

Assessment of LLM-Based Simulated Virtual Patients in Clinical Simulations of Conversations

Sofia Martinelli^{1,2,3,*†}, Daniel Eduardo Taborda^{1,†}, Guido Ivetta^{1,2} and Luciana Benotti^{1,2,3}

¹Universidad Nacional de Córdoba, Argentina

²Fundación Vía Libre, Argentina

³CONICET, Argentina

Abstract

Large Language Models (LLMs) are increasingly employed as simulated virtual patients (SVPs) to train clinical reasoning and communication skills [1]. However, their deployment in diverse clinical settings raises critical questions regarding narrative fidelity, sociocultural robustness, and potential biases in patient representation. This paper presents a mixed-methods pilot study conducted at the Universidad Nacional de Córdoba (Argentina), evaluating the behavior of LLMs in interdisciplinary clinical simulations with 16 healthcare professionals and advanced medical students. We analyze 518 conversational turns across two distinct clinical scenarios: a young adult with a linear pathology and an older adult with complex comorbidities. Our results suggest that while the models maintained fluency in simple scenarios and apparent cultural adaptation to the region, they struggled to sustain clinical realism in complex cases involving older adults and comorbidities, exhibiting excessive narrative organization and limited emotional expressiveness. A recurring information leakage bias was also observed, where models prematurely disclosed diagnostic data, bypassing the clinical reasoning process. While preliminary and based on a limited sample, these findings raise questions about the plug-and-play readiness of current foundation models for medical training and highlight the urgent need for ethical-by-design frameworks that prioritize clinical fidelity and sociocultural nuance over mere conversational fluency.

Keywords

Large Language Models, Simulated Virtual Patients, Medical Education, Clinical Simulation.

1. Introduction

Anamnesis constitutes one of the fundamental pillars of clinical reasoning and represents one of the core competencies in medical training. Various studies have shown that a proper clinical interview can guide a large part of diagnoses even before complementary studies are performed [2, 3]. Traditionally, skills related to clinical communication and the medical interview are trained through interaction with real patients, simulated patients, or standardized patients. Although these strategies have proven useful for improving communication, empathy, and clinical reasoning, they also present significant limitations related to operational costs, the availability of trained actors, and difficulties in scalability [4, 5, 6].

Recent advances in artificial intelligence (AI), particularly LLMs, have opened new possibilities in clinical simulation and medical education. LLM-based conversational agents can act as simulated virtual patients (SVPs) capable of sustaining interactions through natural language, potentially combining the flexibility of simulated patients with the reproducibility of standardized patients [7, 8, 9].

However, important questions remain regarding the clinical consistency, narrative realism, educational transparency, and sociocultural adaptation of these systems. These limitations acquire special relevance in Latin American contexts, where culturally adapted conversational systems remain under-represented [10]. Furthermore, recent studies suggest that LLM performance could vary depending on the clinical complexity and narrative structure of the simulated scenario [7, 11, 12].

In this context, an interdisciplinary workshop was developed using AI-based SVPs to explore their behavior during clinical interviews conducted by physicians and advanced medical students.

First Workshop on Ethics in Clinical AI Applications (ETHICAIA), co-located with the 35th International Joint Conference on Artificial Intelligence (IJCAI-ECAI 2026), Bremen, Germany

*Corresponding author.

†These authors contributed equally.



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Objective

The primary objective of this study is to evaluate the behavior of large language models in the role of simulated virtual patients during interdisciplinary clinical interactions. Specifically, this work analyzes their communication fidelity and potential educational utility across the departments of Internal Medicine, Mental Health, and Family Medicine, as well as for advanced medical students.

3. Methodology

In this section, we first describe the design and technical implementation of the simulated virtual patients (SVPs) and the interaction platform. We then present the methodological design of the pilot study employing SVPs in interdisciplinary clinical simulations.

3.1. Simulated Patients and Chat Interface

The data collection platform was developed using Gradio [13], an open-source Python framework for building interactive machine learning web applications. Prior to initiating the simulation, the interface collected basic demographic and professional data from the users, including name, age, gender, and their designated participant area (Mandatory Final Practice, Internal Medicine, Family Medicine, or Mental Health). The user interface was presented entirely in Spanish, the native language of the participants.

