinical exposure. Aligning closely with the qualitative sample size of Cycle 1, this heterogeneous sample provided a robust context for in-depth evaluation despite varying curricular arrangements across the institutions.

To evaluate the effectiveness of the scaffold in complex diagnostic reasoning scenarios, high-difficulty cases were selected. In Cycle 2, five cases (nephrotic syndrome, Meckel’s diverticulum, acute HIV infection, submassive pulmonary embolism, and acute decompensated heart failure) were utilized from Healer powered by Leturio. These cases were rigorously reviewed by clinical experts to ensure their relevance to the participants’ advanced training stage, variability in diagnostic presentation, and optimal suitability for eliciting complex CR and metacognitive activities.

Stimulated recall interviews, a method highly effective for probing thinking patterns, evaluating instructional strategies, and promoting metacognitive awareness [71], Immediately after completing the every case, each student participated in an individual interview lasting approximately 20 minutes. During the interview, the researcher replayed the screen-recorded video of the student’s case-solving process. The playback was paused at salient moments, including when students sought support from different agents, revised their judgments, added or modified information, or revisited previous agent–student dialogues. At each of these points, students were prompted to describe what they had been thinking, why they had taken a particular action, how the agent’s response had influenced their reasoning, and whether their understanding or confidence had changed..

Cycle 3: advancing toward AODMS

Building upon the findings from Cycle 2, which indicated the need for more flexible metacognitive support and context-aware support, Cycle 3 introduced AODMS. The primary objective of this final iteration was to maximise student agency and provide highly individualised, real-time metacognitive support. To achieve this, the AODMS architecture was restructured around two core design principles: contextual integration [53] and student agency [56]. First, the system was upgraded to continuously synchronize the full student–AI patient dialogue with a dynamically adaptable DRT, providing the AI tutor with a real-time cognitive footprint of the student. Second, the scaffolding mechanism transitioned into a learner-driven metacognitive menu [72]. This paradigm shift allowed participants to autonomously regulate the specific focus and depth of the AI’s interventions, thereby exerting true control over their self-regulation processes.

Between October and December 2025, Cycle 3 engaged 83 fourth-year undergraduate medical students from the primary institution. These participants were randomly recruited from the same 2022 enrollment cohort as in Cycle 1, allowing the study to evaluate the scaffolding’s impact on students at a more advanced stage of their clinical training. These participants were enrolled in a ”Clinical Thinking” course and were already equipped with advanced biomedical knowledge and early CR preparation. Given the participants’ advanced foundational knowledge and concurrent clinical exposure, the diagnostic task design maintained a high level of cognitive challenge. Four high-difficulty cases (nephrotic syndrome, myocarditis, acute HIV infection, and infective endocarditis) were selected from Healer powered by Leturio. These cases were rigorously selected using criteria consistent with Cycle 2.

To capture both representative feedback and deep engagement patterns, a two-stage sampling strategy was employed [73]. First, an initial random sample of 12 students was drawn to establish a baseline of general user experiences. Subsequently, an intensity sample of 3 students was purposively selected for their frequent interactions with the AI tutor. Each interview lasted approximately 30 minutes and was conducted sequentially after the students had completed all practices. This combined approach ensured a comprehensive qualitative understanding of students’ perceptions and experiences.

Ethical Approval

Ethical approval for this design-based research was obtained for all three cycles. Cycles 1 and 3 were approved by the Ethics Research Committee of the primary institution (Approval Numbers SUMC-2024-064 and SUMC-2025-013, respectively). Given the cross-institutional nature of Cycle 2, separate ethical approval was obtained from the Institutional Review Board of the collaborating institution (Approval Number IRB00001052-25168). All participants provided informed consent prior to participation.

Across all cycles, participation was strictly voluntary, and written informed consent was obtained prior to enrolment. Participants were explicitly informed of their right to withdraw from the study at any time and request data deletion without any personal or academic penalty. In Cycles 1 and 2, participants received RMB 20 in financial compensation upon completion of each case. In Cycle 3, to ensure ecological validity, case completion was integrated as a formal course assignment worth 2% of the final grade per case. To rigorously protect students’ academic interests, those who opted out of the Cycle 3 study were provided with alternative equivalent assignments, ensuring their final grades were not compromised.

