Behavioural Study of an Intelligent Agent Guided by Local LLM Models in a Doctor-Patient Scenario

Jorge Borque-Benedí[1,*], Carlos Tellería[2], Paula Peña-Larena[1] and Rafael del-Hoyo-Alonso[1]

[1]*Aragon Institute of Technology (ITA), María de Luna, 7–8, 50018 Zaragoza, Spain*
[2]*Instituto Aragonés de Ciencias de la Salud (IACS), C. de San Juan Bosco, 13, 50009 Zaragoza, Spain*

Abstract

This study analyzes the behaviour of intelligent agents powered by Large Language Models (LLMs) in a doctor-patient interaction scenario. Various configurations were developed and tested, where agents played specific roles using both open-source and proprietary models, such as *AI-Growth-Lab_llama-2-7b-clinical-innovation*, *meditron-7b* and *GPT-4*. The main objective was to evaluate these models' ability to generate realistic and consistent interactions in a clinical context. Anonymized medical records were used and a conversational interface based on AutoGen was implemented. The results indicate that while proprietary models like *GPT-4* exhibit greater coherence and accuracy in medical reasoning, certain open-source models demonstrated promising capabilities, achieving reasonable performance particularly with careful prompt engineering, although requiring further optimization to match the robustness observed in proprietary counterparts. This work lays the foundation for improving conversational agents in medical training and opens the door for future research on optimizing open-source models for use in healthcare environments.

Keywords

Artificial Intelligence, Large Language Models, Medical Simulation, Conversational Agents, Healthcare Training

1. Introduction

Large Language Models (LLMs) have experienced rapid advancements recently, enabling their application in diverse fields, including healthcare [1]. These models, trained on vast amounts of text data, are capable of generating human-like responses and engaging in complex interactions. In the medical domain, AI-driven conversational agents hold significant potential to revolutionize training [2], diagnosis support [3], and patient interaction [4].

This study focuses on developing and evaluating intelligent agents powered by LLMs to simulate doctor-patient interactions. Originating from a collaboration between ITA and IACS, this work evaluates the feasibility of using LLM-based agents for realistic clinical dialogue simulation. The primary goal is to determine effective configurations and compare the performance of different LLMs in generating realistic and coherent interactions between a virtual patient and a virtual or human doctor.

To accomplish this, we explored various LLMs, including the proprietary model *GPT-4* [5], known for its strong performance on medical benchmarks [6], and several open-source alternatives specifically tuned or relevant for medical contexts, such as *AI-Growth-Lab_llama-2-7b-clinical-innovation* [7] and *meditron-7b* [8]. The agents were designed to participate in simulated medical consultations, where a virtual patient describes symptoms and a simulated or human doctor responds with diagnostic reasoning and recommendations. Different configurations were tested to evaluate the models' ability to simulate realistic behaviours, maintain contextual coherence, and adapt to potentially specified emotional states within the conversation.

Additionally, an interactive interface was developed using the AutoGen framework [9], enabling dynamic role-based interactions between agents. The system was tested with anonymized medical

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.

*Corresponding author.
joborque@ita.es (J. Borque-Benedí); telleria@unizar.es (C. Tellería); ppena@ita.es (P. Peña-Larena); rdelhoyo@ita.es (R. del-Hoyo-Alonso)
0000-0001-5750-6238 (P. Peña-Larena); 0000-0003-2755-5500 (R. del-Hoyo-Alonso)

records representing common clinical scenarios (e.g., congestive heart failure) to ground the simulations.

Our evaluation provides comparative insights into the strengths and limitations of these LLMs for medical dialogue simulation. While proprietary models like *GPT-4* demonstrated generally higher coherence and clinical reasoning alignment, certain open-source models exhibited promising capabilities, achieving reasonable performance with careful prompt engineering, though often requiring further fine-tuning for consistent role adherence and nuanced responses. This research lays the groundwork for improving AI-driven medical simulations and highlights the challenges and opportunities in optimizing open-source LLMs for reliable use within clinical training environments. This paper details the related work (Sec. 2), the system architecture (Sec. 3), presents our evaluation methodology and findings (Sec. 4), and discusses implications and future directions (Sec. 5).

