Automated Evaluation of Standardized Patients with LLMs

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

Abstract

Standardized patients (SPs) are essential for clinical reasoning assessments in medical education. This paper introduces evaluation metrics that apply to both human and simulated SP systems. The metrics are computed using two LLM-as-a-judge approaches that align with human evaluators on SP performance, enabling scalable formative clinical reasoning assessments.

1 Introduction

Clinical reasoning (CR) skills are fundamental to accurate diagnosis and effective patient care; accordingly, their systematic instruction and assessment constitute a critical component of undergraduate medical education (Harden, 1988). One of the most widely adopted formats for evaluating clinical reasoning competencies is the Objective Structured Clinical Examination (OSCE). Repeated, structured interactions within OSCEs have been shown to effectively promote the development of clinical reasoning skills (Laschinger et al., 2008). A central feature of OSCEs is the simulated clinical encounter, in which learners engage in a clinical case by interviewing a patient to elicit diagnostically relevant information, including symptoms, medical history, and context. Traditionally, OSCEs employ laypersons trained to portray clinical scenarios—referred to as standardized patients (SPs)—who are tasked with consistently enacting specific patient personas to support teaching, learning, and assessment. SPs adhere to detailed case scripts and standardized response protocols to ensure realism, reliability, and reproducibility across encounters.

While human SPs provide a realistic and safe environment for learners to practice clinical skills, their use is resource-intensive, requiring substantial investments in training, coordination, and examination delivery (Rau et al., 2011). In recent years, large language models (LLMs) have been explored as a way to design simulated standardized patients (SSPs), offering a promising alternative to traditional human SPs. SSPs offer several advantages, including scalability to larger cohorts of learners, increased availability for on-demand practice, and enhanced flexibility in portraying a wide range of patient personas. However, maintaining fidelity to the prescribed patient script and ensuring consistent persona representation remain significant challenges in the deployment of SSPs.

Large language models (LLMs) can be guided to portray specific patient characteristics (e.g., via prompting); however, the reliability and precision of such portrayals remain active areas of research (Cook, 2024; Schmidgall et al., 2024; Shindo and Uto, 2024). Fortunately, existing frameworks for evaluating the performance of SPs can be adapted and automated for use with SSPs (Geathers et al., 2025). The evaluation of SP or SSP performance generally falls into three categories: (1) human evaluation of the responses to physician questions; (2) traditional machine learning methods that are trained on labeled datasets of patient responses; and (3) LLMs that judge the quality of the patient responses with little or no prior training. To ensure the accuracy of methods (2) and (3), human evaluations (Category 1) are typically employed as a reference standard to produce ground-truth labels.

In this paper, we introduce a novel LLM-based evaluation framework to automatically evaluate SPs, applicable to both humans and SSPs. The framework classifies SP responses into one of five performance categories, developed as part of this work: *Correct, Too Much Information, Too Little Information, Incorrect,* or *Not Applicable*. These categories can be used as metrics to monitor the SP or SSP performance over time, across different SPs or SSP systems, and ensure that students are able to engage appropriately with the SP or SSP.

The contributions of this work are as follows:

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 231–238

1. We perform extensive human annotation of a set of 41 transcripts of student-patient interactions from four clinical cases to serve as the ground truth to validate our proposed automated evaluation approaches.
2. We introduce two methods of classifying SP responses:
   In *Method 1*, the LLM uses the case guidelines, the conversation up to this point, and the current physician question to classify the current response into the appropriate category.
   In *Method 2*, the LLM first uses the case guidelines, the conversation up to this point, and the current physician question to generate the prescribed SP response. The LLM then uses the prescribed SP response, current physician question, and conversation up to this point to compare with the actual SP response and classify it into the appropriate category.
3. We validate these proposed methods of SP evaluation by comparing the classification results of each method with the human evaluation, assessing alignment with human expert judgment. We discuss the implications of these results for evaluating both human and simulated SPs.

