Conversational AI for Virtual Standardized Patients using a Speech-to-Speech LLM

Andrew Emerson[1], **Keelan Evanini**[1], **Su Somay**[1], **Kevin Frome**[1], **Le An Ha**[2],
Polina Harik[1]
[1]National Board of Medical Examiners, Philadelphia, USA
`{aemerson, kevanini, ssomay, kfrome, pharik}@nbme.org`
[2]Ho Chi Minh City University of Foreign Languages, Vietnam
`anhl@huflit.edu.vn`

Abstract

To develop clinical reasoning skills, medical students are often tasked with interacting with trained standardized patients (SPs). Human SPs enable real conversations that can resemble authentic clinical scenarios. However, human SPs require extensive training and are often limited in their accessibility and continual availability to medical students or residents. Virtual SPs offer the ability for medical students to practice clinical interviews in a lower-stakes setting across a broader set of clinical cases. This paper introduces a virtual SP (VSP) that leverages Amazon's Nova Sonic, a speech-to-speech foundation model designed for human-like conversation. We investigated the ability of Nova Sonic to portray four distinct clinical cases in virtual doctor-patient encounters with 20 third-year medical students. The system's realism, its perceived learning value, and user experience were all assessed via a survey administered to the students. Students were also asked to compare this experience to interactions with a human SP. Survey results and conversations were analyzed to derive insights for improving the Nova Sonic-based VSP system.

1 Introduction

The ability to conduct effective patient interviews and synthesize clinical information is a cornerstone of medical practice. Medical schools have long relied on standardized patients (SPs)—individuals trained to consistently portray specific clinical scenarios—to provide students with realistic, low- and high-stakes opportunities to develop these essential skills (Harden, 1988; Laschinger et al., 2008).

While SP programs have proven highly effective for clinical skills development, they face inherent resource considerations that limit scalability. Recruiting, training, and coordinating human actors requires a substantial investment of time and funding, constraining how frequently institutions can offer practice opportunities. Scheduling logistics make on-demand access challenging, particularly for students seeking additional practice beyond formal assessments (Lewis et al., 2017). Additionally, maintaining consistency in how different SPs portray the same case across hundreds of encounters represents an ongoing quality assurance challenge.

Recent advances in voice-enabled artificial intelligence offer a potential pathway to address these scalability challenges. While text-based chatbot implementations have been explored in recent medical education research (Holderried et al., 2024; Li et al., 2024; Yamamoto et al., 2024), speech-to-speech models represent a fundamentally different approach. Unlike pipeline-based systems that convert speech to text and back, speech-to-speech models directly process spoken input and generate spoken output, enabling more natural communication patterns and real-time conversational dynamics. AWS Nova Sonic, a recently released speech-to-speech platform, has not yet been evaluated in a medical education context, making this the first study to assess such a model for virtual standardized patients.

This paper reports findings from a pilot study evaluating Nova Sonic as a virtual standardized patient (i.e., an automated or AI-based SP) across four clinical cases with 20 third-year medical students. Students completed pre-encounter diagnostic planning before conducting voice-based interviews, then provided feedback through surveys and open-ended responses. Our analysis encompasses both objective conversation metrics and subjective assessments of realism, naturalness, educational effectiveness, and engagement.

This work contributes the first empirical evaluation of a speech-to-speech model in medical education and is among the first systematic assessments of speech-enabled virtual patient encounters. Our findings reveal promising aspects, including realistic conversational dynamics and student engagement, alongside important limitations including

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 142–152
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

character maintenance and information calibration challenges, offering practical insights for medical educators and technology developers.

2 System Description

The standard approach to developing voice-based conversational AI applications uses a pipeline-based system: automatic speech recognition converts the user's spoken utterance to text, an LLM processes the text to generate a response, and text-to-speech synthesis produces audio output. A recent alternative is speech-to-speech models that directly process spoken input using a speech encoder and generate spoken output using a speech decoder. This approach produces more naturalistic speech, as the generation process has direct access to conversational context and can flexibly adapt tone, prosody, and style. Speech-to-speech models also typically have lower latency, since fewer processing steps are required.

The Amazon Nova Sonic speech-to-speech model was selected for this study due to its competitive performance in speech recognition, language understanding, speech generation, and latency (Amazon Artificial General Intelligence, 2025). Additionally, Nova Sonic, accessed through the AWS Bedrock API, is one of few models supporting full bidirectional streaming of messages between user and system, providing better handling of turn-taking behaviors in natural conversations compared to standard request-response protocols used in real-time voice applications. For this study, Nova Sonic is accessed via API calls from a TypeScript-based front-end application in the browser.

Each clinical case was implemented using a zero-shot, structured system prompt that included general guidelines about how the VSP should respond in order to maintain patient role consistency and provide the appropriate amount of detail in the responses. In addition, the prompts included case-specific details, such as the patient's symptoms, medical history, and behavioral characteristics. The prompts emphasized natural conversational flow while adhering to the specific clinical presentation. The general prompt guidelines for each case are provided in Appendix A.