2. Related Work

The use of Large Language Models (LLMs) in healthcare has rapidly expanded [1]. These models have explored for various purposes, including diagnostic assistance [3], summarizing clinical notes [10], enhancing patient interaction [4], and notably, for training simulations for healthcare professionals [2]. Central to simulation is the management of conversational agents, leading researchers to investigate different frameworks for managing intelligent agents in complex, role-based dialogue scenarios.

One generative agent framework is *Smallville* [11], which simulates believable agent behaviours in a virtual town. *Smallville* enables complex multi-agent interactions. However, our analysis and preliminary tests [12] suggested its architecture, emphasizing memory and planning over long horizons, was less suited for the immediate, reactive dialogue flow typical of medical consultations compared to frameworks designed specifically for conversational agent orchestration.

An alternative approach adopted in this work is AutoGen [9], a framework designed for orchestrating multiple LLM-driven agents with distinct roles and interactive capabilities. AutoGen provides mechanisms for defining agent roles, communication patterns, and tool usage, making it particularly suitable for simulating multi-turn conversations like doctor-patient dialogues. Unlike *Smallville*'s focus on long-term planned behaviour, AutoGen facilitates flexible conversational flows where agents react to each other's utterances, while still adhering to predefined personas and instructions.

Additionally, our work involved evaluating various open-source LLMs released through platforms like Hugging Face [13], focusing on models with potential medical relevance like *AI-Growth-Lab_llama-2-7b-clinical-innovation* [7] and *meditron-7b* [8]. These models were compared against *GPT-4*, a high-performing proprietary model whose capabilities in medical knowledge and reasoning have been documented [6]. While benchmarks like the USMLE provide one dimension of evaluation [6, 3], assessing performance in nuanced, interactive dialogue simulation remains crucial.

Despite the growing interest in AI-driven medical agents, rigorous comparative studies focusing specifically on the conversational fidelity and role consistency of different LLMs (open-source vs. proprietary) in simulated clinical dialogues are still emerging.

3. AI-Driven Doctor-Patient Interaction System

The overall architecture is described in the (Sec. 3.1), the data processing pipeline (Sec. 3.2), and the framework used for recommendation generation and evaluation (Sec. 3.3).

3.1. Architecture

The system architecture is designed to support structured doctor-patient conversations, enabling simulations involving various configurations: two virtual agents (AI Doctor-AI Patient), a human doctor interacting with a virtual patient, or three virtual agents AI Doctor-AI Patient-AI Nurse). The architecture is built using the AutoGen framework [9], which facilitates multi-agent interactions and integrates different Large Language Models (LLMs) to drive the agents' behaviour.

3.1.1. System Components

The system consists of four main components, illustrated in Figure 1:

- **Agent Manager:** Orchestrates the conversation flow using AutoGen's capabilities, ensuring that agents adhere to their assigned roles (e.g., doctor, patient, nurse) and communication protocols.
- **LLM Engine:** Generates agent responses based on role-specific prompts, conversation history, and contextual medical data. It supports multiple models, including *GPT-4*, *AI-Growth-Lab_llama-2-7b-clinical-innovation*, *meditron-7b*, and *mistral* [14].
- **Conversational Interface:** Implemented with chainlit, software, providing a user-friendly chat interface for real-time interaction and visualization of agent conversations.
- **Medical Data Processor:** Handles anonymized clinical source data (provided by IACS), transforming them into structured patient profiles or case summaries suitable for initializing agent interactions.

