"I Wouldn't Really Use It as a Practice Tool": Understanding Medical Students' Perspectives and Needs on LLM-Enhanced Clinical Skills Training

Yuru Huang
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
yhuang760@connect.hkust-gz.edu.cn

Chao Liu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
cliu009@connect.hkust-gz.edu.cn

Yunna Cai
The Hong Kong University of Science and Technology(Guangzhou)
Guangzhou, China
yunna5004@gmail.com

Lina Xu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
lxu582@connect.hkust-gz.edu.cn

Yun Hou
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
yunhou@hkust-gz.edu.cn

Mingming Fan*
The Hong Kong University of Science and Technology (Guangzhou)
Guangdong, China
The Hong Kong University of Science and Technology
Hong Kong, China
mingmingfan@ust.hk

Abstract

Large Language Models (LLMs) are expected to enhance medical education through personalized clinical skills training. However, their practical application from the student user experience perspective remains underexplored. This gap is critical because without understanding students' needs, LLM-based tools risk poor adoption and suboptimal learning outcomes. This study explores medical students' challenges and expectations when using LLM-based clinical skills training through a two-phase investigation involving 14 medical students. We integrated five Type 2 Diabetes cases into a probe platform and conducted probe-based studies followed by co-design workshops. We identified challenges across three categories: dialogue content (lack of realism, insufficient knowledge depth differentiation); dialogue presentation (information overload, single modality limitations); and dialogue interaction (inadequate guidance and feedback). Co-design workshops revealed expectations for enhanced patient modeling, personalized content delivery, structured presentation frameworks, and collaborative features. These findings provide design considerations for developing more effective, user-centered LLM-based medical education systems.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

Medical education, Large language model, User experience

ACM Reference Format:
Yuru Huang, Chao Liu, Yunna Cai, Lina Xu, Yun Hou, and Mingming Fan. 2026. "I Wouldn't Really Use It as a Practice Tool": Understanding Medical Students' Perspectives and Needs on LLM-Enhanced Clinical Skills Training. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3772318.3790322

*Corresponding author.

CHI '26, April 13–17, 2026, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790322

1 Introduction

Clinical skills training is essential in medical education, shaping students' ability to integrate theoretical knowledge with practical application and ultimately improving the quality of healthcare services [4, 11, 29]. Traditionally, clinical training relies on two primary methods: standardized patients (SPs) and clinical internships [13, 59]. While these approaches provide valuable hands-on experience, they also present notable limitations. SP-based simulations often fail to capture the complexity and variability of real-world clinical scenarios [59], while internships are constrained by limited resources and opportunities, making it challenging for students to gain sufficient practical experience [78].

These challenges have motivated Human-Computer Interaction (HCI) researchers to explore technology-enhanced solutions for professional skill development. Early HCI research investigated various approaches, from surgical simulations [47] to gamified learning environments [49, 64]. Virtual Reality (VR) and Augmented Reality (AR) technologies create immersive environments for complex clinical situations, such as virtual surgical procedures or anatomy courses [27, 47]. However, these technologies remain limited in authenticity and adaptability, struggling to fully capture the complexities of real-world clinical encounters while often relying on

fixed training cases that lack flexibility [27]. Large Language Models (LLMs) emerge as a promising solution for medical education, offering unprecedented opportunities for personalized, interactive learning experiences. With their advanced natural language understanding and ability to generate contextually relevant responses, LLMs can dynamically adapt to user inputs, making them well-suited for simulating real-world clinical dialogues [41, 79]. Recent HCI research demonstrates LLMs' potential in educational contexts: supporting collaborative reasoning [86], providing adaptive feedback [45], and serving as intelligent tutoring partners [85].

Despite these promising technical developments, current research primarily focuses on system performance and educational outcomes, often overlooking learners' authentic experiences and specific needs. Without understanding learners' cognitive processes during clinical training, systems risk creating barriers rather than facilitating learning [44]. Clinical education presents unique challenges requiring not only cognitive skills but also empathetic communication and real-time decision-making in high-stakes environments [39, 83]. This necessitates user centered design approaches that systematically explore medical students' challenges and expectations. Given that existing research has not fully explored LLMs' potential and limitations in clinical skills training, this study focuses on the challenges medical students face and their expectations for LLM-based tools. We chose these two aspects because they directly reflect the tools' practical effectiveness and user needs, providing a clear direction for improving educational technologies. Thus, we propose the following research questions (**RQs**):

- **What challenges do medical students encounter when training clinical skills using LLM tools? (RQ1)**
- **What are medical students' expectations for improving LLM-based training tools? (RQ2)**

To explore these questions, we designed a two-phase study in which medical students interacted with an LLM-integrated platform focused on Type 2 Diabetes (T2D) clinical skills development (Figure 1). Through probe-based studies with 14 participants, we identified 11 challenges across three dimensions: 1) Dialogue Content—lack of realism, repetitive responses, and insufficient knowledge differentiation (C1–C4); 2) Dialogue Presentation—linear structures, text-heavy outputs, and non-standardized formats (C5–C8); 3) Dialogue Interaction—absence of guidance, feedback, and scaffolding (C9–C11). Following challenge identification, we conducted co-design workshops where medical students collaboratively developed expectations for LLM-based training tools. For dialogue content, they suggested enhancing authenticity and personalization through enriched patient behavior models, user personas, and strengthened scenario associations. For presentation, they expected improved information delivery efficiency through structured frameworks, information extraction, and multimodal outputs. For interaction, they proposed boosting user engagement and learning effectiveness through increased guidance, feedback, and collaborative features.

Our study offers the following contributions:

- We identify 11 key challenges medical students encounter when using LLM-based clinical training tools, providing insights into current limitations across content, presentation, and interaction dimensions.
- We elicit medical students' expectations through co-design workshops, highlighting needs for enhanced authenticity, personalized experiences, and collaborative mechanisms.
- We propose ten design considerations among three dimensions. We discuss the considerations specific to clinical education from those applicable to general LLM-based learning contexts, and interconnections among design considerations

By addressing these challenges and incorporating user-driven design principles, our research contributes to the development of more adaptable and impactful LLM-powered tools for medical education.

2 Related Work

2.1 The Dilemma and Need for Clinical Skills Training

Clinical skills training is a cornerstone of medical education, essential for developing high-quality healthcare professionals [4, 11, 29]. Medical students require extensive clinical skills training to integrate theoretical knowledge with practical skills, fostering comprehensive understanding of their profession [8, 54, 60]. However, several challenges impede this goal. Traditional clinical skills training typically relies on teacher or peer-simulated patients, which presents limitations [13, 59]. Such approaches fail to accurately replicate real-world clinical scenario complexity and lack standardization, impeding effective skill transfer to actual practice. Additionally, internships provide opportunities to develop clinical skills, but high demand and limited resources severely constrain exposure to real clinical practice and meaningful feedback [78].

Current clinical skills training reveals significant deficiencies in both knowledge transfer and practical skill development. Numerous assessments [41, 63] show that clinicians and medical students fall far below expected standards in diabetes management, highlighting substantial gaps between theoretical knowledge and practical application. This gap drives the need for innovative solutions like LLMs, which can potentially enhance clinical skills training by providing personalized, interactive, and adaptable learning experiences that bridge the theory-practice divide.

2.2 Clinical Skills Training Technologies

Prior to the emergence of LLMs, HCI researchers had already developed various approaches to clinical skills training. From text-based virtual patients [6, 21] to physical simulators [51, 70] to immersive platforms leveraging virtual technologies [43, 64], researchers focused either on reproducing the content of clinical cases, enhancing the immersiveness of delivery media, or emphasizing interactive mechanisms for feedback.

In research on text-based virtual patients, scholars primarily focused on the design of medical case content and interaction workflows. Cook et al.'s meta-analysis [21] demonstrated that carefully constructed cases—with varying complexity, cognitive scaffolding, and timely feedback—achieved effects comparable to traditional instruction. Critically, their research established the principle that *instructional design quality rather than technological sophistication drives learning outcomes*. Meanwhile, Bateman et al. [6] extended

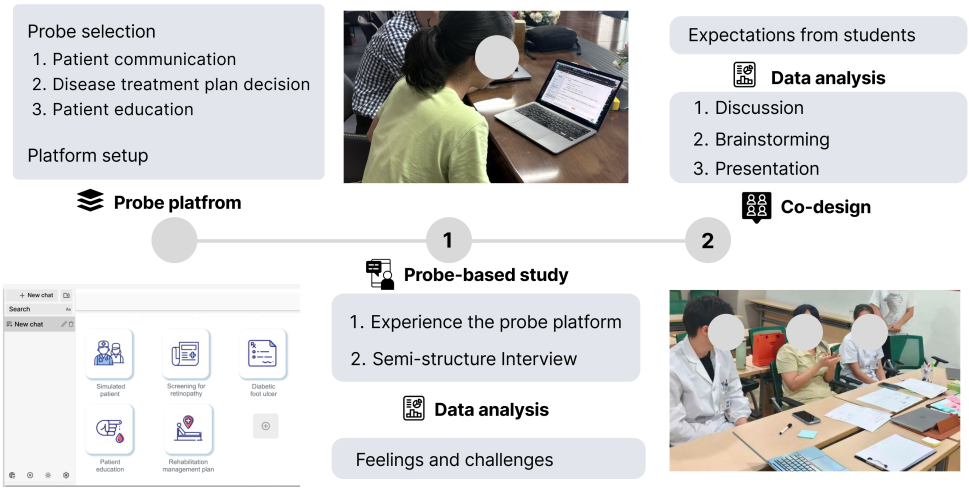


Figure 1: Workflow: We designed a two-phase study in which medical students interacted with an LLM-integrated platform focused on clinical skills development. To build the probe platform, We focused on T2D-related clinical skills, including interactive patient simulations, diagnostic decision-making, and patient education. Phase 1 shows the probe-based study with platform interface featuring five T2D clinical scenarios, participant interaction photo, and data collection process. Phase 2 shows the co-design workshop with participant expectations gathering, collaborative activities (discussion, brainstorming, presentation), and group workshop photo.

this understanding by identifying the components of effective content: making simulated reasoning processes explicit rather than merely presenting diagnostic accuracy. This insight highlighted that novices require visible expert thinking patterns, thereby informing instructional design. This foundation laid the groundwork for subsequent research focusing on the simulation of reasoning processes and experiential learning.

However, the text-only presentation modality constrained skill development. As Kneebone [42] documented, clinical competence requires recognizing non-verbal cues and cultivating procedural fluency—dimensions unattainable through text alone. Moreover, script-bound interaction restricted authentic clinical reasoning practice, as predetermined decision trees inadequately reflected the unpredictability of clinical practice [62].

Recognizing these embodiment limitations, researchers employed physical mannequins, haptic devices, and 3D models to develop simulators [51, 70]. Research on simulators primarily focused on advancing presentation fidelity to enhance clinical skills training. For instance, McGaghie et al. [51] demonstrated that when high-fidelity presentation was combined with deliberate practice principles—namely focused repetition, expert feedback, and progressive difficulty—simulation training produced superior outcomes compared to traditional apprenticeship. However, content adaptability remained constrained. Most simulators targeted specific skills at fixed proficiency levels, creating scaffolding dilemmas [37].

Multimodal platforms integrate virtual reality, augmented reality, and gamification elements, with researchers enhancing user experiences through both interaction modalities and presentation media. For example, Quail et al. [64] demonstrated how narrative presentation enhances motivation through emotional engagement. Similarly, Kneebone et al. [43] employed hybrid simulations combining physical task trainers with trained actors, noting that coordinated presentation modalities simultaneously facilitated technical and communication skill development—neither of which could be independently achieved through single-modality approaches. However, research also highlighted the complexity of coordination: each content modification necessitated synchronized changes across physical props, actor scripts, and system logic. As Kneebone et al. documented, scenarios requiring multidisciplinary team involvement demanded months for development and revision [43]. This pattern exposed a fundamental paradox: rich, adaptive experiences required exponential authoring effort [21].

This evolution reveals both cumulative progress and persistent constraints. Text-based systems established cognitive scaffolding principles [21], embodied simulators validated deliberate practice with tactile feedback [51], and multimodal platforms demonstrated integration of physical, virtual, and social dimensions [43, 64]. These contributions remain foundational for approaching new technologies. However, despite their respective strengths, these systems shared a common limitation: script-bound interactions with predetermined paths that prevented dynamic adaptation to individual learning trajectories. LLMs offer potential to address this challenge through dynamic natural language generation [41, 79].

2.3 LLMs in Practice Education and the Demands of Clinical Training

LLMs demonstrate broad application prospects in practice-based education across various domains. In software engineering and design education, LLMs serve as partners providing on-demand advice and generating practical scenarios, enhancing learner engagement and efficiency [45, 85]. Business and management education research

emphasizes LLM potential in supporting case-based reasoning [86]. However, these applications reveal common challenges, particularly limitations in aligning generated content with domain norms, supporting multi-turn reasoning, and fostering user agency [45, 86].

Clinical skill training faces even more complex challenges. Unlike other practice domains, clinical education requires learners to develop empathetic communication, ethical judgment, and real-time decision-making skills in high-stakes environments beyond reasoning tasks [83]. Simulated patient interactions, as clinical training cornerstones, demand both factual accuracy and emotional authenticity, requiring LLM-driven systems to be highly sensitive to emotional, interpersonal, and contextual dimensions of medical encounters [39, 83].

Although existing research has begun exploring learner interactions with LLM tools in medical education, revealing learner enthusiasm for AI-assisted learning and concerns about trust, transparency, and pedagogical consistency [44], these studies primarily evaluate from technical implementation and educational effectiveness perspectives, lacking deep understanding of medical students' authentic experiences and specific needs as end users. As Yee et al. noted, needs for AI systems in clinical environments are dynamically evolving [76], suggesting necessity to systematically explore medical students' challenges and expectations in different learning contexts from a user experience perspective.

Therefore, designing medical education tools that balance AI generative capabilities with user-centered scaffolding, promoting both realism and pedagogical efficiency, remains an urgent problem. This study specifically targets this gap by focusing on medical students' actual experiences when using LLM-based clinical skill training tools, exploring the challenges they face and their expectations for tool improvement.

To ground our investigation in a concrete clinical context, we focus on Type 2 Diabetes (T2D) as our case study (Appendix A). T2D represents an ideal domain for exploring LLM-assisted clinical skills training due to its global health significance, complex management requirements, and well-documented educational gaps in current medical curricula [41, 78].

