Generative AI for medical education: Insights from a case study with medical students and an AI tutor for clinical reasoning

Amy Wang*, Roma Ruparel*
Anna Iurchenko, Paul Jhun, Julie Anne Seguin, Patricia Strachan, Renee Wong, Alan Karthikesalingam, Yossi Matias, Avinatan Hassidim, Dale R. Webster, Christopher Semturs, Jonathan Krause, Mike Schaekermann
Google Research, Mountain View, USA

Abstract

Generative Artificial Intelligence (AI), particularly Large Language Models (LLMs), have demonstrated significant potential in clinical reasoning skills such as history-taking and differential diagnosis generation—critical aspects of medical education. This work explores how LLMs can augment medical curricula through interactive learning. We conducted a participatory design process with medical students, residents and medical education experts to co-create an AI-powered tutor prototype for clinical reasoning. As part of the co-design process, we conducted a qualitative user study, investigating learning needs and practices via interviews, and conducting concept evaluations through interactions with the prototype. Findings highlight the challenges learners face in transitioning from theoretical knowledge to practical application, and how an AI tutor can provide personalized practice and feedback. We conclude with design considerations, emphasizing the importance of context-specific knowledge and emulating positive preceptor traits, to guide the development of effective AI-powered tools for medical education.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in collaborative and social computing*; • **Computing methodologies** → *Natural language generation.*

Keywords

Education, Medicine, Generative AI, Large Language Models

ACM Reference Format:
Amy Wang*, Roma Ruparel and Anna Iurchenko, Paul Jhun, Julie Anne Seguin, Patricia Strachan, Renee Wong, Alan Karthikesalingam, Yossi Matias, Avinatan Hassidim, Dale R. Webster, Christopher Semturs, Jonathan Krause, Mike Schaekermann. 2025. Generative AI for medical education: Insights from a case study with medical students and an AI tutor for clinical reasoning. In *Proceedings of (CHI '25 Workshop on Envisioning the Future of Interactive Health).* ACM, New York, NY, USA, 5 pages.

*Both authors contributed equally.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI '25 Workshop on Envisioning the Future of Interactive Health, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).

1 Introduction

The long and arduous path of medical education is traditionally broken into several stages. In the US and Canada, medical education is comprised of two distinct phases: the preclinical and clinical years. The first two years of medical school are preclinical, during which students master the foundational knowledge base essential for medical practice. Instruction typically occurs in the classroom with didactic lectures and laboratories covering clinical sciences such as anatomy, physiology, pathology, and pharmacology [18]. In contrast, the last two clinical years emphasize the application of knowledge in real-world healthcare settings. Students gain practical experience across various medical specialties through rotations in hospitals and clinics. Immersed in this environment, they assume responsibility for patient care under the supervision of resident and attending physicians, learning to apply their foundational knowledge to actual clinical situations and mastering clinical reasoning [11]. In their final year, medical students apply to residency training programs in the specialties of their preference. Upon acceptance to a program and following their successful graduation from medical school, students become interns (first-year residents) and then resident physicians, spending anywhere from three to seven years or more engaged in hands-on specialty training before becoming eligible to practice as independent physicians.

With the advent of Generative AI technologies such as Large Language Models (LLMs), recent work has explored their potential in the medical domain and has demonstrated promising capabilities in clinical reasoning ranging from medical question answering [15], to interactive history taking [19] and accurate differential diagnosis development [10]. Existing work has discussed diverse ways this transformative technology could be implemented in medical education, from augmentation of clinical simulations and personalized tutoring, to facilitation in literature reviews, curriculum development, and assessment of students [1]. LLMs have been fine-tuned specifically on medical information (e.g. question-answering data sets and medical dialogues) [10, 15].

While ample opportunities have been identified, there is a lack of human-centered and participatory design research involving the very users of such future technologies to inform how Generative AI can best help address their needs and support their goals. Our study focuses on addressing the major transition between the preclinical and clinical phases of medical education that requires medical learners to quickly pivot and adapt from book to bedside [9]. To do so, we make the following contributions:

- We conducted a design ideation workshop with an interdisciplinary panel of medical education experts and medical students, to identify opportunities for supporting the medical education process with GenAI, and implemented an interactive prototype reflecting key ideas from the workshop.
- We conducted a user study with a separate set of medical students and residents who had not participated in the workshop. We report insights from this study, including learning needs and practices as well as initial impressions and experiences interacting with our clinical reasoning AI tutor.
- We discuss design considerations for the future development of Generative AI-powered systems for medical education, emphasizing the importance of context-specific knowledge and emulating positive preceptor traits, to guide the development of effective AI-powered tools for medical education.

