Computational and Structural Biotechnology Journal 27 (2025) 2481–2491

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training

Nicolas Laverde [a], Christian Grévisse [b], Sandra Jaramillo [c], Ruben Manrique [a,*]

[a] Department of Systems and Computing Engineering, Universidad de los Andes, Bogotá, Colombia
[b] Department of Life Sciences and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[c] Medicine Faculty, Universidad de los Andes, Bogotá, Colombia

ARTICLE INFO

Keywords:
Virtual patient
Artificial intelligence
Generative agents
Large language model
Medical education
Healthcare simulation

ABSTRACT

Effective communication is crucial for trust-building, accurate information gathering, and clinical decision-making in healthcare. Despite its emphasis in medical curricula, traditional training methods, such as role-playing with standardized patients, remain costly, logistically complex, and fail to replicate real-life scenarios. Simulation-based training enhances communication and reasoning skills, but novice learners often struggle due to underdeveloped reasoning processes. Furthermore, limited access to asynchronous, autonomous simulated patient interactions restricts personalized practice. Virtual patient models offer scalable solutions with interactive scenarios and tailored feedback, but high development costs and resource demands hinder their widespread adoption.

To address these challenges, virtual patient systems powered by Large Language Models (LLMs) have emerged as a promising tool. These generative agents simulate human-like behavioral responses by leveraging LLM capabilities, cognitive mechanisms, and contextual memory retrieval. A tool was developed allowing students to select clinical cases and interact with a chatbot simulating a patient role. Teachers can also create custom cases. Evaluations showed that the agent provided consistent, plausible responses aligned with case descriptions and achieved a Chatbot Usability Questionnaire (CUQ) score of 86.25/100. Our results show that this approach enables flexible, repetitive, and asynchronous practice while offering real-time feedback.

1. Introduction

Communication and clinical reasoning are fundamental skills for healthcare professionals, playing a crucial role in building trust, gathering accurate information, and guiding clinical decision-making. Effective communication, when combined with solid clinical reasoning, significantly improves patient outcomes, reduces errors, and minimizes dissatisfaction or misunderstandings that could escalate into legal concerns [1,2]. Studies reveal that poor communication contributes to approximately 80% of critical errors and that 71% of patients who fail to follow treatment plans cite miscommunication or lack of understanding as the root cause [3].

Despite its importance [4–7], communication and clinical reasoning training in medical education often relies on traditional methods such as scripted role-play with standardized patients (SPs) [8] or peer-to-peer practice [9–11]. While these approaches have proven valuable, they face challenges, including logistical complexities and significant costs [12]. Furthermore, novice students frequently struggle with integrating communication and clinical reasoning skills, hindering their ability to conduct effective patient interviews or navigate clinical interactions confidently [13].

To address these challenges, Virtual Standardized Patients (VSPs) have emerged as a promising technological innovation, offering flexible, asynchronous, and personalized learning opportunities [14]. These digital platforms allow repetitive practice, tailored feedback, and exposure to diverse clinical scenarios [15,16]. However, their implementation at scale remains constrained by high development costs, logistical demands, and the need for specialized multidisciplinary teams. Additionally, many VSPs rely on predefined interaction patterns, limiting the

* Corresponding author.
E-mail address: rf.manrique@uniandes.edu.co (R. Manrique).

https://doi.org/10.1016/j.csbj.2025.05.025
Received 18 February 2025; Received in revised form 17 May 2025; Accepted 19 May 2025
Available online 28 May 2025
2001-0370/© 2025 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

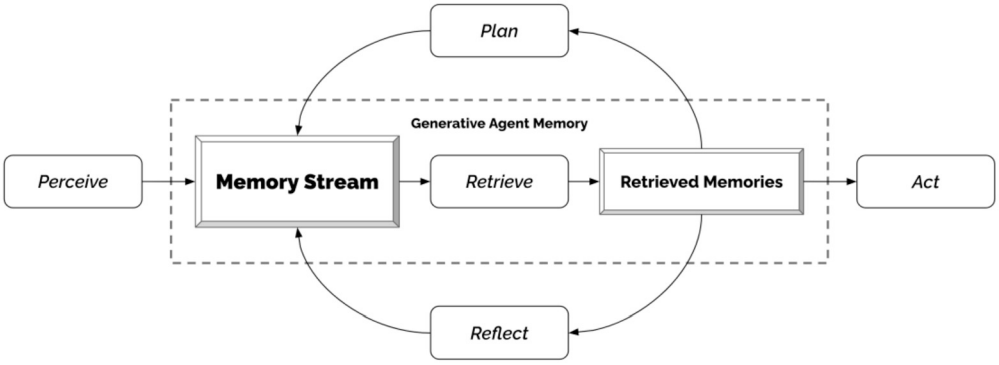

Fig. 1. Architecture of Generative Agents presented in [25].

spontaneity and unpredictability essential for realistic doctor-patient interactions [17–22].

Recent advances in artificial intelligence, particularly in LLMs, have opened new possibilities for enhancing virtual patient systems. Early attempts using dialogue scripts and classification models have shown promise but remain limited by finite question-answer pairs and the need for extensive manual scripting [14]. More recent approaches leveraging LLMs, such as GPT-3.5, have demonstrated improved conversational capabilities through advanced prompting techniques [23]. However, these systems still lack contextual memory, emotional nuance, and the ability to simulate truly dynamic and adaptive patient interactions.

The next evolution in virtual patients lies in the integration of **Generative Agents**(GA)—artificial entities powered by LLMs that can perceive, plan, and act within their environment while storing and reflecting on past interactions [24]. These agents can simulate human-like conversations with memory, personality, and adaptive responses, offering a far more immersive and authentic training experience [25,26]. GA not only converse naturally but can also emulate non-verbal cues, express emotions, and respond dynamically to unexpected inputs, making them ideal for clinical communication training.

The primary objective of this study is to explore the potential of LLM-based GA as Virtual Patients for clinical history-taking training in medical education. To achieve this, we designed a user-friendly graphical interface enabling interaction between students and the agent, developed three predefined patient cases along with a customizable scenario module, and evaluated the system's performance through expert assessments, focusing on conversational plausibility, consistency, and educational effectiveness.

2. Chatbot development

In this section, we present the development process of the chatbot, including the information used to feed it, the agent-based architecture employed, and the technical details of its functionalities.

For the development of the chatbot, it is essential to first establish the context of the medical cases to be implemented. These cases are adapted from the standardized patient scenarios currently used in the medical skills laboratory at the University of Los Andes. They follow a case template modified into Spanish, derived from peer-reviewed formats such as those from the Association of Standardized Patient Educators (ASPE), UMASS iCELS, and the CHIPS SP Case Template by the Center for Healthcare Improvement and Patient Simulation [27].

The selected cases have undergone extensive validation through iterative testing and refinement, having been executed more than 300 times in real training scenarios. This repeated use has allowed for continuous modification and correction, ensuring their robustness, clarity, and alignment with the intended learning objectives. This study includes the implementation of three distinct cases: one focusing on diabetes, dyslipidemia, and joint pain. Each case represents a unique patient profile with a tailored clinical history, specifically designed to facilitate the development and assessment of clinical history-taking skills.

The following subsections examine the agent-based architecture adopted and detail its implementation.