3 Phase 1: Probe-based Study

In order to gather insights into the challenges encountered by medical students when using LLM probes, we conducted a probe-based study, as depicted in phase 1 (Figure2). This phase of the study involved the use of the probes for medical education, followed by semi-structured interviews with 14 medical students. We received ethical approval from our institution's Ethics Committee to conduct all procedures involving human subjects. Throughout the research, we took careful steps to protect participants' rights and privacy. The following sections detail the types of data collected, participant recruitment, and data analysis processes.

3.1 Probe Design

3.1.1 Probe Selection. In T2D clinical skill learning, three core modules—interactive simulated patients, disease decision-making, and patient education—form the foundation for preparing medical students for future clinical work[26, 38]. To understand how LLM-based tools may support skill acquisition across these modules, we designed five probes grounded in prior research.

For **interactive simulated patients**, we implemented **Probe 1** (LLM agent simulate patients) building on Holderried et al. [33], assessing whether students can engage in meaningful history-taking interactions with GPT-based simulated patients. For **disease decision-making** [81], we designed **Probe 2** (Testing diabetic retinopathy screening timing) to examine students' application of evidence-based DR screening recommendations [75], and **Probe 3** (Chatbot in diabetic foot ulcers) to explore LLM-aided clinical action selection for diabetic foot ulcers (DFU) cases [72]. For **patient education** [22], we designed **Probe 4** (Evaluating obesity in T2D according to guidelines) to assess students' critical evaluation of LLM-generated obesity management recommendations [3], and **Probe 5** (Develop rehabilitation plans for the elderly) to explore LLM-supported construction of personalized rehabilitation plans [52]. Table 1 summarizes the design logic, prompt structure, and module alignment for each probe.

3.1.2 Probe Implementation. To facilitate the implementation of the probe try-ons, we integrated all probes into a single platform and developed an easily accessible user website based on JavaScript. This platform is designed to promote seamless user navigation and ensure a consistent experience across all tasks. We designed the five probes as separate scenarios rather than continuous workflows to reflect the current state of LLM-based medical education tools in practice, which typically exist as discrete modules rather than integrated systems. We modeled the user interface on the ChatGPT page[1] and leveraged the ChatGPT-4o API[2] to support LLM-based conversational functionalities (3). Prior to the interviews, we conducted a pilot test with 3 medical students to validate the platform's usability and make necessary adjustments. To ensure a standardized and structured approach, we prepared a detailed interview guide that outlined the specific questions and prompts to be used during the semi-structured interviews. This study is based on custom-designed prompts, which utilize the existing ChatGPT-4o model without any additional fine-tuning or retraining. Therefore, instead of fine-tuning or retraining ChatGPT, we leveraged its existing capabilities to assess its performance in applications such as simulated patient interactions, clinical decision-making, and patient education (Appendix C).

3.2 Participants

We recruited 14 participants (8 female, 6 male) through word of mouth. All had completed five years of clinical medical education, and most (N=12) were about to begin their clinical internship. Participants ranged in age from 22 to 25 years and all had prior experience using LLMs. Detailed demographic information is provided in Table 2. Each study session lasted approximately 2 hours, and participants received $50 compensatio.

3.3 Procedure

Each session was conducted in a one-on-one format, allowing for in-depth exploration of participants' experiences and perceptions.

[1]https://chat.openai.com/
[2]https://platform.openai.com/docs/models

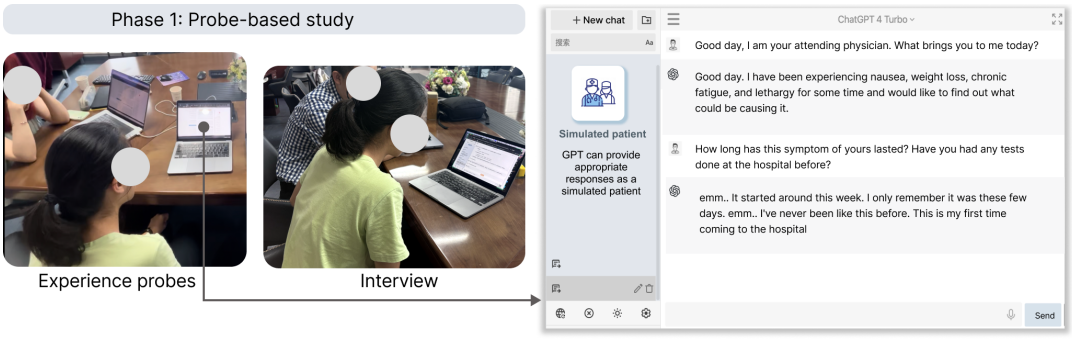


Figure 2: Phase 1:A probe-based study, which included experiencing the probes in clinical skills training, semi-structured interviews with 14 medical students after they experienced probes.On the right side of the picture, we present a screenshot of the user using the probe, showing the content of participant's conversation using Probe 1 (simulating patient).

Table 1: Probe details, references, and their alignment with clinical skill training modules in T2D education.

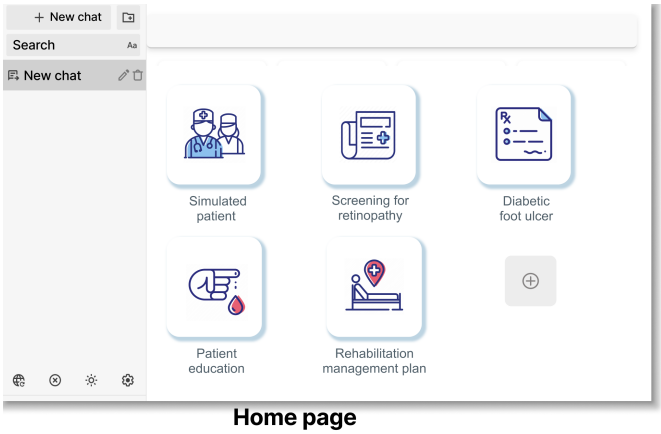
Probe	Description	Target Module	Ref
1. LLM agent simulate patients	This probe aims to test whether GPT can provide appropriate responses as a simulated patient and explore whether doctors can effectively utilize it in medical simulation practices.	Simulated Patients	[33]
2. Testing diabetic retinopathy (DR) screening timing	This probe evaluates whether LLM tools can help learners apply screening guidelines for DR, assessing clinical reasoning under timing and risk trade-offs.	Disease Decision-Making	[75]
3. Chatbot in diabetic foot ulcers (DFUs)	This probe explores whether LLMs can assist in diagnostic and therapeutic decision-making for DFU cases, involving management choices and care prioritization.	Disease Decision-Making	[72]
4. Evaluating obesity in T2D according to guidelines	This probe asks students to critically evaluate whether LLM-provided treatment plans for obesity align with clinical guidelines, emphasizing judgment and credibility assessment.	Patient Education	[3]
5. Develop rehabilitation plans for the elderly	This probe assesses whether LLMs can help co-develop personalized rehab plans considering multimorbidity and lifestyle factors, relevant for patient-centered education.	Patient Education	[52]

Participants were recruited from undergraduate medical students transitioning from textbook learning to clinical practice, as well as graduate students undergoing clinical training. Each session consisted of three main stages:

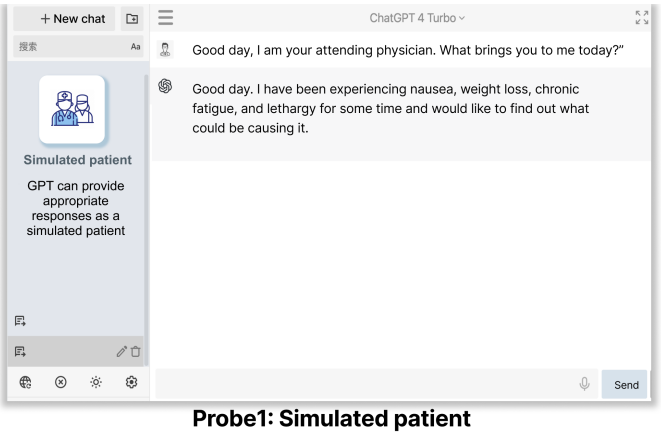
Introduction First, the researcher explained the study's purpose and background to the participant. Second, after brief mutual introductions, participants were asked to describe: 1) the duration of their medical studies; 2) their practical experience with simulation training; and 3) their prior use of LLM tools. Participants' background information, including learning experience, medical discipline, and exposure to medical simulations, is summarized in Table 2.

Experience the Probes Participants were then invited to experience the five probes we designed, one at a time. During this phase, participants were instructed to use the "Think-Aloud" (TA) method [16], in which they verbalize their thoughts and actions while interacting with each probe. This method was used to capture their spontaneous reactions, challenges encountered, and insights gained during task execution. The five probe tasks are described in Appendix D.

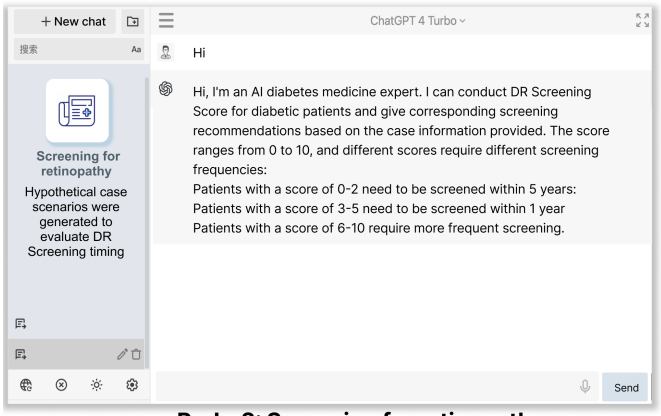
Semi-structured Interview After completing the probe experience, researchers conducted semi-structured interviews. The interview protocol covered four main areas: (1) overall experience with the probes, including perceived advantages and challenges; (2) individual feedback on each probe's strengths and limitations; (3) opinions on the integrated platform's interface design and functionality; and (4) comparisons with existing medical education methods and tools (see Appendix E for the complete interview guide).



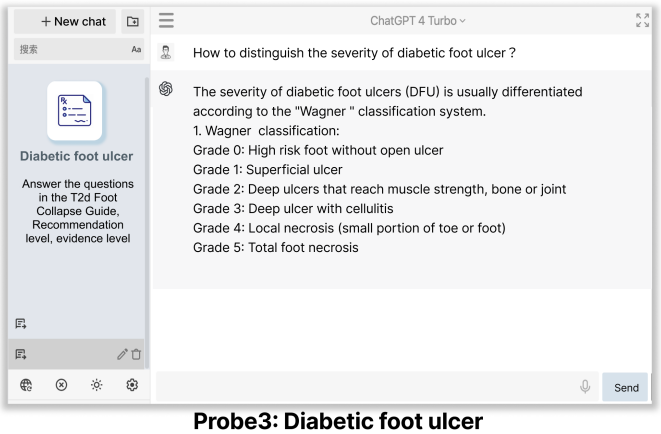
Home page



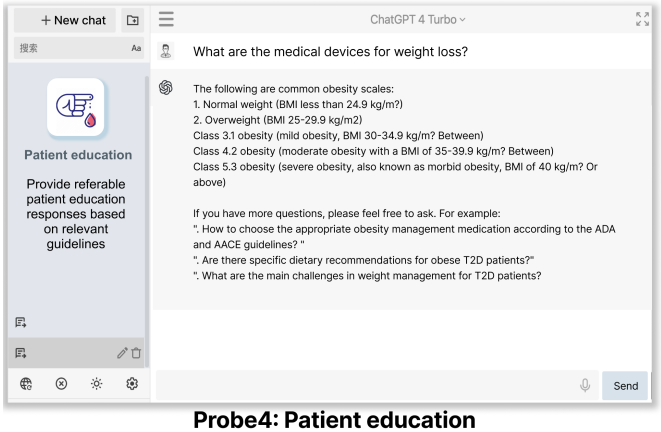
Probe1: Simulated patient



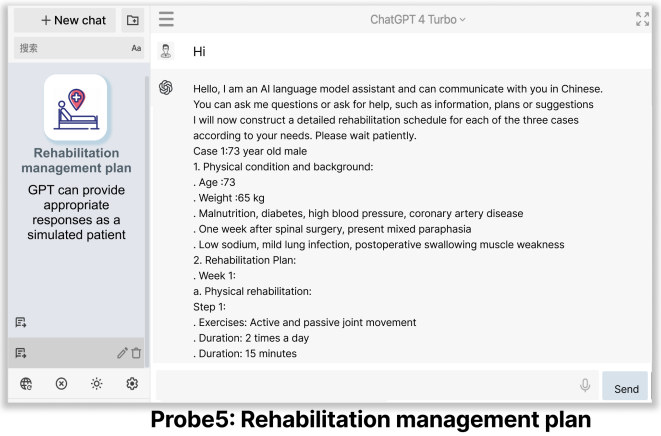
Probe2: Screening for retinopathy



Probe3: Diabetic foot ulcer



Probe4: Patient education



Probe5: Rehabilitation management plan

Figure 3: The interface of probe platform Probe 1: Simulated Patients (This task aims to test whether GPT can provide appropriate responses as a simulated patient and explore whether doctors can effectively utilize it in medical simulation practices); Probe 2: Testing DR Screening Timing (This task focuses on assessing doctors' understanding and mastery of the timing for screening diabetic retinopathy); Probe 3: Diabetic Foot Ulcers (This task aims to explore the application and effectiveness of chatbot technologies, such as ChatGPT, in the diagnosis and treatment of diabetic foot ulcers; Probe 4: Evaluating obesity in T2D (This task requires doctors to analyze the credibility of evaluation results provided by ChatGPT based on guidelines for assessing obesity in T2D); Probe 5: Rehabilitation management plan (In this task, doctors will utilize ChatGPT to develop personalized rehabilitation plans for elderly patients and validate their effectiveness through case studies.)

Table 2: Participants' demographic information

ID	Gender	Age	Education	LLM Experience	Clinical Skills Practice Experience
P1	Male	22	Academic Master	Report Writing	Exam
P2	Female	24	Academic Master	Paper Writing	Exam
P3	Female	24	Academic Master	Report Writing	Exam
P4	Male	24	Professional Master	Report Writing	Exam
P5	Female	24	Academic Master	Report Writing	Exam
P6	Female	25	Academic Master	Report Writing	Exam and Daily Practice
P7	Male	24	Professional Master	Report Writing	Exam and Daily Practice
P8	Male	25	Professional Master	Paper Writing	Exam
P9	Female	22	Academic Master	Papers, Daily Life	Exam
P10	Female	23	Academic Master	Daily Practice	Exam and Daily Practice and Internship
P11	Female	24	Academic Master	Daily Practice	Exam
P12	Female	22	Academic Master	Daily Practice	Exam and Daily Practice
P13	Female	23	Academic Master	Report Writing	Exam and Daily Practice
P14	Female	24	Academic Master	Paper Writing	Exam and Daily Practice

3.4 Data Collection and Analysis

We collected three types of data during the probe-based study: (1) Background information: participants' medical education duration, simulation training experience, and prior LLM tool usage. This contextualized their knowledge base for analysis. (2) Platform interaction data: participants' activities on the probe platform, including Think-Aloud verbalizations of their actions and challenges, captured via audio and video recordings. (3) Interview data: audio and video recordings of all semi-structured interviews.