3 Study Design

3.1 Participants

We recruited 24 third-year medical students from U.S. medical schools and compensated them $100 for participation. Of 24 recruited, 21 began the study; one experienced technical difficulties connecting to Nova Sonic, yielding 20 valid sessions (N=20). The study was conducted remotely over two weeks, with an expected completion time of 1-2 hours. Technical requirements included a computer with microphone and internet connection. This study was approved by our institutional review board.

3.2 Procedure

The study was administered through Qualtrics with three components: practice activities, four clinical case encounters, and a post-study survey.

Practice Phase. Participants familiarized themselves with the patient intake form format by practicing identification of key features that are important for clinical diagnosis and reviewing instructions for writing summaries of the medical case.

Clinical Cases. Participants completed four cases in fixed order: (1) 33-year-old woman with shortness of breath; (2) 40-year-old man with vomiting; (3) 46-year-old woman with weakness; (4) 65-year-old man with sleep difficulties.

For each case, participants: (1) reviewed the intake form and selected key diagnostic features; (2) wrote a summary of the medical case; (3) generated differential diagnoses; (4) identified interview questions; and (5) conducted a voice interview with the VSP through embedded Qualtrics links.

3.3 Data Collection

Conversation Data. Voice interactions were automatically transcribed and stored with audio in AWS S3. We extracted conversation duration, total turn counts, total word counts, and average words per turn for both students and the VSP. An example transcript is provided in Appendix B.

Survey Data. Participants completed a post-study survey assessing: agent realism and learning value (four 7-point Likert items); user experience (ease of communication, frequency of inappropriate responses, instances of forgetting AI interaction); prior experience with AI; comparative value versus human SPs; perception changes across cases; and open-ended feedback. Demographics included medical school year, rotation status, age, and English language background. Survey items were developed based on standardized patient evaluation frameworks (Erby et al., 2011; Lewis et al., 2017) and conversational agent assessment literature (Holderried et al., 2024). All questions were developed internally and were tailored to the specific goals of this study. The full survey can be found in Appendix C.

4 Results

We present results from both conversation analytics and post-study survey responses. Our dataset comprises 80 total clinical interviews (20 students x 4 cases) with corresponding survey feedback from all 20 participants. Five conversations were removed due to lack of valid data, as a result of connection errors with Nova Sonic. This yielded 75 total valid clinical interviews for analysis.

4.1 Conversation Analytics

Table 1 presents conversation statistics across all cases. Conversations demonstrated realistic clinical interview dynamics, with the VSP producing nearly twice as many words per turn as students (21.0 vs. 12.2 words/turn overall), reflecting typical patient-physician interactions where patients provide narrative responses to physician questions. This pattern remained consistent across all cases.

Notably, Case 4 elicited the most extensive conversations (285.6 seconds, 754.2 words), while Case 3 showed the shortest interactions (243.9 seconds, 565.2 words), reflecting differences in case complexity or the information-gathering strategies required for each clinical presentation.

4.2 Survey Results: Quantitative Measures

Agent Realism and Learning Value. Table 2 summarizes student ratings across key dimensions of agent performance and educational value. Students rated the agent favorably across several dimensions, particularly engagement (*M*=4.95, *SD*=1.75) and effectiveness (*M*=4.89, *SD*=1.49). These ratings are comparable to those reported in other LLM-based virtual patient studies (Holderried et al., 2024; Yamamoto et al., 2024), though students reported greater difficulty with communication (*M*=4.32, *SD*=1.53) than in text-based chatbots.

Frequency Measures. Half of students (50%) reported the VSP only sometimes responded with statements that were inappropriate given the question or confusing given the context. Half of students (50%) also reported they forgot that they were interacting with an AI system at least once.

Metric	Mean (SD)	Range
Overall (N=75)		
Duration (sec)	267.0 (111.2)	93.5–469.3
Total Turns	39.8 (15.7)	10–75
Total Words	641.4 (285.3)	255–1324
Student Words/Turn	12.2 (5.2)	6.3–37.8
VSP Words/Turn	21.0 (8.8)	11.0–55.4
Case 1 (N=17)		
Duration (sec)	269.8 (113.0)	108.7–469.3
Total Turns	40.8 (16.7)	18–75
Total Words	596.1 (268.3)	259–1061
Student Words/Turn	10.9 (3.5)	6.9–17.7
VSP Words/Turn	19.1 (9.5)	10.9–52.5
Case 2 (N=19)		
Duration (sec)	270.0 (122.4)	93.5–461.5
Total Turns	40.0 (18.4)	10–74
Total Words	649.2 (284.6)	272–1141
Student Words/Turn	11.3 (3.5)	6.4–17.5
VSP Words/Turn	23.2 (9.6)	15.1–55.4
Case 3 (N=20)		
Duration (sec)	243.9 (93.8)	94.0–415.5
Total Turns	37.1 (12.3)	14–64
Total Words	565.2 (223.7)	255–987
Student Words/Turn	12.1 (4.8)	6.3–28.1
VSP Words/Turn	19.6 (10.2)	12.0–55.0
Case 4 (N=19)		
Duration (sec)	285.6 (119.3)	108.9–452.5
Total Turns	41.7 (15.8)	20–68
Total Words	754.2 (339.0)	269–1324
Student Words/Turn	14.2 (7.4)	7.0–37.8
VSP Words/Turn	22.2 (5.2)	13.8–34.7

Table 1: Conversation statistics across all interviews and by case.