2 Related Work

Medical Education Technology. Technology has already found its way into various facets of medical education. Students report utilizing a varied collection of third-party online study materials such as question banks (e.g., UWorld), videos (e.g., YouTube, Boards and Beyond, digital flash cards (e.g., Anki), and other website-based informational resources (e.g., AMBOSS) [6]. Medical institutions, too, have adapted to the technology era. Examples include virtual lectures and e-learning modules, especially in a post-pandemic world [5], and educational games [4].

Current human-computer interaction (HCI) literature has explored the use of technology in medical training as well. For example, one paper leveraged augmented reality to make a simulation-based learning experience about the topic of stroke in an undergraduate nursing course more immersive [21]. Another study applied virtual reality to introductory obstetrics training for a cohort of fourth year Singaporean medical students to supplement traditional in-person training and ensure all students were exposed to the necessary breadth of education experiences [8].

AI in Medical Education. Various LLMs have demonstrated remarkable capabilities answering medical questions similar to those found on the USMLE Step 1 [15]. Though this has spurred a new wave of excitement and interest in the field, practical applications which integrate Generative AI into the medical learning process are still in their infancy, with medical educators calling for further research into this field [3].

Preliminary studies have explored out-of-the-box LLMs in various medical educational aspects. For example, in a case study with OpenAI's ChatGPT, the LLM was incorporated into the daily rounds of a general internal medicine inpatient service for seven days [16]. The researchers evaluated the usage of ChatGPT as a substitute for other web-based resources (e.g., UpToDate) and found it generated helpful discussion, although its output was found to not be sufficiently thorough on several occasions. Another study [20] showed a statistically significant increase in quiz scores of students before and after engaging in a Socratic-style discussion [12] with GPT-3.5 about information acquired from a medical school textbook, compared to students who only read from the textbook.

In prior work by Li et al. [7], researchers evaluated the performance of several LLM chatbots in the context of Standardized Patients (SPs). SPs are trained actors who are used to portray patient personas in a given clinical scenario and provide medical learners an interactive and standardized environment in which to practice clinical skills. In this study, the LLMs assumed the roles of SPs, and experts proficient in SP education played the role of learners. The LLMs were then used to assess the completed simulated encounters using traditional SP evaluation checklists. Although there were slight differences amongst the various LLMs chatbots, the authors were optimistic of the potential of LLMs as virtual SPs to make medical education more efficient.

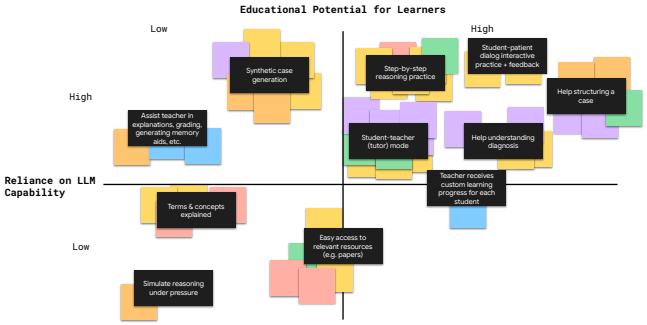

Figure 1: Idea matrix from design ideation workshop.

3 Methods

3.1 Design Ideation Workshop

The design ideation workshop involved an interdisciplinary panel of medical education experts, medical students, practicing clinicians, as well as user experience (UX) designers and HCI and AI researchers (including some co-authors of this work). The in-person workshop focused specifically on the role LLMs may play in the context of learning and teaching clinical reasoning skills. During the workshop, the interdisciplinary team worked step-wise to articulate the teaching and learning processes involved in clinical reasoning, define the needs and goals of educators and learners, and explore the potential of LLMs in clinical reasoning pedagogy. Figure 1 shows a simplified brainstorming artifact from the workshop during which the team first gathered ideas for the use of LLMs in clinical reasoning education, then clustered ideas into groups and categorized the resulting clusters into a matrix of educational value (low vs. high) and reliance on LLMs. This categorization enabled the team to focus on clusters of ideas at the intersection of high educational value and high reliance on LLM-specific capabilities.