Data Collection

A consistent data collection strategy was employed across all three cycles. To comprehensively understand students’ reactions, attitudes, and cognitive processes regarding the GenAI-mediated metacognitive scaffolding, a consistent data collection strategy was employed across all three cycles, incorporating both quantitative and qualitative sources. Quantitatively, all interactions among students, the AI patient, and the AI

tutor were automatically logged via the learning platform to evaluate user interaction patterns and the metacognitive support embedded within the AI tutor’s questions. Qualitatively, experienced researchers facilitated audio-recorded interviews using a predetermined guide (Appendix B) to explore students’ perceptions of the scaffoldings, and the intervention’s influence on their thinking processes. All sessions were transcribed verbatim. To achieve data triangulation, the platform dialogue logs of interviewed students (e.g., the 15 students in Cycle 3) were extracted and analysed alongside their corresponding interview transcripts.

Data analysis

The AI tutor–student dialogue logs were analysed in two primary domains: interaction patterns and metacognitive support. First, to assess interaction patterns, learning platform logs were used to calculate the total number of dialogue exchanges, total interaction episodes, average/maximum exchanges per episode, and the proportion of single-exchange episodes. Second, a deductive content analysis was conducted using a framework adapted from the self-regulated learning scheme developed by Lajoie SP et.al. [74]. It was explicitly modified for CR contexts by consolidating relevant codes and redefining subdimensions to reflect the AI tutor’s metacognitive support (e.g., adapting *“Looking for particular information”* to *“Further history-taking planning”*). The finalised framework is detailed in Table 3.

The unit of analysis was each AI tutor turn. Interaction records were coded to identify both the specific metacognitive activity elicited and its successful triggering (+/-), which was determined by the students’ immediate responses, subsequent AI patient interactions, and revisions made to the Deliberated Reflective Table (DRT). While multiple codes could be assigned within a single turn, each specific code was counted only once per segment. To establish reliability, two authors (NC and XZ) independently coded interaction data from two randomly selected students per design iteration. Interrater reliability was assessed using Cohen’s kappa [75], yielding substantial to almost perfect agreement (0.724–0.909) (Table 4).

Rather than conducting an independent qualitative analysis, the interview transcripts were utilised specifically to triangulate and contextualise the quantitative findings. Relevant student quotes were systematically extracted to corroborate the interaction data, providing granular insights into how and why the AI-supported scaffolding influenced their cognitive and metacognitive processes.

Abbreviations

CR	Clinical Reasoning
GenAI	Generative Artificial Intelligence
DRT	Deliberate Reflection Table
FTMS	Fixed Turn-Triggered Metacognitive Scaffolding
SQMS	Socratic Questioning-Based Metacognitive Scaffolding
AODMS	Adaptive On-Demand Metacognitive Scaffolding

Table 3 Coding framework for the analysis of interaction logs

Metacognitive support	Code	Definition	Examples
Eliciting Monitoring			
Information Gathering Monitoring	M1	Guiding students to reflect on the completeness and accuracy of their current information gathering, determine whether sufficient information has been collected to support diagnosis, and identify potential omissions or gaps.	“...is there any valuable history information that may have been overlooked?” / “...do you notice that you have focused more heavily on certain systems, such as the cardiovascular system?”
Knowledge Limitation Monitoring	M2	Guiding students to reflect on and assess the limits of their understanding of medical concepts, disease mechanisms, treatments, symptoms, signs, or medication effects.	“How can the patient’s dyspnoea and lower limb oedema be explained through the pathophysiological mechanisms of cor pulmonale?” / “Could Biktarvy, used for HIV treatment, potentially affect cardiac function?”
Eliciting Evaluation			
Evidence Linkage Evaluation	E1	Guiding students to evaluate the relevance and coherence between collected symptoms/evidence and diagnostic hypotheses, and to assess the validity of the underlying reasoning chain.	“You proposed peptic ulcer disease as a diagnosis. Considering the nature and location of the abdominal pain, how would you explain the shift of pain from the epigastric region to the periumbilical area?”
Differential Diagnosis Evaluation	E2	Guiding students to evaluate how well differential diagnoses account for the patient’s overall presentation and whether the diagnostic output is reasonable and sufficiently comprehensive.	“Considering the patient’s history and symptoms, what other systems or diseases could potentially explain these findings?” / “Have you considered whether these signs could support alternative diagnoses? Could common causes of fatigue and oedema have been overlooked?”
Eliciting Planning			
Further History-taking Planning	P1	Guiding students to plan subsequent history-taking activities, including identifying additional reasoning cues and determining what information should be collected next.	“To further refine your diagnosis, what additional history information would you like to collect?” / “If your goal is to assess possible portal hypertension and complications of cirrhosis, what specific history questions would you ask next?”
Physical Examination and Test Planning	P2	Guiding students to plan physical examination and diagnostic testing, including selecting relevant examinations, prioritising actions, and determining principles for further investigation.	“Before moving to the next stage, which unresolved issues still require confirmation through physical examination or additional tests?”

