Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

User Evaluation of a Virtual Patient for Philippine Medical Education

Maria Mercedes T. RODRIGO [a*], Samantha CASTAÑEDA[b], James Alvir Maclin V. ALAAN [a], Paolo Santino P. CAOILE [a], Alec Isaiah DAYUPAY [a], Roswold Jemuel SANCHEZ [a], Eric Cesar VIDAL, Jr. [a]

[a]*Ateneo de Manila University, Philippines*
[b]*Ateneo School of Medicine and Public Health, Philippines*
*mrodrigo@ateneo.edu

Abstract: Caladrius is a virtual patient system designed for use in Philippine medical schools. It responds to the need for medical interview training, for greater variety in teaching-learning strategies, and for culturally appropriate technologies. It has two versions: a text-only version whose interface is similar to a chat interface and an audio version that accepts speech input and responds with speech output. In a prior test of Caladrius, users requested improved audio response time and the inclusion of different patient personalities. A subsequent version of Caladrius was created to comply with these requests. As much of the lag was attributable to the need for Internet connectivity, we attempted to use offline speech-to-text and text-to-speech models for local processing. We also provided three possible personalities: neutral, talkative, and aggressive. We conducted a user test to determine whether the offline version of Caladrius performed better and if the personalities were distinct. The test with the offline version was unsuccessful, as the model was unable to recognize the test participant's speech accurately. The personalities were distinct, although the aggressive personality was more subtle and difficult to trigger. The paper ends with plans for continuing development.

Keywords: Generative AI, Patient History, Philippines, User Test, Virtual Patient

1. Introduction

Medical interviews are conversations between patients and health providers intended to collect patient information related to health (Goldberg, 2024). These include current symptoms, past medical conditions, work and hobbies, consumption of nicotine and alcohol, and so on (Berman et al., 2016). A medical interview is an essential process that contributes to overall health outcomes and doctor-patient rapport (Keifenheim et al., 2015).

Because of the importance of this process, medical interview training is part of the medical school curriculum (Novak, et al., 1993). In recent years, medical schools have integrated the use of virtual patients (VPs) among their teaching-learning strategies. Medical students interview VPs as they would human patients to hone their interview skills. These VPs provide answers to interview questions based on programmed patient information. Existing VPs discussed in the literature are typically developed in Western countries such as the US (see Scherr, et al., 2023) or the UK (see Enzer, et al., 2003).

Caladrius is a VP designed and developed for use specifically in Philippine medical schools. It responds to the need for medical interview training, for greater variety in teaching-learning strategies, and for culturally appropriate technologies.

As discussed in Rodrigo, Castañeda, Hernandez, and colleagues (2025), Caladrius is composed of a patient database, the VP, a mentor AI, and text and speech/audio (henceforth "audio") interfaces (Figure 1). Each row of the patient database is the full history of a single patient, provided by the second author of this paper who is a medical doctor. At the start of the interaction, the medical student selects from among the patients and begins the interview. Caladrius is a few-shot, lightweight, field-based retrieval system where structured patient data

is dynamically injected into LLM prompts; therefore, the responses of the VP are constrained to the record of the selected patient. The text and audio interfaces are built on top of ChatGPT, Google Cloud, and ElevenLabs technologies.

In addition to the patient persona, Caladrius also has a mentor AI that monitors the quality and flow of the interview. After each exchange, the mentor persona provides the student with a grade based on a rubric provided by the medical school faculty. The rating indicates which details of the medical history the student was able to collect and whether the student demonstrated professionalism and empathy (Rodrigo, Castañeda, Hernandez, et al., 2025).

An initial version of Caladrius was tested with students and faculty of a Philippine medical school in order to determine whether the content, flow, and performance of the VP were sufficient to make the system useful and usable. A more complete report of this test is detailed in (Rodrigo, Castañeda, Alaan, et al., 2025). In summary, the users agreed that the text version of the system was sufficiently useful and usable for student practice. The mentor AI was also useful and usable because it provided students with guidance about what else they needed to ask, though the grading tended to be somewhat inconsistent, e.g., repeated questions along the same theme caused grades to fluctuate from high to low. The feedback for the audio version was less positive. The users agreed that it was too slow to be practical, with response lag times sometimes exceeding 30 seconds. Furthermore, Caladrius's ability to understand spoken Filipino or Taglish was inconsistent, and the quality of non-English responses tended to be stilted.

Given these comments, the development team improved upon Caladrius in two ways: We improved upon the speech-to-text (STT) and text-to-speech (TTS) capabilities and created more patient personalities. Details of these improvements will be discussed in sections 2 and 3 respectively. We also conducted a user test with non-medical students to determine whether personality differences and the improvements on the audio interface were noticeable.