3.2 Interactive Prototype

Taking into consideration the categorization in Figure 1 in addition to other artifacts and feedback from the design ideation workshop, the UX designers on the team facilitated the creation of initial design mocks incorporating the ideas focused on the intersection of high educational value and high reliance on LLM-specific capabilities. Utilizing the resulting design, an interactive, lightweight web-based prototype was implemented. Figure 2 shows the main elements of the resulting user interface. The resulting prototype provided a default clinical case vignette developed by the authors, with the

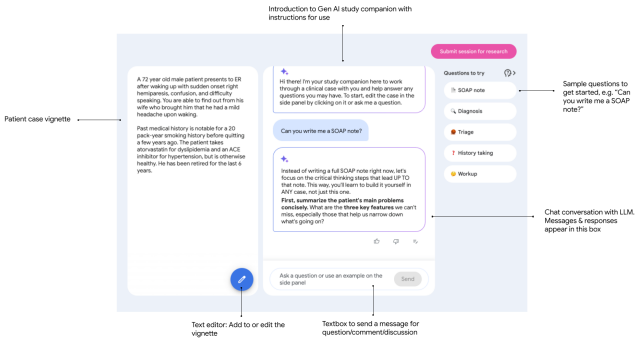

Figure 2: Clinical reasoning AI tutor interactive prototype.

option to change the case by pasting or typing in an alternative vignette text. The case vignette served as the grounding topic of the text chat conversation between a potential user and an LLM-based tutoring agent, and as such stayed visible on the left-hand side of the screen throughout the conversation. Example scaffolding questions to help initiate discussion were displayed in the right-hand column.

The prototype was powered by Med-PaLM 2 [15], i.e., a model fine-tuned on a corpus of medical datasets. The instructional prompt given to the model—a set of instructions passed to the model before any user input in the prompt—was developed together by members of the interdisciplinary team, including researchers and a board-certified physician with academic expertise in medical education. As part of the prompt, the LLM was instructed to play the role of a helpful medical school tutor with the goal of creating a rich learning environment to encourage active learning, reflection, and knowledge integration tailored to the student. The LLM was instructed to guide the student to complete each of the following sequential phases of clinical reasoning [13, 17]: (1) Interpretive summary; (2) Differential diagnosis; (3) Explanation of lead diagnosis; (4) Explanation of alternative diagnoses; (5) Reflection on potential diagnostic errors and bias; and (6) Evaluation and management plan.

3.3 User Study

We conducted a qualitative user study with eight medical learners to elicit their learning needs and challenges, attitudes about AI assistance in the learning process, and to collect initial impressions and experiences interacting with our clinical reasoning AI tutor.

Participants. Participants in the study were US adults recruited through a third-party organization and were provided with monetary incentives to participate in the study. Recruited participants were medical learners, ages 24-35, at different stages in their educational training. Out of a total of N=8 participants recruited, 4 were medical students in their third or fourth year of an accredited doctor of medicine (MD) or doctor of osteopathic medicine (DO) program. The other 4 participants were physicians (residents) undergoing specialty training in residency programs. The participants self-reported their familiarity with LLMs before the interview. Five participants reported having tried at least one AI tool. The remaining 3 participants reported having a basic understanding of AI tools but no hands-on experience.

Expert Interviews and Concept Evaluations. During expert interviews, participants first discussed their own medical learning experiences (20 minutes) and then evaluated the prototype in an interactive session with the interface (40 minutes). All interviews were fully remote and conducted through video conferencing software. Participants had access to the prototype through a web link and shared their screen while interacting with the prototype.

Participants who were unfamiliar with LLMs were provided a brief explanation of this technology to allow them to answer the latter questions. Following the interviews, participants transitioned to the interactive concept evaluation with the prototype. Participants were first asked questions about how they would expect to interact with the prototype; then, they were able to interact freely with the prototype and were invited to verbalize any opinions and running thoughts (also known as a think-aloud protocol).

Data Analysis. We applied thematic analysis [2] to qualitatively analyze participants' open-ended responses to interview questions and their think-aloud feedback from concept evaluation sessions.