Table 4 Inter-rater agreement for metacognitive coding

Measure	Cohen's κ
Information Gathering Monitoring (M1)	0.862
Knowledge Limitation Monitoring (M2)	0.909
Evidence Linkage Evaluation (E1)	0.724
Differential Diagnosis Evaluation (E2)	0.769
Further History-taking Planning (P1)	0.881
Physical Examination and Test Planning (P2)	0.844
Metacognitive elicitation (+/-)	0.780

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to participant privacy, the potentially identifiable nature of student interview data, learning records, and AI interaction logs, and restrictions under the ethics approval and participant consent. De-identified data may be made available from the corresponding author upon reasonable request, subject to institutional approval and appropriate data-use agreements.

References

- [1] Connor, D. M., Durning, S. J. & Rencic, J. J. Clinical reasoning as a core competency. *Academic Medicine* **95**, 1166–1171 (2020).
- [2] Young, M. *et al.* Drawing boundaries: The difficulty in defining clinical reasoning. *Academic Medicine* **93**, 990–995 (2018).
- [3] Norman, G. Research in clinical reasoning: Past history and current trends. *Medical Education* **39**, 418–427 (2005).
- [4] Eva, K. W. What every teacher needs to know about clinical reasoning. *Medical Education* **39**, 98–106 (2005).
- [5] Koufidis, C., Manninen, K., Nieminen, J., Wohlin, M. & Silén, C. Unravelling the polyphony in clinical reasoning research in medical education. *Journal of Evaluation in Clinical Practice* **27**, 438–450 (2021).
- [6] Kosior, K., Wall, T. & Ferrero, S. The role of metacognition in teaching clinical reasoning: Theory to practice. *Education in the Health Professions* **2**, 108–114 (2019).
- [7] de Bruin, A. B., Dunlosky, J. & Cavalcanti, R. B. Monitoring and regulation of learning in medical education: the need for predictive cues. *Medical Education*

51, 575–584 (2017).