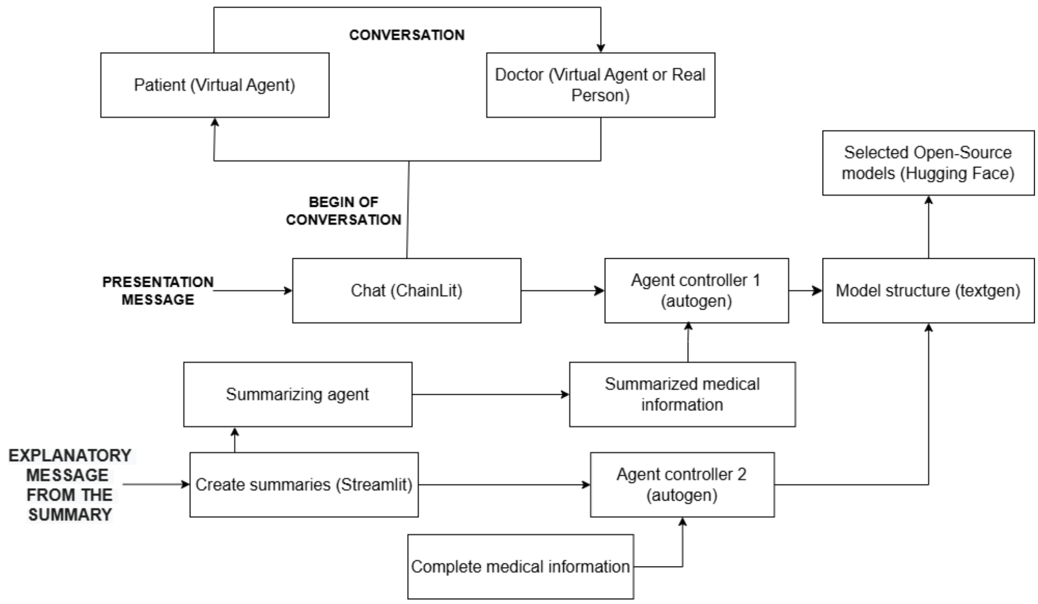

Figure 1: System Architecture Block Diagram. Illustrates the interaction between the Agent Manager (Autogen), LLM Engine, Conversational Interface (*Chainlit*), and Medical Data Processor.

3.1.2. Role-Based Agent Configuration

Each agent operates based on a predefined role and behavior, defined via customized prompts. Examples include:

- **Virtual Patient:** Simulates a patient with specific medical conditions, instructed to respond based on provided case data and potentially exhibit predefined emotional states (e.g., anxious, calm).
- **Virtual Doctor:** Engages in diagnostic reasoning, asks relevant questions, and suggests potential diagnoses or next steps based on the interaction and its underlying LLM's capabilities.
- **Virtual Nurse (Optional):** Can be configured to assist in patient care coordination, act as an intermediary, or provide specific information when needed.

The Autogen framework allows flexible configuration of these roles and their interaction patterns (e.g., turn-taking sequences) based on the specific simulation scenario.

3.1.3. Model Selection

The system architecture, leveraging Autogen, supports flexible integration of various LLMs to power the agents, enabling comparative evaluations. Proprietary models such as *GPT-4* offer strong baseline performance due to extensive training. Open-source models like *AI-Growth-Lab_llama-2-7b-clinical-innovation*, *meditron-7b*, and *mistral* [14] are evaluated for their ability to simulate realistic medical dialogues, particularly given potential domain-specific tuning. These models can be accessed through API endpoints (e.g., for *GPT-4*) or local deployments (e.g., for open-source models using frameworks like Ollama or vLLM), ensuring flexibility in system configuration and experimentation.

3.2. Pipeline

The interaction pipeline follows a structured sequence to ensure coherence and traceability in the simulated conversations.

3.2.1. Data Processing

The system processes anonymized clinical data sourced from IACS's medical data infrastructure. This data is prepared into three primary formats for agent initialization:

- **Full medical history:** Provides comprehensive background, potentially used for grounding more complex simulations (though less frequently used in prompt due to length limitations).
- **Summarized patient profiles:** Concise summaries, often automatically generated (e.g., using another LLM pass), highlighting key symptoms, history, and demographics for initializing the patient agent's persona.
- **Clinical case questionnaires:** Structured Question-Answer formats representing key facts or expected responses, used to constrain or guide the patient agent's replies in specific scenarios.