The conversational engine driving the simulation was GPT-4.1 [14], representing the state-of-the-art LLM at the time the workshop was conducted. This model powered two distinct SVPs, whose behaviors were governed by system prompts designed by medical professionals from the Hospital Nacional de Clínicas. These prompts defined the SVPs' core identities, linguistic styles, current psychophysical states, and internal knowledge bases. The profiles were designed to integrate both their clinical presentation and a specific emotional state to reflect varying levels of clinical structure and diagnostic uncertainty. Furthermore, both simulated patients were contextualized as residents of the province of Córdoba (Argentina) to evaluate the model's capacity for regional and sociocultural adaptation.

The two simulated clinical cases were:

- **Patient A (Javier):** A 35-year-old day laborer with no known prior medical history, attending an outpatient clinic. He presents with several months of extreme fatigue, involuntary weight loss, night sweats, recurring low-grade fever, and enlarged lymph nodes in the neck and armpits. Emotionally, he is downcast, with social withdrawal, sleep difficulties, and increasing alcohol consumption. He does not volunteer sensitive information unless directly asked by the clinician.
- **Patient B (José):** An 84-year-old retired worker arriving at the emergency department with two days of shortness of breath at rest and swollen legs. He has multiple chronic conditions: insulin-dependent diabetes, a previous heart attack with bypass surgery and reduced cardiac function, smoking-related lung disease, and chronic kidney disease. Emotionally, he feels like a burden to his family and fears another hospitalization. Similarly, he does not offer relevant clinical material—an electrocardiogram and a chest X-ray—unless explicitly requested.

To enhance clinical realism, the platform supported multimodal interactions. Specifically, Patient B's (José) prompt included functional links to actual diagnostic images. The model was instructed to provide an electrocardiogram (ECG) and a chest X-ray in the chat if the interviewer explicitly inquired about previous tests or studies. Conversely, Patient A (Javier) was programmed to state that he had no prior medical studies to share.

The chat interface was designed to be straightforward and familiar, mimicking standard clinical messaging applications (see Figure 1 for a screenshot of the tool). Beyond facilitating the consultation, the platform featured an integrated, real-time evaluation system. Users could assess any individual AI-generated message by selecting a *thumbs-up* or *thumbs-down* icon, which subsequently opened a mandatory free-text field for qualitative justification of their feedback. To ensure sufficient evaluative

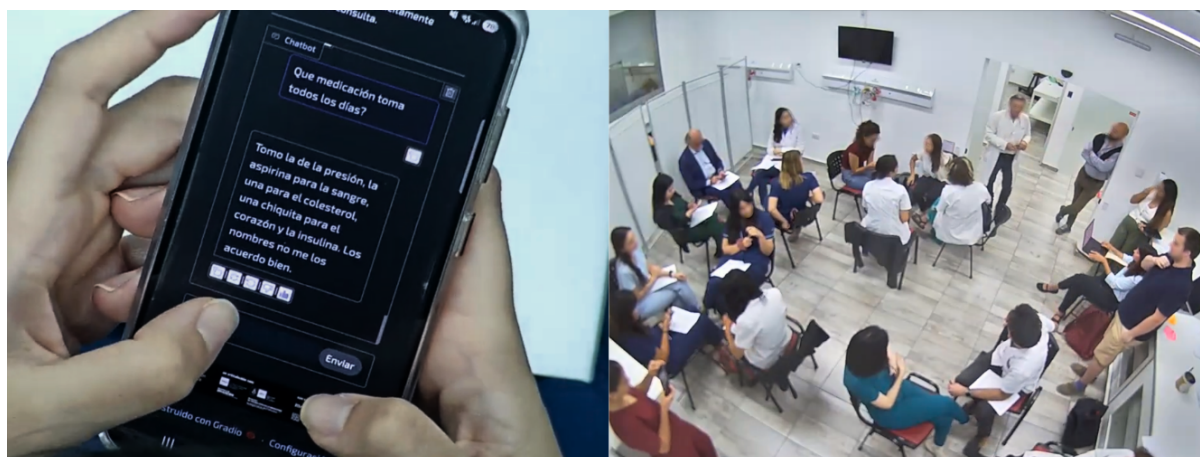


Figure 1: Pilot activity at the Simulation Hospital. Left: A smartphone displaying the chatbot interface in Spanish, where the participant queries the simulated patient about daily medications. Right: Participants during the closing debriefing session.

data, participants were instructed during the prebriefing phase to review and provide feedback on a minimum of five AI messages throughout their individual interaction.