We conducted thematic analysis [14, 58] following an iterative and systematic approach. This method enabled theme development through collaborative coding and consensus-building, ensuring trustworthiness and rigor throughout the process. Two researchers first transcribed and translated all audio and video recordings using Feishu [24]. Transcripts were then coded using open coding methods in FigJam. We used interaction records from the probe platform as supplementary data to cross-validate and refine the codes. The coding process involved four researchers in iterative discussions. Two researchers performed the initial independent coding. The entire team then compared the resulting classifications, discussed differences, and collaboratively developed the final analytical framework. For example, P10's quote "*The patient is too cooperative, which is unrealistic*" was initially coded differently by the two researchers but, through discussion, was classified under the sub-theme **lack of realism in RS scenario** within the theme **dialogue content**. This process revealed that participants regarded Probe 1 as a role simulation (RS) tool and Probes 2–5 as Q&A tools, expressing views on dialogue content, presentation, and interaction for both types.

To ensure the reliability of theme development, the research team engaged in regular discussions to review and refine emerging themes. Quotations were grouped into themes based on their conceptual similarities and differences (see Appendix F). For instance, codes related to initiating conversations, dialogue coherence, and concluding interactions were aggregated into the overarching theme of **interaction**. Similarly, quotations concerning text length, formatting issues, and information overload were consolidated into the theme of **content presentation**.

4 Probe-based Study Findings

While participants generally expressed enthusiasm for LLMs in medical education—with P7 describing the tool as *"a top student who can accurately answer questions"*—the interactions revealed significant usability barriers hindering effective learning. These challenges fundamentally affected participants' willingness to adopt such tools, as P4 noted: *"I wouldn't really use it as a practice tool. It still has some way to go..."* Through systematic analysis, we identified 11 recurring challenges (C1–C11) organized into three categories: **Dialogue Content** (C1–C4), **Dialogue Presentation** (C5–C8), and **Dialogue Interaction** (C9–C11).

4.1 Dialogue Content Issues

4.1.1 C1: Lack of Realism in RS Scenario. In role simulation (RS) scenarios, participants (N=10) reported that LLM-generated patient responses lacked behavioral realism. Simulated patients were overly cooperative and rigid, failing to reflect real clinical interaction variability. P10 noted, *"The patient is too cooperative, which is not realistic."* Symptom descriptions were overly structured and textbook-like, lacking emotional fluctuation. P9 commented, *"The symptom descriptions resemble textbook entries, not real patients."* These limitations weaken educational value, especially for participants to navigate uncertain patient behavior.

4.1.2 C2: Insensitivity to Input Variation in Q&A Scenarios. In Q&A diagnostic exercises, participants (N=12) found LLM outputs failed to adapt treatment plans based on updated input. Despite adding clinical details, models often generated identical responses. P4 remarked, *"Even though I emphasized added patient details, the answer given was the same."* P10 added post-gastrectomy context, yet recovery plans remained generic and inappropriate. This reflects lack of contextual reasoning, limiting usefulness for personalized decision-making training.

4.1.3 C3: Lack of Knowledge Depth Differentiation and Direction Customization. Participants (N=10) reported probes did not differentiate between expertise levels. Content remained basic, lacking scaffolding aligned with academic stages. P3 stated, *"These answers are basic. I could find them using a search engine. It doesn't tell me what I should know now or later."* P5 added, *"This might help undergrads, but for me as an intern, it's not useful."* Without adaptive depth, LLM tools risk disengaging advanced learners while under-supporting beginners. Some participants (N=6) noted LLM content did not reflect specialty-specific needs. Clinical priorities across surgery, internal medicine, or pharmacy were treated homogenously. P10 stated, *"For surgery and internal medicine, case focus differs. I hope for more specialized content."*

4.1.4 C4: Fragmentation Across Clinical Scenarios. Participants expressed frustration that probes operated as isolated tasks rather than coherent diagnostic narratives. P1 remarked, *"The modules for diabetic foot, fundus exam, and rehabilitation have relevant information, but I can't apply them together in clinical analysis."* Students expected cases to evolve across probes, mirroring real-world comorbidities. P10 added, *"Why can't consultation in Probe 1 connect to later cases? Knowledge feels disconnected."*

4.2 Dialogue Presentation Issues

4.2.1 C5: Cognitive Load from Linear Dialogues. In RS scenarios, participants found multi-turn text dialogues increasingly difficult to manage. All participants reported difficulty recalling earlier patient information without excessive scrolling.Without structured summaries or dynamic highlighting, linear presentation increases cognitive load and disrupts diagnostic flow. P3 said, *"I forgot the patient's age or history and had to check again."*

4.2.2 C6: Information Overload from Verbose Outputs. In Q&A contexts, participants (N=12) described LLM responses as excessively long, burying insights within redundant text. P2 noted, *"Only one or two points are useful, I have to read everything to find them."* P6 echoed: *"The response is too long. I lose interest reading it."* This overload leads to fatigue and reduced engagement when responses lack segmentation or prioritization.

4.2.3 C7: Single Modality Output Limits Comprehension. Participants consistently highlighted text-only output limitations. P10 remarked, *"For diabetic foot or retinopathy, it would be better to include videos or images."* Absence of visual or auditory elements hampers understanding of spatial content and diminishes consultation realism. Participants perceived the system as "a smarter search engine," lacking multimodal richness of clinical encounters.

4.2.4 C8: Inconsistent and Non-Standardized Outputs. Participants flagged multiple standardization issues. Lab values lacked reference ranges, making abnormality assessment unclear. P14 noted, *"I cannot tell if it is abnormal."* Inconsistent citation practices (some responses included literature, others didn't) eroded trust. P14 stated, *"Some probes have literature, some don't, and I distrust those without."* These inconsistencies diminish credibility in evidence-based training.

4.3 Dialogue Interaction Barriers

4.3.1 C9: Lack of Role and Task Framing in RS Scenarios. Although participants received experiment instructions, they emphasized role definition should be system-integrated. In clinical settings, patients initiate consultations, yet the system remained passive. P4 remarked, *"The system should proactively inform us of patient symptoms like real patients would."* P7 added, *"In real life, patients open conversations—we shouldn't prompt them first."*

4.3.2 C10: Lack of Feedback in RS Scenarios. In RS interactions, participants reported absent real-time feedback on diagnostic decisions. P12 commented, *"The response seemed wrong, but it didn't tell me I was mistaken or why."* Others struggled judging diagnostic process completion. P2 said, *"It didn't evaluate my diagnosis—just asked if I thought it was right. That's strange."* Without evaluative scaffolding, learners may internalize incorrect reasoning patterns.

4.3.3 C11: Lack of Guidance in Q&A Scenarios. In open-ended inquiries, participants felt systems failed to infer or clarify intent. P8 noted, *"Even when I emphasized key points, the response still deviated."* LLMs often ignored contextual details when generating plans. Participants criticized the tendency for direct answers, bypassing stepwise reasoning opportunities. P5 stated, *"It gives me the answer right away. I don't get to think."* P14 observed system-suggested questions sometimes derailed conversations: *"I forgot my original question after being led away."* These issues reduce user agency and undermine inquiry-based learning development.

5 Phase 2: Co-design workshop

After identifying key challenges through the Phase 1 probe-based study, we conducted Phase 2 within 2-3 days to explore participants' expectations through a co-design workshop. The co-design workshop was grounded in participatory design traditions [67]. Following the generative design research [68] framework, we structured the workshop into four stages:**introduction, discussion, brainstorming, and presentation**. This progression is commonly adopted in participatory design studies [9, 80]. This structure moves participants from reflecting on their experiences, through externalization via tangible artifacts, to collaborative critique and refinement, enabling both individual expression and collective sense making. To ground discussion while supporting creative ideation, we developed structured cards summarizing common Phase 1 challenges (Figure4). These cards scaffolded early discussion and helped participants articulate pain points, though the workshop extended beyond Phase 1 themes through open-ended sketching and collaborative critique.

5.1 Procedure

We invited all 14 medical students from the probe-based study to participate in the co-design workshop, dividing them into four groups of 3-4 students with two researchers each. The aim was to gather insights based on their LLM tool experience and identify expectations for improved probe design and functionality. This 90-minute session was facilitated by trained qualitative research experts. The co-design workshop comprised of four stages (Figure5):

Introduction The moderator introduced workshop objectives and initiated icebreaking activities to facilitate deeper discussion.

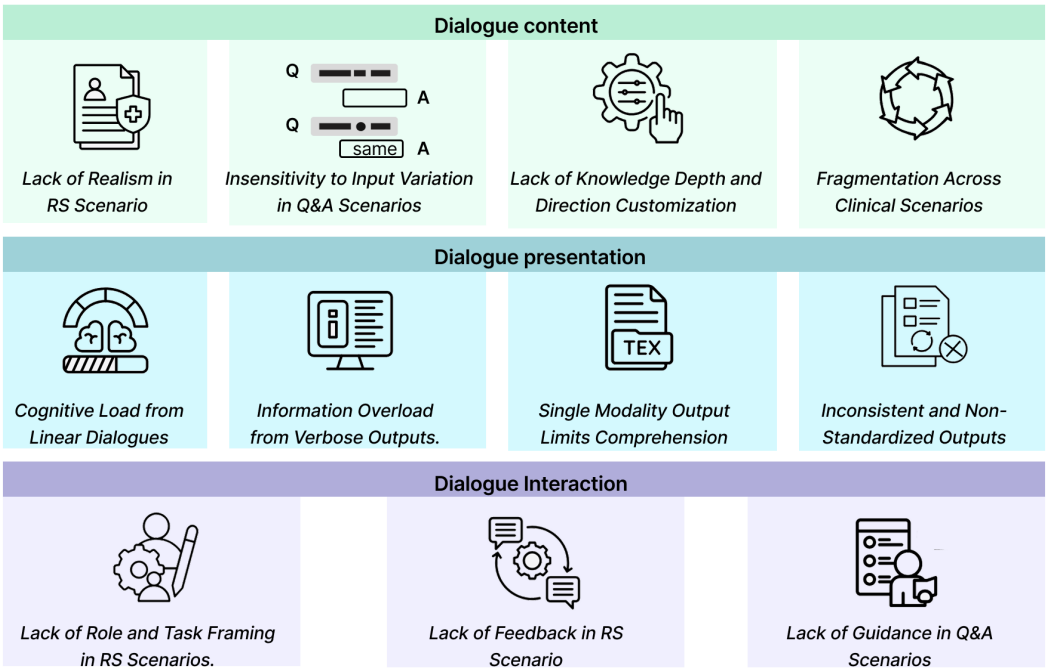


Figure 4: 11 challenges form probe-based study. The top section with a header Dialogue content and contains four items arranged horizontally: Lack of Realism in RS Scenario; Insensitivity to Input Variation in Q&A Scenarios; Lack of Knowledge Depth and Direction Customization; Fragmentation Across Clinical Scenarios.The middle section with the header Dialogue presentation and contains four items: Cognitive Load from Linear Dialogues; Information Overload from Verbose Outputs; Single Modality Output Limits Comprehension; Inconsistent and Non-Standardized Outputs. The bottom section with the header Dialogue Interaction and contains three items: Lack of Role and Task Framing in RS Scenarios; Lack of Feedback in RS Scenario; Lack of Guidance in Q&A Scenarios.

Discussion Participants reviewed their probe platform experiences using prepared materials: 1) Cards: Pre-prepared cards summarized participant experience processes. One type denoted interaction stages (initiation, continuation, conclusion). Another specified content presentation attributes (output format, presentation mode, length, probe rotation). A third indicated content quality reflecting output quality of each probe. 2) Post-It notes: Participants reviewed usage difficulties, challenges, and advantages during their experience and documented them on notes. The facilitator distributed cards and explained their meaning. Participants then reviewed platform experiences, writing challenges or advantages they faced on individual post-it notes. The moderator posted participants results on a blank wall for collective review. Participants shared experiences while answering others questions. This exchange facilitated in-depth retrospectives while generating different perspectives. The process lasted 30-40 minutes.

Brainstorming The facilitator introduced the design task where participants drew sketches of their ideal LLM tool based on discussion elements (1. Interface sketch: Show the visual layout of key functions. 2. Interaction process: Explain how users interact with the system and with whom.) Participants had 20-30 minutes to thoroughly express their ideas. Researchers facilitated by providing materials such as paper and pens. Since participants lacked drawing skills, the host showed common LLM interfaces (ChatGPT). To clarify designs and functions, researchers asked and confirmed participants design intent and presentation forms while they sketched.

Presentation Participants presented sketches and explained prototype designs to the group. After all presentations, participants offered opinions on others designs and discussed them in small groups. Participants could revise their sketches before submitting them to researchers. This process lasted 30-40 minutes.

5.2 Data Collection and Analysis

We collected two categories of data. Participant-generated design artifacts included: 1) 14 lo-fi prototype sketches depicting ideal LLM interfaces, commonly featuring output length controls, structured feedback panels, role-play scenario framing, and multimodal displays; 2) 56 annotated post-it notes documenting pain points and desired features; and 3) 23 collaborative annotations where participants critiqued peers' designs. Researcher-documented data included: 4) full audio and video recordings of all sessions (approximately 6 hours); 5) complete transcripts generated via Feishu platform [24] and manually verified for accuracy; and 6) real-time field notes capturing group dynamics and emergent insights.