2 Related Work

2.1 Evaluation of Human SPs

SPs are typically evaluated on dimensions such as realism, accuracy, consistency, and communication. Realism and communication are commonly assessed through structured observations by faculty and peers, and student feedback immediately after encounters (Gonullu et al., 2023; Erby et al., 2011). Accuracy of performance—the physical, emotional, and cognitive portrayal of the clinical case—is often evaluated using third-party observations and SP self-assessment checklists. Post-encounter self-checklists help SPs reflect on the fidelity of their performance, improving reliability over time (Erby et al., 2011). Consistency of performance, a defining characteristic of SP programs, refers to the uniform delivery of case prompts and behaviors across all student interactions for a given case, and is evaluated through a mix of live observation, video review, and checklists (Lewis et al., 2017; Erby et al., 2011). Overall, the evaluation of human SPs' performance remains mainly manual, involving faculty and peer observation, student feedback, and SP self-assessment, offering a multi-angled view essential for maintaining high standards in simulated clinical environments.

2.2 LLMs as SSPs and their Evaluation

Recent research has explored the potential of LLMs in simulating patient interactions (LLMs as SSPs) to support both clinical skill development and performance evaluation. For instance, Li et al. (2024) examined how SSPs can support clinical inquiry skills, while Holderried et al. (2024) focused on improving medical history-taking skills. Yamamoto et al. (2024) expanded this to encompass general medical interview skills, and Sardesai et al. (2024) applied LLM-based simulations to anesthesia training. Gray et al. (2024) investigated the use of LLMs in guiding prenatal counseling, whereas Tu et al. (2024) worked on advancing AI diagnostic agents to improve their clinical utility.

Human judgment remains the most widely used reference for assessing chatbot-generated interactions. Specialists (Chen et al., 2023; Gray et al., 2024) and students (Fan et al., 2024) have been engaged to evaluate the realism, appropriateness, usability, relevance, rationality, and honesty of chatbot outputs. User surveys—such as Likert-scale questionnaires (Sardesai et al., 2024), the Chatbot Usability Questionnaire (Holderried et al., 2024), and the Simulation-Based Training Quality Assurance Tool (Yamamoto et al., 2024)—have been leveraged to evaluate perceived usability, intuitiveness, accuracy, comfort, and overall user experience. Automated metrics, like algorithmically derived conversational dimensions (Liao et al., 2024), a GPT-4–based chatbot arena framework (Li et al., 2024), and quantitative scoring of chatbot responses (Chen et al., 2023), have been used to enable scalable and objective evaluation of chatbot performance. These metrics assess factors such as accuracy, honesty, focus, passivity, cautiousness, and guidance. Finally, outcome-based evaluations have also been conducted; for example, Yamamoto et al. (2024) compared formal exam performance between students who used SSPs during their preparation and those who did not.

The above studies have identified several limitations of LLMs, including their tendency to produce hallucinated content, overly formal or repetitive responses, and unnaturally polite dialogue (Sardesai et al., 2024). Gray et al. (2024) highlight the importance of expert oversight when using AI-generated content in educational settings. Additionally, current LLM-based systems provide

limited support for nonverbal communication skills, which are essential for effective medical interviewing (Yamamoto et al., 2024).

3 Data

3.1 Clinical Interviews with Standardized Patients

The data were drawn from a prior study in which participating students interacted with four human SPs, each portraying a distinct case. Each scenario was developed along with case-specific guidelines and training protocols designed to elicit observable clinical reasoning behaviors from the students. Students were randomly assigned to begin with one of the four cases and subsequently completed the remaining three cases in order. All encounters were recorded using Recollective[1], a qualitative research platform that supports live and asynchronous (i.e., pre-recorded) video interactions. The software first recorded the conversations and then produced transcriptions in separate files, differentiating the student and SP speech.