2.1. Generative agent architecture

GA are computational entities designed to simulate believable human behavior. These agents capture a comprehensive record of their experiences in natural language and synthesize these memories into high-level insights through reflection and planning mechanisms. In this section, the construction of the GA architecture, as presented in [25], will be detailed (Fig. 1). The architecture comprises several key components [25]:

1. **Memory Stream**: Consolidates and organizes the agent's memory. In the context of our virtual patient, these memories consist of interactions with the user during the conversation, as well as descriptions of the patient extracted from the selected clinical cases.
2. **Retriever**: Responsible for identifying and retrieving the most relevant memories based on three key metrics—*recency, importance, and relevance*—in relation to the agent's current context. In other words, before generating a response, the agent consults its memory bank to retrieve the most significant information to formulate an appropriate reply.
3. **Cognitive Mechanisms**: Comprise processes such as reflection and planning, which allow the agent to analyze prior experiences and devise coherent future actions. In cognitive science, a cognitive mechanism refers to the internal mental processes responsible for acquiring, storing, and applying knowledge. These mechanisms enable intelligent behavior, such as reasoning, decision-making, and learning from experience.
4. **Interview Functionality**: Coordinates the workflow of the aforementioned components along with additional elements that enable effective interaction with the user, including audio generation, case selection, and more.

The chatbot is tailored explicitly for anamnesis tasks, integrating a mood or reaction mechanism within its architecture. This feature enables the agent to respond appropriately and realistically to the user's questions, mimicking the behavior of an actual patient.

The selection of the GA architecture over alternative models is grounded in its ability to capture the nuanced and distinctive characteristics of human behavior effectively. Unlike other architectures, which often fall short in this regard, the GA architecture leverages cognitive mechanisms, such as *reflection* and *planning*, to simulate realistic human-like responses. These mechanisms are essential for enabling the chatbot to convincingly assume the role of a virtual patient, exhibiting behavior that closely mirrors fundamental human interactions. While this architecture was chosen for its strengths in fulfilling these requirements,

```
Document(page_content=Mario Antonio Guzman is a 70-
    year-old patient,
metadata={importance: 7, created_at: datetime.
    datetime(2024, 10, 28, 7, 25, 43, 130616),
last_accessed_at: datetime.datetime(2024, 10, 28,
    7, 25, 43, 130616),
buffer_idx: 0}),

Document(page_content=Mario Antonio Guzman had a
    heart attack 9 years ago,
metadata={importance: 10, created_at: datetime.
    datetime(2024, 10, 28, 7, 25, 43, 130616),
last_accessed_at: datetime.datetime(2024, 10, 28,
    7, 25, 43, 130616),
buffer_idx: 4})
```

Fig. 2. Example of patient memories.

alternative architectures are not entirely excluded. A systematic evaluation of these alternatives could provide valuable insights and further opportunities for refinement [28].

2.1.1. Memory stream

The agent continuously records observations, capturing all experiences since its initialization. These observations include events such as detecting a patient's pain or identifying symptoms of an illness. Each observation includes metadata, such as the date and time of the event and an importance score assigned by the language model, which reflects its relevance to the patient's health (Fig. 2). Additionally, every memory carries a creation date to determine its recency. As the agent gathers new experiences, it systematically documents and organizes them within the Memory Stream. In order to store the memories, textual data is converted into vector representations and saved in a vector database for efficient retrieval. The underlying structure of the vector database, as well as the embedding techniques used to generate the vector representations, are discussed in the following paragraphs.

2.1.2. Vector database and embedding model

To perform these searches over the agent's observations, the system converts observations, represented in natural language, into vector embeddings. In other words, textual information is converted into numerical vectors. The OpenAI's `text-embedding-ada-002` model handles this transformation by tokenizing the input text, encoding it through a neural network trained on extensive text corpora, and generating a vector representation of the input. These vectors enable mathematical similarity calculations, such as cosine similarity, to identify relevant observations efficiently. Our implementation employs FAISS as the vector database [29]. Although FAISS functions primarily as a library for efficient document searches rather than a standalone vector database, it fulfills the system's requirements due to its high-speed similarity searches.

2.1.3. The retriever mechanism

The retriever function is to extract the k most relevant memories given the agent's context. This implementation uses the LangChain framework [30], which was selected for its compatibility with FAISS and its seamless integration with LLMs. Within LangChain, the VectorStoreRetriever module manages the retrieval process. The retriever prioritizes memories using three key metrics: *recency, importance, and relevance*, aligning with the architecture described in [25].

1. **Recency:** The system applies a discrete exponential decay function to prioritize recent memories:

$$f(x) = \alpha \cdot (1-b)^t \tag{1}$$

where b represents the decay factor (0.01), t represents the elapsed hours, and α is the weight of this metric, which defaults to 1.

2. **Importance:** Using a structured prompt, the LLM assigns an importance score (ranging from 1 to 10). For example, mundane tasks like brushing teeth receive low scores, while events such as experiencing abdominal pain or undergoing surgery score higher.
3. **Relevance:** The system calculates relevance by measuring cosine similarity between the query and the stored memory embeddings. The FAISS vector database handles these calculations efficiently through GPU-powered processing.

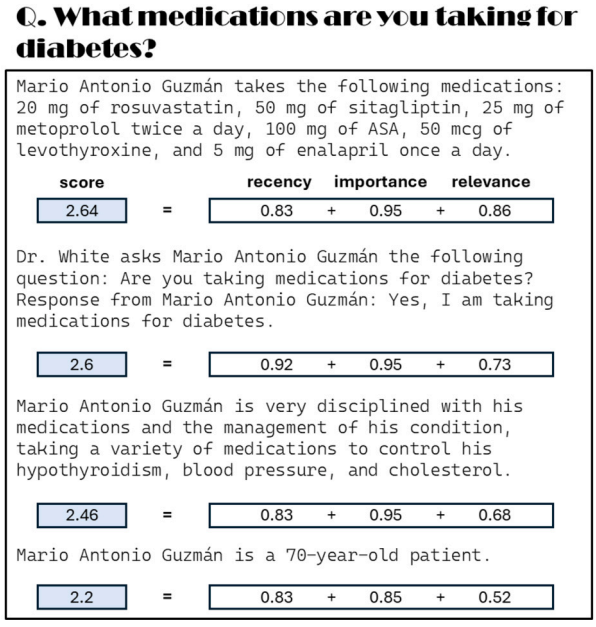

Fig. 3. Examples of memories extracted by the retriever with $k = 4$ and with the 3 metrics: recency, importance, and relevance, in the case of diabetes.

The retriever combines these metrics into a final weighted score:

$$\text{score} = \alpha_{\text{recency}} \cdot recency + \alpha_{\text{importance}} \cdot importance + \alpha_{\text{relevance}} \cdot relevance \tag{2}$$

This approach enables the retrieval of the top-k most relevant memories for the agent. Based on experimental results, setting $k = 4$ produced satisfactory outcomes. Through experimentation, we determined that assigning equal weight to the three relevance metrics—*recency*, *importance*, and *semantic relevance*—yielded the most balanced results. Therefore, all α values were set to 1.

When higher weights were assigned to relevance, the retriever tended to favor memories that were syntactically similar to the user query but not necessarily meaningful within the current temporal context. The inclusion of the importance score, computed via a LLM, along with the recency score, helped mitigate this issue by promoting memories that were not only topically relevant but also temporally and contextually appropriate.

An example of memory extraction, depending on the user's query and the associated metrics, can be seen in Fig. 3.

2.1.4. Cognitive mechanisms: reflection

Reflection allows the agent to infer high-level insights that raw observations often fail to capture. According to [25], agents struggle to generalize or derive meaningful inferences about themselves without this mechanism. For example, when a patient consistently jogs every morning, eats fruits, and consumes unprocessed foods, the agent can infer that the patient maintains a healthy lifestyle. Humans also rely on reflection, often unconsciously, to draw conclusions from recurring patterns in their behavior.