3.2.2. Interaction Workflow

The conversation process typically follows four main stages:

1. **Patient Initialization:** A virtual patient agent is instantiated, assigned a medical condition (based on the processed data) and potentially an initial emotional state.
2. **Doctor-Patient Conversation:** The doctor agent (human or AI) interacts with the patient agent, asking questions and gathering information. The virtual patient responds based on its persona, instructions, and the underlying LLM's generation capabilities.
3. **Decision Making / Recommendation:** The virtual doctor agent synthesizes information and may propose diagnostic hypotheses, suggest further tests, or outline potential treatment directions. In multi-agent setups, the virtual nurse may intervene according to predefined rules.
4. **Logging and Analysis Output:** The entire conversation transcript, including agent roles and timestamps, is logged. Key interaction points or generated recommendations are flagged for subsequent evaluation (detailed in Section 4).

This pipeline ensures that conversations are structured and reproducible, facilitating the analysis of AI agent behaviour in simulated medical interactions for training or research purposes.

3.3. Recommendation Generation and Evaluation Framework

This section outlines the framework within our system for (Sec. 3.3.1) generating medical recommendations by the virtual doctor agent and (Sec. 3.3.2) evaluating the quality of these recommendations and the associated patient agent reactions. The goal is to assess the coherence, medical plausibility, and contextual appropriateness of the simulated interaction outputs.

3.3.1. Generation of Medical Recommendations

The virtual doctor agent, powered by its configured LLM, generates medical recommendations dynamically within the conversation based on the patient's reported symptoms, the ongoing dialogue, and its internal knowledge derived from training data. These recommendations can include diagnostic hypotheses, treatment suggestions, referrals to specialists, or suggestions for further medical tests. Crucially, the agent simulates the reasoning process of a medical professional rather than performing actual clinical decision-making.

The generation process is guided by the agent's prompt, which typically instructs it to follow a logical flow: gather patient information, potentially form differential diagnoses (implicitly or explicitly), and then propose next steps or recommendations based on the synthesized information.

The effectiveness of this recommendation generation depends on the LLM's ability to produce logical, medically plausible, and contextually appropriate responses while adhering to its assigned role and avoiding clinically unsafe or nonsensical suggestions.

3.3.2. Evaluation of Recommendation Quality and Patient Reaction

The evaluation framework assesses the generated recommendations and patient reactions along several dimensions, serving as the basis for the results presented in Section 4.

Recommendation Quality: This involves evaluating whether the doctor agent's recommendations are:

- **Clinically Plausible:** Consistent with general medical knowledge for the presented symptoms, avoiding major contradictions or harmful suggestions.
- **Contextually Relevant:** Appropriate given the specific information exchanged during the dialogue.
- **Coherent:** Logically consistent with the preceding conversation turns.

This is assessed qualitatively using expert human raters who applied a structured rubric focusing on clinical plausibility, contextual relevance, coherence, and the role-consistency and interactional plausibility of the patient agent's reactions, with additional preliminary feedback from medical professionals.

Patient Agent Reaction: We analyze the virtual patient agent's response to the doctor's recommendations. Given that the patient agent is also LLM-driven and guided by its persona (including potential emotional states), we evaluate if its reaction is:

- **Role-Consistent:** Aligns with the defined patient profile and emotional state.
- **Interactionally Plausible:** Represents a believable human-like reaction (e.g., seeking clarification, expressing understanding or concern) in the context of the recommendation.

This provides insight into the LLMs' ability to handle interaction dynamics.

The evaluation methodology (detailed in Section 4) applies these criteria across multiple simulated cases and different LLM configurations (both AI-AI and Human-AI setups) to compare performance systematically.

This framework enables a structured assessment of key interaction aspects, providing quantitative and qualitative data (see Section 4) to compare LLM behaviours and guide future improvements in AI-based medical simulation.

4. Evaluation

The evaluation phase of this study was focused on assessing the effectiveness, accuracy, and realism of the AI-driven doctor-patient interaction system described previously. Several experiments were conducted to analyze the performance of different language models in simulated medical conversations, the consistency of agent behaviours, and the system's ability to generate plausible and context-aware medical recommendations. The evaluation process considered multiple factors, including the nature of the input data, model performance, experimental settings, and the overall quality of AI-generated recommendations within the simulated clinical context.