Participating Students. A total of 76 post-clerkship medical students were recruited from four U.S. medical schools. 32 were in their third year of medical school, 2 were transitional students between their third and fourth years, and 42 were in their fourth year.

Standardized Patients. Standardized patients were recruited from local training programs and partner institutions affiliated with the study sites. Each SP received standard compensation for participation in both training and assessment activities. The medical assessment organization personnel conducted the SP training following established industry protocols, general guidelines, and case-specific requirements. SPs underwent both individual and group training sessions to ensure consistency and reliability across performances. For each clinical case, a minimum of three SPs were trained to serve not only as actors in student encounters but also as peer evaluators, providing feedback and quality assurance for fellow SPs.

Clinical Cases. Four clinical cases were developed as part of a prior study to support the standardized evaluation of medical students' clinical reasoning skills. For each case, both general training protocols and case-specific instructions were designed to guide SP behavior and ensure that student–SP interactions elicited diagnostically productive lines of questioning. SPs were explicitly instructed to refrain from offering suggestions or guidance on how students should conduct the encounter. In instances where students inquired about symptoms not included in the case script, SPs were instructed to deny the presence of such symptoms to preserve case fidelity. Each encounter began with a standardized opening statement delivered by the SP, introducing the primary reason for visiting the clinic. To maintain consistent interactions, boilerplate responses were developed for addressing routine or general questions. For open-ended inquiries, SPs were provided with a sequenced set of acceptable responses, structured to disclose relevant clinical information with the goal of not revealing too much information at once. This approach was intended to support the development of student inquiry skills while preserving the realism and educational value of the simulation. Case contents were designed to be both realistic and engaging for the student. Case 1 consists of a 33-year-old woman who has been experiencing shortness of breath; Case 2 consists of a 40-year-old man who has been continuously vomiting; Case 3 consists of a 46-year-old woman who has been experiencing weakness; and Case 4 consists of a 65-year-old man who has had trouble sleeping.

3.2 Patient Response Annotation

Two human experts who were familiar with the case contents and evaluation criteria annotated transcripts of the conversations in order to evaluate the quality of the SP responses. Each response from the human SPs was labeled with one of the five discrete labels in Table 1. The label categories were derived based on a combination of insights from literature that evaluates human SPs and practical guidance by members of the team who have trained human SPs. The annotation process consisted of several steps to increase agreement between annotators and to ensure high-quality annotations. First, each annotator independently annotated the SP responses in an adjudication set of four transcripts sampled from the same case. Subsequently, the annotators reviewed any discrepant annotations together with other team members and agreed upon adjudicated annotations. Revisions to the annotation guidelines were made accordingly based on these conversations. The final calibration set included 162 *Correct*, 44 *Not Applicable*, 13 *Too Much Information*,

[1] https://www.recollective.com/qualitative-research-recollective

Label	Description
Correct	The response is accurate and appropriate given the instructions contained in the case training guidelines. The response is relevant to the physician's question and contains the appropriate amount of content based on the specific question the physician asked.
Too Much Information	The response is relevant to the physician's question, but it contains more information than is justified based on the specific question that the physician asked. This can occur when the patient provides additional information from the case materials that wasn't prompted by a question from the physician.
Too Little Information	The response is relevant to the physician's question, but it contains less information than is expected based on the specific question that the physician asked. This can occur when the patient omits relevant content from the case materials and provides a generic answer.
Incorrect	The response is not accurate or is inappropriate given the instructions contained in the case training guidelines. This can occur when the patient provides a response that is irrelevant or off-topic, when the patient volunteers made-up information about topics that are not covered in the training guidelines, when the patient provides specific details about their condition that are not specified in the case materials, etc.
Not Applicable	The question is not applicable to the case document and results in a non-clinical or irrelevant response.

Table 1: Annotation guidelines given to annotators for evaluating each SP response.