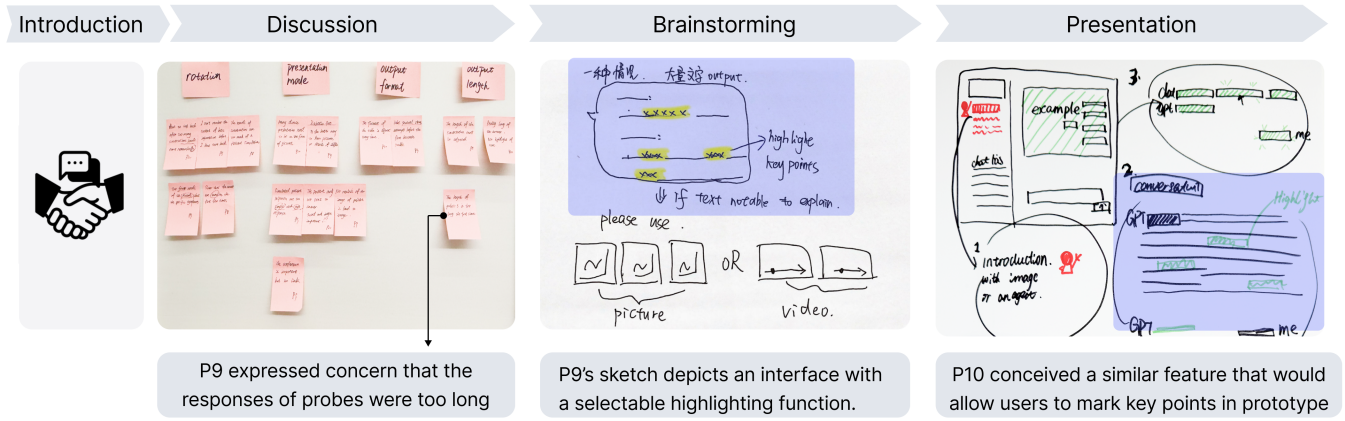


Figure 5: Phase 2 co-design workshop process. (1) Introduction with icebreaking activities, (2) Discussion using cards and post-it notes to capture user experiences, (3) Brainstorming through lo-fi prototype sketching, and (4) Presentation the final design and peer feedback on design concepts.

Our analysis followed an iterative approach, building insights progressively from multiple data sources. Critically, we analyzed both participants' verbal articulations and their design artifacts, treating wireframes as data that could reveal implicit priorities and design tensions not explicitly stated in discussions [25]. Analysis was conducted collaboratively within the research team through regular meetings to discuss emerging themes, refine coding schemes, and ensure methodological rigor. This team-based process allowed triangulation between participants' explicit feedback, design behaviors, and underlying motivations revealed through different data modalities.

We began by linking sketch artifacts with corresponding recorded data to ensure complete feedback capture. Researchers independently conducted initial open coding of transcribed data, generating preliminary codes capturing both explicit statements and implicit design choices. Through iterative team discussions, we refined and consolidated the coding framework, resolving discrepancies and establishing clear criteria. Building on initial codes, we employed affinity diagramming using FigJam to identify recurring patterns and cluster related concepts. We analyzed how participants expressed varying perspectives on challenges, revealing key dimensions along which expectations varied and allowing us to map feedback onto conceptual frameworks illuminating underlying user needs. The final phase involved systematically tracing each emergent theme back to supporting evidence across multiple data sources (verbal feedback, sketch features, and post-it annotations) ensuring robust grounding in participant contributions.

To illustrate our analytical approach, consider feedback about text highlighting functionality (Figure5). During the brainstorming session, one participant (P9) expressed concern that the responses of certain probes were too long and proposed a highlighting mechanism to manage lengthy text. Her sketch depicts an interface with a selectable highlighting function. Another participant (P10) independently conceived a similar feature that would allow users to mark key points in extended responses. The key point is that the analysis of the two sketches revealed that the participants envisioned highlighting: allowing users to annotate important content or enabling the system to pre-highlight key clinical terms. This is a subtle design difference that only emerges through the analysis of artifacts and is not explicitly expressed in oral discussions. Researcher A initially coded this as **Text Management: Highlighting and Length Control** under **Presentation Improvements**. Researcher B, focusing on the underlying learning need, categorized it as **Information Extraction and Key Point Identification**. Through collaborative discussion, the team recognized this reflected broader challenges with information overload and the need for active engagement with lengthy responses. We refined the code to **Key Information Extraction** and positioned it within **Expectations for Content Presentation**.

All codes and themes underwent repeated validation through team discussions and systematic comparison with original data. We operationalized theme saturation through three explicit criteria: (1) prevalence: each theme was supported by evidence from at least 2 participants; (2) triangulation: each theme was corroborated across at least two distinct data modalities (e.g., verbal feedback and sketch artifacts); and (3) stability: no new substantive codes emerged during the final round of transcript review. Themes meeting all three criteria were retained for the final analysis, while those that did not were consolidated into the stronger themes.

6 Co-design Findings

Based on user experience review and researchers' summary, participants expressed expectations in three areas: 1) **expectations for content quality**, 2) **expectations for content presentation**, and 3) **expectations for dialogue interaction**.

6.1 Expectations for Content Quality

Participants outlined four key expectations for improving content quality of medical education LLM tools: **enriching the patient model** (a of Figure6) to better simulate real clinical environments;

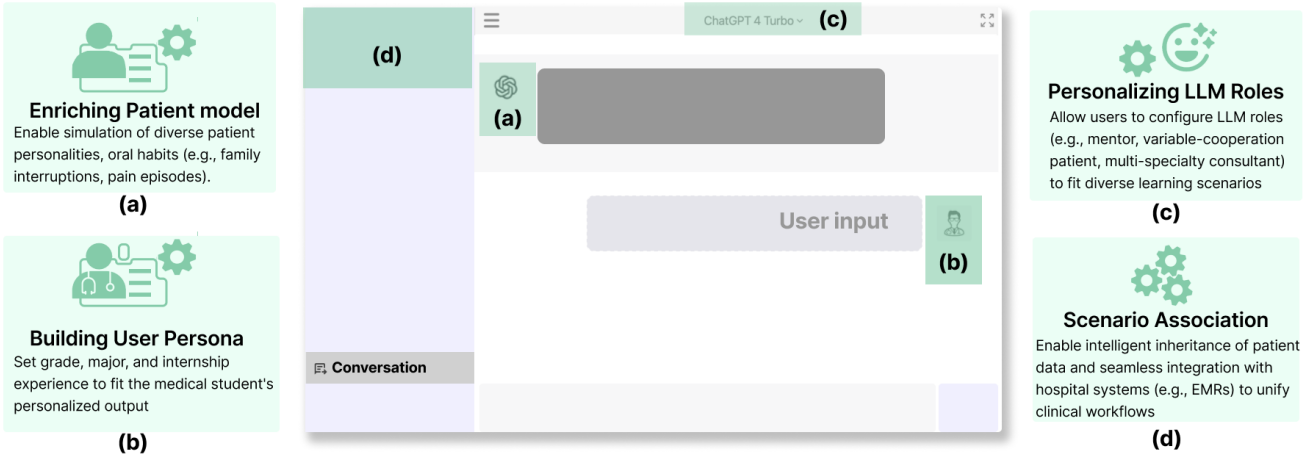


Figure 6: Expectations of content Enriching the patient model (a) to better simulate the clinical environment of the real world ; Building user personas(b) based on learning stages and operational preferences to provide more personalized content ; Strengthening scenario association (c)to improve the efficiency and practical value of tools ; Personalized LLM roles (d)to meet diverse learning.

building user personas (b of Figure6) based on learning stages and preferences for personalized content; **strengthening scenario association** (c of Figure6) to improve tool efficiency and practical value; and **personalized LLM roles** (d of Figure6) to meet diverse learning needs.

6.1.1 Enriching the patient model (for C1). Participants (N=11) suggested enriching patient simulation models to better replicate real clinical environments (a in Figure6). First, adding rich personality details. P10 pointed out: *"The simulated patient can add some details similar to personality, such as being very picky during blood tests."* Second, increasing oral habits of simulated patients to narrow the simulation-reality gap. P9 mentioned: *"The model's responses are overly written or mechanical. I think it can simulate the oral characteristics of real patient conversations."* Third, simulating common clinical interferences such as family interruptions or patient pain attacks to enhance authenticity. P7 suggested: *"This system can simulate family members interjecting or patients experiencing interruptions such as pain."*

6.1.2 Building User Personas and LLM's roles adaptive(for C3). **User Persona Construction:** Building user personas helps systems provide personalized content. Participants (N=14) hoped for modeling from three aspects: learning stage, knowledge graph, and operational preferences. P4 mentioned: *"The system could offer more precise content based on the user's grade and rotated departments."* P9 noted: *"The system should identify my weaknesses and provide relevant knowledge-point strengthening exercises."*

LLM's roles adaptive: Most participants conceptualized probe-like LLMs as active learning instruments assuming multiple roles simultaneously. P9 explained: *"Clinical skill practice is fundamentally a self-training process. LLMs that generate personalized practice scenarios tailored to my specific needs represent a significant improvement over traditional approaches where instructors distribute standardized cases."* However, they emphasized need for human expertise integration, with P7 noting: *"Certain clinical diagnostic procedures require confirmation from experienced instructors, rather than simply accepting LLM recommendations as definitive conclusions."*

Participants articulated their vision for long-term autonomous learning systems centered around three interconnected roles: **simulated patient**, **Socratic tutor**, and **expert consultant**. Rather than sequential transitions, these roles operate concurrently with varying emphasis across learning stages. The simulated patient role provides foundational practice through iterative case-based learning. The Socratic tutor role emerges through knowledge gap identification and strategic questioning, as P7 noted: *"By combining question-and-answer dialogues, I could systematically address my deficiencies in patient consultation techniques."* As students advance, the expert consultant role gains prominence, with P10 explaining: *"As my knowledge progresses, I should independently modify patient complexity levels. At this time, I need the teacher's participation to evaluate me."* P7 also noted that role emphasis varies by learning context: *"During case-based learning, I want the system to be primarily a patient, but when I'm struggling with differential diagnosis, I need more Socratic questioning and expert insights simultaneously."*

6.1.3 Strengthening Scenario Association (for C4). Participants proposed integrating patient data inheritance and pre-integrating hospital system interfaces to enhance tool efficiency and practical value (c in Figure6). Intelligent inheritance of patient data improves experience. P1 suggested: *"Modules for diabetic foot, fundus examination, and rehabilitation management should be integrated into a complete diagnostic and treatment chain"* They believe integration across clinical scenarios improves clinical knowledge learning and connection. Pre-integration of hospital system interfaces is another major improvement direction. P10 pointed out: *"The tool should integrate with hospital electronic medical record systems for easier*

use during internships." Participants emphasized seamless integration with existing hospital systems including electronic medical records, laboratory reporting, and prescription systems. Through such integration, medical students believe they can enhance their *"confidence from student to intern."*(P9)

6.2 Expectations for Content Presentation

Participants identified four expectations for content presentation: **structured content presentation** (e in Figure7) to alleviate cognitive load from linear dialogues; **key information extraction** (f in Figure7) through manual highlighting and learning recommendations; **optional multi-modality outputs** (g in Figure7) such as images, videos, and voice to enhance medical knowledge intuitiveness; and **evidence-based medicine support** (h in Figure7).

6.2.1 Structured Content Presentation (for C5). Participants expressed need for framework-based dialogue displays to address memory burden of linear dialogues (e in Figure7). Framework-based presentation can help better manage content, effectively organizing, reconstructing, and expanding knowledge systems. P9's view was widely agreed upon: *"If LLM can build an effective knowledge framework after information input, I can use it to review or update my understanding of specific knowledge systems. This presentation can integrate content from multiple dialogues like a mind map."*

Specifically, participants hope LLM can automatically divide content into chapters, generate titles, and offer actionable tables of contents for quick access and review. P2 added:

> *"For example, when I start a dialogue and input information, LLM could automatically divide the content into chapters, generate titles, and even provide an expandable and collapsible table of contents. Ideally, I should also be able to adjust these features manually."*

This framework should dynamically update to let them visually track learning progress. If the system allows active user annotations, it would further enhance personalized learning experience. P7 put it: *"Through such interaction, our dialogues can become my personal learning notes."*

6.2.2 Key Information Extraction (for C6). When faced with long dialogues, participants often find it hard to grasp key points. They suggested "manually highlight key points in dialogues" (P9, P2, P5, P7) to break down content (f in Figure7). Manual highlighting can better break down learning tasks, turning complex tasks into actionable sub-goals. P7 hoped LLM could break down long dialogues into smaller, specific parts and recommend learning content based on highlighted parts (P9: *"It could auto-generate learning recommendations based on my highlights, like daily learning plan reminders, to guide my learning process."*) They believe this motivates careful reading of long dialogue outputs.

6.2.3 Optional Multi-modality (for C7). Many participants indicated that besides traditional text output, they hoped for multi-modality output support (g in Figure7). P12 said: *"I hope LLM can output text and support voice so that I can listen directly during clinical internships or emergencies instead of stopping to read."* Additionally, participants proposed multi-modal outputs like charts, images, and videos to enhance medical knowledge intuitiveness and understandability. P11 stated:*"For diseases like retinopathy and diabetic foot ulcers, pictures or videos are essential as text descriptions alone are hard to understand and visualize."*

6.2.4 Evidence-based Medicine Support (for C8). Evidence-based medicine support is crucial for enhancing system credibility (h in Figure7). Participants expected embedded literature systems offering direct PMID links and abstracts. P7 indicated: *"Literature is a must, and it would be better to include publication times, links, and brief abstracts."* Furthermore, participants focused on dynamic comparison functions for examination values. P14 mentioned: *"Examination results should have pop-up prompts with normal ranges to quickly determine abnormalities."* Annotation of treatment plan evidence levels is also expected. P9 pointed out: *"The system should label the evidence levels of treatment plans to help us make more evidence-based decisions."*

6.3 Expectations for the Dialogue Interaction

To address interaction-related limitations from Phase 1, participants proposed four primary design expectations for improving human–LLM collaboration. They sought improved **interaction guidance for RS scenarios** (i in Figure8) with user guides and preset dialogue templates to direct conversations and lower entry barriers; **enhancing feedback in RS scenarios** (j in Figure8) through real-time alerts and detailed analyses to correct diagnostic deviations; **enhancing interaction guidance in Q&A scenarios** (k in Figure8) with more LLM questions and iterative scope narrowing; and **adding collaborative and shared functions** (l in Figure8) like teacher-student collaboration spaces and template markets to boost trust and knowledge exchange.

6.3.1 Enabling Role and Task Guidance (for C9). In response to unclear role framing and task ambiguity in RS scenarios (C10), participants expected explicit, proactive interaction scaffolds. They suggested simulated patients should initiate conversations with self-introductions and symptom descriptions, mirroring real clinical consultations. P3 stated, *"If the simulated patient introduces themselves directly, it feels more realistic and natural."* Moreover, participants proposed function-specific dialogue templates including suggested questions and example interactions to help navigate early-stage uncertainty. P4 mentioned, *"It would be helpful to have a dialogue template that outlines typical questioning paths."*

6.3.2 Providing Timely and Targeted Feedback (for C10). Participants expected feedback mechanisms that could identify reasoning errors and provide real-time corrective signals. This directly linked to Phase 1 findings (C11) where participants expressed concern about making diagnostic mistakes without system prompts. P6 shared, *"If I'm wrong, the system should tell me—not just stay silent."* Several participants proposed visual tools, such as diagnostic progress bars, to help track accuracy and learning stages during case interaction. These suggestions align with deliberate practice principles in medical training, where timely, context-aware feedback improves clinical judgment and confidence.