3.2. Pilot Activity

The activity took place in October 2025 at the Simulation Hospital of the Hospital Nacional de Clínicas, Facultad de Ciencias Médicas, Universidad Nacional de Córdoba, Argentina. A total of 16 participants took part, including 4 advanced medical students in their Mandatory Final Practice (PFO, for its Spanish acronym) and 12 physicians from three distinct clinical specialties: Internal Medicine (4), Family Medicine (4), and Mental Health (4). These disciplines were selected as they represent clinical specialties (non-surgical), each employing a fundamentally different approach to patient interviewing and clinical reasoning. Two clinical scenarios representing common medical situations were designed using an LLM: *Patient A*, a 35-year-old adult presenting with lymphadenopathy, and *Patient B*, an older adult presenting with progressive dyspnea. The 90-minute workshop was conducted in Spanish and structured into four sequential phases: prebriefing, individual interaction, peer cross-validation, and debriefing.

Prebriefing. Participants received methodological guidelines outlining the clinical, mental health, and linguistic evaluation perspectives. The study was approved by the Institutional Health Research Ethics Committee (CIEIS, for its Spanish acronym) of the Hospital Nacional de Clínicas (Res. No. 122-06102025). All participants signed an informed consent form authorizing the academic use of data and video recordings within a confidential framework.

Individual Interaction. Participants conducted a 20-minute individual written chat interview with their randomly assigned simulated virtual patient (A or B). During the interaction, they were required to record a minimum of five real-time feedback entries (positive or negative) regarding the SVP's conversational behavior.

Peer Cross-Validation. Full conversation transcripts were printed. Participants who interviewed Patient A evaluated the transcripts of colleagues who had interacted with Patient B, and vice versa, providing an external peer review.

Debriefing. A final interdisciplinary group discussion was led by a moderator. Prior to this, participants anonymously logged two overall strengths and two weaknesses regarding the general behavior and performance of the SVPs.

4. Results

The results include both quantitative and qualitative data. Quantitative variables include the total number of participant interventions and conversational turns, as well as the frequency distribution of positive and negative feedback across medical specialties and clinical scenarios. A qualitative analysis was also conducted through the initial observations and peer validations.

Regarding participant engagement, a total of 16 participants took part in the study: 12 physicians from three different clinical specialties and 4 advanced medical students. During the workshop, a total of 259 interventions by the participants and 259 corresponding responses by SVPs were registered. Internal Medicine presented the highest total number of interventions (77), followed by PFO (75), Family Medicine (58), and Mental Health (49).

Additionally, 76 feedback observations were collected, of which 49 were positive and 27 were negative. This number was lower than the expected minimum because seven participants did not complete the requested five feedback entries during the interaction phase. All disciplines showed a predominance of positive observations. Mental Health presented the highest relative proportion of favorable observations, while Internal Medicine recorded the highest absolute number of positive observations.

However, relevant differences were observed between both clinical scenarios. Patient A (young adult with lymphadenopathy) received 34 positive and 11 negative observations, while Patient B (older adult with progressive dyspnea) obtained 15 positive and 16 negative observations.

Regarding the qualitative findings, a thematic analysis of the written observations and peer validations mapped the feedback into four core dimensions: realism, emotional expressiveness, narrative structure, and clinical contextualization. Within these dimensions, participants positively highlighted: clinical consistency of the responses, colloquial and sociocultural adaptation of the language, conversational fluency, and ability to sustain interactions similar to a routine medical consultation. In contrast, they identified as negative aspects: excessive narrative organization, limited emotional expressiveness, absence of non-verbal cues, premature delivery of clinical information, and limited contextualization of the scenario. Table 1 summarizes the positive and negative aspects identified by participants, with illustrative examples of their comments.

Table 1

Qualitative findings from participant feedback, with illustrative examples.