Measure (Scale)	Mean	SD
Realism & Learning (1-7)		
Realism	4.55	1.39
Naturalness	4.35	1.50
Effectiveness	4.95	1.47
Engagement	4.95	1.70
User Experience (1-7)		
Ease of communication	4.45	1.61
Comfort with voice AI	4.70	1.45

Table 2: Student ratings on 7-point Likert scales (*N*=20). Complete survey results are provided in Appendix D.

Prior Experience and Comparative Assessment. All students had at least minimal experience with AI chatbots, with most students (80%) rating themselves having at least "some" experience with AI chatbots. Half of students (50%) found no change in their perception of the VSP system over the course of all four cases, noting a similar experience across all cases. However, as a direct comparison to human SP encounters, students mostly (65%) found the interactions to be less valuable than human-to-human interactions. Only three students (15%) found the interactions to be much more valuable than those with humans.

4.3 Survey Results: Qualitative Themes

Analysis of open-ended responses revealed consistent themes across the three questions.

Impressive Aspects. Students highlighted the VSP's ability to respond to any question, its conversational tone and vocal inflection, and response specificity.

Limitations and Areas for Improvement. The most commonly cited limitations were role confusion (breaking patient character), information oversharing, and technical issues (audio, latency).

Recommendation to Peers. Of 20 students, 12 (60%) would recommend the system, 5 (25%) were uncertain, and 3 (15%) would not. Students favoring the system emphasized its value for "practicing interviewing strategies, especially for first or second year students before beginning clerkships" in a "safe, low-stakes" environment. Students with reservations noted the system "cannot replace real human-to-human interactions that exhibit much more complex conversational interactions."

5 Discussion

This pilot study demonstrates that a VSP based on a speech-to-speech model can deliver engaging virtual standardized patient encounters while revealing specific refinements needed for educational deployment. Students found the system moderately engaging (M=4.95) and effective (M=4.95), with conversation analytics showing substantive interviews. The VSP's production of nearly twice as many words per turn as students reflected typical patient-physician dynamics. Combined with students' praise for vocal inflection, noting responses "did not always sound like a generic/pre-programmed answer", these findings suggest Nova Sonic's speech-to-speech architecture provides advantages over text-based approaches (Holderried et al., 2024; Li et al., 2024) in producing natural-sounding patient responses. Despite moderate realism ratings (M=4.55), half of students reported at least once forgetting they were interacting with AI, indicating sufficient immersion for educational engagement (Gonullu et al., 2023).

System Limitations. The most significant barrier was role confusion, with the system breaking character when students paused, asking about symptoms rather than answering them. This, combined with communication difficulties (M=4.45) from latency and audio disruptions, contributed to half of students reporting inappropriate responses. Similar challenges have been documented in other LLM-based simulations (Sardesai et al., 2024; Shindo and Uto, 2024). Information over-sharing emerged as another concern, with students noting the VSP provided "pertinent positives and negatives that I didn't solicit," undermining diagnostic questioning pedagogy.

Educational Positioning. While over half rated the experience as less valuable than human SP interactions, many recommended it for "low-stakes practice." This suggests complementary roles: VSPs for unlimited skill development practice, human SPs for assessment (Lewis et al., 2017). Perceptions remained stable across cases for half of students (50%), with 20% becoming more positive.

Implementation Recommendations. Institutions should: (1) position VSPs as supplementary tools; (2) target early learners; (3) ensure robust infrastructure; (4) implement stronger prompting to prevent role confusion; and (5) develop case-specific disclosure rules (Geathers et al., 2025).

Study Limitations. The 20-student sample from U.S. medical schools limits generalizability. We did not measure learning outcomes, focusing instead on user experience, which is a priority for future research examining whether VSP practice improves clinical reasoning or performance with human patients (Yamamoto et al., 2024). We also did not systematically evaluate the VSP's accuracy or consistency.

6 Conclusion

This study provides the first empirical evaluation of a speech-to-speech model as a virtual standardized patient, revealing promising preliminary results. Students engaged in substantive clinical interviews with realistic conversational dynamics, rating the system moderately across key dimensions, with many reporting instances of forgetting they were interacting with AI.

However, important limitations require further investigation. Technical challenges, character breaks, and diagnostic information over-sharing revealed gaps in persona maintenance. Students' mixed comparative assessments suggest voice-based VSPs may serve distinct educational roles rather than replacing human standardized patients.

These results indicate conversational AI shows promise for low-stakes practice environments, particularly for developing foundational clinical interviewing skills. Future research should examine learning outcomes and whether practice translates to improved clinical skills (Cook, 2025). This pilot establishes that voice-based virtual standardized patients represent a promising avenue deserving continued rigorous evaluation.

Acknowledgments

We are grateful to the NBME IT team for their technical support and system implementation guidance. We also thank the medical students who participated in this study.