6.3.3 Guiding Reasoning in Q&A Scenarios (for C11). To mitigate issues of overly direct responses and lack of guided inquiry (C11), participants expected task-sensitive strategies promoting deeper

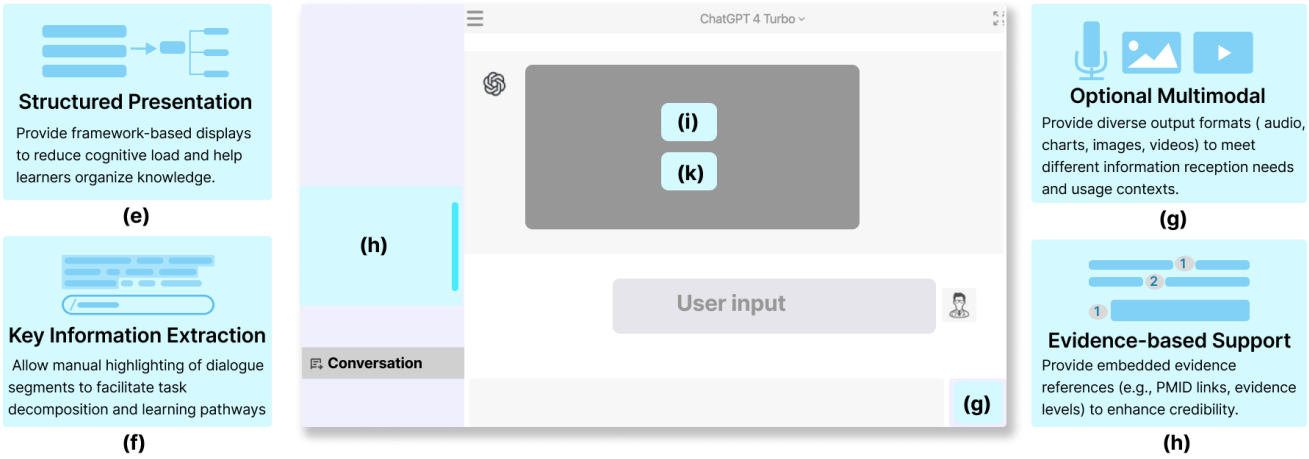


Figure 7: Expectations of presentation Structured Content Presentation (e)to alleviate the cognitive load of linear dialogues; Key Information Extraction(f) through manual highlighting and learning recommendations;Optional Multi-modality Outputs (g) to enhance the intuitiveness of medical knowledge; Evidence-based Medicine Support (h) to enhance the system credibility.

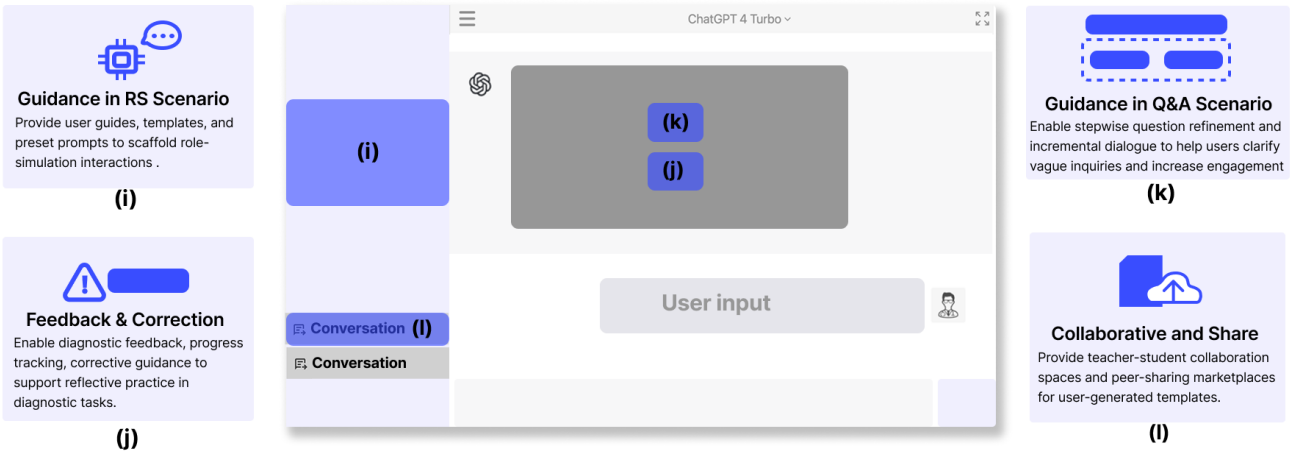


Figure 8: Expectations of interaction Enhancing interaction guidance for RS scenarios (i) to direct conversations and lower entry barriers; Enhancing feedback in RS scenarios (j) to correct diagnostic deviations and improve practice accuracy; Enhancing interaction guidance for Q&A scenarios (k) to increasing interaction frequency and refining output; Increasing Collaborative and Shared Function (l) to boost trust and knowledge exchange.

engagement. Instead of immediate answers, LLMs should ask clarifying or reflective questions helping learners iteratively formulate clinical inquiries. P11 explained, *"I want the system to ask more questions to help me figure out what I really want to ask."* This reflects pedagogical preference for active learning where students are prompted to think critically and refine diagnostic hypotheses.

6.3.4 Supporting Collaborative and Shared Use. Participants emphasized integrating collaborative features to enhance learning beyond individual use. They expected asynchronous teacher-student interaction through case discussion spaces where teachers could offer feedback and review student reasoning post-session. P10 expressed, *"It would be great if I could post my case and get comments from my instructor later."* Additionally, participants hoped to enable peer-based sharing through template repositories where users could upload and access useful prompts or dialogue designs. P2 remarked, *"If I create a good LLM setup, I want to share it with others."*

7 Discussion

Our study explored how LLM tools can enhance medical students' clinical skills in T2D, identifying challenges and desired features from students' perspectives. Through a probe-based study with

14 medical students, we uncovered 11 challenges across dialogue content, presentation, and interaction. Subsequently, co-design workshops clarified specific expectations, leading to ten design considerations (DCs) organized into three dimensions: **content, presentation and interaction** (Table 3).

Rather than discussing each DC in isolation, we reflect on three overarching insights from our findings. We first examine how our learner-centered perspective complements existing research (Section 7.1: **From System Performance to Learner Experience**). We then analyze the interdependencies among DCs across content, presentation, and interaction dimensions (Section 7.2: **Considering Design Considerations as Interconnected**). Next, we distinguish between medical-specific and potentially generalizable considerations (Section 7.3: **Medical-specific Considerations And Potentially Generalizable Considerations**). Finally, we discuss how LLM tools might be integrated into broader medical education contexts (Section 7.4: **How LLM Tools Might Be Integrated Into Broader Medical Education**).

7.1 From System Performance to Learner Experience

Existing LLM research in medical education has focused on two areas: technical performance and educational outcomes. Studies have assessed whether LLMs can pass medical licensing exams [55, 90], generate accurate diagnoses [71], or produce competent clinical communications [74]. Systems like SOPHIE [12] and MedSimAI [31] have advanced emotional modeling for simulated patients. Overall, a substantial body of research has investigated LLMs' capabilities and their impact on learning outcomes within instructional systems.

Our study complements this line of work by foregrounding the learner experience. We observed that even when content was medically accurate, participants encountered barriers that impeded learning. P10, for instance, described simulated patients as "too cooperative"—not because the information was inaccurate, but because the interaction lacked the clinical unpredictability essential for fostering reasoning under uncertainty [57, 69]. P3 reviewed dialogues repeatedly to locate key information, not due to conceptual difficulty, but because the presentation imposed unnecessary cognitive load [50]. Similarly, P12 expressed frustration that feedback "didn't tell me I was mistaken or why": while the system generated correct information, it did not scaffold learners' diagnostic reasoning processes [34].

These observations echo broader findings in technology-enhanced medical education. Cook et al.'s meta-analyses [20, 21] showed that simulation effectiveness depends on instructional design, not just technological sophistication. McGaghie et al. [51] found that simulation yields superior outcomes only when combined with deliberate practice. Our findings extend these insights to LLM contexts: **learner experience deserves attention as a evaluation dimension**, complementing technical capability and learning outcomes.

7.2 Considering Design Considerations as Interconnected

We organized DCs into content, presentation, and interaction for analytical clarity. However, our findings reveal that these dimensions are interdependent. This interdependence has important implications for system design.

Consider DC1 (Dialogue Realism). Though categorized under content, achieving authenticity requires coordinated efforts across all three dimensions. At the content level, it demands complex patient behavior with emotional fluctuation. This extends beyond what current systems like SOPHIE [12] have achieved in emotional modeling. At the presentation level, it requires naturalistic, colloquial language. P9 noted: "The model's responses are overly written or mechanical." This observation aligns with Bateman et al. [6], who found that virtual patient authenticity depends on linguistic naturalness, not just medical accuracy. At the interaction level, participants expected patients to initiate conversations proactively, mirroring real encounters where patients drive dialogue [42]. Addressing any single dimension while neglecting others yields limited improvement. A system with sophisticated emotional content but stilted language still feels artificial [46]. Natural language without proactive patient behavior still lacks clinical realism.

DC4 (Effective Information Presentation) shows similar interdependence. Participants wanted clear, organized information displays. Effective presentation, however, presupposes accurate content and interactive features that support active engagement [18, 50]. P9 requested the ability to manually highlight key points, indicating that presentation effectiveness depends on interactive affordances. This resonates with Selenite's [48] finding that comprehensive overviews are most effective when users can actively explore and restructure information.

DC8 (Feedback in Role-Play Scenarios) illustrates a third pattern of interdependence. Effective feedback requires content accuracy for correct diagnostic evaluation, presentation clarity for comprehensible communication, and appropriate interaction timing [34]. P12's frustration reflects failure across dimensions. The content may have been correct, but presentation lacked specificity and interaction lacked scaffolding. Research on feedback in medical education [53] emphasizes that feedback timing and framing matter as much as accuracy.

This cross-dimensional pattern contrasts with tendencies in existing research to optimize dimensions separately. VR simulators [5, 47, 70] enhanced clinical realism through immersive presentation but often relied on fixed scripts, limiting adaptive content and interaction. Zhou et al.'s [87] projection-based mannequin system improved visual presentation but did not address content adaptation. Multimodal virtual patient systems [7, 64] improved presentation modalities but offered content that could not dynamically adapt to learner levels. Recent work on role-switching pedagogical agents [88, 89] advanced interaction design but focused primarily on agent behavior rather than content-presentation coordination. Thus, we suggest that **explicit attention to cross-dimensional coordination** would benefit future LLM-based educational tool design. LLMs' generative capabilities create new opportunities: unlike script-bound predecessors, LLMs can potentially adjust content complexity, presentation scaffolding, and interaction patterns simultaneously [1, 84]. Realizing this potential requires design frameworks that treat dimensions as interconnected.

Table 3: Overview of Design Considerations

DC	Description	Category	Source
DC1	Dialogue Realism with Patient Authenticity	Content	C1, §6.1.1
DC2	Personalization & Dynamic Role	Content	C2-3, §6.1.2
DC3	Linking Scenarios for Workflow Continuity	Content	C4, §6.1.3
DC4	Effective Information Presentation	Presentation	C5-6, §6.2.1-2
DC5	Multimodal Input & Output	Presentation	C7, §6.2.3
DC6	Output Standardization with EBM	Presentation	C8, §6.2.4
DC7	User Guidance in RS Scenarios	Interaction	C9, §6.3.1
DC8	Feedback in RS Scenarios	Interaction	C10, §6.3.2
DC9	Guidance in Q&A Scenarios	Interaction	C11, §6.3.3
DC10	Collaboration Mechanisms	Interaction	§6.3.4

7.3 Medical-specific Considerations And Potentially Generalizable Considerations

Among our ten DCs, some are medical-specific while others may generalize to other LLM-based learning contexts.

Medical-specific considerations. DC1 (Dialogue Realism with Patient Authenticity), DC3 (Linking Scenarios for Workflow Continuity, and DC6 (Output Standardization with EBM) address needs unique to clinical education. DC1 serves diagnostic training beyond surface-level realism. In real practice, patient hesitation, emotional fluctuation, and communication barriers are themselves diagnostic cues [57, 62]. A patient who avoids eye contact or gives vague answers may signal psychological distress or hidden concerns. Training with overly cooperative simulated patients fails to develop this sensitivity. DC3 addresses the longitudinal nature of clinical care. Unlike learning tasks with clear endpoints, clinical reasoning requires tracking patients across multiple encounters [30]. A T2D patient seen for routine follow-up may present new complications months later. Students need practice managing this continuity. DC6 reflects the high-stakes nature of medical decisions. When participants ask about medication prescription, they need responses grounded in clinical guidelines and current evidence [35]. Trust requirements in medicine exceed those in domains where errors are more easily reversible.

Potentially generalizable considerations. Other DCs resonate with challenges in LLM-based education broadly. Adaptive scaffolding (DC2, DC7, DC9) echoes CodeAid's finding that programming students need different support levels depending on their prior knowledge [40]. Cognitive load management (DC4, DC5) applies whenever learners must process complex information—whether medical cases or data analysis [50]. Feedback design (DC8) parallels MatlabTutee's work on providing constructive responses that promote learning rather than just correctness [65]. Collaboration mechanisms (DC10) connect to the Jigsaw Agent's exploration of peer learning with AI support [23]. These considerations are not unique to medicine, but clinical contexts shape their implementation. Scaffolding in medicine must account for patient safety. Feedback must balance encouraging learners with maintaining diagnostic accuracy.

7.4 How LLM Tools Might Be Integrated Into Broader Medical Education

Our findings raise questions about how LLM tools integrate into broader medical education contexts. Participants articulated expectations across three areas: role functionality, human-AI collaboration, and system integration.

Regarding **Role Functionality**, participants expected LLMs to assume multiple concurrent roles, including simulated patient, Socratic tutor, and expert consultant. They expressed a need for systems that could dynamically shift between these functions based on learning situations. This multi-role expectation challenges designs that position LLMs in singular functions, resonating with Zhu et al.'s [89] findings on dynamic role-switching agents.