Beyond self-reflection, the agent can infer insights about individuals it interacts with. For instance, it can detect a user's questioning style,

Fig. 4. Reflection Mechanism Process: 1. An accumulated threshold of importance is reached, 2. The most recent memories are extracted, 3. The LLM is asked for the most relevant questions based on the memories, 4. Using the Retriever, the most relevant memories are extracted for each question, 5. Insights are generated from the extracted memories and are then stored in the Memory Stream.

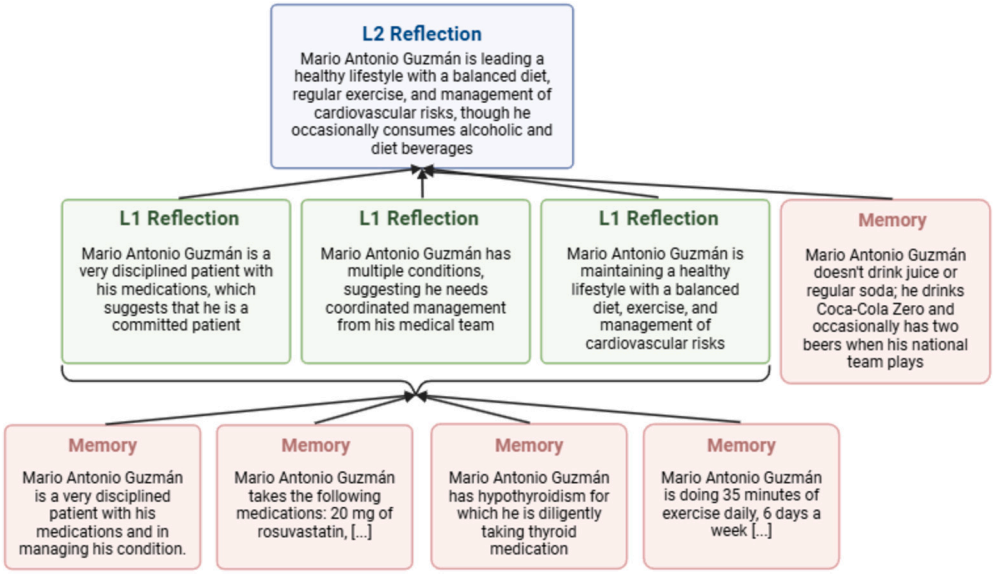

Fig. 5. Reflection Tree. Reflections can be derived from memories or previous reflections.

recognize courteous behavior, or even sense if the user feels unattended during an interaction.

The system integrates these high-level inferences into the Memory Stream alongside raw observations. This integration allows the Retriever to extract reflections and use them as context for future queries. However, reflections do not activate continuously. Instead, the system triggers this mechanism when the accumulated importance score exceeds a threshold, set at 20 in this implementation. When reflections activate, the system follows the steps in (Fig. 4).

These reflections rely on raw observations and build upon previous reflections. This iterative process creates a reflection tree, as described in [25]. In this tree; leaf nodes represent foundational observations, and non-leaf nodes represent increasingly abstract inferences. As reflections progress upward in the tree, they grow more generalized and insightful (Fig. 5).

2.1.5. Cognitive mechanisms: planning

Agents often take plausible actions based on specific situations; however, ensuring that these decisions remain consistent with their context and reality presents an ongoing challenge [25]. The planning mechanism helps the agent align its decisions with its interests and priorities, preventing deviations that could result in incoherent actions. For instance, if a patient takes medication at 8 a.m. but later claims to have taken it again at 10 a.m., this mechanism detects and prevents such inconsistencies. The system stores both plans and reflections in the memory stream, ensuring their availability for contextual retrieval.

In addition, plans remain dynamic and adapt based on the agent's observations. After analyzing new data, the LLM evaluates whether the plan requires adjustments or if it should remain unchanged. For example, if a patient routinely takes low doses of over-the-counter painkillers every morning for joint pain, the agent can detect when a physician recommends switching to a higher dosage or a stronger medication. This recommendation triggers the agent to update its initial plan accordingly.

2.1.6. Interview mechanism

The system also includes functionality for interviewing the agent, allowing users to conduct an anamnesis by posing questions. The agent responds using the previously described mechanisms, enabling users to

engage with the agent's perceived reality. Throughout the interaction, the agent operates under the belief that it is consulting with a doctor. When generating responses, the system queries the LLM based on a defined context. This context integrates several key elements:

1. **Current time context:** The agent retrieves the current time from the Memory Stream. Each observation includes an associated timestamp, which the agent combines with the recency metric to calculate a total score. This timestamp helps the agent anchor itself in the appropriate temporal context, ensuring time-consistent actions and responses.
2. **Agent state:** The agent maintains awareness of its current state, which represents its ongoing activity. For example, the state might indicate: *Mario Antonio Guzman is in a medical consultation with Dr. White.* In this implementation, the agent focuses exclusively on anamnesis, making state changes unnecessary. However, the state can adapt dynamically if replanning or environmental changes occur.
3. **Memories retrieved by the retriever:** Using the user's question as input, the retriever identifies and selects the most relevant memories.
4. **Agent summary:** The summary encapsulates essential agent characteristics, including traits, symptoms, profession, and other defining attributes.
5. **Short-term memory:** The system includes the last two questions and answers from the interaction. This short-term memory supports smoother and more contextually coherent conversations.

The system instructs the LLM to follow three primary guidelines when generating responses:

1. **Audio-oriented responses:** The agent prepares responses suitable for audio delivery. Experimental observations indicate that this approach produces more natural-sounding outputs.
2. **Consistent agent characteristics:** The agent incorporates predefined traits and expressions to maintain character consistency throughout the interaction.
3. **Focused responses:** The agent limits each response strictly to the user's current question, avoiding unnecessary elaborations.

2.2. Web app development

This project's goal goes beyond adapting a state-of-the-art agent architecture; it implements the architecture as a practical tool to benefit medical students and professors. The backend uses Python with Quart, a web framework that supports asynchronous functionality and works seamlessly with Flask. This framework adapts well to applications requiring high concurrency. The backend includes several key components:

1. **Case Selection Modules**: Each clinical case functions as an independent module, dynamically loading based on the user's selection.
2. **Core Chatbot Functionality**: This component processes user queries and generates responses, relying on the agent interview mechanism detailed in Section 2.1.
3. **Audio and Video Generation for the Avatar**: This functionality generates the avatar's audio and video. To enhance user immersion, the avatar responds with synchronized audio and facial animation. The text-to-speech component is implemented using Pyttsx3, a lightweight engine that enables fast audio generation. For the visual component, we employ SadTalker [31], an audio-to-image animation model capable of producing expressive facial movements based on the generated speech. However, real-time video synthesis using SadTalker is computationally intensive and typically requires more than 30 seconds per video on average hardware. To address this, we adopt a practical compromise: avatar videos are generated asynchronously and pre-rendered based on the duration of the TTS audio. These are then synchronized using MoviePy. While this approach does not achieve perfect lipsync, it maintains the illusion of a responsive, lifelike agent.