Qualitative Finding	Representative Participant Feedback
Positive	
Clinical consistency	<i>“Very specific and appropriate to the scenario”</i>
Colloquial and sociocultural adaptation	<i>“Good use of the Cordobian dialect”</i>
Conversational fluency	<i>“Good response, clear”</i>
Consultation realism	<i>“Very similar to how a real patient would describe it”</i>
Negative	
Excessive narrative organization	<i>“The sentence is structured in an unnatural way”</i>
Limited emotional expressiveness	<i>“Response could have been more emotionally expressive”</i>
Absence of non-verbal cues	<i>“More exclamation marks could be used”</i>
Premature information disclosure	<i>“Patients would not usually provide these details without more targeted questioning”</i>
Limited contextualization of the scenario	<i>“It is unclear whether the consultation takes place in the emergency department or an outpatient clinic”</i>

5. Discussion

Our findings suggest that the performance of LLM-based SVPs depends not only on the conversational capabilities of the model but also on the clinical and narrative complexity of the simulated scenario. We observed that as patient comorbidities, age, and the emotional complexity of the case increased, both perceived realism and narrative stability declined.

Consistent with this trend, a marked difference was observed between the two clinical scenarios: the younger patient with lymphadenopathy received a considerably higher proportion of positive observations than the older adult with progressive dyspnea. This finding may suggest that LLMs perform more effectively in relatively linear clinical conditions characterized by fewer simultaneous variables and lower contextual complexity. However, such complexity is likely not limited to the clinical scenario alone but may also encompass communicational and cultural dimensions that could influence the performance of conversational AI systems [10].

The older adult scenario required the integration of multiple simultaneous dimensions, including aging, respiratory failure, frailty, emotional burden, and associated comorbidities. Under this multilayered complexity, participants perceived overly structured responses, limited discourse fragmentation, and reduced emotional expressiveness [12] — patterns that diverged from what would be expected in this patient profile. Beyond clinical complexity, this discrepancy may reflect the model's limited ability to represent older patient demographics or to sustain emotionally complex interactions convincingly. However, because the two SVPs differed in multiple aspects simultaneously (e.g., clinical complexity, age, and emotional burden), this study cannot determine which factor contributed most to the marked differences between them. Future controlled studies should independently manipulate clinical complexity and patient demographic characteristics to disentangle alternative explanations.

These results tentatively suggest that the primary challenge for SVPs may lie not merely in reproducing symptoms or biomedical data, but in sustaining the ambiguity, imprecision, and heterogeneity that often accompany clinically complex patients [1, 15].

In addition to producing overly organized responses, another finding was the model tendency to provide clinical information prematurely. Participants noted that the SVPs revealed diagnostic data too quickly, artificially simplifying the anamnesis process. This phenomenon, which constitutes an information leakage bias, could become problematic in complex scenarios where clinical reasoning relies precisely on the physician's ability to progressively reconstruct scattered, incomplete, or contradictory information. In real-world patients, particularly older adults with multimorbidity, the interview is typically characterized by fragmented narratives, omissions, expressive difficulties, and non-linear associations, elements that current models still struggle to faithfully replicate.

Furthermore, the differences observed across medical specialties suggest that the way clinicians conduct the medical interview may also influence the responses generated by the SVPs. Although this study cannot establish formal associations, these observations highlight the need for future research examining the interaction between interview structure, clinical complexity, and the narrative performance of conversational models. Nevertheless, even when considering this interaction, the limitations became more evident in the clinically more complex scenario, reinforcing the hypothesis that the narrative stability of language models deteriorates as contextual and clinical demands increase.

Participants positively valued the colloquial and sociocultural adaptation of the language used by the model. Nevertheless, this coexisted with limitations regarding the contextualization of the healthcare setting and emotional expressiveness. In multiple interactions, the SVPs failed to specify whether the consultation was taking place in an emergency department, an outpatient clinic, or an inpatient ward. This limitation partially reduced the sense of clinical fidelity, particularly in scenarios where the healthcare context significantly alters the nature of the medical interview.

The findings of this study suggest that the implementation of LLM-based SVPs in medical education should not be evaluated solely by their conversational fluency or technical ease of use. As clinical complexity, age, and comorbidity burdens increase, narrative and clinical limitations emerge that can compromise the educational fidelity of the simulation. This underscores the need to develop validation and design frameworks centered on clinical complexity, narrative coherence, and sociocultural

adaptation, prioritizing the realistic representation of the patient over mere fluent chat generation.