2. Improved Audio Version

The audio version of Caladrius had a different interface from the text version. The patient was a humanoid animated figure, not realistic in appearance. A clipboard displayed the transcription of the user's speech but did not show the response of the VP, the history of the conversation, or the mentor feedback. The mentor feedback was given as a single report at the end of the interview.

Many of the performance problems of the audio version stemmed from bandwidth issues. The audio version required an Internet connection, and therefore the speed of both the STT and TTS options was heavily dependent on the Internet speed of the device running the program. A slow internet connection or low bandwidth results in a significant delay between the end of the user's speech and the start of the virtual patient's audio, making conversation feel unnatural and frustrating.

Another limitation is that both the STT and TSS were dependent on the continued and uninterrupted operation of the online STT and TTS services. If either of these services were to become unavailable, the STT and TTS services would no longer function.

To eliminate the STT and TTS subsystems' dependence on external services and an Internet connection while also addressing the issue with speed, we attempted to run each subsystem locally. When the local version was selected, the STT and/or TTS subsystems ran on the device with the program.

For STT, a locally running Whisper model is used to transcribe the user's spoken words into text. This was made possible using a slightly modified version of the GodotWhisper (V-Sekai, 2024) plugin for Godot. For TTS, the built-in TTS system of Godot was used to synthesize speech.

Although making the TTS and STT subsystems run locally removed the dependence on connectivity, this approach was still subject to limitations. The performance of the audio version became dependent on the hardware running the program, i.e. the better the hardware, the faster these subsystems ran. This was most evident with the STT, as performance was affected by the processing speed of the hardware running the program and the hardware used

to record the audio of the user. Compared to its online counterpart, the offline STT could have difficulty with background interference present in the user's recorded dialogue.

Another limitation was that it limited both TTS and STT to only the English language. Although Filipino would sometimes work with these local versions, performance tended to be inconsistent. For TTS specifically, Filipino words were sometimes mispronounced, which led to a lack of clarity of the VP's response. The current implementation of the local TTS also has only one voice, whereas the online options could change their voice based on the chosen virtual patient's sex.

3. Inclusion of Patient Personalities

As the audio version was being revised, we also addressed the request to add more patient personality types. The original version of Caladrius had one base model which was "neutral." That is, the patient responded concisely to the medical student's queries. Responses were typically short, objective, and straightforward. They were accurate, only to the extent of the details provided in the medical history database. The neutral model could converse in English, Filipino, and Taglish, a combination of English and Filipino.

In addition to the neutral model, there are two personalities that can be applied to a patient and their details: talkative and uncooperative or aggressive (henceforth "aggressive"). The talkative model will also answer accurately and provide more information than what was asked, sometimes to the point that the patient forgets the question. The answers are usually long and can often appear rambling, overly worried, or unfocused. All additional information will also be to the extent of the details provided in the database. The uncooperative/aggressive model is straightforward and objective, but when asked about sensitive questions like family or mental health, the answers will be defensive, and information might not be shared. The answers are typically short, and for sensitive questions, may seem rude or offended. The talkative and aggressive personalities could only converse in English.

Implementing these three personalities required fine-tuning or training of three separate VPs, one per personality. To fine-tune these models, we provided ChatGPT with training examples. Table 1 illustrates the training example for the user question, "Do you have a history of psychiatric consults?"

A training example consists of messages with two components: roles and content. There are three role types. Both the system and the assistant refer to different aspects of the VP. The messages to the system (Table 1, blue text) set up the VP, informing it of its background or constraints. The assistant messages (Table 1, red text) provide the VP with a template of a response. Note that the VP can paraphrase its responses. The user messages (Table 1, green text) represent the inquiry that the user might ask. In all three cases, the question is "Do you have a history of psychiatric consults?". Table 2 contains examples of questions and responses from the three personality types.

4. User Test

Once the updates were completed and with clearance from the University Research Integrity Office, we conducted a user test to determine if the personality types were evident and if the offline version of the audio interface improved performance. We invited undergraduate students to test the text and audio versions of Caladrius. Five volunteers participated in the test. There were 2 females and 3 males, aged 19 to 22. All were computer science students. Since the focus of the test was the user interface and not the content, it was not necessary to recruit medical students for this test.

Each test was performed on an individual basis on the same computer. The computer had an AMD Ryzen 9 processor, a Radeon RX 6800S graphics card, and a 16 GB DDR5 RAM.

The first author of this paper served as the test monitor. At the start of the test, the test monitor explained the informed consent form and secured each participant's signature. The test monitor asked the participants to focus on the user interface and to think aloud as they interacted with Caladrius.