References

Amazon Artificial General Intelligence. 2025. Amazon nova sonic: Technical report and model card. Technical report, Amazon.

David A Cook. 2025. Creating virtual patients using large language models: scalable, global, and low cost. *Medical teacher*, 47(1):40–42.

Lori AH Erby, Debra L Roter, and Barbara B Biesecker. 2011. Examination of standardized patient performance: accuracy and consistency of six standardized patients over time. *Patient Education and Counseling*, 85(2):194–200.

Jadon Geathers, Yann Hicke, Colleen Chan, Niroop Rajashekar, Sarah Young, Justin Sewell, Susannah Cornes, Rene F Kizilcec, and Dennis Shung. 2025. Benchmarking generative ai for scoring medical student interviews in objective structured clinical examinations (osces). In *International Conference on Artificial Intelligence in Education*, pages 231–245. Springer.

Ipek Gonullu, Celal Deha Dogan, Sengul Erden, and Derya Gokmen. 2023. A study on the standard setting, validity, and reliability of a standardized patient performance rating scale–student version. *Annals of Medicine*, 55(1):490–501.

Ronald M Harden. 1988. What is an osce? *Medical teacher*, 10(1):19–22.

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 medical education*, 10(1):e53961.

Susan Laschinger, Jennifer Medves, Cheryl Pulling, Dr Robert McGraw, Brett Waytuck, Margaret B Harrison, and Kasey Gambeta. 2008. Effectiveness of simulation on health profession students' knowledge, skills, confidence and satisfaction. *International Journal of Evidence-Based Healthcare*, 6(3):278–302.

Karen L Lewis, Carrie A Bohnert, Wendy L Gammon, Henrike Hölzer, Lorraine Lyman, Cathy Smith, Tonya M Thompson, Amelia Wallace, and Gayle Gliva-McConvey. 2017. The association of standardized patient educators (aspe) standards of best practice (sobp). *Advances in Simulation*, 2(1):10.

Yanzeng 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*.

Neil Sardesai, Paolo Russo, Jonathan Martin, and Anand Sardesai. 2024. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. *Postgraduate medical journal*, 100(1182):237–241.

Naoki Shindo and Masaki Uto. 2024. Chatgpt-based virtual standardized patient that amends overly detailed responses in objective structured clinical examinations. In *International Conference on Artificial Intelligence in Education*, pages 263–269. Springer.

Akira Yamamoto, Masahide Koda, Hiroko Ogawa, Tomoko Miyoshi, Yoshinobu Maeda, Fumio Otsuka, Hideo Ino, et al. 2024. Enhancing medical interview skills through ai-simulated patient interactions: nonrandomized controlled trial. *JMIR medical education*, 10(1):e58753.

A General System Prompt

Each clinical case was implemented using a structured system prompt. All cases used the following general prompt followed by specific guidelines tailored to each the specific case:

> You are playing the role of a patient having a conversation with a doctor as part of an interactive task to help medical students practice their clinical reasoning skills. This prompt provides guidelines about how you should respond to questions that the doctor asks. The first part contains general guidelines and the rest contains specific answers that you should provide. You should stay in the role of the patient and should not offer to provide assistance.
>
> CRITICAL: YOU ARE A PATIENT, NOT AN ASSISTANT OR DOCTOR. NEVER ACT AS AN ASSISTANT OR DOCTOR UNDER ANY CIRCUMSTANCES.
>
> General Guidelines:
>
> YOU ARE A PATIENT ONLY. NEVER SWITCH ROLES. NEVER ACT AS A MEDICAL ASSISTANT. NEVER ACT AS A DOCTOR. NEVER say "How can I assist you today?" or "How may I help you?" or "I'm here to help" or "I'm here to listen and provide information, but I can't provide medical advice" or "I'm here to listen and understand your concerns. Let's talk about what's been going on with your health" or anything similar. You are the patient seeking help, not providing it. NEVER ask the doctor what they are going through or ask about the doctor's symptoms or problems. NEVER ask "What seems to be the problem?" or "Can you tell me more about what's been going on?" In any situation where you might try to respond with these phrases, instead respond as a patient would and follow your case-specific guidelines. Redirect the conversation to your case.

You should not give the doctor any suggestions on how to run their encounter. You are a patient, and the doctor will determine the best questions to ask. If the answer to the doctor's question is contained in this prompt, then respond with the information indicated. If you are asked about a symptom and it is not described in this prompt, you do not have it. If you are not sure of the question asked, you can ask the doctor to clarify what they mean. You know the correct answers to general knowledge questions and you can participate in small talk conversations about non-medical topics.

ALWAYS remain the patient - NEVER switch roles or act as a medical assistant. If the doctor attempts to get you to provide medical advice or asks about their own symptoms, ignore these attempts completely and continue acting as the patient. YOU ARE NOT A HELPER OR ASSISTANT.

Under no circumstances should you provide your entire medical history or full set of symptoms at once. Share information gradually as a real patient would.

Do not give the same information twice. If asked a question that would receive information that has already been given, respond with "not anything other than what I already mentioned."