7 *Too Little Information*, and 7 *Incorrect* responses. After the adjudication round, the two annotators independently annotated the same 20 transcripts (five randomly sampled from each case). Finally, 17 additional transcripts were single-annotated by the annotators. Transcripts of entire conversations were annotated to allow for contextual information to be available to annotators. Table 2 shows the annotation distribution for each annotator on transcripts that were not in the calibration set. The calibration, double-annotation, and single-annotation sets yielded 41 transcripts with an average of 54.8 annotated question-response pairs (*SD*=16.4). This produced 2248 question-response pairs in total. See Appendix A for exemplar labeled patient responses.

Label	Annotator 1 Count	Annotator 2 Count
Correct	1407 (72%)	748 (65%)
TMI	59 (3%)	14 (1%)
TLI	40 (2%)	2 (0%)
Incorrect	38 (2%)	29 (3%)
NA	414 (21%)	365 (32%)

Table 2: Distribution of annotations by annotator.

The annotation guidelines were developed based on industry practice, SP training guidelines, and conversational agent literature. The Cohen's Kappa value denoting the inter-annotator agreement on the double-annotated set of 20 transcripts was 0.501, and the agreement percentage was 76.5% (*n*=1104). For case 1, the agreement percentage was 70.8% ($\kappa = 0.436$, *n*=281). For case 2, the agreement percentage was 80.2% ($\kappa = 0.481$, *n*=217). For case 3, the agreement percentage was 82.8% ($\kappa = 0.604$, *n*=332). For case 4, the agreement percentage was 71.9% ($\kappa = 0.457$, *n*=274).

4 LLM-as-a-Judge Evaluation

To automatically evaluate SP responses to student questions, we employed a technique that leverages LLMs called LLM-as-a-judge (Gu et al., 2025). For all evaluations, we used OpenAI's GPT-4o (version: *2025-01-01-preview*) as the judge. In this paper, we introduce two methods of using LLMs to judge the SP responses. *Method 1* uses a single request to the LLM to categorize the SP response using the case-specific guidelines, the conversation up to this point, and the current physician question. *Method 2* uses two requests to the LLM, in which the first request generates a prescribed patient response based on the case-specific guidelines, the conversation up to this point, the current physician question, and the second request compares the prescribed and actual SP response to categorize the SP response. *Method 2* was chosen over alternative methods that leverage prior data (e.g., few-shot learning or fine-tuning) to first attempt to solve this problem without the use of labeled examples, which would require a robust set of ground truth labels.

For both methods, the case guidelines are the same instructions that are given to the SPs to portray the patient. The conversations are encoded as transcribed text and each question-response pair is appended up to the current question as context, noting the speaker of the text (i.e., student or SP).

5 Results

To evaluate the performance of the LLM-as-a-judge method relative to human annotations, we used a dataset comprising 2248 human-annotated question-response pairs drawn from 41 encounter transcripts. Table 3 displays the results for this comparison. Both F1 scores and accuracy metrics are

reported to assess the degree of alignment between LLM-generated classifications and human reference annotations in the evaluated methods. Results are reported both in aggregate and disaggregated by individual cases: Cases 1 (n=814), 2 (n=382), 3 (n=571), and 4 (n=481). Across all cases and in the overall analysis, *Method 1* consistently outperforms *Method 2*. A baseline comparison, referred to as the *Majority* baseline, assigns the most frequent class label (which is always *Correct*) to all instances within each case and in the overall dataset. *Method 1* outperforms this baseline in terms of F1 scores, but its performance in terms of accuracy shows mixed results. Given the multi-class nature of the problem and the imbalanced label distributions, F1 score is a more informative metric.

Table 4 displays the results per label, including the distribution of predicted labels, their precision, recall, and F1 score for the entire human-annotated dataset.