- [8] Chew, K. S., Durning, S. J. & van Merriënboer, J. J. G. Teaching metacognition in clinical decision-making using a novel mnemonic checklist: An exploratory study. *Singapore Medical Journal* **57**, 694–700 (2016).
- [9] Mamede, S. & Schmidt, H. G. Deliberate reflection and clinical reasoning: founding ideas and empirical findings. *Medical Education* **57**, 76–85 (2023).
- [10] Fernandes, R. A. F., Malloy-Diniz, L. F., de Vasconcellos, M. C., Camargos, P. A. M. & Ibiapina, C. Adding guidance to deliberate reflection improves medical student’s diagnostic accuracy. *Medical education* **55**, 1161–1171 (2021).
- [11] Yurdal, M. O. & Kınal, R. Y. Effects of digital versus face-to-face peer support on clinical skills: a three-arm randomized trial. *npj Digital Medicine* (2026).
- [12] Preiksaitis, C. & Rose, C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: Scoping review. *JMIR Medical Education* **9**, e48785 (2023).
- [13] Izquierdo-Condoy, J. S., Arias-Intriago, M., Tello-De-la Torre, A., Busch, F. & Ortiz-Prado, E. Generative artificial intelligence in medical education: Enhancing critical thinking or undermining cognitive autonomy? *Journal of Medical Internet Research* **27**, e76340 (2025).
- [14] Boscardin, C. K., Abdulnour, R.-E. E. & Gin, B. C. Macy foundation innovation report part i: current landscape of artificial intelligence in medical education. *Academic Medicine* 10–1097 (2025).
- [15] Tran, M. *et al.* Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. *npj Digital Medicine* **8**, 315 (2025).
- [16] Chen, N. *et al.* Developing and validating a coding scheme for clinical reasoning in history taking using generative ai-based virtual patients: Systematic text condensation approach. *JMIR Medical Education* **12**, e84347 (2026).
- [17] Moser, M., Posel, N., Ganescu, O. & Fleiszer, D. Twelve tips: Using generative ai to create and optimize content for virtual patient simulations. *Medical Teacher* **47**, 1745–1751 (2025).
- [18] Luo, M.-J. *et al.* A large language model digital patient system enhances ophthalmology history taking skills. *npj Digital Medicine* **8**, 502 (2025).
- [19] Xie, Y., Seth, I., Hunter-Smith, D. J., Rozen, W. M. & Seifman, M. A. Investigating the impact of innovative ai chatbot on post-pandemic medical education and clinical assistance: A comprehensive analysis. *ANZ Journal of Surgery* **94**,

68–77 (2024).

- [20] Siah, C.-J. R., Ignacio, M. J. J., Koh, S. L. S., Tong, J. Y. M. & Lau, S.-T. L. Usability and feasibility of a socratic LLM-supported learning tool for clinical reasoning in undergraduate nursing education. *Nurse Education Today* 107092 (2026).
- [21] Li, Y. *et al.* An innovative socratic method-based artificial intelligence platform for healthcare education: A quasi-experimental study. *Nurse education in practice* 104770 (2026).
- [22] Liu, W. *et al.* Large language model-driven socratic questioning for endodontic case analysis: A randomised controlled trial. *International Endodontic Journal* (2026).
- [23] Xin, Y., Yan, D., Shuren, L., Minyang, L. & Liuheng, L. Ai literacy mediates ai assisted diagnosis participation and critical thinking among medical students under supervision. *npj Digital Medicine* (2026).
- [24] Gershov, S., Mahameed, F., Raz, A. & Laufer, S. Towards accurate and interpretable competency-based assessment: enhancing clinical competency assessment through multimodal ai and anomaly detection. *npj Digital Medicine* (2026).
- [25] Indran, I. R., Paranthaman, P., Gupta, N. & Mustafa, N. Twelve tips to leverage ai for efficient and effective medical question generation: A guide for educators using chat gpt. *Medical Teacher* **46**, 1021–1026 (2024).
- [26] Thesen, T. & Park, S. H. A generative ai teaching assistant for personalized learning in medical education. *NPJ Digital Medicine* **8**, 627 (2025).
- [27] Fan, Y. *et al.* Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology* **56**, 489–530 (2025).
- [28] Ke, Y. *et al.* Ai-induced never-skilling in medical education. *Nature Medicine* **32**, 1997–2006 (2026).
- [29] Hoadley, C. & Campos, F. C. Design-based research: What it is and why it matters to studying online learning. *Educational Psychologist* **57**, 207–220 (2022).
- [30] Dolmans, D. H. J. M. & Tigelaar, D. Building bridges between theory and practice in medical education using a design-based research approach: Amee guide no. 60. *Medical Teacher* **34**, 1–10 (2012).
- [31] Bowen, J. L. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine* **355**, 2217–2225 (2006).