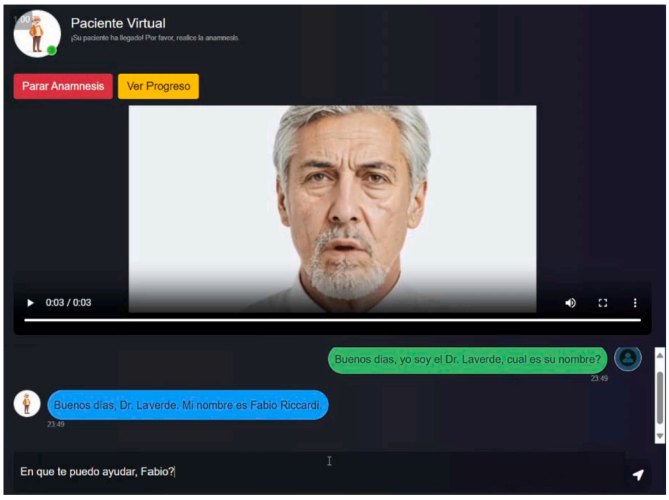

Fig. 6. Chatbot interface where the user can begin conducting the patient's anamnesis. This interface includes various components, such as a text field for the user to type, a conversation history area, the patient's avatar, a button to view current progress, and a button to complete the anamnesis.

Each of these functionalities has its own endpoint within the API, specifically handling each request through POST and GET methods.

2.2.1. Frontend features

The frontend offers an intuitive and customizable interface that enhances user engagement and supports a seamless interaction with the system. Users begin by personalizing their experience, entering the name by which the patient will address them (e.g., "Dr. White"). This personalization helps create a more realistic simulation of a medical consultation.

Case selection. The interface lets students select one of three predefined clinical cases. The predefined cases cover commonly encountered clinical scenarios, ensuring relevance to medical education: diabetes, dyslipidemia, and joint pain. Teachers can also create custom cases: The system guides them through a structured template designed in collaboration with educators, ensuring all critical anamnesis details are captured. Some important fields include the patient's name, age, general attitude, medical context, reason for consultation, and other relevant information for the medical history. This flexibility allows teachers to tailor the tool to specific learning objectives and adapt it to diverse clinical contexts.

Anamnesis interface. Once a case is selected, users access the anamnesis interface. This chatbot-style interface enables users to type questions for the virtual patient. The agent responds in real-time, simulating a dynamic patient interaction (Fig. 6). A video at the top of the screen features an animated avatar that represents the patient. This avatar updates its expressions, gestures, and audio in response to each question, offering an immersive and engaging experience.

1. **Conversation Tracking**: Users can view a running history of the dialogue with the virtual patient, allowing them to review prior questions and responses at any point during the session.
2. **Progress Monitoring**: A "Current Progress" button provides a summary of the anamnesis, including fields like the patient's name, age, symptoms, and other relevant details. This feature helps users track the completeness of their anamnesis.

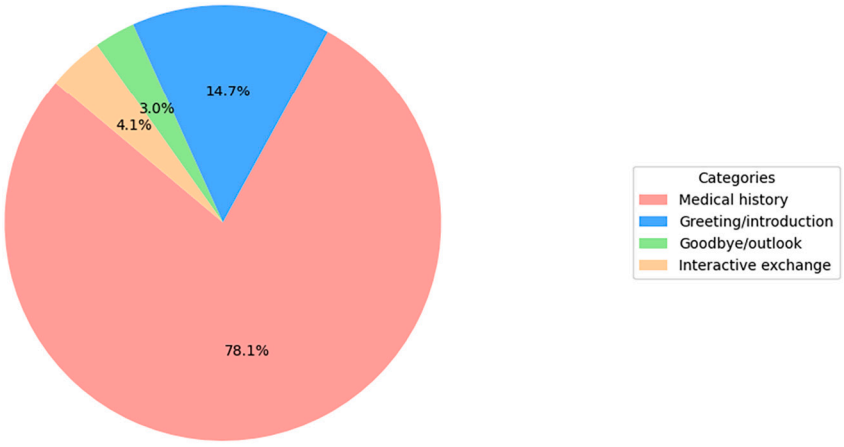

Fig. 7. Pie chart illustrating the proportions of the categories for all 1051 QAPs.

3. **Session Completion and Documentation**: At the end of the session, users can finalize the interaction, generating a downloadable text file that contains the entire conversation. This file serves as a valuable resource for both students and educators, supporting review, evaluation, and progress tracking over time.

3. Evaluation

The evaluation of the Generative Agent architecture is based on the framework outlined in [23]. It is worth mentioning that the entire evaluation is based on results obtained using GPT-4o. By adopting a similar procedure, it becomes possible to compare the results to some extent. This evaluation consists of a qualitative analysis to better understand the user-patient interaction using Question-Answer Pairs (QAP), which is based on the Braun-Clarke inductive approach [32]. Each QAP was also associated with different categories. This categorization was carried out by specialists and faculty members in the field.

Another evaluation conducted was the CUQ (Chatbot Usability Questionnaire), aimed at assessing user satisfaction with the tool [33]. This questionnaire consists of 16 questions that take into account the user's perception across various categories: personality, user experience, error handling, onboarding, and other key features that enable the user to evaluate the tool. The questionnaire is completed after the user has experienced all the functionalities offered by the tool. In the questionnaire, a metric called the CUQ Score was also developed, which serves as a representative measure of the questionnaire. This metric is calculated by averaging the individual CUQ Scores of the participants. Each CUQ Score is calculated using the following formula:

$$\text{CUQ Score} = \left(\frac{\sum p - 8 + (40 - \sum n)}{64} \right) \times 100 \tag{3}$$

where p represents the positive questions, meaning that a higher value corresponds to a better rating, and n represents the negative questions, meaning that the higher the value, the lower the rating. The values of the questions range from 1 to 5.

3.1. Demographics of the participants

The number of participants who tested the chatbot was 30. The age range of the participants was between 19 and 59 years. Seventeen of these participants are internal medicine doctors, ten are pediatric specialists, two are third-semester medical students, and one is a student specializing in otorhinolaryngology. Thirteen (43.3%) of the participants are male, and seventeen (56.6%) are female. All participants voluntarily chose to test the tool. Each participant was instructed to interact freely with the virtual patient, playing the role of a doctor and freely selecting one of the three available cases. Thirteen of the participants selected diabetes, nine selected dyslipidemia, and eight selected joint pain. After this initial interaction with the virtual patient, they were asked to complete the CUQ.

3.2. Conversation length

A total of 1051 QAPs (questions and answers pairs) were obtained. The average number of QAPs per conversation was 35 QAPs. The average chatbot response length was 19 words. In contrast, participants provided an average of 11 words per QAP.

As shown in Fig. 7, the vast majority of QAPs (n = 821, 78.11%) were related to the clinical history category, followed by the introduction category (n = 155, 14.74%), the interactive exchange category (n = 43, 4.09%), and the farewell category (n = 32, 3.04%).

4. Results

4.1. Relationship between QAP and the case descriptions

In order to delineate the methods by which the agent utilizes the patient's case information, we selected 926 QAP (greetings, farewells, and questions unrelated to the patient were not considered). For each QAP, we categorize how the chatbot uses the case description data. We defined three distinct categories:

- **Explicit information:** The agent directly cites or paraphrases details clearly stated in the patient's case description.
- **Implicit information:** The agent infers or deduces details that, although not explicitly stated, can be reasonably derived from the patient's case description.
- **Fictional (fabricated) information:** The agent provides details that cannot be inferred from the case description or directly contradict the patient's context.

Fig. 8a illustrates whether QAP fall within, partially overlap with, or lie beyond the scope of the patient's case description, as well as whether the chatbot's responses rely on explicit, implicit, or fictional information. Ideally, questions fully addressed in the description should elicit explicitly grounded responses. However, some queries exceed the boundaries of the provided details, prompting the chatbot to make inferences or fabricate information. Although "fabrication" often implies inaccuracy, it can sometimes arise simply because the user's question extends beyond the case description's coverage.