6 Discussion

Overall, the findings of this study show positive results for both the human annotation process and the two proposed automated LLM-as-a-judge methods. Human annotation remains a very resource-intensive and cognitively demanding task that requires careful calibration and deliberation among annotators to ensure consistency and validity. LLMs offer a scalable and efficient alternative that can be used in conjuction with human annotations to reduce manual labor involved in annotating. One practical application of this hybrid approach is the selective delegation of annotations to LLMs for labels where model performance is demonstrably high (e.g., the *Correct* label). By pre-filtering such responses, human annotators can allocate their attention to more complex or ambiguous categories that require more nuanced judgment such as *Too Much Information* or *Too Little Information*. In addition, certain LLM-as-a-judge methods were shown to be effective in annotating responses that are *Not Applicable* (i.e., responses associated with non-clinical questions), providing another opportunity for filtering and streamlining the annotation process. Taken together, these findings suggest that LLM-as-a-judge approaches can serve as valuable tools for augmenting human annotation workflows, saving time and effort for human reviewers while preserving annotation quality.

Although human annotations served as the ground truth (reference standard) for this study, notable levels of disagreement were observed among annotators. For the subset that were doubly-annotated (n=1104), the Kappa value for inter-annotator agreement was 0.501. While this reflects only moderate agreement, the task of labeling responses with one of five possible subjective categories can lead to poor agreement. Ironically, the lowest annotator agreement was observed in Case 1, which was also the case for which the initial annotator calibration was performed. This finding underscores the influence of *case-specific* features on both human annotation and LLM labeling. The variability observed suggests the need to standardize the annotation guideline development process to promote consistency across clinical scenarios. Many of the disagreements among annotators were along the threshold of *Correct* and partially correct responses (e.g., *Too Much Information*). These are often very nuanced phrases and require carefully crafted definitions during case generation.

Across all individual cases and in the aggregate analysis, *Method 1* consistently outperformed both the *Majority* baseline and *Method 2*. With the advancement of LLMs and continual refinement of prompting techniques, it is not surprising that an LLM-as-a-judge method can outperform a majority class baseline. What is surprising and noteworthy is the difference in performance between the two LLM-based methods. *Method 1* incorporates the case-specific guidelines directly into the LLM request that conducts the evaluation of each SP response, while *Method 2* references the guidelines only during the initial generation of LLM-recommended patient responses and not during the LLM request that conducts the evaluation. As a result, *Method 2* loses information when making the actual evaluation and classification of the SP response, only comparing it to another response. Our team had hypothesized that this would improve performance by allowing the LLM to split the evaluation task into multiple steps. This result underscores the power of current LLMs in navigating large context windows (e.g., long conversations and long reference documents, simultaneously). It may also suggest that the LLM-generated responses may not be informative enough to serve as a ground-truth response. Despite its superior overall performance, *Method 1* exhibited uneven classification accuracy across labels, with the majority of predicted labels falling under the *Correct* category. Among the minority classes, only *Not*

Method	Overall F1 (Accuracy)	Case 1 F1 (Accuracy)	Case 2 F1 (Accuracy)	Case 3 F1 (Accuracy)	Case 4 F1 (Accuracy)
Majority	0.60 (0.72)	0.56 (0.69)	0.75 (0.83)	0.57 (0.70)	0.59 (0.71)
1	0.72 (0.67)	0.71 (0.70)	0.79 (0.75)	0.69 (0.63)	0.69 (0.64)
2	0.48 (0.42)	0.47 (0.44)	0.59 (0.49)	0.45 (0.38)	0.44 (0.38)

Table 3: LLM-as-a-Judge performance.

Method	Label	Predicted Count	Precision	Recall	F1
1	*Correct*	1396	0.86	0.74	0.80
1	*TMI*	278	0.09	0.33	0.14
1	*TLI*	93	0.02	0.4	0.03
1	*Incorrect*	92	0.09	0.17	0.12
1	*NA*	389	0.74	0.61	0.67
2	*Correct*	1037	0.83	0.53	0.65
2	*TMI*	555	0.11	0.81	0.19
2	*TLI*	501	0.06	0.60	0.10
2	*Incorrect*	151	0.03	0.09	0.04
2	*NA*	4	1.0	0.01	0.02

Table 4: LLM-as-a-Judge performance by label. *TMI=Too Much Information*, *TLI=Too Little Information*, and *NA=Not Applicable*.