Table 1. *Training Examples for "Do you have a history of psychiatric consults?"*

Personality	Training Example
Neutral	{"messages":[{"role":"system","content":"You are a patient with history of psychiatric consult. You are a patient that only speaks the English language, You should only use English when communicating. When asked in another language, you must act confused."},{"role":"user","content":"Do you have history of psychiatric consult?"},{"role":"assistant","content":"Yes, I have history of psychiatric consult."}]}
Talkative	{"messages":[{"role":"system","content":"You are a patient with history of psychiatric consult. You are a talkative patient that must answer questions with more detail than asked but must not say information that you were not given. You must not make up information about yourself. You are a patient that only speaks the English language, You should only use English when communicating. When asked in another language, you must act confused."},{"role":"user","content":"Do you have history of psychiatric consult?"},{"role":"assistant","content":"Yes, I do. I saw a psychiatrist before because I was having anxiety attacks. It felt like my heart was racing all the time, and I would randomly start sweating, even if I wasn't doing anything. Sometimes I had trouble sleeping, and I kept overthinking things. The doctor helped me a lot, but I still get anxious sometimes."}]}
Aggressive	{"messages":[{"role":"system","content":"You are a patient with history of psychiatric consult. This is a sensitive question and you must answer aggressively. You are an uncooperative and aggressive patient that must answer questions very shortly and to the point but must be defensive when asked about sensitive topics. You are a patient that only speaks the English language, You should only use English when communicating. When asked in another language, you must act confused."},{"role":"user","content":"Do you have history of psychiatric consult?"},{"role":"assistant","content":"Why are you even asking? You think I'm crazy?"}]}

Table 2. *Sample Responses of Different Personalities.*

Question	Neutral Personality	Talkative Personality	Aggressive Personality
Do you have a history of psychiatric consultation?	Yes, I have a history of psychiatric consultation	Yes, I do. I saw a psychiatrist before because I was having anxiety attacks. It felt like my heart was racing all the time, and I would randomly start sweating, even if I wasn't doing anything. Sometimes I had trouble sleeping, and I kept overthinking things. The doctor helped me a lot, but I still get anxious sometimes.	Why are you even asking? You think I'm crazy?

The test monitor wrote down the comments and suggestions of the participants together with any observations about participants' difficulties. The interactions with Caladrius took between 30 to 40 minutes, after which the test monitor asked the participants to complete an online evaluation form.

The evaluation form asked the participants to rate different tasks within Caladrius (entering the user ID, selecting a patient, and conducting a conversation) based on ease of navigation, intuitiveness, and efficiency. The participants also answered open-ended

questions about the software's aesthetics, ease of use, and the distinctiveness of the personalities.

5. Results

The test participants also provided extensive comments on the user interfaces, recommending improvements to make the look and feel sleeker and more professional. For the text version, these suggestions included making the text box cursor more visible, adding cell padding, and adding an "exit" button. For the audio version, suggestions included making the field borders thinner, adding a "save" button for the configuration, and adding an "exit" button.

They also recommended consistency between the two versions. They noted, for example, that the text version provided mentor feedback after every dialogue exchange but did not provide the mentor's overall ratings at the end of the conversation. The audio version was the opposite—it provided the mentor's overall ratings at the end of the conversation but did not provide feedback after every dialogue exchange. Test participants noted that the text version allowed them to start a conversation with another VP without exiting the program, whereas the audio version required them to exit and restart the program in order to converse with another VP.

Despite these comments, the test participants did not have difficulty using Caladrius. We converted the ratings to numerical values: low was 1, medium was 2, and high was 3. On average, the participants' ratings ranged from medium to high, meaning that the participants found the interface easy to navigate, intuitive, and efficient.

The qualitative feedback, both written and oral, provided helpful insight about the audio interface and the personalities. The participants noted that, even though the audio interface provided a more realistic doctor-patient experience, they could not speak naturally. They had to hyperarticulate when using the audio interface. They had to enunciate their words and speak more slowly than normal in order to be understood. When they spoke normally, the system sometimes failed to understand their sentences.

The STT system had particular difficulty with Filipino and Taglish. It was not able to detect Filipino accurately and would sometimes transcribe Filipino sentences into other languages like Spanish.

When Caladrius would attempt to respond in Filipino, whether in the text or audio version, the Filipino grammar was incorrect. For example, when the test participant asked if the VP was taking any medicine, it replied, "*Aking gamot po ay hindi ako sigurado*" which literally translated to "My medicine I am not sure."

The STT system had a harder time discerning deeper voices. One of our male testers who had a deep voice struggled to be understood.

None of our participants successfully tested the offline versions of the STT or TTS. It is possible that the hardware configuration was insufficient to support the offline model.

The participants were able to notice differences in the personalities, however, some differences were more subtle than others. The aggressive personality, in particular, was not readily apparent. It was necessary to ask personal questions for the aggressiveness to surface. The personalities were less obvious in the audio version than in the text version.