6. Conclusions

This pilot study evaluated the behavior of GPT-4.1 as an SVP during interdisciplinary clinical interviews conducted by physicians and advanced medical students through a workshop including prebriefing, peer cross-validation, and debriefing activities. Overall, participants perceived the simulated interactions as realistic and highlighted strengths including clinical consistency, conversational fluency, colloquial and socioculturally adapted language, and the ability to sustain interactions resembling medical consultations. At the same time, they identified limitations related to excessive narrative organization, premature disclosure of clinical information, limited emotional expressiveness, insufficient contextualization of the healthcare setting, and the absence of non-verbal cues. A marked difference was also observed between the two clinical scenarios, with positive observations declining from 34 in the linear case to only 15 in the complex scenario. This sharp decline suggests that current models struggle to sustain the narrative stability and discursive fragmentation required to simulate older patients with multiple comorbidities, although the present study cannot determine which specific factors contributed most to these differences. Furthermore, the models' tendency to disclose diagnostic information prematurely, together with their difficulty in clearly establishing the healthcare setting (e.g., distinguishing between an emergency department and an outpatient clinic), may limit their suitability for training advanced clinical reasoning skills.

Future work will involve additional workshops with a larger and more diverse participant pool, evaluating a broader range of patient profiles and clinical scenarios using standardized evaluation rubrics for more systematic comparisons. We also plan to examine how different medical specialties (Internal Medicine, Mental Health, Family Medicine) shape interactions with SVPs, refine the multimodal integration of diagnostic resources such as ECG and chest X-ray links tested here, and extend the framework from closed proprietary models to open ones.

7. Limitations

Several limitations of this work in progress must be acknowledged. The sample size was small ($n=16$), which may limit the generalizability of the findings across healthcare education settings. The study also evaluated only two clinical scenarios (a young adult with linear pathology and an older adult with comorbidities), which does not capture the full spectrum of clinical and demographic complexity. In addition, data collection was partially incomplete, as seven participants did not complete the requested five feedback entries, resulting in a lower-than-expected volume of evaluative data. The study also focused exclusively on written chat interactions, excluding voice-based simulation and non-verbal communication, both essential for comprehensive clinical training. Qualitative feedback from this pilot is still being analyzed and will be presented in future work to further contextualize these findings. Most importantly, the results are tied to the performance of a single model (GPT-4.1) at a specific point in time; other LLMs, including open-source alternatives, may exhibit different behavioral patterns. We plan to address this limitation in the coming months by extending the evaluation to open models.

Acknowledgments

We would like to express our sincere gratitude to the Simulation Hospital of the Facultad de Ciencias Médicas, Universidad Nacional de Córdoba (UNC), for providing the facilities and support necessary to conduct this workshop. In particular, we thank its Director and Technical Secretary, Prof. Dr. Rogelio D. Pizzi, for his institutional support. We also sincerely thank Dr. Lorena Chávez from the Mental Health Department of the Hospital Nacional de Clínicas (HNC) for her invaluable assistance in coordinating the workshop. We are grateful to the Institutional Health Research Ethics Committee (CIEIS) of the HNC for reviewing and approving the research protocol (Res. No. 122-06102025). We also thank the

medical professionals at the HNC who developed the clinical profiles for the simulated virtual patients. Finally, we sincerely thank the 16 healthcare professionals and advanced medical students from the Mandatory Final Practice, Internal Medicine, Family Medicine, and Mental Health departments who participated in this pilot study.

We are also grateful to the High Performance Computing Center (CCAD) of the Universidad Nacional de Córdoba (<https://ccad.unc.edu.ar/>), part of SNCAD-MinCyT, Argentina. Although the present study did not directly use its computational infrastructure, since completing this work we have been migrating the simulations to open-weight LLMs with the support of CCAD's computational resources. This ongoing effort will contribute to the technological sovereignty and sustainability of the project.

The first author is personally grateful to the members of her doctoral advisory committee, Dr. Luis A. Biedma, Dr. Nicolás Wolovick, and Dr. Pedro R. D'Argenio, for their insightful comments and guidance.