Applicable achieved comparable performance. By contrast, *Method 2* demonstrated greater sensitivity to minority classes (i.e., *Too Much Information* and *Too Little Information*), suggesting a stricter approach and potentially greater attention to subtle linguistic nuances. These findings indicate that a hybrid approach, combining the contextual breadth of *Method 1* with the sensitivity of *Method 2*, may yield further improvements in performance.

This study focused specifically on annotation of SP responses to student-initiated questions. The ultimate goal of this work is to develop an automated system for evaluating SSP responses, with the dual goals of monitoring and improving system performance. Importantly, the proposed evaluation framework is equally applicable to human SP responses. Additionally, SP responses offer more natural conversation, more variability in phrasing, and more room for the student to ask variable questions compared to SSP responses. As a result, building an evaluation system by first using SP responses enables a finer-grained view of the types of responses that constitute a realistic patient encounter. This approach has the potential to enhance the formative utility of SSP systems, ultimately supporting more effective development of clinical reasoning skills in students. Future evaluations of SSP responses may find differences in response characteristics compared to SP responses, potentially leading to the adjustment of this evaluation framework.

Limitations

This study represents an initial investigation of both human annotation and automated LLM annotation of SP responses in physician-patient interactions. Due to resource constraints, the annotation process was limited to two annotators. With more annotators, an even larger pool of labeled responses could be annotated. Independent of the number of responses, ensuring high-quality annotations with high agreement is critical. This is especially important for labels that are actionable, such as when the SP provides too much or too little information in response to the student. A larger pool of annotators, with a more rigorous calibration and preparatory period would likely yield improved results.

7 Conclusion

The automatic annotation of SP responses has significant potential for advancing the development of more accurate and effective formative assessments of clinical reasoning. Enhancing the performance of either SPs or SSPs can contribute to more meaningful student-patient interactions. Building such an evaluation system requires high-quality human annotations to serve as the ground truth for what constitutes an effective (e.g., accurate) patient response. This study reported the results of a human annotation effort involving SP responses, guided by a structured rubric comprising five response categories. Building on this foundation, two LLM-as-a-judge methods were introduced as automated approaches to replicate the human annotation pro-

cess. Both methods showed promising agreement with human judgments. Future research should focus on integrating the strengths of each LLM-as-a-judge method into a unified automated annotation pipeline. Ultimately, these methods will be applied to SSP-generated responses, enabling systematic evaluation of both the evaluation engine and the underlying SSP system, and thereby contributing to the iterative improvement of AI-supported clinical reasoning learning tools.

References

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

David A. Cook. 2024. Creating virtual patients using large language models: Scalable, global, and low cost. *Medical Teacher*, 3:1–3.

Lori A. Erby, Debra L. Roter, and Barbara B. Biesecker. 2011. Examination of standardized patient performance: accuracy and consistency of six standardized patients over time. *Patient Education and Counseling*, 85(2):194–200. Epub 2010 Nov 20.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*.

Jadon Geathers, Yann Hicke, Colleen Chan, Niroop Rajashekar, Justin Sewell, Susannah Cornes, Rene F. Kizilcec, and Dennis Shung. 2025. Benchmarking generative ai for scoring medical student interviews in objective structured clinical examinations (osces). *Preprint*, arXiv:2501.13957.

Ipek Gonullu, Cansu Derya Doğan, Serap Erden, and Derya Gökmen. 2023. A study on the standard setting, validity, and reliability of a standardized patient performance rating scale - student version. *Annals of Medicine*, 55(1):490–501.