Regarding **Human-AI collaboration**, participants emphasized the irreplaceable value of human expertise despite their enthusiasm for LLM capabilities. They noted that certain clinical procedures require validation from experienced instructors rather than LLM recommendations alone. Participants also desired asynchronous spaces where instructors could review student reasoning (DC10). This aligns with findings that instructor involvement remains critical for effective simulation [19, 35].

Regarding **System Integration**, participants proposed connecting LLM tools with hospital systems, particularly EMRs. This vision positions LLM tools as embedded workflow components, consistent with situated learning theory [15]. Such integration offers opportunities for authentic case complexity but raises challenges around privacy and maintaining learning scaffolds in production environments. While participants articulated clear expectations, realizing them involves navigating several tensions that warrant future research attention.

Student preferences versus pedagogical principles Participants often preferred receiving direct answers, yet learning theory suggests that guided discovery may be more effective for developing clinical reasoning [37, 82]. This tension invites deliberate decisions about when to provide direct information versus Socratic questioning, which likely vary by learning objective and student competence [40, 61].

Potential risks of AI in clinical education Students may develop over-reliance on AI feedback, potentially affecting independent reasoning over time. LLM-simulated patients may also reflect demographic biases present in training data [66], and uneven access

to sophisticated tools could contribute to educational disparities [10].

The authenticity paradox Participants desired both greater realism and greater scaffolding, yet these expectations may conflict. Authentic clinical encounters typically offer limited scaffolding [43]. Progressive authenticity, in which systems gradually introduce complexity as learner competence develops [51], offers one possible resolution.

7.5 Limitations

Our research examines medical students' use of LLM tools for T2D clinical skills through formative probes and co-design workshops. Findings illuminate user challenges and expectations, while limitations suggest future research directions. First, recruitment was limited to medical students from Chinese universities. Participants' backgrounds—varied by specialization and prior experience—may influence outcomes. This study observed students from five-year undergraduate programs; broader sampling across training stages is needed. The focus on diabetes also calls for extension to other clinical domains. Second, though observations and discussions yielded rich insights, study duration was relatively short. Long-term difficulties and adaptive behaviors remain unexamined. Future work should develop LLM-based tools informed by our findings and evaluate them through longitudinal deployment. Third, methodological choices shaped the findings in specific ways. The probe-based approach enabled exploratory insight but limited observation of naturalistic interaction with fully functional systems. Participants engaged with researcher-designed prompts and scenarios rather than self-directed queries, potentially narrowing the range of interaction patterns. Co-design workshops elicited expectations and design ideals but may have emphasized envisioned rather than actual use. Fourth, interpreting observed limitations requires distinguishing among interacting factors: LLM output quality (e.g., occasional oversimplification or hallucination), probe design decisions (case complexity and scaffolding), and user characteristics.

Future research should address these limitations through several directions. First, deploying functional systems based on our design considerations and investigating actual long-term usage patterns would provide ecological validity. Second, expanding participant populations across different medical specializations, disease types, and cultural contexts would test generalizability. Third, controlled experiments isolating the effects of content, presentation, and interaction dimensions could empirically validate the observed interdependence patterns. Fourth, investigating when direct answers facilitate versus hinder learning would help navigate the tension between student preferences and pedagogical principles. Finally, future designs should incorporate mechanisms for promoting learner independence and auditing for representational biases to mitigate identified risks.

8 Conclusion

This study explores the main challenges and corresponding attitudes faced by medical students as well as expectations regarding using LLM medical education tools. Our research method comprised two main phases. First of all, we developed a technology probe platform integrating five relevant studies from previous research. Through this platform, we conducted user experiments, interviews, and surveys with 14 medical students. Our analysis revealed insights into three key areas: conversation interaction, content presentation, and content itself. In the second phase, we organized a co-design workshop to explore optimizations for the design of LLM tools. Participants expressed needs for LLM as personalized customized LLM tools, enhanced multimodal and stable content presentation, and structured knowledge frameworks. Finally, based on the insights from the user study using the probe platform and co-design workshop, we proposed interconnected design considerations for LLM medical education tools that shift focus from system performance to learner experience, distinguish medical-specific from generalizable elements, and explore integration into broader educational contexts.

Our work contributes to the HCI and medical education communities in several ways. We provide a user-centric perspective on the potential of leveraging LLM in clinical skill learning scenarios. Additionally, we highlight the challenges and opportunities of using LLM-powered tools to enhance medical students' T2D clinical skills learning. Furthermore, we offer design implications for researchers and designers interested in developing and promoting LLM-based clinical learning for medical students.

Acknowledgments

The authors thank Gang YUAN (The First Affiliated Hospital, Sun Yat-sen University), Lin MA (The Seventh Affiliated Hospital, Sun Yat-sen University), and Jinying LI (Shunde Hospital of Jinan University) for their guidance and support.

This work is partially supported by the Guangzhou-HKUST(GZ) Joint Funding Project (No. 2024A03J0617), Education Bureau of Guangzhou Municipality Funding Project (No. 2024312152), Guangzhou Higher Education Teaching Quality and Teaching Reform Project (No. 2024YBJG070), Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007), the Project of DEGP (No. 2023KCXTD042), and the Guangzhou Science and Technology Program City-University Joint Funding Project (No. 2023A03J0001).

References

[1] Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, Javaid Sheikh, et al. 2023. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Medical Education* 9, 1 (2023), e48291. https://doi.org/10.2196/48291

[2] Sedat Arslan. 2023. Exploring the potential of Chat GPT in personalized obesity treatment. *Annals of biomedical engineering* 51, 9 (2023), 1887–1888.

[3] Tugba Barlas, Alev Eroglu Altinova, Mujde Akturk, and Fusun Balos Toruner. 2024. Credibility of ChatGPT in the assessment of obesity in type 2 diabetes according to the guidelines. *International Journal of Obesity* 48, 2 (2024), 271–275.

[4] S Barry Issenberg, William C Mcgaghie, Emil R Petrusa, David Lee Gordon, and Ross J Scalese. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical teacher* 27, 1 (2005), 10–28.

[5] Sandra Barteit, Lucia Lanfermann, Till Bärnighausen, Florian Neuhann, Claudia Beiersmann, et al. 2021. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. *JMIR serious games* 9, 3 (2021), e29080.

[6] J Bateman, M Allen, D Samani, J Kidd, and D Davies. 2013. Virtual patient design: exploring what works and why. A grounded theory study. *Medical Education* 47, 6 (2013), 595–606. https://doi.org/10.1111/medu.12151

[7] Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. 2025. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences* 118 (2025), 102601. https://doi.org/10.1016/j.lindif.2024.102601

[8] Lourdes García Blasco, Pedro J Pinés Corrales, Felicia Hanzu, Alberto Fernández Martínez, Irene Bretón Lesmes, Javier Escalada San Martín, et al. 2023. A survey on the perception of the specialty of Endocrinology and Nutrition among students preparing for the entrance exam for medical specialty training in Spain. *Endocrinología, Diabetes y Nutrición (English ed.)* 70, 4 (2023), 240–244.

[9] Susanne Bødker, Christian Dindler, and Ole Sejer Iversen. 2017. Tying Knots: Participatory Infrastructuring at Work. *Computer Supported Cooperative Work (CSCW)* 26, 1-2 (2017), 245–273. https://doi.org/10.1007/s10606-017-9268-y

[10] Vikas L Bommineni, Sanaea Bhagwagar, Daniel Balcarcel, Vishal Bommineni, Christos Davazitkos, and Donald Boyer. 2023. Performance of ChatGPT on the MCAT: the road to personalized and equitable premedical learning. *MedRxiv* (2023), 2023–03. https://doi.org/10.1101/2023.03.24.23287713

[11] William F Bond, Richard L Lammers, Linda L Spillane, Rebecca Smith-Coggins, Rosemarie Fernandez, Martin A Reznek, John A Vozenilek, and James A Gordon. 2007. The use of simulation in emergency medicine: a research agenda. *Academic Emergency Medicine* 14, 4 (2007), 353–363.

[12] Mehdi Boukhechba and Ehsan Hoque. 2025. SOPHIE: An AI-Driven Virtual Patient for Serious Illness Communication Skills Training. *arXiv preprint arXiv:2505.02694* (2025). https://arxiv.org/abs/2505.02694

[13] Paul Bradley. 2006. The history of simulation in medical education and possible future directions. *Medical education* 40, 3 (2006), 254–262.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[15] John Seely Brown, Allan Collins, and Paul Duguid. 1989. Situated cognition and the culture of learning. *1989* 18, 1 (1989), 32–42.

[16] Elizabeth Charters. 2003. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal* 12, 2 (2003).

[17] Seung Min Chung and Min Cheol Chang. 2024. Assessment of the information provided by ChatGPT regarding exercise for patients with type 2 diabetes: a pilot study. *BMJ Health & Care Informatics* 31, 1 (2024).

[18] Ruth C. Clark and Richard E. Mayer. 2016. *e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning* (4th ed.). John Wiley & Sons, Hoboken, NJ. https://doi.org/10.1002/9781119239086

[19] DA Cook and MM Triola. 2009. Virtual patients: a critical literature review and proposed next steps. *Medical Education* 43, 4 (2009), 303–311. https://doi.org/10.1111/j.1365-2923.2008.03286.x

[20] D. A. Cook, R. Brydges, S. J. Hamstra, et al. 2012. Comparative effectiveness of technology-enhanced simulation versus other instructional methods: a systematic review and meta-analysis. *Simulation in Healthcare* 7, 5 (2012), 308–320.

[21] D. A. Cook, R. Hatala, R. Brydges, et al. 2011. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *JAMA* 306, 9 (2011), 978–988.

[22] A. Coppola, Loredana Sasso, A. Bagnasco, A. Giustina, and C. Gazzaruso. 2016. The role of patient education in the prevention and management of type 2 diabetes: an overview. *Endocrine* 53 (2016), 18–27. https://doi.org/10.1007/s12020-015-0775-7

[23] Kevin Doherty, Jocelyn Fernandez, Leanne Hirshfield, et al. 2025. Piecing Together Teamwork: A Responsible Approach to an LLM-based Educational Jigsaw Agent. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3706598.3713349

[24] Feishu. 2023. The Introduction of Feishu. https://www.feishu.cn/en/ (Accessed on 16/05/2023).

[25] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: Cultural probes. *interactions* 6, 1 (January 1999), 21–29. https://doi.org/10.1145/291224.291235 Influential paper on cultural probes methodology, cited 2,168+ times.

[26] A. Gayef. 2019. Using simulated patients in medical and health professions education. *SHS Web of Conferences* (2019). https://doi.org/10.1051/SHSCONF/20196601016

[27] Masaki Goda, Goshiro Yamamoto, Chang Liu, Kazumasa Kishimoto, Sho Mitarai, Yukiko Mori, and Tomohiro Kuroda. 2024. Virtual Patientization: A Playable Design for Clinical Ultrasound Training by Embedding Virtual Lesions. In *Companion Proceedings of the 2024 Annual Symposium on Computer-Human Interaction in Play* (Tampere, Finland) *(CHI PLAY Companion '24)*. Association for Computing Machinery, New York, NY, USA, 104–108. https://doi.org/10.1145/3665463.3678788

[28] Nikhil Gopalakrishnan, Aishwarya Joshi, Jay Chhablani, Naresh Kumar Yadav, Nikitha Gurram Reddy, Padmaja Kumari Rani, Ram Snehith Pulipaka, Rohit Shetty, Shivani Sinha, Vishma Prabhu, et al. 2024. Recommendations for initial diabetic retinopathy screening of diabetic patients using large language model-based artificial intelligence in real-life case scenarios. *International Journal of Retina and Vitreous* 10, 1 (2024), 11.

[29] Ronald M Harden. 1999. AMEE Guide No. 14: Outcome-based education: Part 1-An introduction to outcome-based education. *Medical teacher* 21, 1 (1999), 7–14.

[30] I. Hege, A. A. Kononowicz, N. B. Berman, B. Lenzer, and J. Kiesewetter. 2018. Advancing clinical reasoning in virtual patients – development and application of a conceptual framework. *GMS Journal for Medical Education* 35, 1 (2018).

[31] Kevin Hicke, Michael Kane, Mohammed Alabduljabbar, and Ehsan Hoque. 2025. MedSimAI: An AI-Powered Communication Skills Training Tool for Healthcare Professionals. *arXiv preprint arXiv:2503.05793* (2025).

[32] Janice D. Ho and Vincent C. Woo. 2016. A Study of Diabetes Teaching in Canadian Medical Schools. *Canadian journal of diabetes* 40, 2 (April 2016), 149–151. https://doi.org/10.1016/j.jcjd.2015.08.017

[33] Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, and Moritz Mahling. 2024. A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study. *JMIR Med Educ* 10 (16 Jan 2024), e53961. https://doi.org/10.2196/53961

[34] Jonathan S Ilgen, Irene W Y Ma, Rose Hatala, and David A Cook. 2013. Giving feedback in medical education: verification of recommended techniques. *Journal of general internal medicine* 28, 10 (2013), 1360–1366. https://doi.org/10.1007/s11606-013-2451-1

[35] S. B. Issenberg, W. C. McGaghie, E. R. Petrusa, et al. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher* 27, 1 (2005), 10–28.

[36] Joel Jaskari, Jaakko Sahlsten, Paula Summanen, Jukka Moilanen, Erika Lehtola, Marjo Aho, Elina Säpyskä, Kustaa Hietala, and Kimmo Kaski. 2024. DR-GPT: A large language model for medical report analysis of diabetic retinopathy patients. *PLOS ONE* 19 (2024). https://api.semanticscholar.org/CorpusID:267018930

[37] Slava Kalyuga. 2007. Expertise reversal effect and its implications for learner-tailored instruction. *Educational psychology review* 19, 4 (2007), 509–539.

[38] Jessica Kaplonyi, K. Bowles, D. Nestel, D. Kiegaldie, S. Maloney, T. Haines, and Cylie M. Williams. 2017. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. *Medical Education* 51 (2017). https://doi.org/10.1111/medu.13387

[39] Harleen Kaur, Jina Huh-Yoo, and Tawanna Dillahunt. 2022. The Emotional Labor of AI in Health Professional Training. In *Proceedings of the 2022 ACM Conference on Computer-Supported Cooperative Work (CSCW)*. 1–25.

[40] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 650, 20 pages. https://doi.org/10.1145/3613904.3642773

[41] Tamkeen Khan, Gregory D Wozniak, and Kate Kirley. 2019. An assessment of medical students' knowledge of prediabetes and diabetes prevention. *BMC medical education* 19 (2019), 1–7.

[42] R. Kneebone. 2005. Evaluating clinical simulations for learning procedural skills: a theory-based approach. *Academic Medicine* 80, 6 (2005), 549–553.