- [32] Ng, I. K., Tham, S. Z., Loh, C. W. & Teo, D. B. When i say... socratic questioning. *Med Educ* **58**, 1029–1031 (2024).
- [33] Xi, L., Zhang, Y. & Wang, Q. Investigating the effects of an LLM-based Socratic conversational agent on students’ academic performance and reflective thinking in higher education. *Computers & Education* **241**, 105494 (2026).
- [34] Kholis, N., Al Hakim, V. G., Heung, Y. M. E., Gumelar, A. S. & Mubarak, H. Ai chatbot as Socratic Butler: effects on metacognitive awareness, cognitive load, and learning performance in introductory programming for engineering education. *European Journal of Engineering Education* 1–19 (2026).
- [35] Abou-Hanna, J. J., Owens, S. T., Kinnucan, J. A., Mian, S. I. & Kolars, J. C. Resuscitating the socratic method: student and faculty perspectives on posing probing questions during clinical teaching. *Academic Medicine* **96**, 113–117 (2021).
- [36] Rucker, L. *et al.* Medical faculty and medical student opinions on the utility of questions to teach and evaluate in the clinical environment. *Medical Science Educator* **33**, 669–678 (2023).
- [37] Wang, F., Liu, D. & Zhang, M. Metacognitive processes, situational factors, and clinical decision-making in nursing education: a quantitative longitudinal study. *BMC Medical Education* **24**, 1530 (2024).
- [38] Detsky, A. S. The art of pimping. *JAMA* **301**, 1379–1381 (2009).
- [39] Oyler, D. R. & Romanelli, F. The fact of ignorance revisiting the socratic method as a tool for teaching critical thinking. *American Journal of Pharmaceutical Education* **78**, 144 (2014).
- [40] Chen, S. *et al.* The landscape of virtual simulation in undergraduate dental education with implications for artificial intelligence incorporation. *npj Digital Medicine* (2026).
- [41] Stoddard, H. A. & O’Dell, D. V. Would socrates have actually used the “socratic method” for clinical teaching? *Journal of general internal medicine* **31**, 1092–1096 (2016).
- [42] Vrikki, M. & Evagorou, M. An analysis of teacher questioning practices in dialogic lessons. *International Journal of Educational Research* **117**, 102107 (2023).
- [43] Jonassen, D. H. Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development* **45**, 65–94 (1997).

- [44] Thomann, H. & Deutscher, V. Scaffolding through prompts in digital learning: A systematic review and meta-analysis of effectiveness on learning achievement. *Educational Research Review* **47**, 100686 (2025).
- [45] Li, X. *et al.* An advanced ai-powered engine to facilitate hybrid human-ai regulated learning. *Computers & Education* 105527 (2025).
- [46] Croskerry, P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine* **78**, 775–780 (2003).
- [47] Schmidt, H. G. & Mamede, S. How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Medical Education* **49**, 961–973 (2015).
- [48] Norman, G. R. *et al.* The causes of errors in clinical reasoning: Cognitive biases, knowledge deficits, and dual process thinking. *Academic Medicine* **92**, 23–30 (2017).
- [49] Cloude, E., Wiedbusch, M., Dever, D., Torre, D. & Azevedo, R. *The Role of Metacognition and Self-regulation on Clinical Reasoning: Leveraging Multimodal Learning Analytics to Transform Medical Education*, 105–129 (Springer, 2022).
- [50] Cutrer, W. B., Sullivan, W. M. & Fleming, A. E. Educational strategies for improving clinical reasoning. *Current problems in pediatric and adolescent health care* **43**, 248–257 (2013).
- [51] Visser, C. L., Wouters, A., Croiset, G. & Kusurkar, R. A. Scaffolding clinical reasoning of health care students: a qualitative exploration of clinicians’ perceptions on an interprofessional obstetric ward. *Journal of medical education and curricular development* **7**, 2382120520907915 (2020).
- [52] Nicol, D. J. & Macfarlane-Dick, D. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education* **31**, 199–218 (2006).
- [53] van der Graaf, J. *et al.* How to design and evaluate personalized scaffolds for self-regulated learning. *Metacognition and Learning* **18**, 783–810 (2023).
- [54] Kasneci, E. *et al.* Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences* **103**, 102274 (2023).
- [55] Dabbagh, N. & Kitsantas, A. Personal learning environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. *The Internet and Higher Education* **15**, 3–8 (2012).
- [56] Nieminen, J. H., Tai, J., Boud, D. & Henderson, M. Student agency in feedback: beyond the individual. *Assessment & Evaluation in Higher Education* **47**, 95–108

(2022).