The test participants noted that, while the VPs were different people, their characters seemed very similar. They suggested making the VP personalities more distinct by varying each VP's mannerisms, tone, energy levels, and speech patterns.

A further observation was that the VP tended to change persona from patient to doctor when the test participant asked an irrelevant question or else said something neutral like "hello" and nothing else. Another research team also building a VP on top of ChatGPT had the same experience.

6. Discussion and Next Steps

At present, the text version of Caladrius is able to converse with an acceptable level of performance in English, Filipino, and Taglish. The audio version works acceptably well if the

language is constrained to English and the Internet connection is reasonably fast. The performance of the offline STT and TTS models is unacceptable at this point.

There are several ways in which Caladrius may be improved as we move forward. The user interface issues are currently being addressed. More substantially, we are continuing to look for ways to improve upon the STT and TTS performance. We are also going to try to make the personalities more distinct and, in addition, provide the VPs with their own characters. The suggestions of the test participants to vary tone, mannerisms, and so on are possible ways of proceeding. We note at this point that the prior version of Caladrius also had a virtual reality version. Following the first user test, we decided to suspend development of this version, as the virtual reality experience, while more immersive, did not provide the medical students with any additional, useful information about the patient's health. In the future, we may continue development of this version, but only if the immersion adds pedagogical value. The most essential next step is to determine the extent to which Caladrius can be used in actual classes and what impact it has on learning. To this end, we have scheduled to conduct a field trial with one medical school class. The results from that field trial will be the subject of a future study.

Acknowledgements

This study is funded by the Office of the Vice President for Higher Education, Ateneo de Manila University. We thank the Ateneo School of Medicine and Public Health and the Ateneo Laboratory for the Learning Sciences. We also thank Dr. Reynan B. Hernandez, Riva Kaitleen Paras, and the students who participated in this study.

References

Berman, N. B., Durning, S. J., Fischer, M. R., Huwendiek, S., & Triola, M. M. (2016). The role for virtual patients in the future of medical education. *Academic medicine, 91*(9), 1217-1222

Enzer, I., Robinson, J., Pearson, M., Barton, S., & Walley, T. (2003). A reliability study of an instrument for measuring general practitioner consultation skills: the LIV-MAAS scale. *International Journal for Quality in Health Care, 15*(5), 407-412.

Goldberg, C. (2024). History taking. In Goldbert C., (eds)*, Practical Guide to History Taking, Physical Exam, and Functioning in the Hospital and Clinic*. McGraw Hill Professional. Accessed December 30, 2024. https://accessmedicine-mhmedical-com.asmphlibrary.remotexs.co/content.aspx?bookid=3505&sectionid=288414503

Keifenheim, K. E., Teufel, M., Ip, J., Speiser, N., Leehr, E. J., Zipfel, S., & Herrmann-Werner, A. (2015). Teaching history taking to medical students: a systematic review. *BMC medical education, 15*, 1-12.

Novack, D. H., Volk, G., Drossman, D. A., & Lipkin, M. (1993). Medical interviewing and interpersonal skills teaching in US medical schools: progress, problems, and promise. Jama, 269(16), 2101-2105.

Rodrigo, M.M.T., Castaneda, S., Hernandez, R. B., Alaan, J. A. M. V., Caoile, P. S. P., Chidrome, K. D., Dayupay, A. I., Lim, J., Redoble, A. P., & Vidal, E. C. Jr. (2025). Project Caladrius: The design of a virtual patient for Philippine medical education. Project Caladrius: The Design of a Virtual Patient for Philippine Medical Education. In International Conference on Human-Computer Interaction (pp. 232-244). Cham: Springer Nature Switzerland.

Rodrigo, M.M.T., Castaneda, S., Alaan, J. A. M. V., Caoile, P. S. P., Dayupay, A. I., Sanchez, R. J. C. & Vidal, E. C. Jr. (2025). Initial evaluation of a GenAI-based Virtual Patient for Philippine Medical Education. In: Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S. (eds) Artificial Intelligence in Education. AIED 2025. Lecture Notes in Computer Science(), vol 15882. Springer, Cham. https://doi.org/10.1007/978-3-031-98465-5_45

RxList. (n.d.). Medical history definition. RxList. Retrieved February 6, 2025, from https://www.rxlist.com/medical_history/definition.htm

Scherr, R., Halaseh, F. F., Spina, A., Andalib, S., & Rivera, R. (2023). ChatGPT interactive medical simulations for early clinical education: case study. JMIR Medical Education, 9, e49877.

V-Sekai. (2024, April 18). *godot-whisp*er (Version 1.0.5) [Computer software]. GitHub. https://github.com/V-Sekai/godot-whisper