[43] R. L. Kneebone, D. Nestel, C. Vincent, and A. Darzi. 2007. Complexity, risk and simulation in learning procedural skills. *Medical Education* 41 (2007), 808–814.

[44] Sung Ju Lee, Juho Kim, and Rena Kizilcec. 2024. Student Perceptions of AI Tutors: Trust, Transparency, and Pedagogical Alignment. In *Proceedings of the 2024 ACM Conference on Learning at Scale (L@S)*. 1–13.

[45] Q. Vera Liao, Shazeda Ahmed, Jessica Hullman, Jeff Hancock, and Daniel Weld. 2023. Designing Human-AI Interaction for Supporting Reflective Practices in Education. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[46] Chao Liu, Mingyang Su, Yan Xiang, Yuru Huang, Yiqian Yang, Kang Zhang, and Mingming Fan. 2025. Toward Enabling Natural Conversation with Older Adults via the Design of LLM-Powered Voice Agents that Support Interruptions and Backchannels. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 163, 22 pages. https://doi.org/10.1145/3706598.3714228

[47] Chang Liu, Felicia Fang-Yi Tan, Shengdong Zhao, Abhiram Kanneganti, Gosavi Arundhati Tushar, and Eng Tat Khoo. 2024. Facilitating Virtual Reality Integration in Medical Education: A Case Study of Acceptability and Learning Impact in Childbirth Delivery Training. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 458, 14 pages. https://doi.org/10.1145/3613904.3642100

[48] Michael Xieyang Liu, Tianying Chen, Franklin Mingzhe Li, Tongshuang Wu, Brad A. Myers, and Aniket Kittur. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York,

NY, USA, Article 174, 20 pages. https://doi.org/10.1145/3613904.3642149
[49] Viveta Lobo, Andrew Q Stromberg, and Peter A Rosston. 2017. The Sound Games: Introducing Gamification into Stanford's Orientation on Emergency Ultrasound. *Cureus* 9 (2017). https://api.semanticscholar.org/CorpusID:2316839
[50] Richard E. Mayer. 2009. *Multimedia Learning (2nd Edition)*. Cambridge University Press.
[51] W. C. McGaghie, S. B. Issenberg, E. R. Cohen, et al. 2011. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Academic Medicine* 86, 6 (2011), 706–711.
[52] Kartik Mittal and Minakshi Dhar. 2023. Use of ChatGPT by physicians to build rehabilitation plans for the elderly: a mini-review of case studies. *Journal of the Indian Academy of Geriatrics* 19, 2 (2023), 86–93.
[53] Carol-Anne E Moulton, Adam Dubrowski, Helen MacRae, Beverley Graham, Elliot Grober, and Richard K Reznick. 2006. Teaching surgical skills: what kind of practice makes perfect? A randomized, controlled trial. *Annals of surgery* 244, 3 (2006), 400–409. https://doi.org/10.1097/01.sla.0000234808.85789.6a
[54] Sarah E Myers, Nicholas R Bender, Marina A Seidel, and Ruth S Weinstock. 2021. Diabetes SPECIAL (Students Providing Education on Chronic Illness and Lifestyle): a novel preclinical medical student elective. *Perspectives on Medical Education* 10 (2021), 312–315.
[55] Takahiro Nakao, Soichiro Miki, Yuta Nakamura, Tomohiro Kikuchi, Yukihiro Nomura, Shouhei Hanaoka, Takeharu Yoshikawa, and Osamu Abe. 2024. Capability of GPT-4V(Ision) in the Japanese National Medical Licensing Examination: Evaluation Study. *JMIR medical education* 10 (March 2024), e54393. https://doi.org/10.2196/54393
[56] Carine M Nassar, Robert Dunlea, Alex Montero, April Tweedt, and Michelle F Magee. 2023. Feasibility and preliminary behavioral and clinical efficacy of a diabetes education chatbot pilot among adults with type 2 diabetes. *Journal of Diabetes Science and Technology* (2023), 19322968231178020.
[57] G Norman. 2006. Building on experience—the development of clinical reasoning. *New England Journal of Medicine* 355 (2006), 2251–2252. https://doi.org/10.1056/NEJMp068134
[58] Lorelli S. Nowell, Jill M. Norris, Deborah E. White, and Nancy J. Moules. 2017. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods* 16, 1 (Dec. 2017), 1609406917733847. https://doi.org/10.1177/1609406917733847
[59] Yasuharu Okuda, Ethan O Bryson, Samuel DeMaria Jr, Lisa Jacobson, Joshua Quinones, Bing Shen, and Adam I Levine. 2009. The utility of simulation in medical education: what is the evidence? *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine* 76, 4 (2009), 330–343.
[60] Pilar Ortega, Christian González, Itzel López-Hinojosa, Yoon Soo Park, and Jorge A Girotti. 2022. Medical Spanish endocrinology educational module. *MedEdPORTAL* 18 (2022), 11226.
[61] Jin Park, Yue Wang, Vijay Natarajan, and Divya Kumar. 2024. Guiding Students in Using LLMs in Supported Learning Environments: Effects on Interaction Dynamics, Learner Performance, Confidence, and Trust. In *Proceedings of the ACM on Computer-Supported Cooperative Work and Social Computing (CSCW)*. https://doi.org/10.1145/3687038
[62] V. L. Patel and G. J. Groen. 1991. Knowledge based solution strategies in medical reasoning. *Cognitive Science* 15, 1 (1991), 91–116.
[63] Anastasia Pozdnyakova, Michael Andersen, Sebastian Cruz, Hannah Wilson, Mikhail Pakvasa, and Julie Oyler. 2019. Assessing quality of diabetes care and medical student volunteer knowledge of diabetes care at the University of Chicago Community Health Clinic. *American Journal of Medical Quality* 34, 6 (2019), 621–621.
[64] NPA Quail and JG Boyle. 2023. Twine virtual patient games as an online resource for undergraduate diabetes acute care education. *BMC Medical Education* 23, 1 (2023), 417. https://doi.org/10.1186/s12909-023-04398-3
[65] Kantwon Rogers, Anh Truong, Marissa Radensky, Sonia Chernova, Zahra Ashktorab, Qian Li, Farhana Mohsen, and James Witschey. 2025. Playing Dumb to Get Smart: Creating and Evaluating an LLM-based Teachable Agent within University Computer Science Classes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3706598.3713644
[66] Malik Sallam. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, Vol. 11. MDPI, 887.
[67] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.
[68] Elizabeth B.-N. Sanders and Pieter Jan Stappers. 2012. *Convivial Toolbox: Generative Research for the Front End of Design*. BIS Publishers, Amsterdam. Comprehensive guide to generative design research methods.
[69] H. G. Schmidt, G. R. Norman, and H. P. A. Boshuizen. 1990. A cognitive perspective on medical expertise: theory and implications. *Academic Medicine* 65, 10 (1990), 611–621.
[70] N. E. Seymour, A. G. Gallagher, S. A. Roman, et al. 2002. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of Surgery* 236, 4 (2002), 458–464.
[71] Kiyoshi Shikino, Taro Shimizu, Yuki Otsuka, Masaki Tago, Hiromizu Takahashi, Takashi Watari, Yosuke Sasaki, Gemmei Iizuka, Hiroki Tamura, Koichi Nakashima, Kotaro Kunitomo, Morika Suzuki, Sayaka Aoyama, Shintaro Kosaka, Teiko Kawahigashi, Tomohiro Matsumoto, Fumina Orihara, Toru Morikawa, Toshinori Nishizawa, Yoji Hoshina, Yu Yamamoto, Yuichiro Matsuo, Yuto Unoki, Hirofumi Kimura, Midori Tokushima, Satoshi Watanuki, Takuma Saito, Fumio Otsuka, and Yasuharu Tokuda. 2024. Evaluation of ChatGPT-Generated Differential Diagnosis for Common Diseases With Atypical Presentation: Descriptive Research. *JMIR medical education* 10 (June 2024), e58758. https://doi.org/10.2196/58758
[72] Makoto Shiraishi, Haesu Lee, Koji Kanayama, Yuta Moriwaki, and Mutsumi Okazaki. 2024. Appropriateness of artificial intelligence chatbots in diabetic foot ulcer management. *The International Journal of Lower Extremity Wounds* (2024), 15347346241236811.
[73] Nicole Slater, Anthony Todd, and Abby Grimm. 2020. Pharmacy students as educators: An interprofessional approach to insulin management education. *Currents in pharmacy teaching and learning* 12, 6 (2020), 689–693.
[74] Kannan Sridharan and Reginald P. Sequeira. 2024. Evaluation of Artificial Intelligence-Generated Drug Therapy Communication Skill Competencies in Medical Education. *British journal of clinical pharmacology* (July 2024). https://doi.org/10.1111/bcp.16144
[75] Haonan Sun, Kai Zhang, Wei Lan, Qiufeng Gu, Guangxiang Jiang, Xue Yang, Wanli Qin, and Dongran Han. 2023. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. *Journal of Medical Internet Research* 25 (2023), e51300.
[76] Sing Yee Toh, Chang Cai, Li Rong Wang, Xiaoyin Bai, Joanne Ngeow, and Xiuyi Fan. 2025. The Effect of Explainable AI and Uncertainty Quantification on Medical Students' Perspectives of Decision-Making AI: A Cancer Screening Case Study. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 515, 13 pages. https://doi.org/10.1145/3706599.3719791
[77] Eva Tseng, Raquel C. Greer, Paul O'Rourke, Hsin-Chieh Yeh, Maura M. McGuire, Jeanne M. Clark, and Nisa M. Maruthur. 2017. Survey of Primary Care Providers' Knowledge of Screening for, Diagnosing and Managing Prediabetes. *Journal of General Internal Medicine* 32, 11 (Nov. 2017), 1172–1178. https://doi.org/10.1007/s11606-017-4103-1
[78] L Varadhan, C Rowley, and GI Varughese. 2010. Medical student evaluation of clinical teaching sessions in diabetes and endocrinology: a quantitative analysis based on formatted feedback over 1 year. *Diabetic medicine* 27, 11 (2010), 1329–1331.
[79] Jacqueline Vaughn, Shannon H Ford, Melissa Scott, Carolyn Jones, and Allison Lewinski. 2024. Enhancing Healthcare Education: Leveraging ChatGPT for Innovative Simulation Scenarios. *Clinical Simulation in Nursing* 87 (2024), 101487.
[80] Josina Vink, Katarina Wetter-Edman, Bo Edvardsson, and Bård Tronvoll. 2016. Understanding the Influence of the Co-Design Process on Well-Being. In *Service Design Geographies: Proceedings of the ServDes 2016 Conference*, Nicola Morelli, Amalia de Götzen, and Francesco Grani (Eds.). Linköping University Electronic Press, Copenhagen, Denmark, 390–402.
[81] P. Vlasses. 1990. Medical Decision Making. *Annals of Pharmacotherapy* 24 (1990), 103 – 103. https://doi.org/10.1177/106002809002400128
[82] Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.
[83] Alan Xu, Abraham Akinyemi, and Michael Bernstein. 2022. CareAI: Designing Human-AI Collaboration in Empathetic Care Training Simulations. In *Proceedings of the 2022 ACM Conference on Computer-Supported Cooperative Work (CSCW)*. 1–29.
[84] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* 18, 6 (2024), 1–32.
[85] Ming Yin, Ruotong Sun, and Jeff Hancock. 2023. Teacher or Teammate? Exploring AI's Role in Practice-Based Learning Environments. In *Proceedings of the 2023 ACM Conference on Computer-Supported Cooperative Work (CSCW)*. 1–27.
[86] Amy X. Zhang, Jennifer Wortman Vaughan, and Hanna Wallach. 2023. Co-Prompting Strategies for Supporting Domain-Specific Reasoning with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.
[87] Guoyang Zhou, Amy Nagle, George Takahashi, Tera Hornbeck, Ann Loomis, Beth Smith, Bradley Duerstock, and Denny Yu. 2022. Bringing Patient Mannequins to Life: 3D Projection Enhances Nursing Simulation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 565, 15 pages. https://doi.org/10.1145/3491102.3517562
[88] Xiuqi Tommy Zhu, Heidi Cheerman, Minxin Cheng, Sheri R Kiami, Leanne Chukoskie, and Eileen McGivney. 2025. Designing VR Simulation System for Clinical Communication Training with LLMs-Based Embodied Conversational

Agents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 181, 9 pages. https://doi.org/10.1145/3706599.3719693

[89] Zihao Zhu, Xiyun Hu, Lijun Zhu, and Karthik Ramani. 2025. Exploring LLM-Powered Role and Action-Switching Pedagogical Agents for History Education in Virtual Reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3706598.3713109

[90] Hui Zong, Jiakun Li, Erman Wu, Rongrong Wu, Junyu Lu, and Bairong Shen. 2024. Performance of ChatGPT on Chinese National Medical Licensing Examinations: A Five-Year Examination Evaluation Study for Physicians, Pharmacists and Nurses. *BMC medical education* 24, 1 (Feb. 2024), 143. https://doi.org/10.1186/s12909-024-05125-7

A Background of Clinical Skills in Type 2 Diabetes

T2D is a global health challenge with high incidence and long-term management needs, and its management requires a range of clinical skills applicable to other chronic diseases [32]. However, T2D clinical skills training in current medical education has significant deficiencies in knowledge imparting and practical skills development [41, 78]. Recent surveys indicate that primary care physicians have major knowledge gaps in pre-diabetes screening, diagnosis, and management, with fewer than 20% answering correctly [77].

A cross-sectional study also found that medical students answered less than 50% of questions correctly regarding diabetes prevention and management [41]. These gaps are closely tied to the current medical education framework [78]. About 73% of diabetes specialists believe the existing education inadequately prepares graduates for effective clinical diabetes management [63]. The Society for Endocrinology has further noted that many medical schools do not require diabetes coursework, and curricula rarely cover practical aspects such as insulin types, dose adjustments, infusion management.