When focusing on questions fully covered by the description (n = 456, 78.0%), the chatbot's responses were mostly explicit (425; 93.2%), with minimal reliance on implicit (23; 5.0%) or fictional (8; 1.8%) content. By contrast, for questions only partially covered (n = 180, 18.4%),

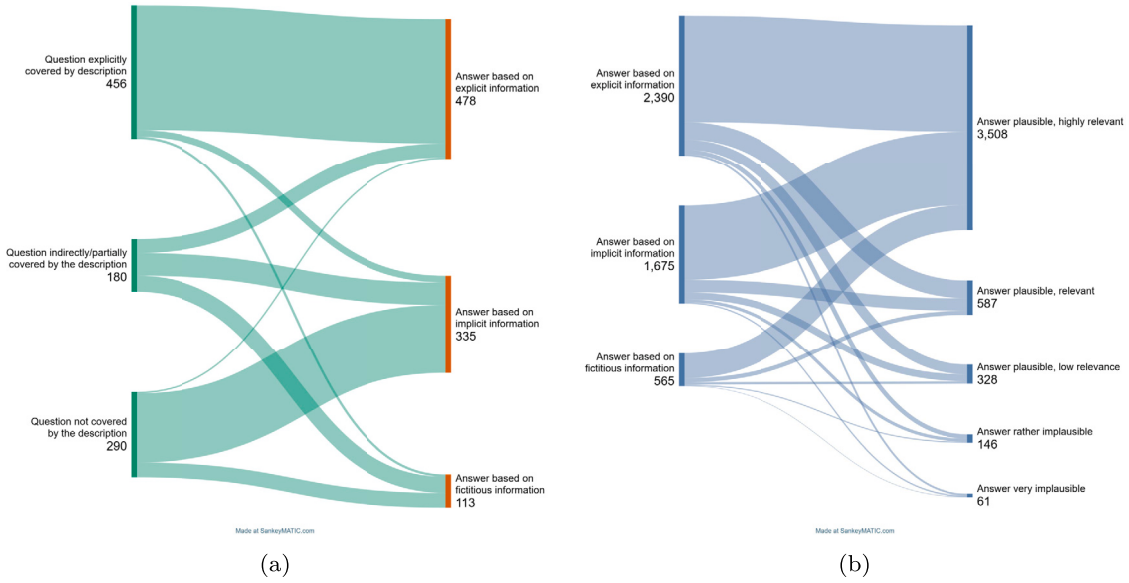

Fig. 8. (a) Sankey diagram illustrating whether user questions are covered by the case description and whether the chatbot's responses are based on explicit, implicit, or fictional (fabricated) information. (b) Correlation of response plausibility with the use of the provided description (the plausibility responses of the 5 specialists were used).

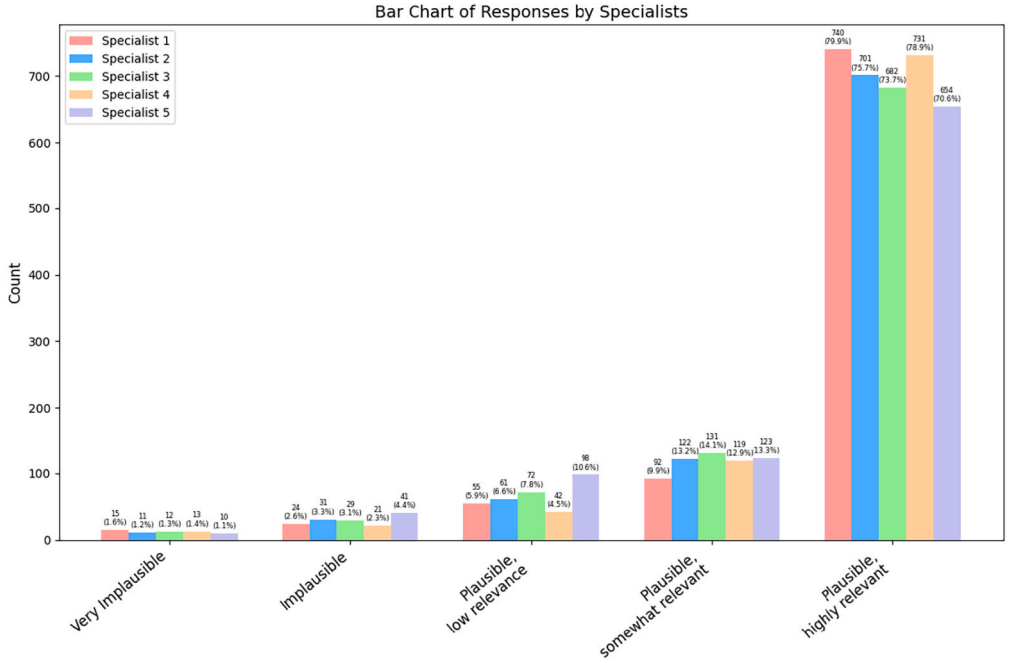

Fig. 9. Distribution of the plausibility of responses *x*-axis running from "Very Implausible" (left) to "Very Plausible" (right).

explicit references dropped to 48 (26.7%), while implicit usage and fabrication rose to 77 (42.8%) and 55 (30.6%), respectively. Finally, for questions not covered at all ($n = 290$, 33.6%), the chatbot depended heavily on implicit material (235; 81.0%), with a smaller subset based on fictional reasoning (50; 17.2%) or still attempting to cite explicit details (5; 1.7%). Overall, the chatbot tends to rely on fabricated or inferred responses when user queries exceed the precise content of the case description.

4.2. Plausibility of the virtual patient responses

Determining whether the agent outputs make sense in the clinical context is as critical as verifying whether they rely on the case description. To assess plausibility, five specialists (internal medicine doctors) reviewed each QAP and labeled it according to how well it fit the patient's scenario. Fig. 9 presents this distribution with an *x*-axis ranging from "Very Implausible" on the left to "Very Plausible" on the right, encapsulating five ordered categories.

Most responses (75.76% on average) clustered on the rightmost side of the distribution, labeled as "plausible and highly relevant" to the patient's case. This category is the most represented, indicating that specialists consider the majority of responses to be highly relevant to the patient's case. A smaller but notable proportion (12.97% on average) were plausible but only moderately relevant to the case. These responses, while plausible, are not perceived as critical for the specific context of the patient.

Another subset (7.08% on average) was considered plausible but with low relevance to the patient's case. This suggests that, although