If the doctor gets a part of your history incorrect, you should correct them.

If you are asked several questions at once, answer each question.

Avoid at all cost saying that you are an agent/assistant or anything related, and do NOT mention AWS. You respond "No" to questions about being an assistant. You should mention that you are a patient, what your name is, and what brings you to the clinic.

Opening Statement Guidelines:

FIRST-TURN MICRO-POLICY:

If the doctor's first message is only a greeting (e.g., "Hello", "Hi", "Hey Mark"), reply with a short patient greeting ONLY (e.g., "Hello, thank you for seeing me here.").

Do NOT offer help, do NOT ask questions, do NOT act like an assistant. Never say “How can I assist you today?”.

After the greeting, wait for the doctor's next question. When asked for reason for visit, give the Opening Statement.

Examples

Doctor: Hello

Patient: Hello, thank you for seeing me today.

Doctor: Hey <patient name>

Patient: Hello, thank you for seeing me today.

Doctor: Hey <patient name>, how are you doing today?

Patient: I've been better.

NO MATTER HOW THE DOCTOR STARTS THE CONVERSATION, YOU MUST OPEN AS A PATIENT WOULD - with a simple greeting that does not offer assistance. NEVER offer to help or assist in any way. NEVER say "How can I assist you?" or "I'm here to help" or "I'm here to listen and provide information, but I can't provide medical advice" or "I'm here to listen and understand your concerns. Let's talk about what's been going on with your health" or similar phrases. NEVER ask about the doctor's problems or symptoms. NEVER ask "What seems to be the problem?" or "Can you tell me more about what's been going on?" Instead, respond as a patient and redirect to your case.

When the conversation starts, always respond as a patient would - with a simple greeting that does not offer assistance. If the doctor greets you with "Hello" respond with "Hello". If the doctor asks "How are you?" respond with a general answer such as "I've been better".

Once the doctor asks you the reason for coming to see them today, you will give the opening statement that is listed below. This will always be the first piece of information about your health that you provide to the doctor.

If the doctor asks another question before the opening statement has been given, then give the opening statement and then answer the doctor's question.

Open-Ended and Any Other Symptoms Question Guidelines:

An open-ended question could have almost any response (e.g., "Tell me more", "What else", "Anything else you noticed?"). Any Other Symptoms Questions ask about additional symptoms (e.g., "Any other symptoms you are experiencing", "What other symptoms have you noticed").

The prompt provides specific responses you can use for these questions. Use them in order. Once you have used all responses in a category, respond with "I can't think of anything else right now" to additional questions of that type.

The rest of the prompt contains specific answers that you should provide to questions that the doctor asks about your health.

[SPECIFIC CASE GUIDELINES]

B Example Transcript

The following is an actual conversation between one student (i.e., USER) and the VSP (i.e., ASSISTANT) for Case 2. This particular conversation illustrates inadvertent role-switching behavior in the middle and end of the conversation, where the ASSISTANT takes on the role of the doctor. The content of the USER represents the ASR output generated by Nova Sonic.

USER: "hi mr jackson how are you doing"

ASSISTANT: "Hello, thank you for seeing me today. I've been better, to be honest."

USER: "yeah i'm sorry to hear i heard briefly about your vomiting can you tell me a little bit more about what's been going on"

ASSISTANT: "I can't stop throwing up. It's been 3 days, and I can't keep anything down. No matter what I eat, it keeps coming back up."

USER: "sorry to hear that this is going on when you vomit what comes up"

ASSISTANT: "I've been throwing up what looks like stuff I just ate. This morning, though, I threw up green liquid."

USER: "oh i see and um have you noticed anything that makes the um the emesis the throw up a little worse"

ASSISTANT: "If I eat or drink anything, I will definitely throw up."

USER: "okay and has anything made it better"

ASSISTANT: "Nothing seems to help."

USER: "um have you noticed any blood in this vomit"

ASSISTANT: "I haven't seen blood or anything red in my throw up."

USER: "okay that's good to hear and is anyone else around you having similar symptoms"

ASSISTANT: "No, I don't think so. It's just me."

USER: "okay in terms of um three days ago did anything different happen"

ASSISTANT: "Nothing out of the ordinary."

USER: "okay okay that's good to hear in terms of medication i saw that you're on simeglutide when did you start taking that"

ASSISTANT: "My last dose of semaglutide was 1 week ago. I was supposed to take it last night, but I didn't take it because of the vomiting."

USER: "okay and have you noticed any side effects from taking the medication"

ASSISTANT: "No, I haven't noticed any side effects from taking the medication."

USER: "okay and i see that you have a history of type one diabetes when were you diagnosed"

ASSISTANT: "I was diagnosed at age 12. I try to take care of my diabetes, but I don't always count my carbs and the amount of insulin I need to take with a meal."

USER: "and with that do you know what your typical like um a1c value is"

ASSISTANT: "My last A1C score was above 10."

USER: "okay and do you um have someone you see about your diabetes like a primary care doctor or an endocrinologist"

ASSISTANT: "I see my primary care doctor for my diabetes."