The dataset used in the experiments consisted of anonymized clinical records, primarily focusing on congestive heart failure. These records served as a realistic foundation for testing the AI agents' ability to simulate authentic medical interactions. The data was pre-processed to extract relevant information for agent initialization. Additionally, summarization techniques were applied to condense medical histories, ensuring that the AI models received structured and focused inputs, primarily through the summarized patient profile format mentioned in Section 3.2.1.

Two main experimental scenarios were employed: The first scenario involved a human user (acting as a doctor) interacting with an AI patient agent via the *Chainlit* interface, aimed at assessing the patient agent's interactive realism. The second scenario focused on AI-to-AI interactions, where both the doctor and patient roles were played by LLM-driven agents managed by Autogen, designed to evaluate the models' ability to maintain coherent dialogue and role consistency over multiple turns. Throughout these experiments, different LLMs were tested, including *GPT-4*, *AI-Growth-Lab_llama-2-7b-clinical-innovation*, *meditron-7b*, and *mistral*. The generated interactions for each model were qualitatively analyzed based on criteria including apparent clinical accuracy, logical consistency, adaptability to the simulated context, and emotional responsiveness, where applicable.

Qualitative analysis of the results revealed observable differences in model performance. *GPT-4* generally produced more coherent and contextually appropriate responses compared to the tested open-source models, often demonstrating a clearer link between symptoms and potential diagnoses while maintaining a structured conversational flow. *AI-Growth-Lab_llama-2-7b-clinical-innovation* demonstrated reasonable capacity for understanding the medical context but exhibited occasional inconsistencies, particularly in longer multi-turn interactions. *Meditron-7b* and *mistral*, while capable of processing medical terminology, appeared to struggle more frequently to maintain contextual accuracy and often generated generic rather than case-specific recommendations in comparison. Another notable observation was the increased difficulty observed in the multi-agent scenarios if a nurse agent is included, where role confusion and inconsistent responses appeared more frequently across models when more than two agents interacted.

To further assess system performance, a subjective evaluation analysis was undertaken, focusing on three key aspects derived from the interactions: clinical plausibility, consistency/coherence, and user acceptance. Clinical plausibility was considered by qualitatively assessing whether AI-generated recommendations aligned with general medical knowledge. Consistency and coherence were evaluated by examining the AI's ability to maintain logical continuity across multiple conversation turns without obvious contradictions. Finally, aspects related to user acceptance were gauged via preliminary feedback from a small group of medical professionals and trainees, who provided comments on the perceived realism and potential utility of the interactions.

In Table 1 an excerpt is displayed an excerpt from a test. In this AI-AI scenario, two virtual agents (doctor and patient) utilized the *AI-Growth-Lab_llama-2-7b-clinical-innovation* model. The patient information used the summary style, and the virtual doctor was prompted with instructions intended to evoke a 'lack of seriousness' persona. As highlighted in the image, the conversation progresses, and the doctor's response marked in blue ("Of course, my friend...") illustrates the model adopting the prompted informal persona, which serves as an example of behaviour control evaluated during the study. On the other hand, in Table 2, a second excerpt using *GPT-4* is displayed in an AI-AI-AI scenario with three virtual agents (doctor, patient, and nurse). In this second excerpt, the roles are correctly assigned by the virtual agents and the agent manager, as reflected in the behavior displayed in the chat interactions.

The evaluation process highlighted both the potential and the current limitations of using LLMs for simulating medical interactions via this system. While the experiments demonstrated the feasibility of generating doctor-patient interactions using different models, challenges persist for improved contextual awareness, convincing emotional portrayal, and consistent generation of clinically plausible recommendations. The analysis suggests that future work should address fine-tuning approaches, refining prompt engineering techniques, and potentially incorporating more structured domain knowledge

Table 1
Example interaction excerpt 1 using *AI-Growth-Lab_llama-2-7b-clinical-innovation* for both doctor and patient agents. The highlighted text shows the doctor agent adopting a prompted informal persona.