At present, the application of LLM in T2D clinical skills training is still in its infancy and mainly focuses on diagnosis rather than medical education applications [36, 73, 75]. For example, while some studies have developed AI dietitians using ChatGPT to make personalized diet recommendations [73, 75] or to evaluate LLM performance in screening for diabetic retinopathy [36], further optimization of LLM tools is needed to better meet the combined clinical skills and practical experience requirements for T2D management in medical education. The potential opportunities LLM technology shows in medical education make T2D an ideal entry point to explore LLM-assisted clinical skills training. Clinical skills training for T2D encompasses three basic modules as follow [22, 26, 38, 81]:

- Interactive Simulated Patients [26, 38] is a key component of T2D clinical skill learning. These patients are trained to act as real patients, simulating symptoms and conditions that are relevant to T2D. This allows students to practice their clinical skills, such as taking medical histories, performing physical examinations, and developing treatment plans. Simulated patients also provide a safe and controlled environment for students to make mistakes and learn from them without compromising patient care.
- Disease Decision-Making [81] is another critical module in T2D clinical skill learning. Students need to understand the pathophysiology of T2D, its complications, and the current treatment guidelines. They should be able to diagnose and manage the disease effectively, taking into account individual patient factors, such as lifestyle, comorbidities, and personal preferences. Disease decision-making involves making informed decisions about treatment options, monitoring, and follow-up care.
- Patient Education [22] is an essential module in T2D clinical skill learning. Medical students need to understand the importance of patient education in managing T2D. They should be able to communicate effectively with patients, providing them with clear and concise information about their condition, treatment options, and self-management strategies.

B Research paper

Author	Title	Key takeaway	Ref
Sun et al.	An AI Dietitian for Type 2 Diabetes Mellitus Management Based on Large Language and Image Recognition Models: Preclinical Concept Validation Study	The project integrates ChatGPT and GPT 4.0, a deep learning-based food recognition model, and a user-friendly application to improve diet-related care for people with type 2 diabetes by addressing knowledge gaps and providing evidence-based nutrition advice.	[75]
Nikhil et al.	Recommendations for initial diabetic retinopathy screening of diabetic patients using large language model-based artificial intelligence in real-life case scenarios	Use hypothetical case scenarios generated by clinicians and AI applications to assess DR Screening timing. Explore the role of artificial intelligence in DR Screening for patients with newly diagnosed diabetes.	[28]
Jaskari et al.	DR-GPT: A large language model for medical report analysis of diabetic retinopathy patients	This study proposes a large language model, DR-GPT, for classifying the severity of DR In unstructured medical reports and shows that LLM can be applied to unstructured medical report databases to classify diabetic retinopathy and have multiple applications.	[36]
Barlas et al.	Credibility of ChatGPT in the assessment of obesity in type 2 diabetes according to the guideline	ChatGPT was asked questions in segments by the endocrinologist about the assessment of obesity and different treatment options based on guidelines. The study assessed the credibility of ChatGPT against obesity assessment guidelines for Type 2 diabetes (T2D).	[3]
Arslan et al.	Exploring the Potential of Chat GPT in Personalized Obesity Treatment	This paper focuses on the potential application of Chat GPT in the treatment of obesity. Chat GPT can provide personalized advice on topics such as nutrition programs, exercise programs, psychological support, etc.	[2]
Shiraishi et al.	Appropriateness of Artificial Intelligence Chatbots in Diabetic Foot Ulcer Management	Evaluate the accuracy of the DFU information provided by ChatGPT according to established guidelines. According to the DFU guidelines, seven AI chatbots were asked clinical questions (CQ) and the accuracy of the answers were analyzed.	[72]
Nassar et al.	Feasibility and Preliminary Behavioral and Clinical Efficacy of a Diabetes Education Chatbot Pilot Among Adults With Type 2 Diabetes	A diabetes education and support chatbot was deployed through the system to recruit adults with type 2 diabetes who had recently completed a 12-week diabetes care management program, collecting data to assess satisfaction, engagement, and initial glucose outcomes.	[56]
Chung et al.	Assessment of the information provided by ChatGPT regarding exercise for patients with type 2 diabetes: a pilot study	Diabetes care and rehabilitation physicians discuss and identify the 14 most common questions patients ask in clinical practice about how to manage type 2 diabetes through exercise	[17]
Mittal et al.	Use of ChatGPT by Physicians to Build Rehabilitation Plans for the Elderly: A Mini-Review of Case Studies	The potential of using the AI language model ChatGPT to develop personalized rehabilitation plans for elderly patients was explored	[52]
Holderried et al.	A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study	An interactive chatbot interface and a specific prompt were developed using GPT-3.5, including chatbot optimized disease scripts and behavioral components.	[33]

C Prompt

C.1 Prompt 1: Simulated Patients

Role: Patient (Ferdinand Wunderlich) Content: "Hello Mr. Wunderlich, in the following scenario, you will assume the role of a patient presenting with specific symptoms. Please respond based solely on the provided medical information. You should answer questions in a manner consistent with a real patient's experience.Your name is Ferdinand Wunderlich. You are a 55-year-old male presenting with the following symptoms: Chief complaints: Nausea and significant weight loss of 10 kg over the past 6 weeks. Chronic fatigue, exhaustion, and a general lack of drive. Muscle cramps, primarily localized in the legs.Medical history: [Further case details...] Additional details: You have experienced these symptoms progressively worsening over the past few months. You have tried various over-the-counter remedies with little to no relief. When responding, please stay in character as Ferdinand Wunderlich. Provide concise and relevant answers to the questions asked, reflecting your personal experience and symptom progression."

C.2 Prompt 2: Testing DR Screening Timing

Role: Diabetes Medical Expert Content: "Hello, I am a diabetes medical expert. You are tasked with providing a DR Screening Score based on the following cases, using the Diabetic Retinopathy (DR) screening guidelines. The DR Screening Score ranges from 0 to 10, which helps determine the urgency and frequency of diabetic retinopathy screening for each patient. The guidelines for screening timing are as follows: Score 0-2: Screen within 5 years. Score 3-5: Screen within 1 year. Score 6-10: Frequent screening may be required. Please provide a DR Screening Score and the corresponding screening time for each of the following cases. Also, give a summary of your findings.

Example Cases: Case 1: Age: 46; Type of DM: Type 2; DM control: Poor; Systemic co-morbidities: Yes; Family History of DM: No; DR Screening Score: 8; Recommendation: Immediate screening. Case 2: Age: 30; Type of DM: Type 1; DM control: Good; Systemic co-morbidities: No; Family History of DM: Yes; DR Screening Score: 2; Recommendation: Within 5 years. Case 3: Age: 55; Type of DM: Type 2; DM control: Poor; Systemic co-morbidities: Yes; Family History of DM: No; DR Screening Score: 8; Recommendation: Immediate screening."

C.3 Prompt 3: Chatbot in Diabetic Foot Ulcers

Role: Diabetes Specialist Content: "Hello, I am a diabetes specialist. Your task is to evaluate responses from a chatbot regarding diabetic foot ulcers (DFU) based on current clinical guidelines. You will assess the chatbot's responses based on the following aspects: The accuracy of clinical problem answers. The grading of recommendations and evidence level. Consistency with established clinical guidelines. The authenticity of the references provided by the chatbot.

Here is an example case to evaluate: Case Example: A 55-year-old male with Type 2 diabetes presents with a non-healing foot ulcer. The patient reports numbness in the foot and occasional burning pain. The wound has been present for 6 weeks and is not improving despite over-the-counter treatments. Please assess the chatbot's response to this case, focusing on the above criteria."

C.4 Prompt 4: Assessing Obesity in T2D According to Guidelines

Role: Endocrinologist Content: "Hello, I am an endocrinologist. You are tasked with providing obesity management recommendations for Type 2 diabetes (T2D) based on current guidelines from the American Diabetes Association (ADA) and the American Association of Clinical Endocrinologists (AACE). The guidelines include: Recommendations on obesity pharmacotherapy, including GLP-1 receptor agonists and GIP receptor agonists. Emphasis on individualized, sustained weight management goals. Updated screening for comorbidities such as heart failure, peripheral arterial disease (PAD), and non-alcoholic fatty liver disease (NAFLD). Please answer the following questions based on these guidelines and provide examples where applicable.

Example Question: A 55-year-old male with Type 2 diabetes and obesity presents with poorly controlled blood sugar despite medication. He is also at high risk for cardiovascular disease due to his hypertension. Based on the ADA and AACE guidelines, what obesity management plan would you recommend for this patient?"

C.5 Prompt 5: Using ChatGPT to Develop Rehabilitation Plans for the Elderly Role: Rehabilitation Specialist

Content: "Hello, I am a rehabilitation specialist. Your task is to design a detailed rehabilitation plan for an elderly patient based on their specific conditions. The plan should cover physical, occupational, and speech rehabilitation, including exercise dosage, timing, frequency, and progression over a 3-week period. Please use the case information provided to build a table with specific rehabilitation plans. Example Case 1: A 73-year-old male with sarcopenia, diabetic neuropathy, and hypertension, who is 1 week post-cervical spine surgery. He has mild chest infection, low oral intake due to pharyngeal muscle weakness, and is clinically improving. Create a detailed 3-week rehabilitation plan that includes exercises, frequency, and progression. Example Case 2: A 73-year-old female with severe diabetic neuropathy of the feet and deforming rheumatoid arthritis of the hands. She has good disease control but limited financial and social support. Develop a detailed

3-week rehabilitation plan, focusing on physical and occupational therapy. Please ensure the plans are tailored to the patient's unique needs, including their medical conditions, functional status, and available support."

D Probe-design user task

Probe 1: Simulating a Patient Participants were asked to role-play as a clinician diagnosing and treating a virtual patient simulated by GPT-3.5. They were instructed to conduct a thorough history-taking, physical examination, formulate a treatment plan, and issue medical orders. *Probe 2: DR Screening Timing* Participants were presented with 20 AI-generated hypothetical case scenarios of DR and were tasked with determining the appropriate timing for DR screening based on the latest guidelines. *Probe 3: Chatbot in DFUs* Participants were assessed on their ability to use a chatbot, including ChatGPT, to provide accurate information on DFUs based on established guidelines. *Probe 4: ChatGPT Credibility in T2D and Obesity* Participants evaluated ChatGPT's credibility in providing guidance on the management of T2D and obesity by answering 20 questions based on guidelines from the American Diabetes Association and American Association of Clinical Endocrinologists. *Probe 5: Rehabilitation Planning for Elderly* Participants were tasked with creating a comprehensive and personalized rehabilitation plan for an elderly patient with T2D using ChatGPT. The plan needed to address the patient's medical, physical, and psychosocial needs.

E Semi-structure Interview Question Outline

General experience inquiry

1. General impression: Can you briefly describe the overall experience of using the probe platform? 2. Major pain points: What major challenges or inconveniences have you encountered during use? Do these pain points seriously affect your experience? 3. Significant Benefits: Conversely, what do you find most appealing or helpful about the probe platform? Please give an example.

Specific feedback on each probe

1. Experience: How do you feel about probe 1 ? Is there anything particularly satisfying to you? 2. Function evaluation: Does the function of probe 1 meet your expectations? If there are unmet needs, please be specific. 3. Improvement suggestions: For probe 1, do you have any suggestions or ideas to improve it? (Repeat the above question for each probe)

Views on integrated probes

1. Interface design: Do you find the integrated probe interface design intuitive and easy to use? Are there any design elements that stand out to you or that you feel need improvement? 2. Interface functions: Do the functions provided by the integrated probe (such as data integration, analysis reports, etc.) meet your study or work needs? Are there any missing features you would like to add? 3. Interactive experience: How smooth and easy is it to switch between multiple probes or work together using integrated probes?

Comparison with existing learning methods and tools for medical education

1. Efficiency comparison: What do you think are the advantages or disadvantages of integrated probes in improving learning efficiency compared to the learning methods and tools you have previously used in medical education? 2. Learning effect: How do you feel about your learning effect (depth of understanding, persistence of memory, etc.) after using the integrated probe? 3. Applicable scenarios: What specific learning or teaching scenarios do you think integrated probes are more suitable for? 4. Willingness to adopt: Based on your current experience, are you considering adopting more integrated probes in your future study or work? Why?

F Codebook for Phase 1: Probe-based Study

Theme	Sub-theme	Example Quote
Dialogue Content Issues	Lack of Realism in RS Scenario	"The patient is too cooperative, which is not realistic." (P10)
	Similar Treatment Answers in Q&A Scenario	"Even though I emphasized the added patient details, the answer given was the same…" (P4)
	Lack of Knowledge Depth Differentiation	"The current responses are basic. I could also find these answers using a search engine." (P3)
	Scenario Isolation	"The modules convey related medical knowledge in the dialogue, but I cannot truly apply them to my clinical analysis." (P1)
Dialogue Presentation Issues	Linear Dialogue Issue	"I forgot the patient's age, medical history, or details of the description. I need to check it again." (P3)
	Information Overload Issue	"Only one or two points are useful. I have to read through the long text to analyze it." (P2)
	Single Modality Issue	"For conditions like foot ulcers and retinopathy, it would be better to provide specific case responses, such as dynamic videos or images." (P10)
	Lack of Standardization	"I cannot tell if it is abnormal." (P14)
Dialogue Interaction Barriers	Lack of Role and Task Guidance	"The system should proactively inform us of the patient's symptoms and feelings like a real patient." (P4)
	Lack of Feedback	"When I was diagnosing, the response seemed different from what I expected, but it did not remind me that my thought process might be wrong." (P12)
	Lack of Targeted Guidance	"Even when I emphasize the key points of the question, the response still deviates." (P8)

G Codebook for Phase 2: Co-design Workshop

Theme	Sub-theme	Example Quote
Expectations for Content Quality	Enriching the Patient Model	"The simulated patient can add some details similar to personality, such as being very picky during blood tests." (P10)
	Building User Personas	"The system could offer more precise content based on the user's grade and rotated departments." (P4)
	Strengthening Scenario Association	"Modules for diabetic foot, fundus examination, and rehabilitation management should be integrated into a complete diagnostic and treatment chain." (P1)
Expectations for Content Presentation	Structured Content Presentation	"If LLM can build an effective knowledge framework after information input, I can use it to review or update my understanding." (P9)
	Key Information Extraction	"It could auto-generate learning recommendations based on my highlights, like daily learning plan reminders." (P9)
	Optional Multi-modality Outputs	"For diseases like retinopathy and diabetic foot ulcers, pictures or videos are essential as text descriptions alone are hard to understand." (P11)
	Evidence-based Medicine Support	"Literature is a must, and it would be better to include publication times, links, and brief abstracts." (P7)
Expectations for Dialogue Interaction	Enhancing Interaction Guidance for RS Scenarios	"If a simulated patient feature is used, it would be easier for everyone if it introduced itself directly." (P3)
	Enhancing Feedback in RS Scenarios	"When I was diagnosing, the response seemed different from what I expected, but it did not remind me that my thought process might be wrong." (P12)
	Enhancing Interaction Guidance for Q&A Scenarios	"Starting with a vague question, then letting the system guess what I might want by asking me more questions." (P11)
	Increasing Collaborative and Shared Function	"If I set up a good LLM, I'd be happy to share it with my classmates. The system could offer a template area for user sharing." (P2)