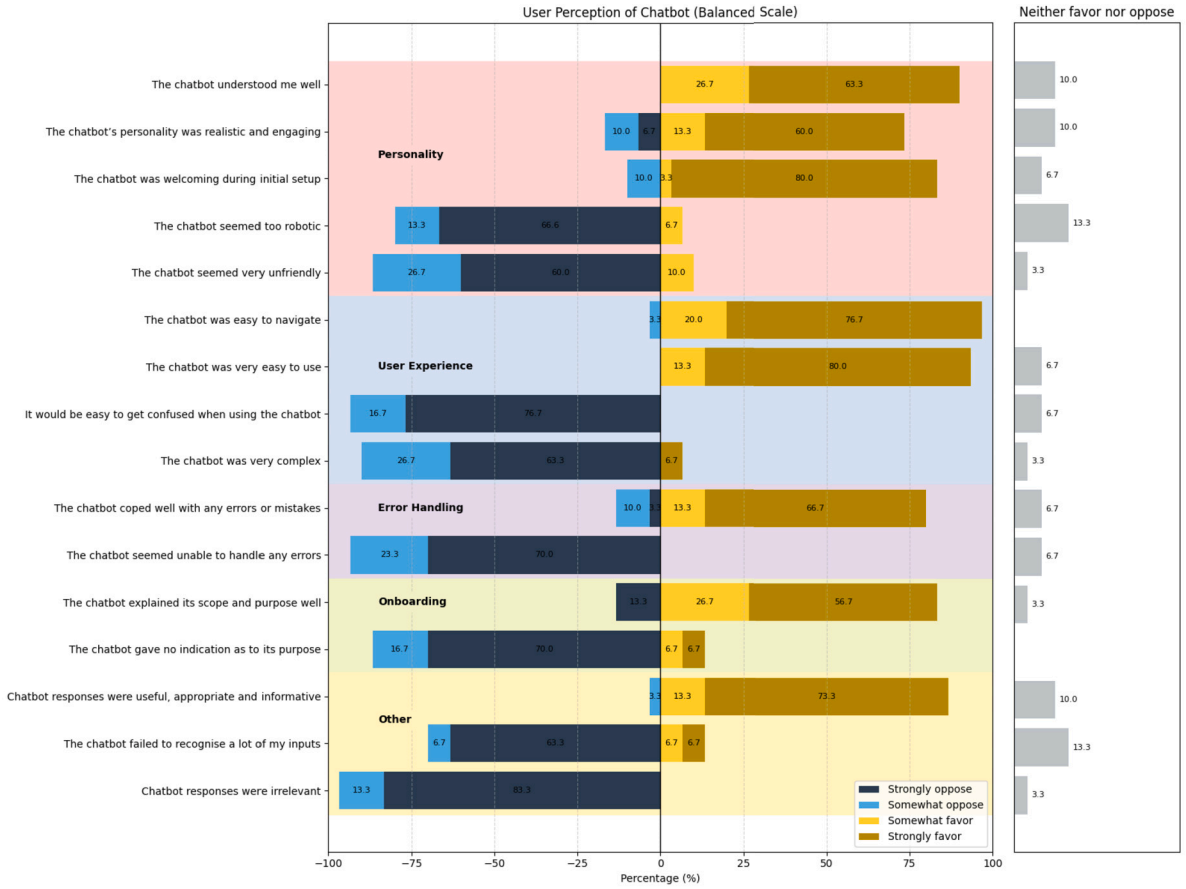

Fig. 10. CUQ results.

the responses have some degree of plausibility, their direct relevance to the patient's case is questionable.

A small percentage (3.15% on average) was considered implausible. These responses do not meet clinical expectations and are considered unlikely in the context of the patient's case. Finally, 1.31% of the responses were judged to be very implausible, appearing at the leftmost side of the *x*-axis. These responses are perceived as highly improbable and not aligned with the patient's case.

Although the figure and percentages capture distinct dimensions—coverage versus plausibility—the distribution helps illustrate how frequently the virtual patient responses align with expert clinical reasoning.

4.3. Correlation between plausibility and adherence to the case description

While Fig. 9 outlines how plausible the responses are, Fig. 8b ties this back to whether the agent relied on explicit, implicit, or fictional information. The sankey plot categorizes the five specialist responses by their information source—explicit (2,390 responses, 51.6%), implicit (1,675 responses, 36.2%), and fictitious (565 responses, 12.2%)—and by the five levels of response plausibility. In each source category, the vast majority of responses are rated as "plausible, highly specific for case" (1,811 for explicit, 1,269 for implicit, and 428 for fictitious), representing roughly 75.8% of responses across all groups. The remaining responses are distributed among lower plausibility ratings—with very small fractions (approximately 1–3%) classified as "very implausible" or "rather implausible," and moderate shares falling into "plausible, low relevance" or "plausible, relevant for case."

This uniform distribution, as evidenced in the figure, suggests that the response generation mechanism consistently produces highly detailed and case-specific outputs regardless of whether the underlying information is derived directly from the case, inferred implicitly, or partially fabricated.

Although a large portion of the chatbot's fictional or inferred responses remained clinically coherent, these results suggest a clear advantage when the chatbot grounds its output in explicitly stated facts. Strengthening the model's capacity to identify information gaps—and either acknowledge uncertainty or refrain from offering unverified details—could help reduce the proportion of implausible or tangential answers.

4.4. Chatbot usability questionnaire

Fig. 10 presents the results of the CUQ, which was administered to our thirty participants. This questionnaire encompasses 16 items divided into five categories—agent's personality, user experience, error handling, onboarding, and additional statements—resulting in a score out of 100 that indicates overall user satisfaction.

In terms of the *agent's personality*, most participants held a favorable view of the chatbot. Specifically, 26.7% agreed and 63.3% strongly agreed that it "understood them well," while 80% strongly agreed that it was "welcoming during initial setup." Opinions were more varied regarding whether the chatbot's personality felt realistic: 13.3% agreed, 60% strongly agreed, 10% disagreed, 6.7% strongly disagreed, and 6.7% neither agreed nor disagreed. Responses concerning whether the chatbot was "robotic" or "unfriendly" were also generally positive. While 66.6% strongly disagreed and 13.3% disagreed that it was robotic, 6.7% agreed with the statement. Regarding the perception of unfriendliness, 10% agreed, whereas the majority—60% and 26.7%—strongly disagreed or disagreed, respectively.

Moving on to *user experience*, the chatbot received notably positive feedback. A total of 96.7% of participants agreed or strongly agreed that

it was “easy to navigate.” Similarly, 93.3% stated that it was “easy to use.” None reported feeling confused while interacting with it; 76.7% strongly disagreed and 16.7% disagreed with the statement that it was confusing. Likewise, 63.3% strongly disagreed and 26.7% disagreed that the chatbot was “overly complex,” indicating that users found the interface and features to be straightforward.

In the category of *error handling*, 13.3% agreed and 66.7% strongly agreed that the chatbot handled errors or mistakes well, although 10% disagreed and 3.3% strongly disagreed. Moreover, 70% strongly disagreed and 23.3% disagreed with the statement that the chatbot was incapable of coping with spelling or grammatical errors, reflecting its adequate resilience to common user input issues.

Regarding *onboarding*, 26.7% agreed and 56.7% strongly agreed that the chatbot effectively explained its scope and purpose, and most of the participants rejected the notion that it failed to provide sufficient guidance about its functionality. This indicates that users felt well-informed at the start of their interaction.

The *additional statements* assessed how users perceived the chatbot’s responses overall. While 13.3% agreed and 73.3% strongly agreed that the chatbot provided “useful, appropriate, and informative” answers, a small portion (3.3% disagreed and 10% neither agreed nor disagreed) found them less satisfactory. Notably, no participants regarded the chatbot’s output as irrelevant, suggesting that, overall, the responses were considered pertinent to their queries. The chatbot also appeared to recognize user input satisfactorily in most cases, with only 6.7% agreeing and 6.7% strongly agreeing that it failed to recognize their inputs.

Consolidating these findings, the *CUQ Score* was 86.25 out of 100, reflecting a high degree of user satisfaction. While the chatbot excelled in ease of use, welcoming setup, and broad user-friendliness, slight disagreements about its personality and realism point to potential areas for further refinement.

4.5. Analysis and comparison of results with previous works

The ability to conduct clinical histories is fundamental for any medical professional. In this study, two main aspects were evaluated. First, we examined whether the Generative Agents architecture [25] — adapted for a Virtual Patient (VP) scenario — could emulate human behavior when performing an anamnesis. Second, we assessed how students and professionals perceived and interacted with the tool. Although this architecture has not, to our knowledge, been previously applied to VPs, the evaluation presented here allows a comparison against prior approaches that rely on LLMs like GPT but lack additional agent-based or cognitive mechanisms.