4 Results

Expert Interviews. During the expert interviews, participants described clinical reasoning as a combination of theoretical, didactic knowledge and practical, hands-on application. They characterized "good" clinical reasoning as being comprised of a holistic understanding of the patient and their presentation leading to a step-wise, evidence-based approach to diagnosis. **P6** painted clinical reasoning as an "art" that is developed and perfected over time. A major challenge in developing strong clinical reasoning skills was noted to be in acquiring the ability to quickly synthesize all of the relevant information necessary (foundational knowledge, labs, history, etc.) to establish a reasonable working hypothesis for a differential diagnosis while flexibly updating this hypothesis as new information becomes available. **P4** emphasized *"a gap between the resources we use [in the first two years] and your time in the hospital–third and fourth year, then residency."*

When surveyed about how LLMs could enhance their process of acquiring clinical reasoning skills, participants noted many unique ways they could use the technology at various points in their learning journey. A few of these included using the technology to work through a differential diagnosis or sample questions, providing extra practice for clinical situations through a virtual setting, and summarizing notes and other patient information succinctly. In one anecdote, **P3** described a peer tutor program at their medical school that paired more senior medical students with their juniors. However, peer tutors could often be occupied with rotations or residency interviews, so they felt an "AI companion" could step in to fill that role while the peer was unavailable. They also proposed that such an AI companion could offer supplementary materials or recommendations based off information they, the student, had misunderstood during the interactive learning session.

Interactive Concept Evaluations. Interactive concept evaluations suggested that the prototype could serve as a helpful tool in early medical education, with students emphasizing the tool's clean interface and built-in prompts. This was especially true for clinical reasoning, which often requires input and interaction with clinicians who are experienced in their practice. **P1** stated, *"I think*

this [prototype] will be really good for early learners when they don't know what type of questions to ask." Participants liked that the prototype was interactive and that they could probe the model about the clinical case and ask it questions. **P2** said of the demo, *"it could help fine-tune your critical reasoning skills and expand your knowledge base"* and potentially *"correct any misunderstanding or even present an alternative diagnosis that may not have come to mind."*

During the chat interaction, several participants attempted to request from the model additional information about the patient presentation that was not provided in the clinical vignette. Participants intuitively assumed or expected that the model would be able to make additional hidden information about the case available. While this was not the case in our prototype, this kind of interaction seemed natural to several participants. This was reflected in comments suggesting the AI tutor could guide a student through a case (**P4**), play a mentor role and suggest alternative diagnoses a student might not have considered (**P5**), and break down management steps (**P8**).

5 Discussion

Addressing Learners' Needs and Practices. Our study revealed key challenges medical learners face in developing clinical reasoning skills, particularly the difficulty in synthesizing vast amounts of information and the gap between theoretical learning and practical application (as highlighted by **P4**'s comment). The development of clinical reasoning skills is a lifelong process that starts with mastering the basic and clinical sciences in the preclinical years. During the clinical years, learners synthesize and apply their textbook knowledge to begin recognizing classic disease patterns and developing mental frameworks, such as diagnostic algorithms and illness scripts, to organize and assess likelihoods of potential diagnoses based on clinical findings. Participants envisioned Generative AI as a valuable tool to iteratively develop and refine these mental frameworks, dealing with uncertainty and atypical presentations and suggesting use cases such as working through differential diagnoses, providing virtual practice scenarios, and summarizing patient information. This aligns with the expressed need for support during the transition from preclinical to clinical training.

Design Considerations for AI-Powered Tutors. The interdisciplinary nature of our design ideation workshop proved invaluable in generating diverse perspectives and grounding the design in real-world needs. Specifically, the workshop helped us identify key features for an effective AI tutor, such as the need for structured interactive guidance through each of the six phases of clinical reasoning and the importance of adapting to each user's unique learning styles and needs (e.g., finding the balance between providing guidance and fostering independent thinking). There was also an expectation that the model should provide context-specific recommendations, highlighting the need for AI tutor localization and adapting its knowledge and recommendations to the specific context of the learner's institution and geographic location. Further, participants' intuitive expectation that the model would reveal hidden information points towards a design that supports progressive disclosure of case details, mimicking the iterative nature of real-world clinical reasoning. The desire for the AI tutor to act as a mentor reinforces the importance of incorporating *preceptor-like behaviors*, such as managing cognitive load, providing constructive feedback, and encouraging questions and reflection.