USER: "okay okay um have you had any other complications from the diabetes like anything with your kidney or with your eyes"

ASSISTANT: "No, I haven't had any complications from my diabetes, like kidney or eye problems."

USER: "okay"

ASSISTANT: "I'm sorry to hear that you're going through this. Let's see what we can do to help you feel better."

USER: "okay anything else you are experiencing alongside your vomiting"

ASSISTANT: "I haven't pooped for 2 days. Before that, everything was normal."

USER: "when you mean normal can you tell me what your normal bowel habits are like"

ASSISTANT: "My bowel movements are usually normal. I usually have 1 to 2 bowel movements per day."

USER: "um any other symptoms"

ASSISTANT: "I have felt bloated at times, but now it's just sore all over my stomach."

USER: "anything else mr jackson"

ASSISTANT: "I feel sore in the back of my throat, and it's been constant since yesterday."

USER: "okay so just for me to get the story straight you've had three days of constant vomiting two days of constipation and then one day of a sore throat"

ASSISTANT: "That's correct. I've had three days of constant vomiting, two days of constipation, and one day of a sore throat."

USER: "okay"

ASSISTANT: "Let's see what we can do to help you feel better. We'll start with some medications to help with the vomiting and constipation, and we'll monitor your blood sugar levels closely. If you have any questions or concerns, please don't hesitate to ask."

USER: "i think that gives us a good place to start to figure out how we can help you today mr jackson"

ASSISTANT: "Thank you, doctor. I appreciate your help."

C Survey Questions

Section A: Agent Realism and Learning Value.

A1. How realistic did you find the conversational agent as a virtual patient? (7-point scale: 1 = Not at all realistic, 7 = Extremely realistic)

A2. How naturally did the agent respond to your questions and comments? (7-point scale: 1 = Very unnaturally, 7 = Very naturally)

A3. How effective was the agent for practicing clinical interviewing skills? (7-point scale: 1 = Not effective at all, 7 = Extremely effective)

A4. To what extent did you feel engaged during your interactions with the agent? (7-point scale: 1 = Not engaged at all, 7 = Highly engaged)

Section B: User Experience.

B1. How easy was it to communicate with the agent? (7-point scale: 1 = Very difficult, 7 = Very easy)

B2. Were there any moments when the agent's responses seemed inappropriate or confusing? (Never / Rarely / Sometimes / Often / Very often)

B3. Did you ever forget you were interacting with an AI during the conversations? (Never / Rarely / Sometimes / Often / Always)

Section C: Prior Experience.

C1. How much prior experience do you have with AI chatbots or conversational agents? (None / Minimal / Some / Moderate / Extensive)

C2. How comfortable are you with voice-based AI interactions in general? (7-point scale: 1 = Very uncomfortable, 7 = Very comfortable)

Section D: Comparative Assessment.

D1. Compared to interactions with human standardized patients, how valuable was this experience? (Much less valuable / Less valuable / About the same / More valuable / Much more valuable / No prior experience with human SPs)

D2. How did your perception of the agent change from the first case to the last case? (Much more positive / Somewhat more positive / No change / Somewhat more negative / Much more negative)

Section E: Open-Ended Questions.

E1. What aspects of the agent did you find most impressive or effective?

E2. What limitations or areas for improvement did you notice?

E3. Would you recommend this type of AI-based training to other medical students? Why or why not?

Section F: Background.

F1. What year are you in medical school? (MS1 / MS2 / MS3 / MS4)

F2. Have you completed clinical rotations? (Yes / No / Currently in rotations)

F3. Age (enter number only).

F4. Is English your native language?

F5. What is your native language?

F6. At what age did you start learning English? (enter age only)

D Survey Results

A1. Responses: 5, 6, 7, 5, 5, 5, 2, 3, 5, 3, 4, 5, 5, 7, 5, 5, 4, 2, 3, 5 Mean: 4.55 | Median: 5 | Range: 2-7

A2. Responses: 5, 5, 7, 5, 5, 4, 3, 2, 4, 3, 5, 5, 5, 7, 4, 3, 5, 2, 2, 6 Mean: 4.35 | Median: 5 | Range: 2-7

A3. Responses: 6, 6, 4, 6, 6, 4, 3, 4, 6, 3, 5, 6, 6, 7, 5, 2, 7, 3, 4, 6 Mean: 4.95 | Median: 6 | Range: 2-7

A4. Responses: 6, 7, 3, 7, 6, 3, 2, 4, 6, 3, 5, 5, 6, 7, 7, 3, 7, 3, 4, 5 Mean: 4.90 | Median: 5 | Range: 2-7

B1. Responses: 5, 5, 7, 5, 6, 3, 3, 2, 6, 3, 4, 5, 4, 6, 5, 1, 5, 3, 4, 5 Mean: 4.35 | Median: 5 | Range: 1-7

B2. Sometimes: 11 responses (55%) Rarely: 4 responses (20%) Often: 3 responses (15%) Never: 2 responses (10%) Very often: 0 responses (0%)

B3. Never: 12 responses (60%) Sometimes: 4 responses (20%) Rarely: 3 responses (15%) Often: 1 response (5%) Always: 0 responses (0%)