This work is supported by the project "Responsible AI in Healthcare: Integrating Language Models, Clinical Knowledge, and Ethical Evaluations" (FMV_1_2025_1_186830), funded by the Maria Viñas Fund of the National Agency for Research and Innovation (ANII), Uruguay.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini in order to translate the original Spanish manuscript into English. After using this tool, the bilingual authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] J. W. Song, M. Kim, C. Hong, Y. S. Kim, J. Cho, J. H. Kim, J. Myung, A. Choi, H. Yoon, S. G. W. Lee, S. C. You, C. Park, Development and validation of cpx-mate: An end-to-end medical education platform integrating voice-based virtual patient simulation and automated real-time evaluation, medRxiv preprint, 2026. doi:10.64898/2026.02.21.26346803.
- [2] K. E. Keifenheim, M. Teufel, J. Ip, N. Speiser, E. J. Leehr, S. Zipfel, A. Herrmann-Werner, Teaching history taking to medical students: a systematic review, BMC Medical Education 15 (2015) 159. doi:10.1186/s12909-015-0443-x.
- [3] X. Meng, M. Zhang, W. Ma, X. Cheng, X. Yang, A clinical medicine level test at Jinan University School of Medicine reveals the importance of training medical students in clinical history-taking, PeerJ 11 (2023) e15052. doi:10.7717/peerj.15052.
- [4] J. H. Aranda, S. M. Monks, Roles and Responsibilities of the Standardized Patient Director in Medical Simulation, in: StatPearls, StatPearls Publishing, Treasure Island (FL), 2026.
- [5] D. Bauer, F.-M. Lahner, S. Huwendiek, F. M. Schmitz, S. Guttormsen, An overview of and approach to selecting appropriate patient representations in teaching and summative assessment in medical education, Swiss Medical Weekly 150 (2020) w20382. doi:10.4414/smw.2020.20382.
- [6] J. Kaplonyi, K.-A. Bowles, D. Nestel, D. Kiegaldie, S. Maloney, T. Haines, C. Williams, Understanding the impact of simulated patients on health care learners' communication skills: a systematic review, Medical Education 51 (2017) 1209–1219. doi:10.1111/medu.13387.
- [7] F. Holderried, C. Stegemann-Philipps, L. Herschbach, J.-A. Moldt, A. Nevins, J. Griewatz, M. Holderried, A. Herrmann-Werner, T. Festl-Wietek, M. Mahling, A Generative Pretrained Transformer (GPT)-Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study, JMIR Medical Education 10 (2024) e53961. doi:10.2196/53961.
- [8] J. J. Webb, Proof of Concept: Using ChatGPT to Teach Emergency Physicians How to Break Bad News, Cureus 15 (2023) e38755. doi:10.7759/cureus.38755.
- [9] C. Wang, S. Li, N. Lin, X. Zhang, Y. Han, X. Wang, D. Liu, X. Tan, D. Pu, K. Li, G. Qian, R. Yin, Application of Large Language Models in Medical Training Evaluation—Using ChatGPT as a Standardized Patient: Multimetric Assessment, J Med Internet Res 27 (2025) e59435. doi:10.2196/59435.

- [10] D. Peters, F. Espinoza, M. da Re, G. Ivetta, L. Benotti, R. A. Calvo, Towards culturally-appropriate conversational ai for health in the majority world: An exploratory study with citizens and professionals in latin america, arXiv preprint arXiv:2507.01719, 2025.
- [11] K. Masters, H. MacNeil, J. Benjamin, T. Carver, K. Nemethy, S. Valanci-Aroesty, D. C. M. Taylor, B. Thoma, T. Thesen, Artificial Intelligence in Health Professions Education assessment: AMEE Guide No. 178, *Medical Teacher* 47 (2025) 1410–1424. doi:10.1080/0142159X.2024.2445037.
- [12] Y. Liu, C. Shi, L. Wu, X. Lin, X. Chen, Y. Zhu, H. Tan, W. Zhang, Development and Validation of a Large Language Model-Based System for Medical History-Taking Training: Prospective Multicase Study on Evaluation Stability, Human-AI Consistency, and Transparency, *JMIR medical education* 11 (2025) e73419. doi:10.2196/73419.
- [13] A. Abid, A. Abdalla, T. Ali, D. Khan, A. Alfozan, J. Zou, Gradio: Hassle-free sharing and testing of ml models in the wild, arXiv preprint arXiv:1906.02569, 2019.
- [14] OpenAI, Gpt-4.1, 2025. URL: <https://openai.com/index/gpt-4-1/>.
- [15] Y. Hicke, J. Geathers, K. Vu, J. Sewell, C. Cardie, J. Talwalkar, D. Shung, A. G. Jack, S. Cornes, M. Preston, R. Kizilcec, Medsimai: Simulation and formative feedback generation to enhance deliberate practice in medical education, in: *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference, LAK '26*, Association for Computing Machinery, New York, NY, USA, 2026, p. 415–424. doi:10.1145/3785022.3785092.