Broader Potential and Considerations. While our focus was on clinical reasoning, participants also identified broader applications for LLMs in medical education such as peer tutoring support and educational resource recommendations. This aligns with the growing interest in using Generative AI for creating learning materials and assessments, personalizing learning pathways, simulating patient interactions, and supporting medical writing, as highlighted in the literature [14]. However, as the field moves forward, careful attention must be paid to ensuring accuracy, mitigating bias, and maintaining the crucial role of human interaction in medical education. Responsible development and use of Generative AI has the potential to augment educator roles and learner journeys.

Acknowledgments

We thank Mathias Fleck, Laura Vardoulakis and Meredith Ringel Morris for their thoughtful reviews of the manuscript.

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.

[2] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association, Washington DC, USA.

[3] Gunther Eysenbach et al. 2023. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Medical Education* 9, 1 (2023), e46885.

[4] Phyllis A Guze. 2015. Using technology to meet the challenges of medical education. *Transactions of the American clinical and climatological association* 126 (2015), 260.

[5] Rachel Hilburg, Niralee Patel, Sophia Ambruso, Mollie A Biewald, and Samira S Farouk. 2020. Medical education during the coronavirus disease-2019 pandemic: learning from a distance. *Advances in chronic kidney disease* 27, 5 (2020), 412–417.

[6] Emily CN Lawrence, C Jessica Dine, and Jennifer R Kogan. 2023. Preclerkship medical students' use of third-party learning resources. *JAMA Network Open* 6, 12 (2023), e2345971–e2345971.

[7] Yaneng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024. Leveraging Large Language Model as Simulated Patients for Clinical Education. *arXiv preprint arXiv:2404.13066* (2024).

[8] 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 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

[9] Bunmi S Malau-Aduli, Poornima Roche, Mary Adu, Karina Jones, Faith Alele, and Aaron Drovandi. 2020. Perceptions and processes influencing the transition of medical students from pre-clinical to clinical training. *BMC Medical Education* 20 (2020), 1–13.

[10] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2023. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164* (2023).

[11] Association of American Medical Colleges. [n. d.]. What to Expect in Medical School. https://web.archive.org/web/20240303000307/https://students-residents.aamc.org/choosing-medical-career/what-expect-medical-school. Accessed: 2024-08-20.

[12] Douglas R Oyler and Frank Romanelli. 2014. The fact of ignorance revisiting the Socratic method as a tool for teaching critical thinking. *American Journal of Pharmaceutical Education* 78, 7 (2014), 144.

[13] Jennifer M Pascoe, James Nixon, and Valerie J Lang. 2015. Maximizing teaching on the wards: review and application of the One-Minute Preceptor and SNAPPS models. *Journal of hospital medicine* 10, 2 (2015), 125–130.

[14] Carl Preiksaitis and Christian Rose. 2023. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR medical education* 9 (2023), e48785.

[15] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).

[16] Anthony Skryd and Katharine Lawrence. 2024. ChatGPT as a Tool for Medical Education and Clinical Decision-Making on the Wards: Case Study. *JMIR Formative Research* 8 (2024), e51346.

[17] Satid Thammasitboon, Joseph J Rencic, Robert L Trowbridge, Andrew PJ Olson, Moushumi Sur, and Gurpreet Dhaliwal. 2018. The Assessment of Reasoning Tool (ART): structuring the conversation between teachers and learners. *Diagnosis* 5, 4 (2018), 197–203.

[18] Joan Carles Trullàs, Carles Blay, Elisabet Sarri, and Ramon Pujol. 2022. Effectiveness of problem-based learning methodology in undergraduate medical education: a scoping review. *BMC medical education* 22, 1 (2022), 104.

[19] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654* (2024).

[20] Cai Ling Yong, Mohammad Shaheryar Furqan, James Wai Kit Lee, Andrew Makmur, Ragunathan Mariappan, Clara Lee Ying Ngoh, and Kee Yuan Ngiam. [n. d.]. The Use of Large Language Models Tuned with Socratic Methods on the Impact of Medical Students' Learning: A Randomised Controlled Trial. ([n. d.]).

[21] 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