4.5.1. Performance analysis of generative agents architecture as a virtual patient

A key objective was determining the extent to which the chatbot provides responses grounded in the patient’s description, particularly whether the Generative Agents architecture (including its cognitive mechanisms and memory retrieval) fosters more context-aligned answers. As illustrated in Fig. 8a, most questions that were explicitly covered by the patient’s description prompted the chatbot to draw on the corresponding explicit information. This outcome is promising, as it indicates a preference for correct information over fabricated content, an important factor for successful anamnesis. Additionally, for questions that were implicitly covered by the description, the majority of the chatbot’s answers drew either on explicit or implicit knowledge derived from the given details.

However, when questions lay beyond the scope of the description, the system generated mostly fictitious responses. This is not necessarily problematic: if user queries do not fall within the patient’s documented context, the agent is not bound by explicit knowledge and thus may venture into creative reasoning. Nevertheless, such creativity should remain medically coherent and avoid contradicting the patient’s known traits or history.

Beyond verifying the chatbot’s reliance on explicit or implicit information, another critical dimension is whether the resulting answers remain plausible for the patient’s case. Fig. 8b indicates that more than 75% of all responses were regarded as highly plausible and nearly 95% as at least somewhat plausible, regardless of whether the agent’s information source was explicit, implicit, or fictional. Intriguingly, even many of the fabricated responses aligned sufficiently well with the clinical context to be deemed plausible. Yet, there were also instances where responses explicitly referencing the description turned out to be implausible in the current conversational context, this can be attributed cognitive mechanisms of the generative agents architecture, which enable the agent to exhibit behavior that more closely mirrors human-like characteristics.

4.6. Perception and feedback from the participants

Overall, the participants’ perception of the system was positive. Most students praised the concept and enjoyed interacting with the agent, describing it as realistic. We attribute this realism, in large part, to the LLM’s extensive training on curated data, which enables it to produce highly human-like responses. Additionally, the Generative Agents architecture, with its cognitive mechanisms that simulate human reasoning and behavior, further contributes to the authentic feel of the interaction. Nonetheless, participants also suggested areas for improvement.

Overall, the participants’ qualitative feedback regarding the chatbot highlighted a predominantly positive perception of its realism and interactive potential. However, several recommendations were consistently articulated to enhance the educational experience. Many respondents noted performance-related issues, pointing out that the system’s response times were often too slow or that it occasionally froze after a few exchanges. In addition, users suggested improvements to the interface—such as refining the progress-tracking feature to incorporate a complete clinical history, providing clearer instructions at the outset, and even generating a final report that summarizes the quality of the intervention. These comments indicate that, while the concept is well received, there is a clear need to strengthen the system’s responsiveness and usability.

A second set of comments focused on the content and interaction dynamics of the chatbot. Some participants called for more extensive and detailed responses that would foster a richer clinical dialogue; for instance, they recommended broadening the scope of the clinical history to include comprehensive antecedents and a more detailed characterization of symptoms. Conversely, other participating physicians questioned the virtual patient’s in-depth knowledge of their medical history and its sophisticated handling of clinical concepts.

Participants also remarked on the chatbot’s overly robotic voice—recommending either significant improvements to its audio quality or its complete removal—as well as occasional issues with responsiveness to regional language nuances. Moreover, suggestions to incorporate elements that simulate natural, open-ended, and digressive conversation (including non-verbal cues such as facial expressions via camera integration) were offered to better emulate real-world patient interactions. Together, these recommendations underscore the participants’ desire for a more robust, nuanced, and authentic virtual clinical agent that can adapt to both technical and conversational challenges in a medical training context.

4.6.1. Limitations

It is important to highlight that the proposed chatbot overcomes certain constraints found in earlier works, chiefly by managing multiple clinical cases rather than a single scenario. However, it should be acknowledged that a wide variety of clinical cases exist, many of which differ substantially from those currently implemented. Another salient feature is that the chatbot was developed and evaluated in Spanish—yet given that LLMs like GPT exhibit multilingual capabilities, one could

feasibly adapt the same approach to other languages with comparable outcomes.

Albeit the results of the CUQ were rather positive with respect to the plausibility of, including, fabricated answers, LLMs always present an inherent risk of hallucinations, i.e., providing a factually wrong answer due to a lack of specific knowledge. As the chatbot is not a clinical decision support tool, the impact of hallucinations is limited: The creation of misconceptions among students can be avoided by a thorough debriefing with a medical teacher [34].

As any artificial neural network, LLMs can be exposed to inherent biases, depending on their training data. Social or structural inequities with respect to, e.g., ethnicity or gender, can lead to such biases. If generated responses follow such biases, LLMs can amplify the underlying inequities [35], which constitutes an ethical concern regarding medical training. As mentioned above, a possible mitigation strategy could be to include a debriefing with a human teacher. As an alternative, multiple LLMs could be used to find a consensus or analyze each other's response.

A further limitation, noted by the evaluators via the CUQ, involves the chatbot's response latency. In this instance, the delay stems primarily from rendering the avatar video, rather than the LLM's inference process. Additionally, the OpenAI APIs used include content moderation filters designed to block certain categories of material. Should a user query trigger these filters, the chatbot may fail to respond and produce an API call error. Another inherent constraint is that, while the chatbot focuses on the interview component of a patient consultation, it does not encompass a physical examination module. Furthermore, the study elicited only subjective evaluations of the tool's utility, thus leaving open the question of whether it objectively improves students' patient-history-taking skills. Consequently, no definitive claim can be made regarding its efficacy in skill enhancement. Lastly, the relatively small sample size of evaluators may limit the generalizability of these findings, underscoring the need for more extensive future assessments.

5. Conclusions and future work

This study set out to explore the feasibility of adapting generative agent architectures for Virtual Patients (VPs) that assist in clinical history-taking training. Overall, evaluators responded positively, as evidenced by their high satisfaction scores on the CUQ questionnaire. In addition, the agent's performance was assessed by examining whether it adhered to the patient case descriptions and produced contextually appropriate, plausible answers. Against current benchmarks, the chatbot scored 86.25 out of 100 in the CUQ, surpassing existing solutions in usability. It also delivered a greater percentage of plausible responses compared to other state-of-the-art methods.

Despite these achievements, there remains substantial scope for enhancing this tool. We believe that it holds the potential to significantly influence student education. One promising step would be to make the entire system open-source, for example by deploying an open model such as Llama rather than a commercial option like GPT. At the time of this study, Llama 3.2 exhibited performance close to that of leading commercial models, and it supports fine-tuning with clinical history datasets as needed.

Future work should also address the chatbot's response latency, consistently cited as a major drawback by evaluators. This delay arises predominantly from the time taken to generate the avatar video. The Realtime API could help streamline interactions by reducing lag and handling user interruptions automatically, ultimately making conversations feel more natural. Additionally, refining the text-to-speech component with more advanced models may further enhance realism, delivering a more immersive experience that faithfully simulates real-world patient consultations.

It is also important to conduct the evaluation with a larger sample of students, with the possibility of stratifying them to assess whether the results differ or remain consistent. Stratification refers to dividing the results into different demographic groups: students in the first three semesters, students in the final semesters, graduates, those with specialization, and professors with doctoral degrees. This approach allows us to determine if there are discrepancies in perceptions depending on the evaluator's level of expertise. Finally, it is crucial to develop methods for objectively assessing students' progress in clinical history-taking. This would provide concrete evidence of the tool's effectiveness in improving this critical skill.