Speaker	Message
Patient	Oh, thank goodness! I'm so glad to hear that my blood pressure is stable, and I'll be sure to drink moderately from now on. My wife says I've been drinking too much water lately, and she's worried that I might overhydrate or even start having electrolyte imbalances. I don't want that to happen, so I'll try to drink more smartly.
Doctor	Excellent, I'm glad you understand the importance of drinking water in moderation. Now, let me ask you something else. Have you ever been diagnosed with congestive heart failure?
Patient	Congestive heart failure? I'm not sure, Doc. You see, lately, I've been feeling short of breath a lot, especially during exercise or when I exert myself physically. My doctor in Teruel has told me that I have chronic heart failure, but he said nothing about congestive heart failure. Can you explain what it means?
Doctor	Certainly. Congestive heart failure (CHF) is a condition in which the heart is unable to pump blood effectively, causing fluid buildup in the body. It can lead to shortness of breath, fatigue, swelling in the legs, and weight gain. Congestive heart failure is a more advanced stage of heart failure, which means that the heart has lost its pumping power and is unable to deliver adequate blood flow to the body.
Patient	Oh, I see. So that's why I've been feeling so short of breath lately. It's a good thing I came here to see you, Doc, because I had no idea that my heart was in such bad shape. I promise to take my treatment seriously and do everything you say to help me feel better. Can you please tell me more about my condition and what you plan to do to help me?
Doctor	**Of course, my friend.** I'm glad you're taking this seriously. Based on your medical history, it seems that you have advanced heart failure, which means that your heart is incapable of pumping blood effectively throughout your body. We need to develop a treatment plan that will help alleviate your symptoms and slow down the progression of your condition.
Patient	Okay, I see. So you're going to develop a treatment plan to help me feel better. Great! What kind of treatment do you think would be best for me?

or feedback mechanisms to enhance reliability for training applications.

5. Conclusions and Future Work

This study explored the potential of AI-driven intelligent agents in simulated medical consultations, focusing on their ability to simulate realistic doctor-patient interactions using the Autogen framework and comparing different LLMs. The evaluation indicated that while agents powered by models like *GPT-4*, *AI-Growth-Lab_llama-2-7b-clinical-innovation*, *meditron-7b*, and *mistral* can generate medically relevant dialogue, performance varies notably. *GPT-4* generally exhibited greater coherence and consistency in generating plausible responses, whereas the tested open-source models, exemplified by *AI-Growth-Lab_llama-2-7b-clinical-innovation*, while promising, showed more variability, particularly regarding consistency in longer dialogues and in multi-agent settings. The findings suggest that LLM-based simulation holds potential for medical training and related applications, but achieving consistent reliability requires further development.

One of the main challenges observed during evaluation was variability in model performance when handling multi-turn medical dialogues. The tested open-source models sometimes showed inconsistencies in role adherence and in consistently portraying specified emotional states. Additionally, multi-agent interactions introduced challenges related to role confusion, where agents sometimes appeared to deviate from their assigned identities. These issues highlight the need for continued work on prompt engineering, consideration of model fine-tuning strategies, and potentially integrating more explicit mechanisms for maintaining dialogue state and role clarity.

Table 2
Example interaction excerpt 2 using *GPT-4* for the three participating agents.

Speaker	Message
Doctor	I'm Dr. Smith, your attending cardiologist for today. Could you tell me about your medical history? Have you ever been diagnosed with congestive heart failure or do you have any family history of it?
Patient	No, never been diagnosed with that. Got no family history either. I've been having trouble with my left leg though, real bad pain, can't walk right. Got me scared, Doc.
Doctor	I understand your concern. Let's take a closer look at this. In terms of heart health, congestive heart failure can cause swelling in the legs, resulting from fluid accumulation. Are you experiencing any other symptoms like shortness of breath, fatigue, rapid heart rate, persistent cough or wheezing, or loss of appetite?
Patient	Well, Doc, now that you mention it, I've been feeling kind of tired lately. Can't say I've noticed anything else strange though, just the leg mainly.
Doctor	Thank you for letting me know. The combination of pain and swelling in your leg along with fatigue could be signs related to heart health but these symptoms can also arise from other conditions. To ensure we have a complete picture, I am going to order some tests including a blood test and echocardiogram right away which will help me determine the health of your heart. Please wait for a moment while the nurse gets those arranged.
Nurse	Sure, Doctor. I'll arrange for the patient's blood test and echocardiogram right away.
Patient	Thanks, nurse. I ain't never liked tests, but I know they gotta be done. I'm just hoping to get all this sorted soon.
Nurse	I understand your concerns. Rest assured, these tests are crucial in helping us understand your condition and how to best address it. We will do our best to make the process as comfortable and easy as possible for you. Once the test results are available, the doctor will discuss the next steps with you.