- [57] Yu, F. *et al.* Heterogeneity and predictors of the effects of ai assistance on radiologists. *Nature Medicine* **30**, 837–849 (2024).
- [58] Everett, S. S. *et al.* From tool to teammate in a randomized controlled trial of clinician-ai collaborative workflows for diagnosis. *npj Digital Medicine* **9**, 409 (2026).
- [59] Spitzer, P. *et al.* The effect of medical explanations from large language models on diagnostic accuracy in radiology. *npj Digital Medicine* **9**, 333 (2026).
- [60] Ebihara, H. *et al.* Development of a clinical clerkship mentor using generative ai and evaluation of its effectiveness in a medical student trial compared to student mentors: 2-part comparative study. *JMIR Medical Education* **11**, e76702 (2025).
- [61] Baillifard, A., Gabella, M., Lavenex, P. B. & Martarelli, C. S. Effective learning with a personal ai tutor: A case study. *Education and Information Technologies* **30**, 297–312 (2025).
- [62] Reinhold, F., Leuders, T., Loibl, K. *et al.* Learning mechanisms explaining learning with digital tools in educational settings: A cognitive process framework. *Educational Psychology Review* **36**, 14 (2024).
- [63] Kozma, R. B. Will media influence learning? reframing the debate. *Educational Technology Research and Development* **42**, 7–19 (1994).
- [64] Abdunour, R.-E. E., Gin, B. & Boscardin, C. K. Educational strategies for clinical supervision of artificial intelligence use. *New England Journal of Medicine* **393**, 786–797 (2025).
- [65] Goh, E. *et al.* Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open* **7**, e2440969 (2024).
- [66] Teng, D. *et al.* Impact of ai misinformation on diagnostic accuracy and confidence calibration in novice medical students. *npj Digital Medicine* **9**, 356 (2026).
- [67] Li, X. *et al.* The flora engine: Using analytics to measure and facilitate learners’ own regulation activities. *Journal of Learning Analytics* (2025).
- [68] Ho, Y.-R., Chen, B.-Y. & Li, C.-M. Thinking more wisely: using the socratic method to develop critical thinking skills amongst healthcare students. *BMC medical education* **23**, 173 (2023).
- [69] Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P. & Glaser, R. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science* **13**, 145–182 (1989).

- [70] Neal, M.-A. Engaging students through effective questions. *Education Canada* **51**, n1 (2011). ISSN:0013-1253.
- [71] Recall SR, S. A systematic review of stimulated recall (sr) in educational research from 2012 to 2022. *Humanit Soc Sci Commun* 489 (2024).
- [72] Fleur, D. S., Bredeweg, B. & van den Bos, W. Metacognition: ideas and insights from neuro-and educational sciences. *npj Science of Learning* **6**, 13 (2021).
- [73] Nyimbili, F. & Nyimbili, L. Types of purposive sampling techniques with their examples and application in qualitative research studies. *Humanit Soc Sci Commun* (2024).
- [74] Lajoie, S. P., Zheng, J., Li, S., Jarrell, A. & Gube, M. Examining the interplay of affect and self regulation in the context of clinical reasoning. *Learning and Instruction* **72**, 101219 (2021).
- [75] McHugh, M. L. Interrater reliability: the kappa statistic. *Biochemia Medica* **22**, 276–282 (2012).

Funding

This project is supported by National Natural Science Foundation of China, Grant (Award Number: 62407001), Guangdong Provincial Undergraduate Teaching Quality and Teaching Reform Project for 2025 (Document No. 4 of the Department of Higher Education of Guangdong Province, 2026, 678), and China Scholarship Council, number 202306140034, awarded to Xiaoyu Zhao.

Acknowledgements

The authors thank the participating students and instructors for their contributions to the study.

Author Contributions

NC and XZ contributed equally to this work. NC conceptualised and designed the study, implemented the intervention, analysed the data, and wrote the manuscript. XZ designed the study, analysed the data, and wrote the manuscript. LT designed the study, implemented the intervention, supported data interpretation and manuscript writing. MX, ZL, JX, XL, MX, LL collected data and supported implementation. OX, CL, MCR, MCF, Dragan G, Danijela G coordinated resources, supervised the study, provided design guidance, and reviewed the manuscript. YF acquired funding, supervised the project, coordinated resources, contributed to study design, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

All authors declare no financial or non-financial competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)