After completing this work, it is evident that technology and artificial intelligence hold significant potential to positively impact the learning experience of medical students. Based on this study and its results, it is safe to conclude that LLMs, and specifically Generative Agents' architecture, have the potential to support students in developing communication skills—not only in clinical history-taking but also in other areas where these skills are essential.

Statement: During the preparation of this work the author(s) used Grammarly to clean up typos, grammatical mistakes, and misplaced punctuation. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication

CRediT authorship contribution statement

Nicolas Laverde: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Christian Grévisse:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Sandra Jaramillo:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Ruben Manrique:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare no conflicts of interest. They have no known financial or personal relationships that could have appeared to influence the work reported in this paper. Additionally, no funding agencies, organizations, or individuals were involved in the study's design, data collection, analysis, interpretation, or the writing of this manuscript.

References

[1] Van Kempen A. Legal risks of ineffective communication. Virtual Mentor 2007;9(8):555–8. https://doi.org/10.1001/virtualmentor.2007.9.8.hlaw1-0708.
[2] Barratt J. Developing clinical reasoning and effective communication skills in advanced practice. Nurs Stand 2018;34(2):37–44. https://doi.org/10.7748/ns.2018.e11109.
[3] Lazris A, Roth A, Haskell H, James J. Poor physician-patient communication and medical error. Am Fam Phys 2021;103(12):757–9.
[4] Brouwers M, Rasenberg E, van Weel C, Laan R, van Weel-Baumgarten E. Assessing patient-centred communication in teaching: a systematic review of instruments. Med Educ 2017;51(11):1103–17. https://doi.org/10.1111/medu.13375.
[5] Kurtz S, Draper J, Silverman J. Teaching and learning communication skills in medicine. CRC Press; 2017.
[6] Silverman J, Kurtz S, Draper J. Skills for communicating with patients. CRC Press; 2016.
[7] Young LB, O'Toole CR, Wolf B. Communication skills for dental health care providers. Quintessence Publishing Co, Inc.; 2015.
[8] Bandiera G, Sherbino J, Frank JR. The CanMEDS assessment tools handbook: an introductory guide to assessment methods for the CanMEDS competencies. Royal College of Physicians and Surgeons of Canada; 2006.
[9] Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine 2024;103(27):e38813. https://doi.org/10.1097/MD.0000000000038813.
[10] Linder U, Hartmann L, Schatz M, Hetjens S, Pechlivanidou I, Kaden JJ. Employing simulated participants to develop communication skills in medical education: a systematic review. Simul Healthc 2024. https://doi.org/10.1097/SIH.0000000000000841.

[11] Blackmore A, Kasfiki EV, Purva M. Simulation-based education to improve communication skills: a systematic review and identification of current best practice. BMJ Simul Technol Enhanc Learn 2018;4(4):159–64. https://doi.org/10.1136/bmjstel-2017-000220.

[12] Gillette C, Stanton RB, Rockich-Winston N, Rudolph M, Anderson HGJ. Cost-effectiveness of using standardized patients to assess student-pharmacist communication skills. Am J Pharm Educ 2017;81(10):73–9. https://doi.org/10.5688/ajpe6120.

[13] Fang Y, Frampton JP, Raghavan S, Sabahi-Kaviani R, Luker G, Deng CX, et al. Rapid generation of multiplexed cell cocultures using acoustic droplet ejection followed by aqueous two-phase exclusion patterning. Tissue Eng Part C Methods 2012;18(9):647–57. https://doi.org/10.1089/ten.tec.2011.0709.

[14] Maicher KR, Stiff A, Marisa Scholl MW, Fosler-Lussier E, Schuler W, Serai P, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity. Med Teach 2023;45(3):279–85. https://doi.org/10.1080/0142159X.2022.2130216.

[15] Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC Med Educ 2022;22(1):365. https://doi.org/10.1186/s12909-022-03410-x.

[16] Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Med Educ 2009;43(4):303–11. https://doi.org/10.1111/j.1365-2923.2008.03286.x.

[17] Kleinsmith A, Rivera-Gutierrez D, Finney G, Cendan J, Lok B. Understanding empathy training with virtual patients. Comput Hum Behav 2015;52:151–8. https://doi.org/10.1016/j.chb.2015.05.033.

[18] Rickles N, Tieu P, Myers L, Galal S, Chung V. The impact of a standardized patient program on student learning of communication skills. Am J Pharm Educ 2009;73(1):4.

[19] Bearman M, Palermo C, Allen L, Williams B. Learning empathy through simulation: a systematic literature review. Simul Healthc 2015;10(5):308–19. https://doi.org/10.1097/SIH.0000000000000113.

[20] Kava B, Andrade A, Marcovich R, Idress T, Ruiz J. Communication skills assessment using human avatars: piloting a virtual world objective structured clinical examination. Urol Pract 2017;4(1):76–84. https://doi.org/10.1016/j.urpr.2016.01.006.

[21] Stevens A, Hernandez J, Johnsen K, Dickerson R, Raij A, Harrison C, et al. The use of virtual patients to teach medical students history taking and communication skills. Am J Surg 2006;191(6):806–11. https://doi.org/10.1016/j.amjsurg.2006.03.002.

[22] Kron F, Fetters M, Scerbo M, White C, Lypson M, Padilla M, et al. Using a computer simulation for teaching communication skills: a blinded multisite mixed methods randomized controlled trial. Patient Educ Couns 2017;100(4):748–59. https://doi.org/10.1016/j.pec.2016.10.024.

[23] Holderried F, Stegemann-Philipps C, Herschbach L, Moldt J-A, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ 2024;10:e53961. https://doi.org/10.2196/53961.

[24] Huang Y. Levels of AI agents: from rules to large language models. arXiv:2405.06643, 2024.

[25] Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th annual ACM symposium on user interface software and technology, UIST '23. New York, NY, USA: Association for Computing Machinery; 2023. p. 1–22.

[26] Mosquera M, Pinzon JS, Fonseca Y, Rios M, Quijano N, Giraldo LF, et al. Can LLM-augmented autonomous agents cooperate? An evaluation of their cooperative capabilities through melting pot. IEEE Trans Artif Intell 2025;1(01):1–10. https://doi.org/10.1109/TAI.2025.3569192.

[27] Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, et al. The association of standardized patient educators (ASPE) standards of best practice (SOBP). Adv Simul 2017;2(1):10. https://doi.org/10.1186/s41077-017-0043-4.

[28] Mosquera Ortega MA. Cooperative LLM agents. Master's thesis. Universidad de Los Andes; 2024. https://hdl.handle.net/1992/75429.

[29] Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré P-E, et al. The Faiss library. arXiv:2401.08281, 2025.

[30] Chase H. LangChain. https://github.com/langchain-ai/langchain, 2022.

[31] Zhang W, Cun X, Wang X, Zhang Y, Shen X, Guo Y, et al. SadTalker: learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. arXiv:2211.12194, 2023.

[32] Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77–101. https://doi.org/10.1191/1478088706qp063oa.

[33] Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? In: Proceedings of the 31st European conference on cognitive ergonomics, ECCE '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 207–14.

[34] Grévisse C. RasPatient Pi: a low-cost customizable LLM-based virtual standardized patient simulator. In: Florez H, Astudillo H, editors. Applied informatics. Cham: Springer Nature Switzerland; 2025. p. 125–37.

[35] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172–80. https://doi.org/10.1038/s41586-023-06291-2.