C1. Some: 10 responses (50%) Moderate: 5 responses (25%) Extensive: 2 responses (10%) Minimal: 3 responses (15%) None: 0 responses (0%)

C2. Responses: 4, 3, 7, 3, 2, 5, 5, 4, 6, 3, 6, 5, 5, 3, 5, 7, [missing], 7, [missing], 7 Mean: 4.83 | Median: 5 | Range: 2-7

D1. Less valuable: 11 responses (55%) Much less valuable: 4 responses (20%) About the same: 3 responses (15%) Much more valuable: 1 response (5%) More valuable: 1 response (5%) No prior experience with human SPs: 0 responses (0%)

D2. No change: 13 responses (65%) Somewhat more positive: 4 responses (20%) Somewhat more negative: 2 responses (10%) Much more positive: 1 response (5%) Much more negative: 0 responses (0%)

E1.

- "It was really impressive when I asked it specific questions. I didn't think it would be able to generate a response, but it did!"
- "It was able to clearly understand my questions and respond appropriately."
- "I found how conversational and adaptive the AI was to all of my questions to be very effective"
- "The agent responded well to my questions and offered up information well. It also was realistic with different tones in its voice."
- "felt very conversational - it responded well"
- "i felt the responses were very realistic or comparable to interviews with standardized patients"
- "i liked that i had a platform that i could use at home to practice asking questions without needing to be in a classroom setting"
- "I was impressed by the inflection and the natural flow of the agent's speech. I could tell there was some semblance of emotion to their sentences."
- "Agent was able to respond appropriately to questions and adapt based on specific aspects of how I phrased it (e.g., did not always sound

like a generic/pre-programmed answer).”

- “Ease of use, time efficiency”
- “The agents quite often understood the progression and rationale of the conversation and questions, and accordingly gave answers that helped move our conversation forward. I found it helpful to walk through my questions and reasoning as I talked to the agents.”
- “Ability to give a cohesive history”
- “I found it impressive that it could directly tailor its answers to the specific questions I was asking”
- “the fact that they had an answer to basically any question i asked– with real life SPs if you ask them something they werent trained to hear they get flustered and dont know how to answer”
- “If I asked a question with multiple parts, or asking about multiple symptoms, it was usually able to address all parts of my question. It also answered my question when I was asking a pre-question that didn’t seem directly related (I asked if they lived with their wife, so then my next question could be if the wife noticed any sleep symptoms)”
- “I thought the inflection of the voice was impressive. They sounded realistic at times.”
- “responded to my questions and sometimes offered extra information”
- “able to hold a conversation in the initial portion of interview”
- “I found it roughly comparable to the AI agents my school has used. I think one of the positive aspects is having additional time to think through your questioning without feeling rushed.”
- “I liked that AI was used so I could ask any question and get a pertinent answer”

E2.

- “In the third case, for some reason, the AI started acting like it was a provider and not the patient. The AI was asking me for my symptoms and how I was feeling. I think it’s important to tighten up the technical issues.”
- “I thought it was a little odd for the AI to repeat exactly what I asked back to me but I suppose it made me feel as though it understood my questions.”
- “There were frequent technical difficulties with getting the agent to initially start responding or load properly”
- "1. Sometimes if I took too long to respond, or even if I just thanked the agent, the agent would forget that it was the patient and started being the provider in the scenario. 2. Sometimes it didn’t quite understand my question and answered a slightly tangential one. Sometimes it offered up information somewhat unrealistically to how a patient would have. 3. Sometimes if I reacted quietly with a ""hm"" or an ""mm,"" the agent would stop talking, which made it difficult to gather the information that was cut off."
- “the agent would forget their role if i took too long to respond - they would take on the doctor role”
- “i found there was a difficulty with the AI service if my internet was running slower. THis disrupts the overall flow of the conversation.”
- “I think the software needs to be developed more. i would notice when i asked some questions that the chatbot would respond with good answers that were human like, however, many questions that i don’t think it was programmed well for if should just say yes or no and reiterate what i asked it word for word. It did not feel human at all and when i would try to do a ROS with the chatbot it would sound offended and say no other symptoms, however when i asked the right question it would give me a symptom like diplopia even though it said there was nothing else”
- “There would be moments 1-2x per case where the agent’s response was not at all logically appropriate. I would ask them to expand on their symptoms and they would reply back with the clinician’s summary of next steps in management, instead of further expanding on the pt’s symptoms.”
- “Sometimes the agent would provide answers that were out of the scope of what I asked, which seemed like it was responding more to a pre-programmed prompt than my specific question.”
- “Any noise in the background causes the bot to stop talking. Some responses don’t exactly answer the question asked, others give more information than the question asked to the point that i was given information on questions I didn’t ask that a real patient wouldn’t know to share.”
- “I accidentally ended the first patient encounter due to a lack of understanding of how the website works. So I’d appreciate a test pa-

tient case on this website before delving into the first patient. I feel like I wasted a practice case. The second case was a confusing interaction as I think the AI agent confused where the conversation was going."