Megan Gray, Austin Baird, Taylor Sawyer, Jasmine James, Thea DeBroux, Michelle Bartlett, Jeanne Krick, and Rachel Umoren. 2024. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of chatgpt: Exploratory observational study. *JMIR Medical Education*, 10:e50705.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

R.M. Harden. 1988. What is an OSCE? *Medical Teacher*, 10(1):19–22.

Florian Holderried, Carolin Stegemann-Philipps, Lisa Herschbach, Jan-Alexander Moldt, Alexander Nevins, Jan Griewatz, Marc Holderried, Anne Herrmann-Werner, Thomas Festl-Wietek, and Martin Mahling. 2024. A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR Medical Education*, 10:e53961.

S. Laschinger, J. Medves, C. Pulling, R. McGraw, B. Waytuck, M.B. Harrison, and K. Gambeta. 2008. Effectiveness of simulation on health profession students' knowledge, skills, confidence and satisfaction. *International Journal of Evidence-Based Healthcare*, 24:278–302.

Kelly L. Lewis, Carrie A. Bohnert, William L. Gammon, Henrike Hölzer, Layla Lyman, Cheryl Smith, Tara M. Thompson, and Georgia Gliva-McConvey. 2017. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). *Advances in Simulation*, 2:10.

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.

T. Rau, J. Fegert, and H. Liebhardt. 2011. How high are the personnel costs for OSCE? A financial report on management aspects. *GMS Journal for Medical Education*, 28(1):16.

Neil Sardesai, Peter Russo, James Martin, and Anjali Sardesai. 2024. Utilizing generative conversational artificial intelligence to create simulated patient encounters: A pilot study for anaesthesia training. *Postgraduate Medical Journal*, 100(1182):237–241.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Naoki Shindo and Masaki Uto. 2024. ChatGPT-based virtual standardized patient that amends overly detailed responses in objective structured clinical examinations. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 263–269, Cham. Springer Nature Switzerland.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi,

Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, and 6 others. 2024. Towards Conversational Diagnostic AI. *arXiv preprint arXiv:2401.05654*.

Akira Yamamoto, Masahide Koda, Hiroko Ogawa, Tomoko Miyoshi, Yoshinobu Maeda, Fumio Otsuka, and Hideo Ino. 2024. Enhancing medical interview skills through AI-simulated patient interactions: Nonrandomized controlled trial. *JMIR Medical Education*, 10:e58753.

A Examples of Human-Labeled SP Responses

A.1 Correct

Student Question: "Consistently how long has your breathing difficulty been going on? Has it gotten worse recently?"
SP Response: "Yeah, uh, it's only been going on for about two years. Um, it's been the same from the beginning."
Explanation: The SP answered both questions correctly and did not provide additional detail.

A.2 Too Much Information

Student Question: "So, when you're having a really bad shortness of breath, do you have feelings like you're going to die or like it's like impending doom?"
SP Response: "No, but I'll just have that pounding in my chest and neck kind of on and off sometimes when I feel like I can't breathe."
Explanation: The student is asking only about the SP's thoughts or feelings during shortness of breath. The SP provided "pounding in my chest and neck" in addition to denying feelings of impending doom.

A.3 Too Little Information

Student Question: "So, I'm just going to go through a bunch of questions, and just tell me if you've had any of them in the last two weeks, okay? Have you had any headaches?"
SP Response: "Well, I do get stress headaches at work sometimes."
Explanation: The complete response for headaches is "When I'm at work, I get this feeling like a band around my head."

A.4 Incorrect

Student Question: "Can you tell me a bit more about how your periods are, like how many pads you use, kind of like what the experience is with those?"
SP Response: "I'll go through maybe four to five pads each day while I'm on my period."
Explanation: Her periods are heavy only during the first three days, not throughout her period.

A.5 Not Applicable

Student Question: "Well, thank you so much for talking to me and coming in when you did. I will be with you to support you throughout this process."
SP Response: "All right, thank you. I appreciate everything."
Explanation: The question and the response are not related to the case content.