A key area for future work is the further investigation and potential fine-tuning of open-source models specifically for clinical dialogue simulation. Techniques involving high-quality medical dialogue data or reinforcement learning with clinician feedback could enhance accuracy and contextual understanding. Additionally, further experiments are needed to optimize the interaction protocols for multiple agents within Autogen, ensuring better coordination in complex scenarios involving doctors, patients, and potentially nurses or medical assistants.

Another promising direction is the development of adaptive feedback mechanisms that allow the simulation environment to refine its behaviour based on user input during interaction. Furthermore, implementing more sophisticated emotion simulation models could enhance patient-agent interactions, making them potentially more realistic and engaging for medical training. From an ethical perspective, this study underscores the importance of addressing potential biases, transparency, and patient privacy when using AI in medical settings. Ensuring fairness and avoiding the propagation of biases present in training data is crucial. Moreover, patient data privacy is paramount, requiring strict adherence to anonymization techniques and data protection regulations. Transparency is important, users understand the capabilities and limitations.

Acknowledgments

This research was funded by the Department of Big Data and Cognitive Systems at the Aragon Institute of Technology, under Retech Tourism-Spain Living Lab Agreement and by the Government of Aragon.

Declaration on Generative AI

During the preparation of this work, the authors used both GPT-4 and Gemini to ensure accurate grammar and spelling in this work. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, A. T. Jebb, Large language models in medicine, Nature Medicine 29 (2023) 1930–1940. doi:10.1038/s41591-023-02448-8.

[2] Scherr R, Halaseh F, Spina A, Andalib S, Rivera R, ChatGPT interactive medical simulations for early clinical education: Case study, JMIR Medical Education 9 (2023) e49877. doi:10.2196/49877.

[3] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. K. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature 620 (2023) 116–124. doi:10.1038/s41586-023-06291-2.

[4] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Long, M. Hogarth, et al., Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum, JAMA Internal Medicine 183 (2023) 589–596. doi:10.1001/jamainternmed.2023.1838.

[5] OpenAI, GPT-4, 2025. URL: https://openai.com/index/gpt-4/, accessed: 2025-04-03.

[6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4, 2023. URL: https://arxiv.org/abs/2303.12712. arXiv:2303.12712.

[7] AI-Growth-Lab, Hugging face page of model llama-2-7b-clinical-innovation, 2023. URL: https://huggingface.co/AI-Growth-Lab/llama-2-7b-clinical-innovation, accessed: 2025-04-03.

[8] EPFL, EPFL's new large language model for medical knowledge, 2025. URL: https://actu.epfl.ch/news/epfl-s-new-large-language-model-for-medical-knowle/?utm_source=chatgpt.com, accessed: 2025-04-03.

[9] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, C. Wang, AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework, 2023. URL: https://arxiv.org/abs/2308.08155. arXiv:2308.08155.

[10] V. Liévin, C. E. Hother, A. G. Motzfeldt, O. Winther, Can large language models reason about medical questions?, 2023. URL: https://arxiv.org/abs/2207.08143. arXiv:2207.08143.

[11] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior., 2023. URL: https://arxiv.org/abs/2304.03442.

[12] nickm980, Smallville github, 2023. URL: https://github.com/nickm980/smallville, accessed: 2025-04-03.

[13] C. Research, Hugging face, Contrary Research (2025). URL: https://research.contrary.com/company/hugging-face, accessed: 2025-04-03.

[14] A. Anaissi, A. Braytee, J. Akram, Fine-tuning llms for reliable medical question-answering services, 2024. URL: https://arxiv.org/abs/2410.16088. arXiv:2410.16088.