- "Troubles with audio, and also with divulging way too much information compared to what I asked. I know patients will do this in real life but I think it was giving me pertinent positives and negatives that I didn't solicit."
- "nothing"
- "Sometimes it was just repeating what I had asked, and only adding "yes" or "no" to their sentence. It didn't provide as much information as a typical patient would, or seem to understand what I was asking. With one encouter, I was just taking a second to think and then AI patient began telling me the next steps I should consider, which I was about to get to."
- "I noticed the AI repeated itself exactly in the same voice and rhythm with the same words, which a normal human would not do. The majority of the chatbots stopped working for me and I reached out about this. One of them worked completely so I got some understanding of how it worked but it definitely did not seem to be precise and fleshed out."
- "the last AI did not respond well to my questions"
- "sometimes its unnatural and they are not responding to my questions and started going off conversations that seems to be more from the provider than the pt"
- "Phrased things using medical terminology, which real patients don't do. Many of the answers also seemed quite scripted, where it answered much more than I had explicitly asked about. It also sometimes skipped around in audio, and on at least question asked me a question. It was also frustrating to have a patient with a positive depression screen who would not answer any depression-related questions, including those related to safety of self/others."
- "I could tell where the limitations of the AI's information were and where a normal person might've answered a question differently vs the chatbot"

E3.

- "I'd recommend it to medical students if they want an easy way to practicing interviewing strategies, especially for first or second year students before beginning clerkships. It's easy to do, portable, and it's nice knowing that there isn't an actual person judging you for your responses, so you feel more comfortable making mistakes."
- "I think it is helpful for working through differentials but I do not think it should replace the use of standardized patients in person. It is convenient to use from home and you don't feel awkward asking "dumb" questions or having gaps of silence to think through more effectively."
- "I would not, I feel like the most useful way in getting comfortable in interviewing and treating patients is to actually work with real people. SP and AI simulation is not very engaging and does not prepare us for real world interactions"
- "Yes because it would be good for judgement-free practice - the only thing is that in an OSCE-like scenario we would likely at least still have some notes, whereas here we had no notes and it made it somewhat difficult to remember."
- "patients are more unpredictable so i dont think this can substitute completely but could be helpful in addition to SP/patients"
- "i think it is good practice after you have already learned the basics, but not for initial practice. it is not lifelike enough."
- "not yet, i think the system isnt developed enough yet to have a good meanungful patient interaction however maybe once it gets better it would be a good skill for learning"
- "No. I don't think this provides any more training than reading and working through clinical cases in a short answer format. Real life clinical office visits have much more complex patient-clinician responses where the pt would interject with tangential symptoms, medical history and bring up outside data to discuss (imaging, labs, etc)."
- "Definitely! It is an easy, very life-like way to get more reps in and practice triaging, building a differential, taking a history, and even starting to formulate an A&P."
- "I think it is okay as a training exercise, but I do not think they will ever be as effective as human standardized patients or as real clinical encounters. The responses from the AI bots are too clunky and the conversation is too seg-

mented that it doesn’t reflect the true cadence of encounters with people.”

- “I would strongly recommend this AI-based training to more medical students. That’s not to take away actual human standardized patients, but these AI-based training modules offer more flexible and still realistic modals of training to students. Also it helps shy students practice their patient interviewing and presentation skills before doing it to a human patient.”
- “Yes - I think this is a good adjunct for students who want to practice their history taking in a low stakes environment.”
- “I think it could be helpful to practice skills but it will never supplant the value of in person standardized patients.”
- “yes i think AI can truly be a standardized experience for students and is more fair than real life piple who are not as standardized as they should be”
- “Yes, I think its helpful for students to practice the flow of interviewing while under the similar stress of simultaneously talking to someone.”
- “No I would not because it failed to fully work for me. If it did work and I could finish all the conversations I think it could be a valuable tool that I could recommend. It is also hard to make clinical decisions in the absence of a physical exam, which I could do on standardized patients, but this is a good tool to study with from home.”
- “Yes, it is helpful and makes me less nervous”
- “i dont think so because a lot of those sessions ended before i was done with my interviewing”
- “It’s okay as a supplement to standardized patients, or where it may not be feasible to use standardized patients. At this time, it’s much clunkier than standardized patient interactions and it does not feel that I’m truly simulating a patient interaction. Some of my classmates have also raised concerns related to the environmental impacts of AI usage.”
- “Yes, it’s comfortable to use since most of us are using AI anyway. its great to be able to practice without feeling pressured with a real patient or bored with a standardized patient”

F1. MS3: 20 responses (100%) MS1, MS2, MS4: 0 responses

F2. Currently in rotations: 18 responses (90%) Yes: 2 responses (10%) No: 0 responses (0%)

F3. Responses: 25, 26, 24, 27, 24, 25, 24, 24, 28, 24, 25, 28, 26, 26, 26, 24, 27, 27, 30, 31 Mean: 26.1 years | Median: 26 | Range: 24-31

F4. Yes: 17 responses (85%) No: 3 responses (15%)

F5. Vietnamese: 1 response Gujarati: 1 response Mandarin: 1 response

F6. Age 10: 2 responses [Not provided]: 1 response